UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Porovnání dekompozičního a ARMA přístupu při analýze časové řady

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **RNDr. et PhDr. Ivo Müller Ph.D.**
Vypracoval: **Bc. Marcela Pokorná**
Studijní program: N1103 Aplikovaná matematika
Studijní obor Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2017

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Marcela Pokorná

Název práce: Porovnání dekompozičního a ARMA přístupu při analýze časové řady

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: RNDr. et PhDr. Ivo Müller Ph.D.

Rok obhajoby práce: 2017

Abstrakt: Tato práce se zabývá zpracováním časových řad pomocí dekompozičního a ARMA přístupu. Cílem práce je nastudovat danou problematiku, porovnat tyto dva přístupy a některé výpočty zpracovat ve statistickém softwaru R.

Klíčová slova: Časové řady, dekompoziční přístup, ARMA modely, R software

Počet stran: 74

Počet příloh: 0

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Marcela Pokorná

Title: Comparison of decomposition and ARMA approach in analyzing a time series

Type of thesis: Master's thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: RNDr. et PhDr. Ivo Müller Ph.D.

The year of presentation: 2017

Abstract: This thesis is studying processing of time series, using decomposition and ARMA method. The goal of this thesis is to study the issue and to compare mentioned methods and some calculations in statistic software R.

Key words: Time series, decomposition approach, ARMA model, R software

Number of pages: 74

Number of appendices: 0

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana RNDr. et PhDr. Iva Müllera Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu diplomové práce RNDr. et PhDr. Ivovi Müllerovi Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích.

Obsah

Úvod

Tématem diplomové práce je *Porovnání dekompozičního a ARMA přístupu při analýze časové řady*, kde práce se především zaměřuje na zpracování sezónních časových řad.

V praxi se často setkáváme s daty vypozorovanými za určité období v čase, které je potřeba nadále zpracovat a vyhodnotit. Nejrozšířenějším přístupem, který se volí pro zpracování těchto dat, je dekompoziční a Boxův-Jenkinsův přístup. Cílem mé práce je tedy najít vhodná data a na nich porovnat oba tyto přístupy a posoudit, který přístup je vhodnější. S narůstajícím množstvím dat a sofistikovanějšími výpočty při analýze dat, zvláště u přístupu Boxe-Jenkinse, se jen stěží obejdeme bez výpočetní techniky. Proto jsem ve své práci pro zpracování dat volila statistického softwaru R.

Práce je rozdělena do 7 kapitol. První tři kapitoly se věnují teoretické části, kde jsou popsány základní principy fungování obou výše zmíněných přístupů. V těchto kapitolách jsou uvedeny také postupy výpočtů a vzorečky nezbytné pro praktické zpracování dat.

Čtvrtá kapitola se věnuje praktickému využívání výpočetního statistického softwaru. Důraz je zde kladen na zpracování dat v Boxově-Jenkinsově přístupu pomocí softwaru R.

V následujících třech kapitolách jsou postupně zpracovány zvolené datové sady. První zvolená datová sada - Vývoj teploty v ČR, je popsána a analyzována v kapitole 5. Další zvolenou datovou sadou je vývoj tržeb maloobchodů (kapitola 6) a vývoj počtu nakažených virem HIV v Jihoafrické republice za posledních 20

let (kapitola 7). Všechny tyto datové sady jsou modelovány pomocí přístupu Boxe a Jenkinse i dekompozičního. Na konci každé kapitoly následuje shrnutí výsledků z obou přístupů.

Kapitola 1

Úvod do teorie časových řad

V úvodní kapitole nejprve popíši základní myšlenky časových řad, které jsem čerpala především z literatury [1], [2] uvedené na konci mé práce. Teprve v následujících dvou kapitolách se budu věnovat zpracováním časových řad pomocí dvou konkrétních přístupů - dekompozičního a Boxově-Jenkinsově přístupu.

Časové řady jsou posloupnosti srovnatelných pozorování (věcně i prostorově) uspořádaných chronologicky od minulosti k přítomnosti. Tyto posloupnosti můžeme hojně najít v ekonomii, ve fyzikálních, biologických a společenských vědách a v neposlední řadě také v meteorologii.

Analýzou těchto pozorování můžeme zkonstruovat odpovídající matematický model. To umožňuje porozumět procesu, jak se posloupnost jednotlivých pozorování vyvíjí, a následně můžeme odhadnout s určitou pravděpodobností, jak se bude vyvíjet.

Tato posloupnost pozorování (dále jen časová řada) čelí různým specifickým problémům. Prvním významným specifickým problémem časové řady je výběr okamžiků pozorování. Některé diskrétní časové řady mohou vznikat diskretizací spojitých časových řad, anebo akumulací hodnot za dané časové období (někdy se provádí též průměrování). Druhý problém časových řad se týká délky. U dlouhých časových řad může dojít ke změně průběhu modelu. To může vést k nesnadno popsatelné časové řadě. Na druhou stranu, některé metody vyžadují minimální délku

časové řady. Musíme tedy hledat kompromis mezi dlouhou a krátkou časovou řadou. Významným problémem časových řad je také kalendář. Nestejný počet dní v měsíci nám může posloupnosti zkreslovat. Tento problém můžeme řešit přepočtením časových řad na měsíce se stejnou délkou dní v měsíci, a to buď budeme uvažovat

1) měsíce se standardní délkou 30 dnů, nebo

2) měsíce s délkou $\frac{365}{12}$.

Na podobný problém můžeme narazit i při změně času nebo v přestupném roce.

Existují tři základní metody a postupy pro analýzu časových řad. Prvním z nich je dekompozice časové řady. Tato metoda rozkládá časovou řadu do jednotlivých složek. Druhým novějším přístupem, který se začal objevovat ve 20.století, je Boxova-Jenkinsova metoda. Základním prvkem modelování je náhodná složka, která může být tvořena závislými náhodnými veličinami. A třetí metodou je modelování časových řad pomocí lineárních dynamických modelů. Tyto modely dokážou pracovat s hodnotami časových řad, které jsou vysvětlovány pomocí vysvětlujících hodnot časové řady. Nadále se ve své práci budu podrobněji zabývat dekompoziční a Boxovou-Jenkinsovou metodou.

Kapitola 2

Dekompoziční přístup v časových řadách

Hlavní myšlenkou dekompozičního přístupu je rozložit časovou řadu do jednotlivých složek tak, abychom jednotlivá pozorování časové řady mohli popsat následující rovnicí

$$y_t = Tr_t + Sz_t + C_t + \epsilon_t, \tag{2.1}$$

kde y_t je pozorovaná hodnota v čase t, Tr_t značí trendovou složku, Sz_t sezónní složku, C_t cyklickou složku, ϵ_t reziduální složku a t čas, kde $t = 1, ..n$ a n je počet pozorování. Modelování a vznik všech těchto složek bude podrobně rozepsán dále.

2.1. Analýza trendové složky

Trendová složka nám popisuje dlouhodobou tendenci vývoje. Uvažujme časovou řadu ve tvaru

$$y_t = Tr_t + \epsilon_t. \tag{2.2}$$

Trendovou složku můžeme popsat matematickými křivkami. Typ nejvhodnější matematické křivky usuzujeme na základě grafického zobrazení napozorovaných hodnot, anebo dle výpočtů, které nám ukazují míru vhodnosti zvolené křivky. Míry vhodnosti modelu jsou rozebrány na konci v této kapitole. Nyní níže uvedu vybrané příklady funkcí, které dále využiji při praktické analýze.

2.1.1. Konstantní trend

Nejjednodušším modelem trendové složky je konstantní trend tvaru

$$Tr_t = \beta_0, \qquad t = 1, ...n. \tag{2.3}$$

Naším cílem je odhadnout hodnotu β_0 tak, aby byla minimalizována rezidua. Tuto hodnotu odhadujeme metodou nejmenších čtverců. Minimalizujeme výraz

$$min \sum_{t=1}^{n} (y_t - \beta_0)^2. \tag{2.4}$$

Výraz zderivujeme podle β_0 a položíme rovno nule

$$-2 \sum_{t=1}^{n} (y_t - \beta_0) = 0. \tag{2.5}$$

Jednoduchou úpravou dostaneme odhad parametru $\hat{\beta}_0$

$$\sum y_t - n\hat{\beta}_0 = 0, \tag{2.6}$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_{t=1}^{n} y_t, \tag{2.7}$$

$$\hat{\beta}_0 = \bar{y}. \tag{2.8}$$

Bodový odhad pozorovaných hodnot y_t získáme jako

$$\hat{y_t} = \hat{\beta}_0 = \bar{y}. \tag{2.9}$$

Analogický získáme i bodovou předpověď pro budoucí hodnoty pro čas $T > n$

$$\hat{y_T^P} = \hat{\beta}_0 = \bar{y}. \tag{2.10}$$

Interval spolehlivosti pro budoucí bodové odhady vypočteme následovně

$$\hat{y_T} \pm t_{n-1,(1-\alpha/2)} S \sqrt{1 + \frac{1}{n}}, \tag{2.11}$$

kde S vychází z reziduí a spočítá se podle vzorce

$$S = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \hat{y_t}^2)}. \tag{2.12}$$

Číslo t_{n-1} je $(1-\alpha/2)$ kvantil Studentova rozdělení o $n-1$ stupních volnosti. U intervalového odhadu hledáme interval, ve kterém se bude budoucí skutečná hodnota y_T vyskytovat s předem danou pravděpodobností.

2.1.2. Lineární trend

U lineárního trendu budeme mít trendovou složku tvaru

$$Tr_t = \beta_0 + \beta_1 t, \tag{2.13}$$

kde t je čas a β_0, β_1 jsou neznámé parametry modelu. Parametry opět odhadneme metodou nejmenších čtverců, tj. budeme minimalizovat rezidua výrazu

$$min\sum_{t=1}^{n}[y_t - (\beta_0 + \beta_1 t)]^2. \tag{2.14}$$

Odhadnuté parametry β_0 a β_1 získáme ve tvaru

$$\hat{\beta_0} = \bar{y} - \bar{t}\hat{\beta_1}, \tag{2.15}$$

$$\hat{\beta_1} = \frac{\sum_{t=1}^{n} t y_t - n\bar{y}\bar{t}}{\sum_{t=1}^{n} t^2 - n\bar{t}^2}, \tag{2.16}$$

kde $\bar{t} = \frac{1}{n}\sum_{t=1}^{n} y_t$. Lineární trend použijeme v případě, jsou-li první diference dat přibližně konstantní. Bodovou předpověď pro čas t a budoucí bodovou předpověď pro čas $T > n$ vypočteme následovně

$$\hat{y}_t = \hat{\beta_0} + \hat{\beta_1} t, \quad \hat{y}_T^P = \hat{\beta_0} + \hat{\beta_1} T. \tag{2.17}$$

Intervalový odhad budoucí hodnoty vypočteme pomocí vzorce

$$\hat{y_T}^P \pm t_{n-2,(1-\alpha/2)} S \sqrt{1 + \frac{1}{n} + \frac{(T-\bar{t})^2}{\sum_{t=1}^{n} t^2 - n\bar{t}^2}}, \tag{2.18}$$

kde S spočítáme

$$S = \sqrt{\frac{1}{n-2}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}. \tag{2.19}$$

2.1.3. Kvadratický trend

U kvadratického trendu bude trendová složka ve tvaru

$$Tr_t = \beta_0 + \beta_1 t + \beta_2 t^2, \tag{2.20}$$

kde t je čas a β_0, β_1, β_2 jsou neznámé parametry modelu. Parametry stejně jako v předchozích dvou případech odhadneme metodou nejmenších čtverců, tj. budeme minimalizovat výraz

$$min \sum_{t=1}^{n}[y_t - (\beta_0 + \beta_1 t + \beta_2 t^2)]^2. \tag{2.21}$$

Odhadnuté parametry ve zjednodušeném maticovém zápisu budou následující

$$\hat{\beta} = (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'Y}, \tag{2.22}$$

kde

$$\mathbf{X} = \begin{pmatrix} 1 & t_1 & t_1^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 \end{pmatrix},$$

$\mathbf{X'}$ pak značí matici transponovanou k matici $\mathbf{X}$ a

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Kvadratický trend použijeme v případě, jsou-li druhé diference dat přibližně konstantní. Bodovou předpověď pro čas t a budoucí bodovou předpověď pro čas $T > n$ využiji vzorec

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2, \hat{y}_T^P = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 T^2. \tag{2.23}$$

Intervalovou předpověď vypočítám dle vzorce

$$\hat{y_T^P} \pm t_{n-3,(1-\alpha/2)} S \sqrt{1 + (1, T, T^2)(\mathbf{X'X})^{-\mathbf{1}} \begin{pmatrix} 1 \\ T \\ T^2 \end{pmatrix}}, \tag{2.24}$$

kde

$$S = \sqrt{\frac{1}{n-3} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2}.$$

2.1.4. Logistický trend

Logistický trend má tvar S-křivky vyjádřený následující rovnicí

$$y_t = \frac{k}{1 + \beta_0 \beta_1^t}, \quad t = 1, ...n \quad \beta_1 > 0, k > 0, \tag{2.25}$$

kde k, β_0, β_1 jsou neznámé parametry logistického trendu. Logistický trend je asymptoticky omezen parametrem k a má inflexi (tj. průběh křivky konkávní se mění na konvexní a naopak) v bodě $t = -\frac{ln\beta_0}{ln\beta_1}$. Trend je vykreslen na obrázku 2.1.

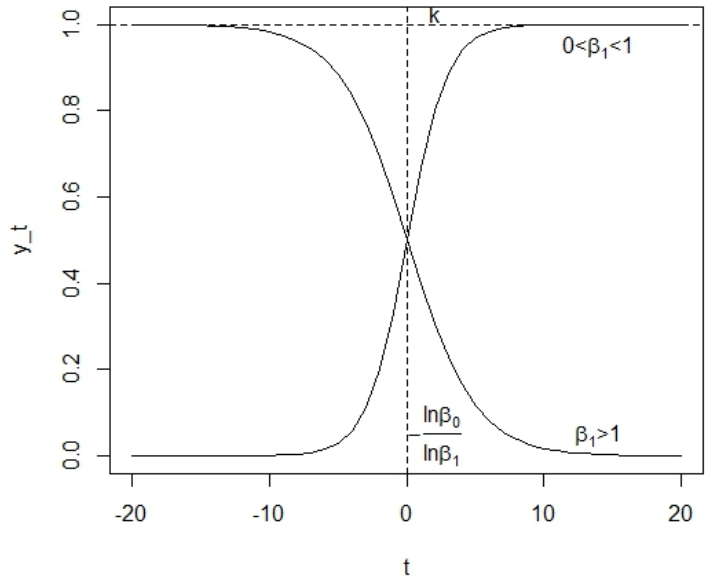
Abychom mohli vypočítat neznámé parametry, převedeme si nejprve metodou inverzní transformace logistický trend do lineární podoby

$$\frac{1}{y_t} = \frac{1 + \beta_0 \beta_1^t}{k} = \frac{1}{k} + \frac{\beta_0}{k} \beta_1^t. \tag{2.26}$$

Nahradíme-li $\frac{1}{y_t} = y_t^*$, $\frac{1}{k} = k^*$ a $\frac{\beta_0}{k} = \beta_0^*$, získáme následující tvar

$$y_t^* = k^* + \beta_0^* \beta_1^t. \tag{2.27}$$

Z výše upraveného tvaru odhadneme neznámé parametry k^*, β_0^*, β_1 pomocí metody částečných součtů. U této metody si nejprve rozdělíme celkový počet pozorování n na tři stejně velké třetiny o délce m a sečteme pozorování v jednotlivých



Obrázek 2.1: Tvar logistického trendu

třetinách. Pro celkový počet pozorování platí, že $n = 3m$. Obdržíme následující tři součty

$$V_1 = \sum_{t=1}^{m} y_t, \quad V_2 = \sum_{t=m+1}^{2m} y_t, \quad V_3 = \sum_{t=2m+1}^{3m} y_t. \tag{2.28}$$

Řešením následujících soustav obdržíme odhady parametrů k^*, β_0^*, β_1

$$\hat{\beta}_1 = \sqrt[m]{\frac{V_3 - V_2}{V_2 - V_1}}, \tag{2.29}$$

$$\hat{\beta}_0^* = (V_2 - V_1)\frac{\hat{\beta}_1 - 1}{\hat{\beta}_1(\hat{\beta}_1^m - 1)^2}, \tag{2.30}$$

$$\hat{k^*} = \frac{V_1 - \frac{\hat{\beta}_0^* \hat{\beta}_1 (\hat{\beta}_1 - 1)}{(\hat{\beta}_1 - 1)}}{m}. \tag{2.31}$$

Odhady původních parametrů $\hat{\beta}_0$ a $\hat{k}$ získáme následovně

$$\hat{k} = \frac{1}{\hat{k^*}}, \quad \hat{\beta}_0 = \hat{\beta}_0^* \hat{k}. \tag{2.32}$$

Odhadované hodnoty jednotlivých pozorováních $\hat{y}_t$ pro čas $t = 1, \ldots, n$ získáme pomocí následujícího vzorce

$$\hat{y}_t = \frac{\hat{k}}{1 + \hat{\beta}_0 \hat{\beta}_1^t}. \tag{2.33}$$

Logistický trend použijeme v případě, mají-li data tvar S-křivky a křivka prvních diferencí je tvarem podobná křivce hustoty normálního rozdělení. Teorii logistického trendu jsem čerpala z literatury [8].

2.2. Analýza periodické složky

Periodická složka obsahuje sezónní a cyklickou složku. Do sezónní složky zahrnujeme nepravidelnosti v rámci roku, tzn. sezónní složka má periodu 1 rok anebo menší než 1 rok. Naopak do cyklické složky náleží kolísání kolem trendu, které nelze popsat pouhou sezónností. Délka cyklu může být nepravidelná a v ekonomických pozorováních je spojena s hospodářským cyklem. Ve své práci se zaměřím na modelování sezónní složky. Pro modelování sezónnosti budu využívat model s konstantní sezónností

$$y_t = Tr_t + Sz_t + \epsilon_t, \tag{2.34}$$

kde sezónní složku Sz_t budu uvažovat ve tvaru

$$Sz_t = \alpha_2 x_{2t} + \ldots + \alpha_v x_{vt}, \tag{2.35}$$

kde v značí počet sezón v roce a x_{kt} je umělá proměnná pro k-tou sezónu, kde $k = 2, \ldots, v$. Jestliže $x_{kt} = 1$, pak čas t odpovídá k-tému období v roce. Jinak $x_{kt} = 0$. $\alpha_2, \ldots, \alpha_v$ jsou neznámé parametry sezónní složky, které odhadneme metodou nejmenších čtverců. Z odhadnutých parametrů $\hat{\alpha}_2, \ldots, \hat{\alpha}_v$ se již snadno určí sezónní efekty sz_j, kde $j = 1, \ldots, v$ a j značí danou sezónu. Sezónní efekt v j-té sezóně vypočteme

$$\hat{sz}_j = \alpha_j - \bar{\alpha}, \tag{2.36}$$

kde průměr sezónních parametrů $\bar{\alpha}$ vypočteme

$$\bar{\alpha} = \frac{\hat{\alpha}_1 + \hat{\alpha}_2 + \ldots + \hat{\alpha}_v}{v}. \tag{2.37}$$

Odhadované hodnoty $\hat{y}_t$ v případě konstantního trendu dostaneme dle vzorce

$$\hat{y_t} = \hat{\beta}_0 + \bar{\alpha} + sz_j \tag{2.38}$$

a pro případ lineárního trendu

$$\hat{y_t} = \hat{\beta}_0 + \hat{\beta}_1 t + \bar{\alpha} + sz_j. \tag{2.39}$$

Předpovědi pro čas $T > n$ v případě lineárního trendu získáme následujícím výpočtem

$$\hat{y_T} = \hat{\beta}_0 + \hat{\beta}_1 T + \bar{\alpha} + sz_j. \tag{2.40}$$

V případě konstantního trendu budou předpovědi pro čas $T > n$ rovny $\hat{y_t}$.

Jak bylo zmíněno výše neznámé parametry trendové a sezónní složky odhadneme MNČ. Konkrétně budeme-li uvažovat model s konstantním trendem a s měsíční sezónností, budeme minimalizovat rezidua modelu tvaru

$$y_t = \beta_0 + \alpha_2 x_{2t} + \ldots + \alpha_{12} x_{12t} + \epsilon_t, \tag{2.41}$$

kde $x_{2t}, \ldots, x_{12t}$ jsou umělé proměnné pro 12 sezón v roce. Maticově bychom mohli psát

$$\mathbf{y} = \mathbf{X}\gamma + \varepsilon, \tag{2.42}$$

kde

$$\mathbf{X} = \left(\begin{array}{ccccc}
1 & 0 & 0 & 0 & \ldots\ 0 \\
1 & 1 & 0 & 0 & \ldots\ 0 \\
1 & 0 & 1 & 0 & \ldots\ 0 \\
\vdots & \vdots & \vdots & \ddots & \\
1 & 0 & 0 & 0 & \ldots\ 1 \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots\ \ \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots\ \ \vdots \\
\hline
1 & 0 & 0 & 0 & \ldots\ 0 \\
1 & 1 & 0 & 0 & \ldots\ 0 \\
1 & 0 & 1 & 0 & \ldots\ 0 \\
\vdots & \vdots & \vdots & \ddots & \\
1 & 0 & 0 & 0 & \ldots\ 1
\end{array}\right),$$

γ značí vektor parametrů

$$\gamma = \begin{pmatrix} \beta_0 \\ \alpha_2 \\ \vdots \\ \alpha_{12} \end{pmatrix}.$$

Neznámé parametry $\beta_0, \alpha_2, \ldots, \alpha_{12}$ odhadneme z následující rovnice

$$\hat{\gamma} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}, \tag{2.43}$$

kde $\mathbf{X}'$ je transponovaná matice $\mathbf{X}$. Do modelu nezahrnujeme 1 sezónu z důvodu kolinearity. Kolinearita vzniká právě tehdy, jsou-li sloupce matice $\mathbf{X}$ závislé nebo pokud je matice $\mathbf{X}$ singulární, tj. $\det(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = 0$. Matice $\mathbf{X}'\mathbf{X}$ by poté nešla invertovat. Proto jsme nuceni obětovat první sezónu. Odhad parametru pro první sezónu volíme $\hat{\alpha}_1 := 0$. Výpočet sezónních efektů dopočítáme následovně. Vypočteme si průměr sezónních parametrů

$$\bar{\alpha} = \frac{\hat{\alpha}_1 + \hat{\alpha}_2 + \ldots + \hat{\alpha}_{12}}{12}. \tag{2.44}$$

Sezónní efekt v j-té sezóně vypočítáme

$$\hat{sz}_j = \hat{\alpha}_j - \bar{\alpha}, \tag{2.45}$$

kde $j = 1, 2, \ldots 12$. Přitom součet sezónních efektů nám musí dát 0. Touto úpravou dostaneme nový absolutní člen $\beta_0 + \bar{\alpha}$. Odhadované hodnoty $\hat{y}_t$ vypočteme pomocí následujícího vzorce

$$\hat{y}_t = \hat{\beta}_0 + \bar{\alpha} + (\hat{\alpha}_1 - \bar{\alpha})x_{1t} + (\hat{\alpha}_2 - \bar{\alpha})x_{2t} + \ldots + (\hat{\alpha}_v - \bar{\alpha})x_{12t}, \tag{2.46}$$

nebo také můžeme napsat

$$\hat{y}_t = \hat{\beta}_0 + \bar{\alpha} + \hat{sz}_j + \epsilon_t, \tag{2.47}$$

kde $j = 1, \ldots, 12$. Pro budoucí bodovou předpověď pro čas $T > n$ bude platit

$$\hat{y}_T^P = \hat{y}_t = \hat{\beta}_0 + \bar{\alpha} + \hat{sz}_j + \epsilon_t. \tag{2.48}$$

Obdobně budeme-li uvažovat model s lineárním trendem a s 12 sezónami v roce, budeme minimalizovat rezidua modelu tvaru

$$y_t = \beta_0 + \beta_1 t + \alpha_2 x_{2t} + \ldots + \alpha_{12} x_{12t} + \epsilon_t. \tag{2.49}$$

Maticově bychom mohli psát

$$\mathbf{y} = \mathbf{X}\gamma + \varepsilon, \tag{2.50}$$

kde

$$\mathbf{X} = \left(\begin{array}{ccccccc} 1 & 1 & 0 & 0 & 0 & \ldots & 0 \\ 1 & 2 & 1 & 0 & 0 & \ldots & 0 \\ 1 & 3 & 0 & 1 & 0 & \ldots & 0 \\ 1 & 4 & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & & \\ 1 & 12 & 0 & 0 & 0 & \ldots & 1 \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \hline 1 & \vdots & 0 & 0 & 0 & \ldots & 0 \\ 1 & \vdots & 1 & 0 & 0 & \ldots & 0 \\ 1 & \vdots & 0 & 1 & 0 & \ldots & 0 \\ 1 & \vdots & 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & & \\ 1 & n & 0 & 0 & 0 & \ldots & 1 \end{array}\right),$$

γ značí vektor parametrů

$$\gamma = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \alpha_2 \\ \vdots \\ \alpha_{12} \end{pmatrix}.$$

Neznámé parametry γ odhadneme opět MNČ

$$\hat{\gamma} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.51}$$

Stejně jako v předchozím případě spočítáme efekty sezón. Odhadnuté hodnoty $\hat{y}_t$ vypočteme podle vzorce

$$\hat{y}_t = \hat{\beta}_0 + \bar{\alpha} + \hat{\beta}_1 t + (\hat{\alpha}_1 - \bar{\alpha})x_{2t} + (\hat{\alpha}_2 - \bar{\alpha})x_{2t} + \ldots + (\hat{\alpha}_{12} - \bar{\alpha})x_{12t}, \tag{2.52}$$

nebo

$$\hat{y}_t = \hat{\beta}_0 + \bar{\alpha} + \hat{\beta}_1 t + \hat{sz}_t. \tag{2.53}$$

Bodová předpověď se vypočte analogicky dle bodového odhadu trendové složky dle vzorce 2.23, tj.

$$\hat{y}_t^P = \hat{\beta}_0 + \hat{\beta}_1 t_t + \bar{\alpha} + s\hat{z}_j, \tag{2.54}$$

a budoucí bodová předpověď

$$\hat{y}_T^P = \hat{\beta}_0 + \hat{\beta}_1 T + \bar{\alpha} + s\hat{z}_j, \tag{2.55}$$

kde $T > n$ je budoucí čas.

2.3. Reziduální složka

Reziduální složka je náhodná složka, která zůstane v modelu po odstranění trendové a periodické složky. Předpokládá se, že náhodné veličiny ϵ_t jsou mezi sebou nekorelované a mají stejný rozptyl. Tvoří tzv. bílý šum (WN). Nejznámějšími testy pro otestování předpokladů náhodné veličiny je znaménkový test, test založený na bodech zvratu, test založený na Spearmanově korelačním koeficientu a test založený na Kendallově koeficientu.

2.4. Hodnocení modelu

Jedním z ukazatelů pro posouzení vhodnosti zvoleného modelu (podle zdroje citehron) je reziduální součet čtverců (RSČ). RSČ je založen na porovnávání skutečných hodnot s odhadovanými. RSČ vypočteme následujícím vzorcem

$$\text{RSČ} = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2, \tag{2.56}$$

kde t značí čas, n je počet pozorování, y_t je skutečně pozorovaná hodnota v čase t a $\hat{y}_t$ je odhadnutá hodnota v čase t podle námi zvoleného modelu. RSČ nám udává, jak velké čtvercové chyby jsme se dopustili při odhadování hodnot.

Dalším hojně používaným ukazatel pro posouzení vhodnosti modelu je koeficient determinace (R^2), který vypočteme jako jedna mínus podíl RSČ a celkové variability v pozorováních (S_t). Tedy

$$R^2 = 1 - \frac{\text{RSČ}}{S_t}, \tag{2.57}$$

kde celkovou variabilitu modelu vypočteme

$$S_t = \sum_{t=1}^{n} (y_t - \bar{y})^2, \tag{2.58}$$

kde $\bar{y}$ značí průměrnou pozorovanou hodnotu za čas t.

Koeficient determinace nabývá hodnot z intervalu [0,1]. Čím je hodnota větší, tím je model vhodnější. RSČ ani R^2 nám ale nezohledňuje počet parametrů v modelu. Proto byl navržen upravený koeficient determinace

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p}, \tag{2.59}$$

kde n je celkový počet pozorování a p je počet parametrů. Upravený koeficient determinace se používá pro menší počet pozorování. Pomocí R^2 mohu testovat hypotézu, zda jsou všechny parametry nevýznamné, oproti alternativě, že alespoň jeden parametr je nevýznamný, tj.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad H_A : \text{alespoň jeden parametr je nenulový.}$$

Testová statistika

$$F = \frac{R^2}{(1 - R^2)} \frac{n-p-1}{p} \tag{2.60}$$

má Fisherovo rozdělení o p a $n - p - 1$ stupních volnosti. Nulovou hypotézu zamítáme na hladině α, je-li $F > F_{p,n-p-1,1-\alpha}$.

Významnost jednotlivých parametrů v modelu můžeme testovat také pomocí t- testu, kde

$$H_0 : \beta_i = 0 \quad \text{oproti alternativě} \quad H_a : \beta_i \neq 0.$$

Testová statistika je tvaru

$$T = \frac{\hat{\beta}_i}{S\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}}, \tag{2.61}$$

kde $(\mathbf{X}'\mathbf{X})^{-1}_{jj}$ je prvek na j-tém řádku v j-tém sloupci matice $(\mathbf{X}'\mathbf{X})^{-1}$. Například pro testování parametru β_1 budeme brát prvek ve 2. řádku v 2. sloupci matice $(\mathbf{X}'\mathbf{X})^{-1}$, pro testování parametru kvadratického členu β_2 budeme brát prvek ve 3. řádku v 3. sloupci matice $(\mathbf{X}'\mathbf{X})^{-1}$ apod. Za platnosti H_0 bude testová statistika T t_{n-p}.

Odhadnuté předpovědi budu posuzovat podle střední čtvercové chyby a střední absolutní chyby. Střední čtvercová chyba předpovědi ($MSEP$) se vypočítá pomocí následujícího vzorce

$$MSEP = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y_t}^P)^2}{n}}. \tag{2.62}$$

Střední absolutní chyby ($MAEP$) vypočteme pomocí vzorce

$$MAEP = \frac{\sum_{t=1}^{n}|y_t - \hat{y_t}^P|}{n}. \tag{2.63}$$

Kapitola 3

Modelování pomocí přístupu Boxe a Jenkinse

Boxův a Jenkinsův přístup modeluje časovou řadu pomocí reziduální složky, která je tvořena vzájemně závislými veličinami. Důležitým předpokladem pro tyto modely je stacionarita daného procesu. Požadujeme tedy, aby se daný náhodný proces v čase ustálil. Stacionaritu rozlišujeme striktní a slabou. Striktní stacionarita znamená, že pravděpodobnostní chování procesu je invariantní vůči posunům v čase. To je v praxi ale obtížné ověřit. V praxi si naštěstí vystačíme se slabou stacionaritou. Slabě stacionární proces je taková časová řada, která má konstantní střední hodnotu, konstantní rozptyl a její kovariance mezi dvěma libovolně zvolenými pozorováními zavisí jen na časovém rozestupu, a ne na jejich skutečném časovém umístění v řadě. V praxi se ovšem můžeme poměrně často setkat s nestacionárními časovými řadami. V těchto případech můžeme stacionarity obvykle dosáhnout diferencováním daného procesu anebo pomocí různých transformací. Vykreslením dat můžeme subjektivně odhalit některé typy porušení stacionarity.

Analýzu časových řad podle Boxe a Jenkinse můžeme rozdělit do tří fází. V první fází identifikujeme model, poté odhadneme parametry a nakonec je zapotřebí posoudit vhodnost modelu. Nejprve si ale zavedeme základní pojmy a vlastnosti, které jsou nezbytné pro analýzu, zvlášť pro identifikaci modelu.

3.1. Základní pojmy a vlastnosti

V této podkapitole se zaměřím především na základní pojmy a charakteristiky používané v rámci Boxovy a Jenkinsovy metody, které jsou nezbytné k praktickým výpočtům.

Budeme-li mít stochastický proces stacionární, pak můžeme střední hodnotu procesu odhadnout jako výběrový průměr, tj.

$$\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} y_t = \bar{y}, \tag{3.1}$$

a rozptyl daného procesu odhadnout jako výběrový rozptyl

$$C_0 = S^2 = \frac{\sum_{t=1}^{n}(y_t - \bar{y})^2}{n}. \tag{3.2}$$

Další důležitou charakteristikou stacionárního procesu, která se v modelování podle Boxe a Jenkinse využívá, je odhad autokovariancí procesu v čase t a $t-k$

$$C_k = \frac{1}{n}\sum_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y}), \tag{3.3}$$

kde $k = 1, ..., n-1$. Pomocí odhadnutého rozptylu daného procesu a autokovariancí odhadneme autokorelační funkci (ACF)

$$r_k = \frac{C_k}{C_0} = \frac{\sum_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}. \tag{3.4}$$

Grafem těchto korelací je korelogram. Existuje-li takový bod, od kterého jsou všechny korelace nulové, pak můžeme tvrdit, že vymizí závislost mezi pozorováními. Tento bod nazýváme identifikačním bodem a testujeme ho pomocí korelací následující testovou statistikou

$$|r_k| \sim \sqrt{\frac{1}{n}(1 + 2\sum_{j=1}^{k_0} r_j^2)}, \tag{3.5}$$

která testuje $H_0 : r_k = 0$ vs. $H_A : r_k \neq 0$ pro $k > k_0$. Dalším důležitým ukazatelem jsou odhady parciálních autokorelací (PACF), které se vypočítají rekurentně pomocí vzorce

$$r_{11} = r_1, \tag{3.6}$$

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_j}, \tag{3.7}$$

$$r_{kj} = r_{k-1,j} - r_{kk} r_{k-1,k-j}, \tag{3.8}$$

kde $k > 1$ a $j = 1, 2, \ldots, k-1$. Nulovost odhadovaných hodnot PACF r_{kk} testujeme pomocí statistiky

$$|r_{kk}| \sim \sqrt{\frac{1}{n}}. \tag{3.9}$$

Aby výše uvedené odhadované charakteristiky byly spolehlivé, požaduje se dle zdroje [1], aby počet pozorování $n > 50$ a $k < n/4$. Tyto uvedené vlastnosti jsou nezbytné pro identifikaci modelu. Dle ACF, PACF stanovujeme, zda se jedná o proces MA, AR, ARMA, dle identifikačního bodu určujeme řád procesu.

3.2. Identifikace modelu

Boxův a Jenkinsův přístup pracuje se speciálními případy lineárního stacionárního procesu, jako je proces klouzavých součtů MA, autoregresní proces AR a smíšený proces ARMA.

Autoregresní proces AR

Autoregresivní proces AR řádu p je model tvaru

$$y_t = \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t, \tag{3.10}$$

kde $\varphi_1, \ldots, \varphi_p$ jsou neznámé parametry modelu. V literatuře se často můžeme setkat se zápisem pomocí operátoru zpětného posunutí, který se definuje jako

$$B y_t = y_{t-1}, \tag{3.11}$$

a lze jej aplikovat několinásobně pro $j = 1, ..., p$

$$B^j y_t = y_{t-j}. \tag{3.12}$$

Pomocí operátoru zpětného posunutí můžeme model zapsat ve zkrácené formě

$$(1 - (\varphi_1 B + \ldots + \varphi_p B^p))y_t = \varepsilon_t, \tag{3.13}$$

neboli

$$\varphi(B)y_t = \varepsilon_t, \tag{3.14}$$

kde $\varphi(B) = 1 - (\varphi_1 B + \ldots + \varphi_p B^p)$ je tzv. autoregresní operátor. Model bude stacionární, pokud bude platit, že kořeny polynomu $\varphi(B)$ budou ležet mimo jednotkovou kružnici v rovině komplexních čísel. Autoregresní proces je vždy invertibilní. V praxi se nejčastěji setkáváme s autoregresním procesem AR(1) nebo AR(2).

Autoregresní proces AR(1) je model tvaru

$$y_t = \varphi_1 y_{t-1} + \varepsilon_t \tag{3.15}$$

s podmínkou stacionarity

$$|\varphi_1| < 1 \tag{3.16}$$

Příslušná ACF procesu AR(1) má tvar

$$r_k = \varphi_1^k \quad \text{pro } k \geq 0 \tag{3.17}$$

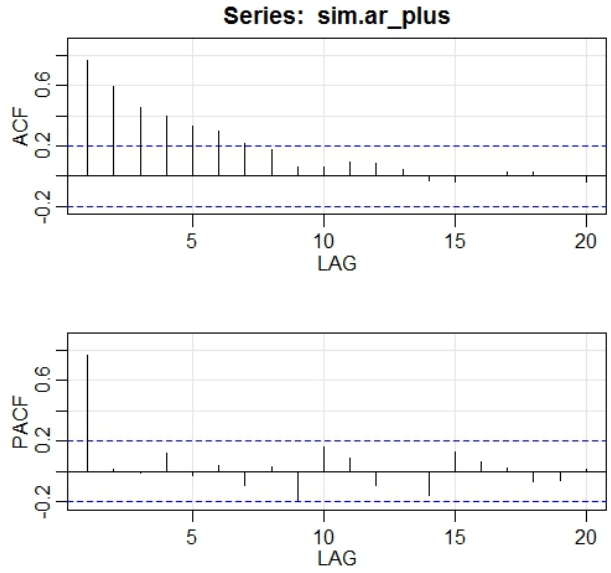
a PACF je tvaru

$$r_{11} = \varphi_1, r_{kk} = 0 \quad \text{pro } k > 1, \tag{3.18}$$

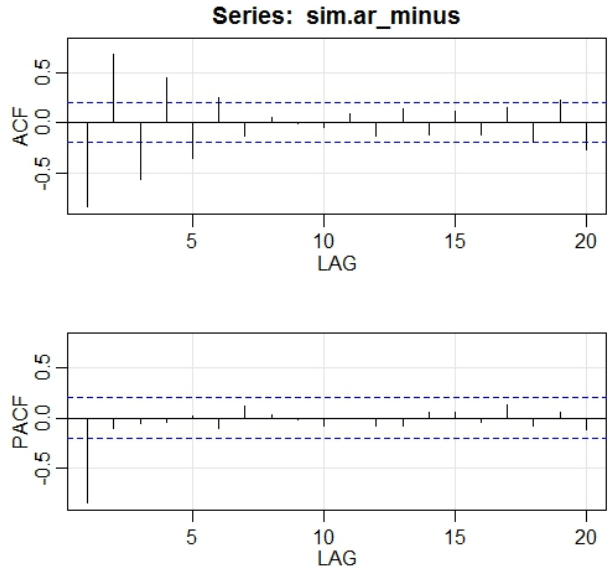
identifikačním bodem je $k_0 = 1$. Na Obrázkách 3.1 a 3.2 jsou ukázany odhady ACF a PACF simulovaného procesu s kladným parametrem φ_1 $y_t = 0,8y_{t-1} + \varepsilon_t$ a se záporným parametrem $y_t = -0,8y_{t-1} + \varepsilon_t$.

Autoregresní proces AR(2) je model tvaru

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varepsilon_t \tag{3.19}$$



Obrázek 3.1: ACF a PACF procesu AR(1) s parametrem 0,8

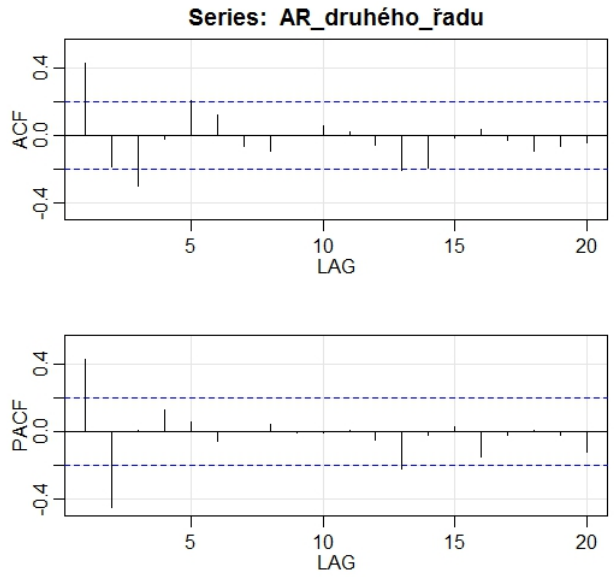


Obrázek 3.2: ACF a PACF procesu AR(1) s parametrem -0,8

s podmínkami stacionarity

$$\varphi_2 + \varphi_1 < 1, \varphi_2 - \varphi_1 < 1, -1 < \varphi_2 < 1 \tag{3.20}$$

PACF má identifikační bod $k_0 = 2$. Na Obrázku 3.3 je ukázana odhadovaná ACF a PACF simulovaného procesu φ_1 $y_t = 0,8y_{t-1} - 0,6y_{t-2}\varepsilon_t$.



Obrázek 3.3: ACF a PACF procesu AR(2) s parametrem 0,8 a -0,6

Proces klouzavých součtů MA

Proces klouzavých součtů řádu q (MA(q)) má tvar

$$y_t = \varepsilon_t + \phi_1\varepsilon_{t-1} + \ldots + \phi_q\varepsilon_{t-q}, \tag{3.21}$$

kde ε_t je bílý šum a $\phi_j, j = 1, \ldots, q$, jsou neznámé parametry. Model můžeme opět zapsat ve zkráceném tvaru pomocí operátoru zpětného posunutí. Zkrácený tvar bude

$$y_t = (1 + \phi_1 B + \ldots + \phi_q B^q)\varepsilon_t, \tag{3.22}$$

$$y_t = \phi(B)\varepsilon_t, \tag{3.23}$$

kde

$$\phi(B) = 1 + \sum_{j=1}^{q} \phi_j B^j \tag{3.24}$$

je operátor klouzavých součtů. Oproti autoregresnímu procesu je proces klouzavých součtů vždy stacionární, ale ne vždy invertibilní.

V praxi se nejčastěji můžeme setkat s procesem $MA(1)$. Tento proces je tvaru

$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} \tag{3.25}$$

s podmínkou invertibility

$$|\phi_1| < 1. \tag{3.26}$$

ACF procesu $MA(1)$ má tvar

$$r_1 = \frac{\phi_1}{1+\phi_1^2}, r_k = 0 \quad \text{pro } k > 1. \tag{3.27}$$

Identifikačním bodem této funkce je $k_0 = 1$. Dle podmínky invertibility (3.26) pro libovolný proces $MA(2)$ musí platit

$$|r_1| < \frac{1}{2}. \tag{3.28}$$

PACF je funkce omezená geometricky klesající posloupností

$$|r_{kk}| < |\phi_1|^k. \tag{3.29}$$

Na Obrázcích 3.4 a 3.5 jsou ukázany odhadované ACF a PACF simulovaného procesu s kladným parametrem ϕ_1 $y_t = \varepsilon_t + 0,8\varepsilon_{t-1}$ a se záporným parametrem $y_t = \varepsilon_t - 0,8\varepsilon_{t-1}$.

Druhým často využívaným procesem je $MA(2)$ tvaru

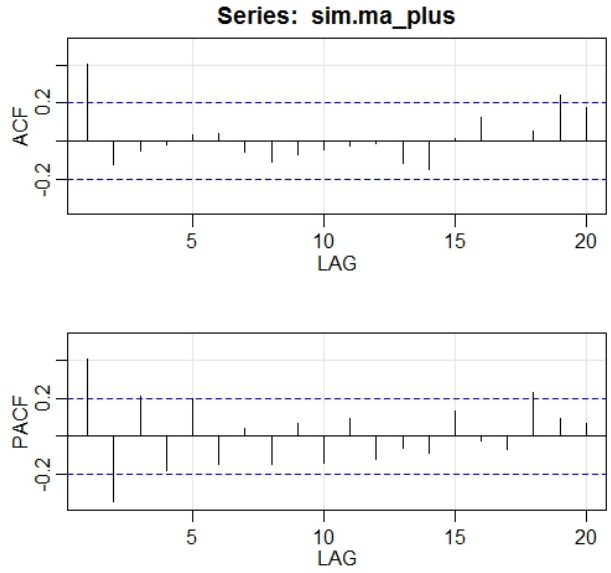
$$y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} \tag{3.30}$$

s podmínky invertibility

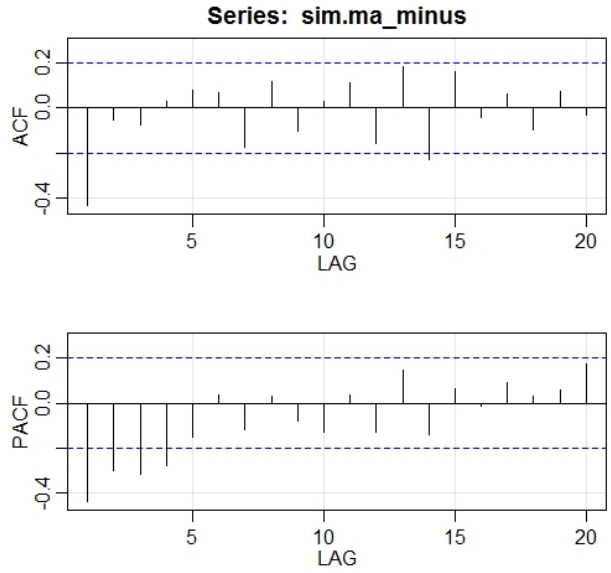
$$\phi_2 + \phi_1 > 1, \phi_2 - \phi_1 > 1, -1 < \phi_2 < 1 \tag{3.31}$$

ACF procesu $MA(2)$ má tvar

$$r_1 = \frac{\phi_1(1+\phi_2)}{1+\phi_1^2+\phi_2^2}, r_2 = \frac{\phi_2}{1+\phi_1^2+\phi_2^2}, r_k = 0 \quad \text{pro } k > 2. \tag{3.32}$$



Obrázek 3.4: ACF a PACF procesu MA(1) s parametrem 0,8

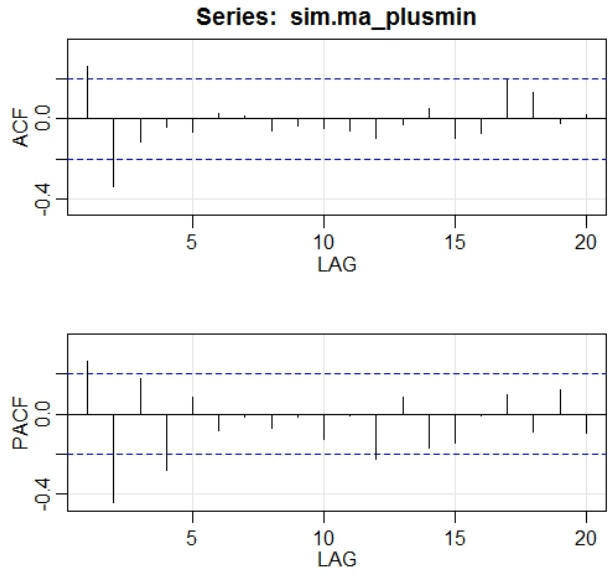


Obrázek 3.5: ACF a PACF procesu MA(1) s parametrem -0,8

Identifikačním bodem této funkce je $k_0 = 2$. Pro libovolný proces $MA(2)$ platí

$$|r_1| \leq \frac{1}{\sqrt{2}}, |r_2| \leq \frac{1}{2} \tag{3.33}$$

PACF je funkce omezená geometricky klesající posloupností a nebo sinusoidou s geometricky klesající amplitudou. Na Obrázku 3.6 je vyobrazena odhadovaná ACF a PACF simulovaného procesu $y_t = \varepsilon_t + 0,6\varepsilon_{t-1} - 0,4\varepsilon_{t-2}$.



Obrázek 3.6: ACF a PACF procesu MA(2) s parametrem 0,6 a -0,4

Smíšený proces ARMA

Smíšený proces ARMA(p, q) je kombinací autoregresního procesu řádu p a procesu klouzavých součtů řádu q. Tvar tohoto modelu je

$$y_t = \varphi_1 y_{t-1} + \ldots + \varphi_p y_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \ldots + \phi_q \varepsilon_{t-q}. \tag{3.34}$$

Ve zkrácené podobě pomocí operátoru zpětného posunutí B můžeme zapsat model ARMA(p, q)

$$\varphi(B) y_t = \phi(B) \varepsilon_t. \tag{3.35}$$

Nejčastěji se můžeme setkat s modelem $ARMA(1, 1)$ tvaru

$$y_t = \varphi_1 y_{t-1} + \varepsilon_t + \phi_1 \varepsilon_{t-1} \tag{3.36}$$

s podmínkou stacionarity $|\varphi_1| < 1$ a podmínkou invertibility $|\phi_1| < 1$. ACF procesu $ARMA(1,1)$ má tvar

$$r_1 = \frac{(1+\varphi_1\phi_1)(\varphi_1+\phi_1)}{1+\phi_1^2+2\varphi_1\phi_1}, r_k = \varphi_1 r_{k-1}, k > 1. \tag{3.37}$$

PACF procesu $ARMA(1,1)$ je stejně jako u procesu $MA(1)$ omezena geometricky klesající posloupností. Identifikační bod pro ACF a pro PACF neexistuje.

Jak již bylo zmíněno identifikaci výše popsaných modelů provádíme na základě tvaru autokorelační a parciální autokorelační funkce. V tabulce 3.1 je dodatečně shrnut tvar ACF a PACF pro model AR(p), MA(q) a ARMA(p,q).

Model	ACF	PACF
$AR(p)$	exponenciální nebo exponenciálně sinusoidní pokles	$r_{kk} = 0$ pro $k > p$
$MA(q)$	$r_k = 0$ pro $k > q$	omezená exponenciálním a/nebo exponenciálně sinusoidním poklesem
$ARMA(p,q)$	od zpoždění $(q-p)$ pro $q > p$ exponenciální nebo exponenciálně sinusoidní pokles	od zpoždění $(p-q)$ pro $p > q$ omezená exponenciálním nebo exponenciálně sinusoidním poklesem

Tabulka 3.1: Tvar ACF a PACF pro dané modely (převzato z [1])

V praxi se snažíme najít vždy co nejjednodušší model, proto se nejčastěji setkáváme z výše podrobněji rozepsanými modely AR(1), AR(2), MA(1) a MA(2), popřípadě se smíšeným procesem ARMA(1,1).

V následujících dvou odstavcích jsou uvedeny další dva zobecněné případy modelů a jejich využití.

První často používanou třídou modelů je model ARIMA(p,d,q) tvaru

$$\varphi_p(B)(1-B)^d y_t = \phi_q(B)\varepsilon_t, \tag{3.38}$$

kde d značí řád diference. Tento model se hojně používá v případě porušení předpokladu stacionarity procesu. Diferencováním můžeme převést nestacionární

proces na stacionární. ARIMA model je tedy integrovaný systém. Nejčastěji používáme diferencování 1. nebo 2. řádu. Řád diferencování volíme na základě porovnání odhadovaných rozptylů. Pokud nám rozptyl oproti původním pozorování klesne, bylo diferencování vhodné. Nejmenší odhadovaný rozptyl nám určí řád diferencování.

Druhou zobecněnou třídou modelů jsou modely SARMA a SARIMA. Tyto modely využijeme pro sezónní časové řady. Sezónní časové řady mají periodu o délce s, po které se cyklus opakuje. Například pro měsíční časové řady $s = 12$, pro čtvrtletní $s = 4$. Sezónní časovou řadu můžeme posoudit dle grafu, anebo také vypozorovat pomocí korelogramu. Model SARMA$(p, q)(P, Q)$ má tvar

$$\Phi_P(B^s)\varphi(B)y_t = \Phi_Q(B^s)\phi(B)\varepsilon_t, \tag{3.39}$$

kde p je řád procesu AR, q řád procesu MA, P řád sezónního procesu AR, Q je řád sezónního procesu MA Jsou-li data stacionární, neznamená to ještě, že budou stacionární i v sezóně. Nestacionaritu v sezóně odstraníme sezónním diferencováním. V tomto případě použijeme model SARIMA$(p, d, q)(P, D, Q)$ tvaru

$$\Phi_P(B^s)\varphi_p(B)(1-B)^d(1-B^s)^D y_t = \Phi_Q(B^s)\phi_q(B)\varepsilon_t, \tag{3.40}$$

kde D značí řád sezónní diference.

3.3. Ověření modelu

Kontrolu zvoleného modelu provedeme pomocí posouzení reziduí a významnosti paramatrů v modelu. Cílem je dosáhnout, aby rezidua v modelu ε_t tvořila bílý šum a zároveň abychom měli co nejjednodušší model. Proto je vhodné se zabývat i testováním významnosti parametrů, aby se zjistila nenadbytečnost parametrů.

Významnost parametrů otestujeme dle směrodatné odchylky odhadnutých parametrů. Směrodatnou odchylku parametrů v modelu AR(1) vypočítáme

$$\hat{\sigma}(\varphi_1) \doteq \sqrt{\frac{1-\hat{\varphi_1^2}}{n}}. \tag{3.41}$$

Pro model MA(1) bude výpočet analogický, tj.

$$\hat{\sigma}(\phi_1) \doteq \sqrt{\frac{1-\hat{\phi_1^2}}{n}}. \tag{3.42}$$

Sestavením t-testu

$$t_{\varphi_1} = \frac{\hat{\varphi_1}}{\hat{\sigma}(\varphi_1)}, t_{\phi_1} = \frac{\hat{\phi_1}}{\hat{\sigma}(\phi_1)}, \tag{3.43}$$

budeme testovat $H_0 : \varphi_1 = 0$ vs. $H_A : \varphi_1 \neq 0$ a analogicky pro ϕ_1 $H_0 : \phi_1 = 0$ vs. $H_A : \phi_1 \neq 0$ významnost a nenadbytečnost parametrů.

K prověření náhodnosti reziduí v modelu využijeme ACF reziduí (dle zdroje [2]), Ljungův-Boxův test a Q-Q plot. Q-Q plot nám graficky zobrazuje, zda rezidua jsou normálně rozdělená. U ACF reziduí požadujeme, aby žádná korelace nebyla významná. Chceme, aby rezidua byla nezávislá. Autokorelaci reziduí testujeme pomocí výběrové autokorelační funkce

$$r_k(\hat{\varepsilon}) = \frac{\sum_{t=k+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon_{t-k}}}{\sum_{t=1}^{n} \hat{\varepsilon_t^2}}. \tag{3.44}$$

V případě neautokorelovanosti reziduí budou hodnoty výběrové autokorelační funkce ležet uvnitř intervalu $|r_k(\hat{\varepsilon})| < \frac{2}{\sqrt{n}}$.

Ljungův-Boxův test dle zdroje [6] i [2] testuje autokorelace reziduí (ρ_k) modelu pro zpoždění k, kde $k = 1 \ldots K$, tj.

$$H_0 : \rho_1 = \rho_2 = \ldots = \rho_K = 0 \quad H_a : \text{alespoň jedno } \rho \neq 0$$

Testová statistika je

$$Q = n(n+2) \sum_{k=1}^{K} \frac{\hat{r}_k^2(\hat{\varepsilon})}{n-k}, \tag{3.45}$$

kde K je počet zpoždění pro které budeme testovat. Nulovou hypotézu zamítáme, jestliže $Q > \chi^2_{K,1-\alpha}$, kde $\chi^2_{K,1-\alpha}$ je $(1-\alpha)$-kvantil χ^2 rozdělení s K stupni volnosti.

Kapitola 4

Zpracování časových řad v softwaru R

Modelování časových řad dat je časově a početně náročné. Proto je vhodné ke zpracovávání dat zapojit i výpočetní techniku. Pro modely dekompozičního přístupu si vystačíme s Excelem, modely Boxe a Jenkinse vyžadují již více sofistikovanější software. Na trhu existuje celá řada statistických softwarů. K jednomu z nejpopulárnějších a nejdostupnějších softwarů patří bezesporu software R.

Základní knihovnou v softwaru R pro modelování časových řad pomocí přístupu Boxe a Jenkinse je knihovna `astsa` a `tseries`. Ke knihovně `astsa` byla vydána publikace [3], z které jsem čerpala pro tuto kapitolu. Autokorelační a parciální autokorelační funkce vypočtena dle vzorce 3.4 a 3.8 se pomocí této knihovny spočte jednoduchým příkazem

```
> acf(x),
> pacf(x),
```

kde `x` značí vektor jednotlivých časově uspořádaných pozorování. Výstupem příkazů je přímo grafické vyobrazení korelací, tzv. korelogram. PACF počítá a vyobrazuje hodnoty od první parciální korelace, zatímco ACF od nulté korelace. Naprogramováním si vlastního výpočtu hodnot autokorelací a poté jejim vykreslením do grafu, se můžeme vyvarovat předchozímu korelogramu. Následující navrženým

kódem se můžeme zavádějící první hodnotě v korelogramu vyvarovat

```
> c0=sum((x-mean(x))*(x-mean(x)))/length(x) # rozptyl
>
> acf=NULL
> for (i in 1:(length(x)-1)){
+   c=(sum((x-mean(x))[1:length(x[-(1:i)])] *(x[-(1:i)]-mean(x)))/length(x))/c0
+   acf=c(acf,c)
+ }
> round(acf,2),
```

kde v první řádku v proměnné `c0` je výpočten rozptyl dané časové řady a příkazem `mean(x)` se napočte střední hodnota. Do proměnné `acf` se postupně počítají a ukládají korelace vypočtené dle vzorce 3.4 pro $i = 1, \ldots, n-1$, kde n je délka časové řady `x`. Výstupem této funkce jsou hodnoty ACF zaokrouhlené na dvě desetiná místa. Příkazem `plot` si můžeme poté tyto hodnoty ACF vykreslit do grafu.

Výpočet neznámých parametrů probíhá na základě příkazu

```
> sarima(x,p,d,q,P,D,Q,Sz),
```

kde za první parametr `x` vyplníme název časové řady, parametr `p` určuje řád procesu AR, parametr `d` diferencování, za čtvrtý parametr `q` vyplníme řád procesu MA. Poslední 4 parametry slouží k vypočtům pro sezónní modely. Za parametr `P` vyplňujeme řád sezónního procesu AR, parametr `D` značí sezónní diferencování a parametr `Q` značí sezónní proces MA. Za poslední parametr vyplňujeme počet sezón v roce.

Pro předpovědi lze využít příkazu

```
> sarima.for(k,x,p,d,q,P,D,Q,Sz),
```

kde přidaný parametr `k` značí počet předpovědí.

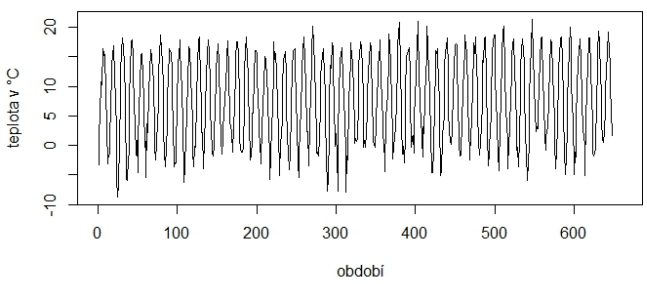
Kapitola 5

Analýza vývoje teploty v ČR

Naměřené teploty v °C v ČR jsou zveřejňovány na internetových stránkách Českého hydrometeorologického ústavu a jsou získávány jako průměr z jednotlivých měřících stanic rozmístěných na území ČR. Pro modelování teploty využiji měsíční pozorování od roku 1961 do roku 2015.

K dispozici budu mít tedy časovou řadu o délce $N = 660$ pozorování. Pozorování jsou vykreslena na obrázku 5. Pro modelování využiji pozorování z let 1961-2014, tj. $n = 648$ pozorování. Posledních 12 pozorování za rok 2015 využiji ke kontrole a zhodnocení predikcí.

Naměřenou teplotu budu nejprve analyzovat pomocí dekompozičního přístupu a poté pomocí metod Boxe a Jenkinse, kde budu vycházet ze zdroje [4], [7], [4] a



Obrázek 5.1: Měsíční pozorování vývoje teploty na území ČR v letech 1961-2015

z teoretických poznatků a vzorců uvedených v kapitole 1-3. Na konci porovnám výsledky z obou přístupů a vyhodnotím, který přístup je vhodnější.

5.1. Modelování vývoje teploty pomocí dekompozičního přístupu

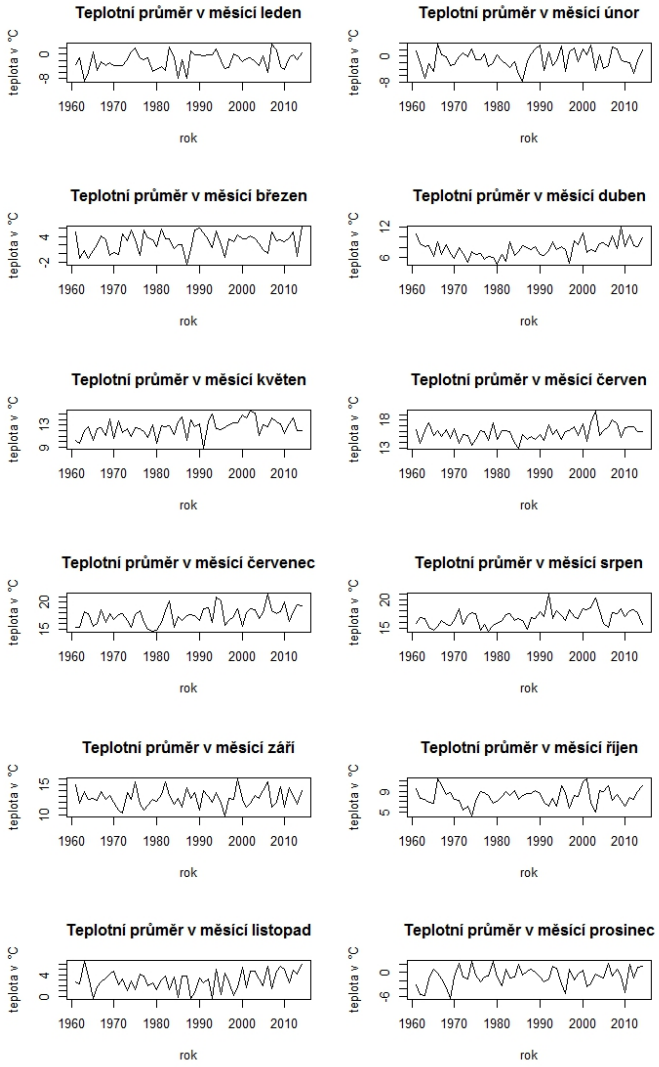
Z grafu na obrázku 5 je vidět, že důležitou roli bude hrát sezónní složka. Jsou zde vidět určité opakující se pravidelné výkyvy s periodou 12. Dlouhodobý vývoj se zdá být konstantní, pro upřesnění si vykreslím data dle jednotlivých měsíců (obrázek 5.2).

Z vývoje teploty v jednotlivých měsící můžeme pozorovat kolísání kolem konstantní úrovně. V případě letních měsících můžeme pozorovat od roku 1990 mírný růst. Zvláště v 7. a 8. měsíci, popřípadě i 5. a 6. měsíci. V datech budu proto modelovat sezónnost spolu s konstatním trendem, ale podívám se na modelování sezónnosti i s lineárním trendem. Poté oba modely srovnám a vyhodnotím, který je lepší.

5.1.1. Modelování sezónnosti s konstantním trendem

Využitím vzorce 2.41 pro model s konstatním trendem a s umělými sezónními proměnnými vypočítám MNČ odhady parametrů uvedené v tabulce 5.1.

Parametr $\hat{\beta}_0$ mi popisuje trendovou složku, zatímco parametry $\hat{\alpha}$ jsou sezónní parametry. Úpravou odhadnutých parametrů dle vzorce 2.44 a 2.45 dopočteme efekty jednotlivých sezón sz_j, tj. vliv jednotlivých měsíčních pozorování. Výsledek je uveden v tabulce **??**. Pro ověření správnosti výpočtu sečtu jednotlivé efekty sezón. Součtem dostávám 0. Efekty sezón jsou správně nanormovány. Po výpočtu neznámých parametrů a nanormování efektu jednotlivých sezón mohu odhadnout jednotlivá pozorování $\hat{y}_t$ dle vzorce 2.46. Srovnání skutečných naměřených hodnot a odhadnutých hodnot ukazuje následující graf na obrázku 5.3. Z důvodu přehlednosti je vykresleno na obrázku 5.3 jen 72 měscíců za posledních 6 let.



Obrázek 5.2: Teplotní průměry v jednotlivých měsicích od roku 1961-2014

Parametr	Odhad	Upraveny odhad
$\hat{\beta}_0$	$-2,21$	7,72
$\hat{\alpha}_1$	–	-9,92
$\hat{\alpha}_2$	$1,24$	-8,68
$\hat{\alpha}_3$	$4,90$	-5,02
$\hat{\alpha}_4$	$9,89$	-0,03
$\hat{\alpha}_5$	$14,80$	4,88
$\hat{\alpha}_6$	$17,95$	8,03
$\hat{\alpha}_7$	$19,63$	9,71
$\hat{\alpha}_8$	$19,06$	9,14
$\hat{\alpha}_9$	$15,02$	5,1
$\hat{\alpha}_{10}$	$10,18$	0,25
$\hat{\alpha}_{11}$	$5,16$	-4,77
$\hat{\alpha}_{12}$	$1,23$	-8,69

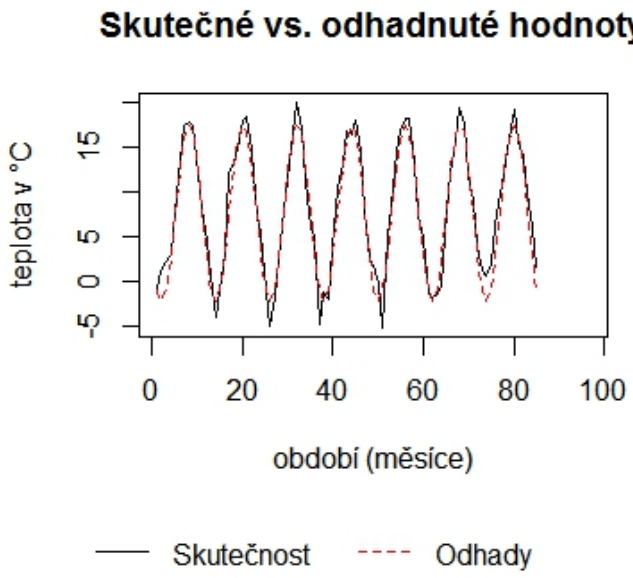
Tabulka 5.1: Odhadnuté parametry modelu proporcionální sezónnosti s konstantním trendem a normované efekty sezón

Vhodnost zvoleného modelu prověřím výpočtem reziduálního součtu čtverců (vzorec 2.56) a koeficientem determinace (dle vzorce 2.57) Hodnota RSČ je 2256,71. Hodnota celkové variability (dle vzorce 2.58) je 34089,13 a hodnota koeficientu determinace je 0,93, což svědčí ve prospěch zvoleného modelu.

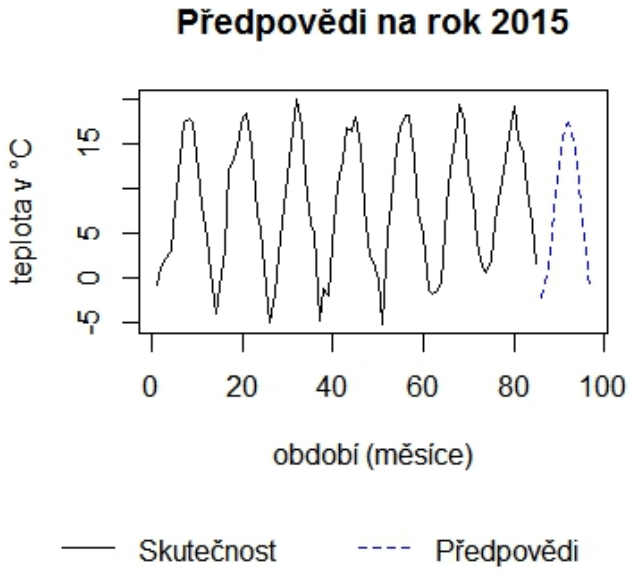
Nyní, když jsem zjistila, že model je správně zvolen, zkonstruuji předpovědi. Předpovědi budu konstruovat pro rok 2015, které budu poté zpětně srovnávat se skutečností.

Budoucí předpověď pro rok 2015 sestavím na základě dříve odhadnutých koeficientů, tedy dosazením do vzorce 2.48. Bodová předpověď pro rok 2015 v jednotlivých měsících je uvedena v tabulce 5.2 a příslušný graf s předpovědmi je uveden na obrázku 5.4.

Střední čtvercová chyba předpovědi dle vzorce 2.62 je MSEP$= 5,81$. Střední absolutní chyba předpovědi dle vzorce 2.63 je MAEP$= 1,75$. Obě chyby nám říkají, jaké chyby jsme se dopustili v průměru na každé předpovědi.



Obrázek 5.3: Odhadnuté hodnoty teplot (období 2008-2014



Obrázek 5.4: Naměřené hodnoty do roku 2014 společně s předpověďmi na rok 2015

Měsíc	Odhad	Skutečnost
leden	$-2,21$	$0,9$
únor	$-0,96$	$-0,1$
březen	$2,69$	4
duben	$7,68$	$7,8$
květen	$12,59$	$12,4$
červen	$15,75$	$16,1$
červenec	$17,43$	$20,2$
srpen	$16,86$	$21,3$
září	$12,81$	$13,1$
říjen	$7,97$	$7,9$
listopad	$2,95$	$5,8$
prosinec	$-0,98$	$3,7$

Tabulka 5.2: Měsíční předpovědi na rok 2015 a jejich posouzení

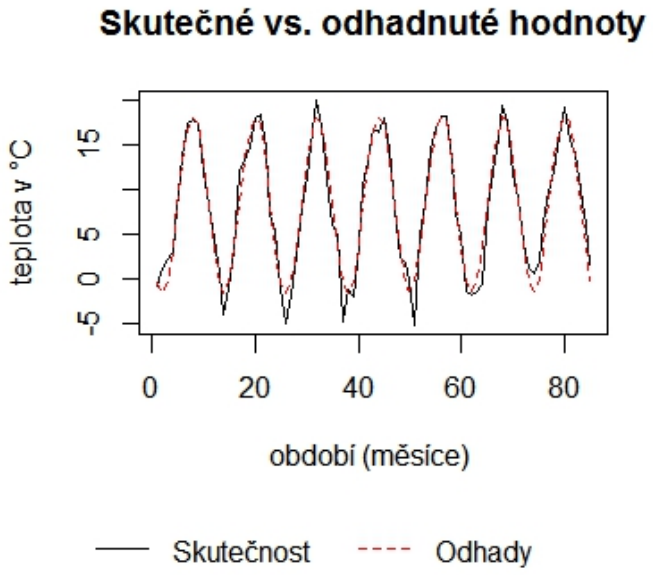
5.1.2. Modelování sezónnosti s lineárním trendem

Pro modelování sezónnosti s lineárním trendem budu uvažovat model tvaru 2.49. Neznámé parametry (β_0, β_1, α_j, $j = 1, \dots 12$) vypočtu obdobně jako u sezónního modelu s konstantním trendem, tj. pomocí MNČ. Odhadnuté parametry jsou uvedené v tabulce 5.3.

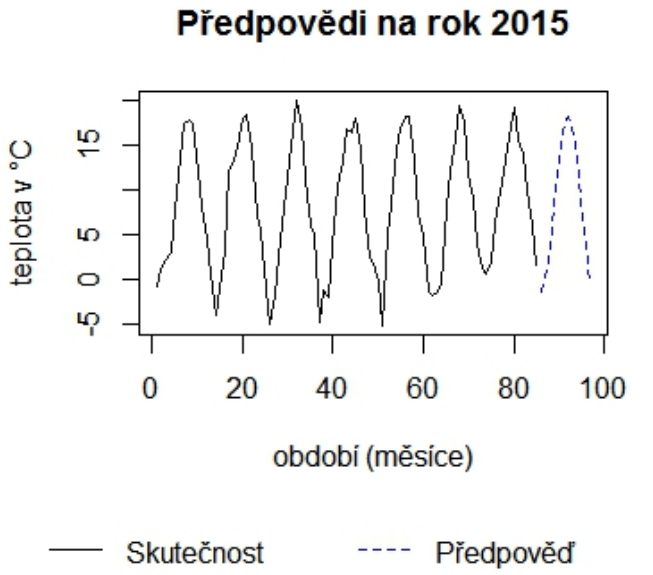
Parametry α dále nanormuji, nanormované parametry jednotlivých efektů v j-té sezónně jsou uvedeny v tabulce 5.3. Opět součtem normovaných efektů sezón dostávám 0. Efekty sezón jsou správně nanormovány. Graf odhadnutých hodnot a skutečných hodnot za posledních 7 let je uveden na obrázku 5.5.

Vhodnost zvoleného modelu prověřím opět výpočtem reziduálního součtu čtverců a koeficientem determinace, RSČ= $2124,3$, $R^2 = 0,938$. Odhadnuté hodnoty jsou blízké skutečným hodnotám.

Pro bodovou budoucí předpověď využiji vzorce 2.55. Bodová předpověď pro rok 2015 je uvedena v tabulce 5.4, předpovědi jsou vykresleny na obrázku 5.6. Hodnota MSEP (dle vzorce 2.62) je $3,71$ a MAEP (2.63) vychází na $1,49$.



Obrázek 5.5: Srovnání skutečných hodnot a odhadnutých



Obrázek 5.6: Předpovědi na rok 2015

Parametr	Odhad	Normování
$\hat{\beta}_0$	$-2,98$	-
$\hat{\beta}_1$	$0,002$	-
$\hat{\alpha}_1$	0	-9,91
$\hat{\alpha}_2$	$1,24$	-8,67
$\hat{\alpha}_3$	$4,90$	-5,01
$\hat{\alpha}_4$	$9,88$	-0,03
$\hat{\alpha}_5$	$14,79$	4,88
$\hat{\alpha}_6$	$17,94$	8,03
$\hat{\alpha}_7$	$19,62$	9,71
$\hat{\alpha}_8$	$19,05$	9,14
$\hat{\alpha}_9$	$15,00$	5,1
$\hat{\alpha}_{10}$	$10,15$	0,25
$\hat{\alpha}_{11}$	$5,13$	-4,78
$\hat{\alpha}_{12}$	$1,20$	-8,71

Tabulka 5.3: Odhadnuté parametry modelu a příslušné normované parametry

5.1.3. Srovnání sezónního modelu s konstantním trendem a s lineárním trendem

Oba modely důvěryhodně popisují a předpovídají napozorované hodnoty. Pro srovnání, který model je ale lepší, využiji statistické ukazatele pro posouzení vhodnosti zvoleného modelu popsané v kapitole 2. Nejprve oba modely srovnám pomocí RSČ dle vzorce 2.56 a koeficientu determinace. Hodnoty RSČ a koeficientu determinace obou modelů jsou uvedeny v následující tabulce 5.1.3.

RSČ i R^2 svědčí ve prospěch sezónního modelu s lineárním trendem. U tohoto modelu se dopouštíme menších chyb v jednotlivých odhadnutých pozorování od skutečností (RSČ je nižší). A naopak R^2 je u tohoto modelu vyšší. Vyšší hodnota R^2 vypovídá o menší variabilitě pozorováních v modelu a tedy i o přesnějším popisu skutečných naměřených dat. To může být ale důsledek složitějšího modelu. Čím složitější model máme, tím menší reziduální chyby se zpravidla dopustíme. Pro objektivnější posouzení provedu statistický t-test (dle vzorce 2.61) na významnost parametru β_1, hodnota t-testu je 6,35. Srovnáním s kvantilem t-rozdělení o 635 stupních volnosti na hladině $\alpha = 0,05$, tj. $t_{635,(0,05)} \doteq 1,96$,

měsíc	odhad	Skutečnost
leden	$-1,41$	$0,9$
únor	$-0,17$	$-0,1$
březen	$3,49$	4
duben	$8,48$	$7,8$
květen	$13,39$	$12,4$
červen	$16,54$	$16,1$
červenec	$18,23$	$20,2$
srpen	$17,66$	$21,3$
září	$13,61$	$13,1$
říjen	$8,77$	$7,9$
listopad	$3,75$	$5,8$
prosinec	$-0,15$	$3,7$

Tabulka 5.4: Měsíční předpovědi na rok 2015 a jejich posouzení

Model	RSČ	R^2	R^2_{adj}	MSEP	MAEP
Sezonní+lineární trend	$2124,31$	$0,938$	$0,937$	$3,71$	$1,49$
Sezonní+konstantní trend	$2256,71$	$0,934$	$0,933$	$5,81$	$1,75$

Tabulka 5.5: Porovnání sezónního modelu s konstantním trendem se sezónním modelem s lineárním trendem

nulovou hypotézu o nulovosti parametru β_1 zamítáme. Parametr β_1 je v modelu významný. Výpočtem upraveného koeficientu determinace, který také zohledňuje parametry v modelu, nepatrně lepšího výsledku dosáhnu pomocí lineárního sezónního modelu. Veškeré výsledky jsou shrnuty v tabulce 5.1.3.

Chyby v odhadech předpovědí obou modelů jsou nepatrné. Přesnější odhady předpovědí nám ale dle ukazatelů MSEP a MAEP uvedené také v tabulce dává sezónní model s lineárním trendem. MAEP a i MSEP je pro tento model nižší než u sezónního modelu s konstantním trendem.

Budeme-li modelovat vývoj teploty pomocí dekompozičního přístupu, upřednostníme spíše sezónní model s lineárním trendem. Tento model nám dává nepatrně lepší výsledky. Nicméně budeme-li chtít jednodušší model, můžeme použít i sezónní model s konstatním trendem.

5.2. Modelování vývoje teploty pomocí přístupu Boxe a Jenkinse

Boxovy-Jenkinsovy modely jsou početně náročnější, proto se v praxi často pro výpočet využívá nástroj výpočetní techniky. Já pro svoji práci zvolila k výpočtům software R. Základní knihovnou pro modelování ARMA modelů je knihovna astsa a tseries.

Jak už bylo zmíněno v teoretické části, Boxovy a Jenkinsovy modely se zaměřují na modelování náhodné složky. Důležitým předpokladem je stacionarita. Na základě obrázku 5 uvedeného v úvodu této části můžeme tvrdit, že řada je stacionární. Data teploty nevykazují v čase žádný velký nárust a ani pokles. Pouze opakující se pravidelný cyklus kolísání kolem přibližně konstantní úrovně. Pravidelný cyklus kolísání identifikuje sezónní data. Sezónnost v datech nám identifikuje také autokorelační funkce (obrázek 5.7), kde můžeme pozorovat opakující se periodu s délkou 12. Perioda v datech je tedy roční. Podle obrázku 5.7 jsou dále hodnoty autokorelační funkce významné a klesají v sezóně velice pomalu, přičemž první autokorelace je blízká 1. To značí neustálenost procesu a tedy nestacionaritu v sezóně.

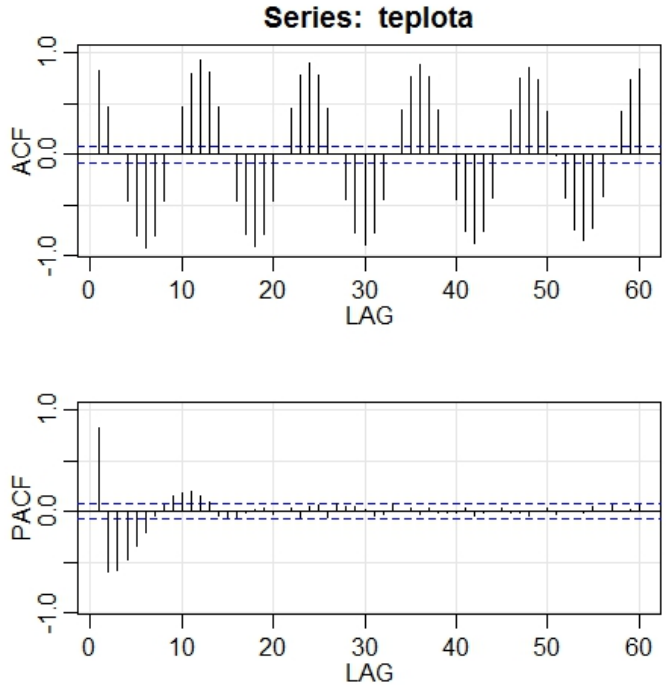
Hodnoty autokorelační a parciální autokorelační funkce vyobrazené na obrázku 5.7 vypočtu pomocí softwaru R dle vzorce 3.4 a 3.8.

V první řadě potřebuji nyní dosáhnout stacionarity v sezóně. Stacionarity dosáhneme diferencováním. Protože ACF značí nestacionaritu v sezóně, provedu sezónní diferencování 1. řádu.

Podle zdroje [1] vyzkouším ještě klasické diferencování 1. řádu a kombinaci obou diferencí, které porovnám podle odhadovaných rozptylů. Odhady rozptylů původní řady y_t, klasicky diferencované řady $y_t - y_{t-1}$, sezóně diferencované řady $y_t - y_{t-12}$ a kombinací klasicky a sezóně diferencované řady y_{komb} jsou

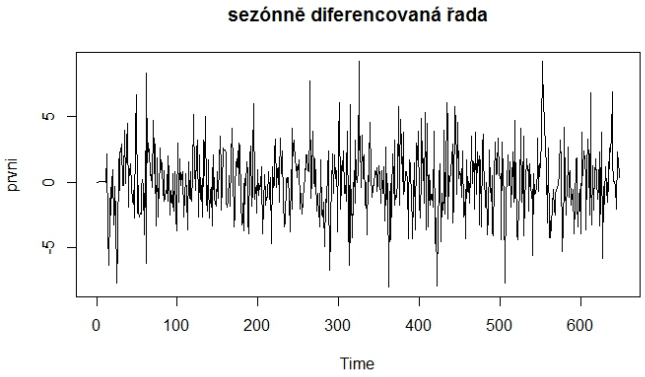
$$\hat{\sigma}^2_{y_t} = 52,61, \quad \hat{\sigma}^2_{y_t - y_{t-1}} = 18,86, \quad \hat{\sigma}^2_{y_t - y_{t-12}} = 6,36, \quad \hat{\sigma}^2_{y_{komb}} = 11,13.$$

Nejmenší odhadovaný rozptyl dostanu pomocí sezónního diferencovaní. Sezónní



Obrázek 5.7: ACF a PACF

diferencovaní je nejvhodnější. Graf sezónně diferencovaných dat je uveden na obrázku 5.8.



Obrázek 5.8: Sezónně diferencované data

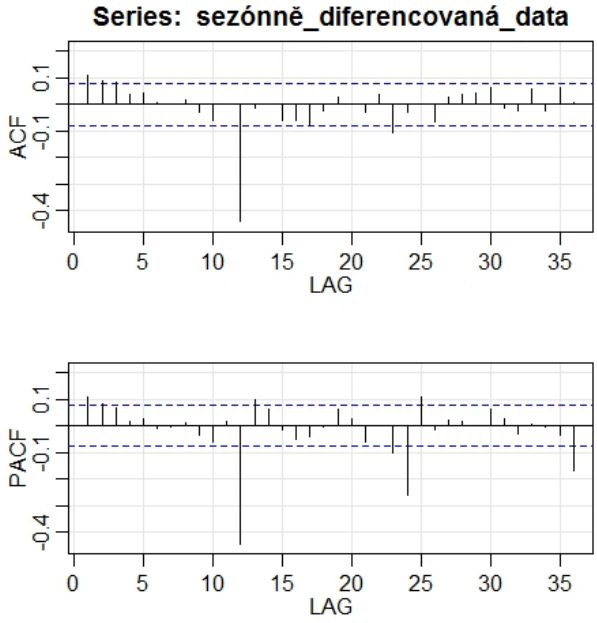
K identifikaci vhodného modelu si opět vykreslím ACF a PACF. Tentokrát pro sezónně diferencované data (obrázek 5.9).

ACF na obrázku 5.9 má významnou autokorelaci v 12. hodnotě. PACF má významnou každou dvanáctou hodnotu, která postupně klesá. To identifikuje sezónní model MA(1). ACF i PACF (obrázek 5.10) modelu SARIMA$(0,0,0)(0,1,1)_{12}$ stále vykazuje významné hodnoty. Budu-li model testovat pomocí Ljungovy-Boxovy statistiky dle vzorce 3.45, hypotézu o náhodnosti reziduí zamítnu. To poukazuje na nedostatečně zvolený model.

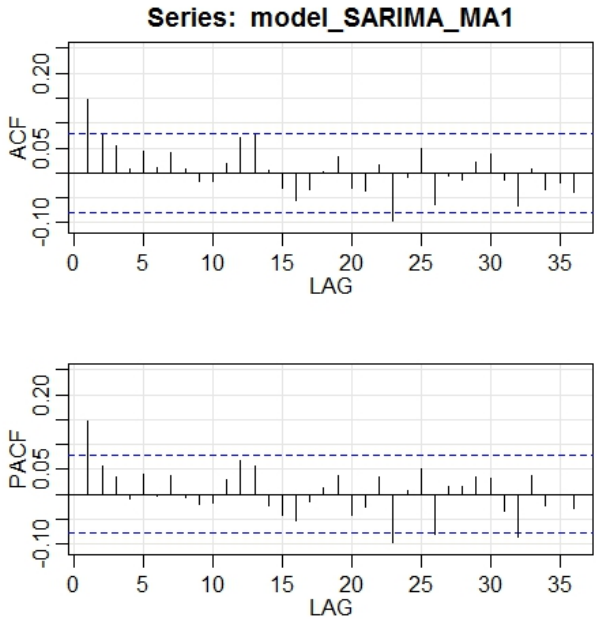
ACF modelu SARIMA$(0,0,0)(0,1,1)_{12}$ má klesající charakter s mírně významnou 1. hodnotou. PACF má první hodnotu významnou a další mírně významnou. To může odpovídat modelu MA(1), popřípadě MA(2) nebo AR(1). Při zvolení sezónního modelu MA(1) se dle zdroje [1] nedoporučuje dále kombinovat s parametrem AR(1) a ani ARMA(1,1). Proto dále budou rozebrány pouze varianty s přidaným parametrem MA(1) a MA(2).

Model SARIMA$(0,0,1)(0,1,1)_{12}$

Odhadnuté parametry modelu SARIMA$(0,0,1)(0,1,1)_{12}$ jsou:



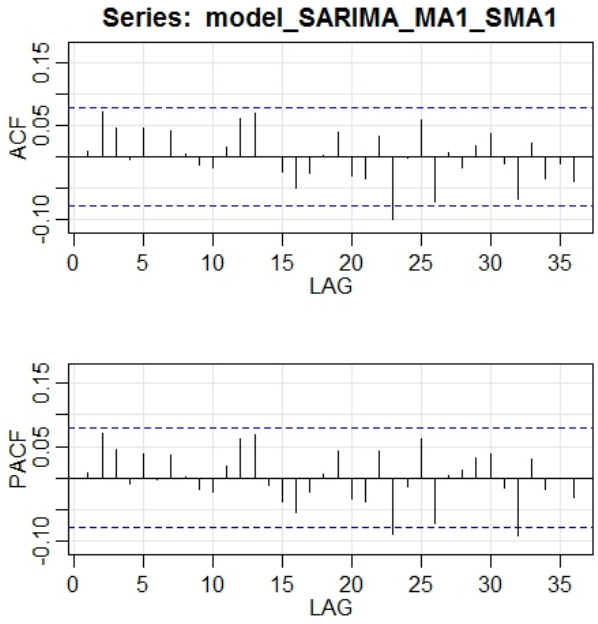
Obrázek 5.9: ACF a PACF pro sezóně diferencované data



Obrázek 5.10: ACF a PACF pro model SARIMA$(0,0,0)(0,1,1)_{12}$

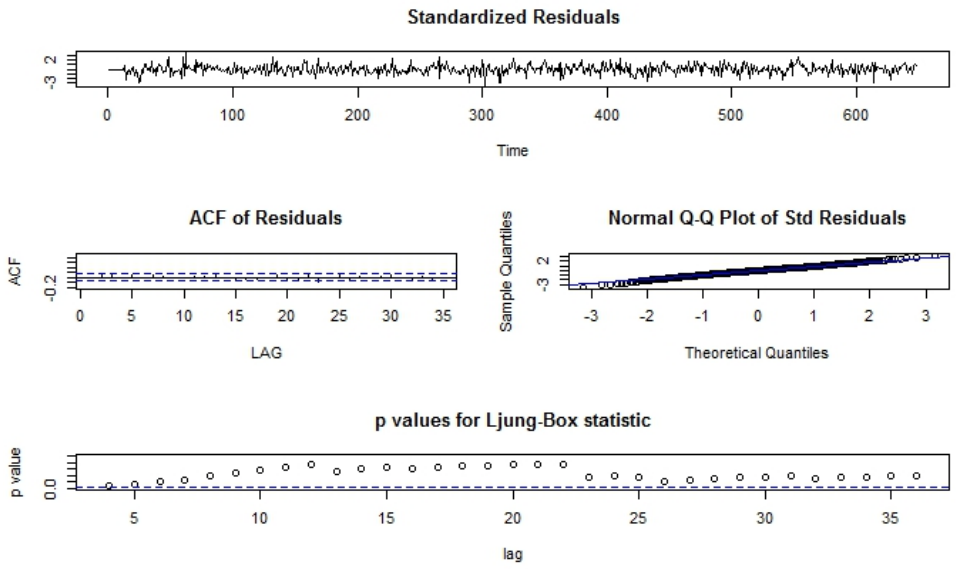
```
Coefficients:
          ma1     sma1  constant
      0.1329  -1.0000    0.0024
s.e.  0.0375   0.0413    0.0004

sigma^2 estimated as 3.337
```

S pomocí směrodatných odchylek odhadnutých parametrů (druhý řádek, kde s.e.=standard deviation), otestujeme významnost parametrů podle vzorce 3.43. Na hladině $\alpha = 0,05$ z parametrů nezamítáme. V modelu nemáme žádný nadbytečný parametr. ACF a PACF tohoto modelu je uvedena na obrázku 5.11.



Obrázek 5.11: ACF a PACF pro model SARIMA$(0,0,1)(0,1,1)_{12}$

Diagnostiku vhodnosti daného modelu provedu na základě ACF reziduí a pomocí Ljungova-Boxova testu (**??**). ACF reziduí nevykazuje žádnou významnou hodnotu, rezidua mají nahodný charakter. Ve prospěch náhodilosti reziduí svědčí i vysoké hodnoty p-value pro testovou statistiku Ljunga-Boxe. Hodnota Ljunova-Boxova testovacího kritéria je 13,16. Na hladině významnosti 0,05 s kvantilem



Obrázek 5.12: Diagnostika modelu SARIMA$(0,0,1)(0,1,1)_1 2$

χ^2 rozdělení(28,8693) nulovou hypotézu nezamítáme, rezidua jsou tedy náhodně rozdělená.

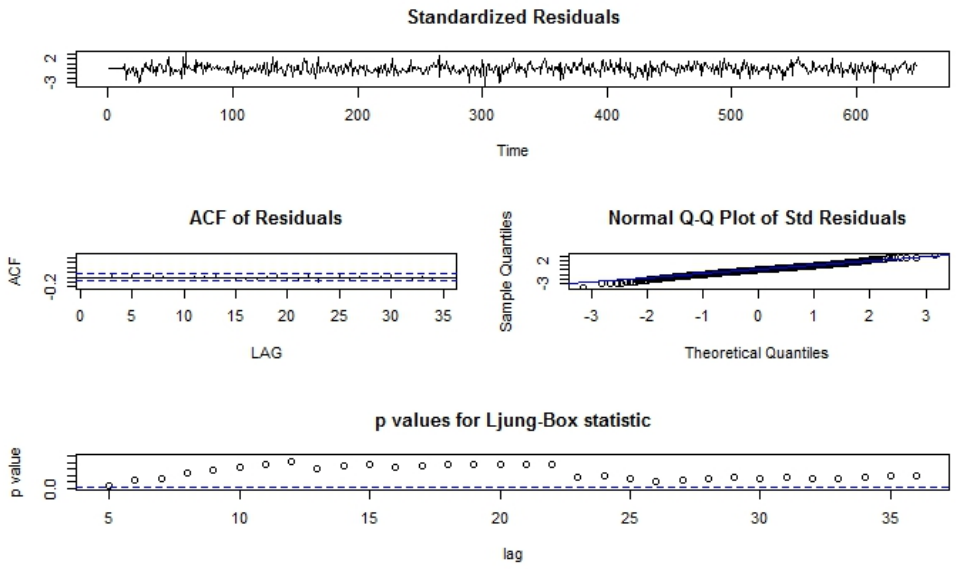
Model SARIMA$(0,0,2)(0,1,1)_{12}$

Odhadnuté parametry modelu SARIMA$(0,0,2)(0,1,1)_{12}$ jsou:

```
Coefficients:
          ma1     ma2     sma1  constant
      0.1348  0.0687  -1.0000    0.0024
s.e.  0.0399  0.0401   0.0448    0.0005

sigma^2 estimated as 3.322
```

T-testem 3.43 nezamítám hypotézu o nulovosti sezónního parametru $MA(2)$. Tento parametr je v modelu nadbytečný.

Diagnostiku modelu (na Obrázku 5.13) provedu stejným způsobem jako u předchozího modelu. ACF reziduí nevykazuje žádnou závislost, podle Q-Q grafu jsou rezidua normálně rozdělená. P-value pro Ljung-Boxovu statistiku svědčí ve



Obrázek 5.13: Diagnostika modelu SARIMA$(0,0,2)(0,1,1)_12$

prospěch nezávislosti reziduí. Vyčíslením si hodnoty pro LJung-Boxovu statistiku (11.988) a srovnáním s 0,05 kvantilem χ^2 rozdělení, nulovou hypotézu o nulové korelaci reziduí nezamítáme. Přesto, že u modelu jsme dosáhli nezávislé, normálně rozdělené náhodné složky, model z důvodu nadbytečnosti parametru $MA(2)$ není vhodný pro modelování dat vývoje teploty. Vždy se snažíme najít co nejjednodušší model. Vhodnější model pro modelování vývoje teploty je SARIMA$(0,0,1)(0,1,1)_{12}$.

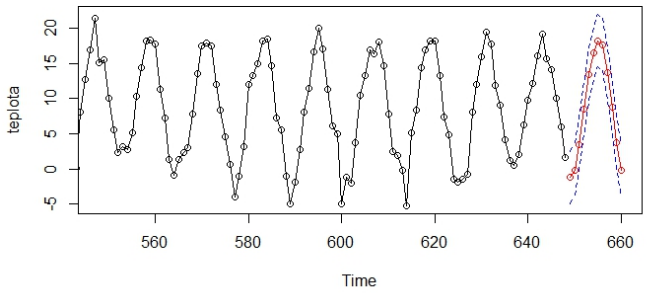
Po identifikování vhodného modelu a odhadnutí parametrů mohu přejít k předpovědím. Měsíční předpovědi pro rok 2015 jsem konstruovala pro model SARIMA$(0,0,1)(0,1,1)_{12}$, ale také jsem se podívala na předpovědi modelu SARIMA$(0,0,2)(0,1,1)_{12}$. Odhady předpovědí obou modelů pro rok 2015 a jejich měsíční absolutní chyba je shrnuta v tabulce 5.6.

Celkové nejmenší chyby při předpovídání jsme se dopustili s modelem $SARIMA(0,0,2)(0,1,1)_1$ s absolutní chybou $17,55$. Vzhledem k tomu, že rozdíl oproti druhému modelu je nepatrný, volili bychom pro modelování předpovědi nejjednoduší model, tj. $SARIMA(0,0,1)(0,1,1)_{12}$. Tento model měl i dle t-testu na hladině $\alpha = 0,05$ všechny parametry významné. Předpovědi tohoto modelu jsou vykresleny na

skutečnost roku 2015	SARIMA$(0,0,2)(0,1,1)_{12}$		SARIMA$(0,0,1)(0,1,1)_{12}$	
	odhad	abs. rozdíl	odhad	abs. rozdíl
$0,9$	$-1,06$	$1,96$	$-1,20$	$2,10$
$-0,1$	$-0,06$	$0,04$	$-0,16$	$0,06$
4	$3,49$	$0,51$	$3,49$	$0,51$
$7,8$	$8,48$	$0,68$	$8,48$	$0,68$
$12,4$	$13,39$	$0,99$	$13,39$	$0,99$
$16,1$	$16,55$	$0,45$	$16,55$	$0,45$
$20,2$	$18,23$	$1,97$	$18,23$	$1,97$
$21,3$	$17,66$	$3,64$	$17,66$	$3,64$
$13,1$	$13,61$	$0,51$	$13,61$	$0,51$
$7,9$	$8,77$	$0,87$	$8,77$	$0,87$
$5,8$	$3,75$	$2,05$	$3,75$	$2,05$
$3,7$	$-0,17$	$3,87$	$-0,18$	$3,88$
		$17,55$		$17,72$

Tabulka 5.6: Předpovědi modelů a jejich zhodnocení

obrázku 5.7, kde modrou přerušovanou čárou je vyznačen i interval spolehlivosti.



Obrázek 5.14: Předpovědi pro model SARIMA$(0,0,1)(0,1,1)_{12}$

5.3. Srovnání a Závěr

Model dekompozičního přístupu a SARIMA modely budu srovnávat pomocí vybraných charakteristik pro model a pro předpovědi.

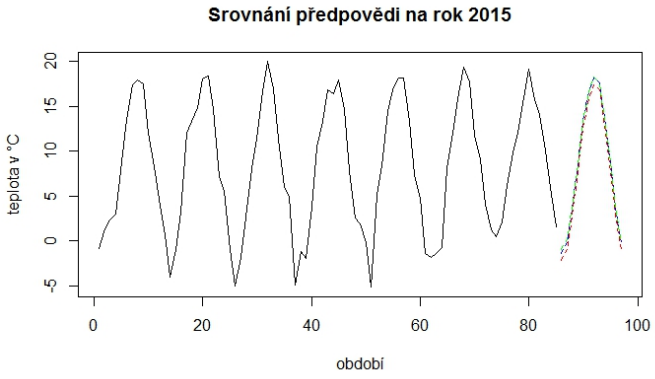
Chyb, kterých jsem se dopustila odhadováním původních dat, mají menší

Model	MSEP	MAEP	RSČ	R^2
Sezonní+konstatní trend	943, 33	21, 04	2256, 71	0, 934
Sezónní+lineární trend	907, 54	17, 90	2124, 31	0, 938
SARIMA$(0, 0, 2)(0, 1, 1)_{12}$	893, 60	17, 55	2070, 38	0, 939
SARIMA$(0, 0, 1)(0, 1, 1)_{12}$	895, 74	17, 72	2078, 18	0, 939

Tabulka 5.7: Předpovědi modelů a jejich zhodnocení

modely SARIMA. Boxovy-Jenkinsovy modely nám lépe popisují původní data. Nejlépe nám popisuje původní data model SARIMA$(0, 0, 2)(0, 1, 1)_{12}$, kde absolutní chyba odhadu je 893,6. Také RSČ má tento model nejmenší. Tento model nám vychází nejlépe, protože má více parametrů než model $SARIMA(0, 0, 1)(0, 1, 1)_{12}$.

Podle grafu na obrázku 5.15 jsou předpovědi modelů vyrovnané, téměř totožné. Pro podrobnější pohled na předpovědi se podívám do tabulky 5.7, kde druhý sloupec udává absolutní chybu předpovědi. Srovnáním absolutní chyby



Obrázek 5.15: Srovnání předpovědí

předpovědí jsem nejpřesnější budoucí odhady pro rok 2015 získala pomocí modelu $SARIMA(0, 0, 1)(0, 1, 1)_{12}$.

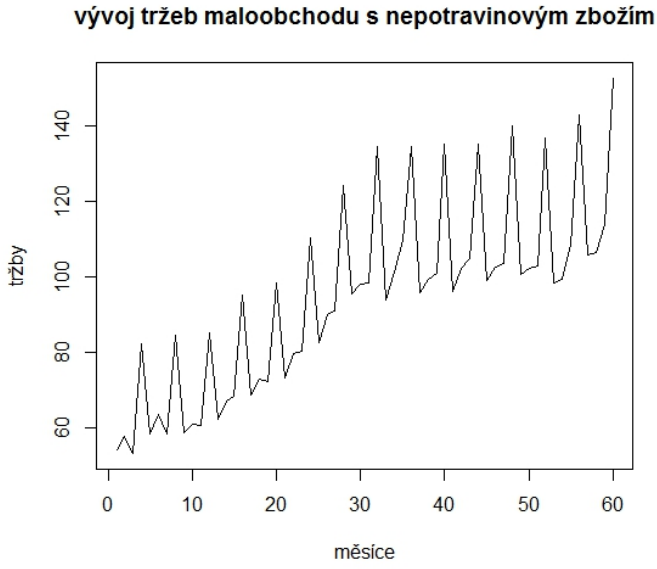
Závěrem lze poznamenat, že Boxovy-Jenkinsovy modely nám lépe popisují vývoj teploty a udávají přesnější předpovědi průměrných měsíčních teplot. Rozdíly jsou ale oproti dekompozičnímu přístupu nepatrné. Pro modelování vývoje teploty bychom mohli použít oba způsoby.

Kapitola 6

Analýza vývoje tržeb maloobchodů s nepotravinovým zbožím

Druhým příkladem, kterým se budu zabývat, je čtvrtletní vývoj tržeb maloobchodů s nepotravinovým zbožím. Celkem mám k dispozici čtvrtletní pozorování od roku 2000-2015, tj. $n = 64$ pozorování. Tato data jsem získala ze stránek České národní banky. Pro modelování využiji pozorování od roku 2000-2014, tj prvních $n = 60$ pozorování. Grafické znázornění tržeb od roku 2000-2014 je uvedeno na obrázku 6.1. Pozorování za rok 2015 využiji ke kontrole a zhodnocení predikcí.

Tržby maloobchodních prodejen zaměřených na nepotravinové zboží v průběhu 14 let se stále zvyšovaly. Stagnaci tržeb v letech 2008-2010 je způsobena ekonomickou krizí a jejími následky. V průběhu roku můžeme pozorovat výkyvy. Zvláště vždy ke konci roku tržby zaznamenávají prudký růst. Naproti tomu začátkem každého roku můžeme pozorovat prudký pokles. Mírný růst tržeb nastává opět až během období léta.



Obrázek 6.1: Vývoj tržeb maloobchodu s nepotravinovým zbožím v letech 2000-2014

6.1. Analýza vývoje tržeb maloobchodů pomocí dekompoziční přístup

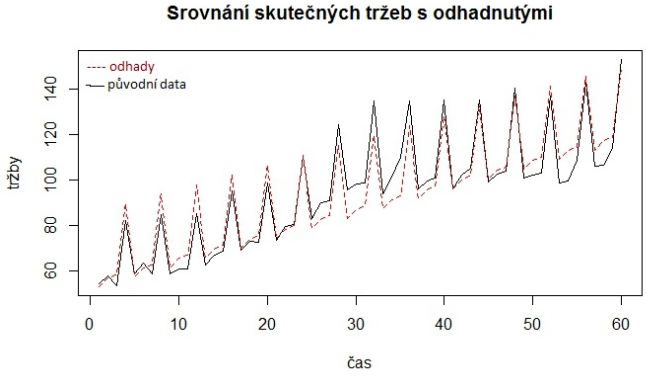
Vzhledem k dlouhodobému růstu tržeb v průběhu 14 let, budu na data aplikovat lineární trend. Bude také potřeba vyřešit kolísání během roku (namodelovat sezonní složku). Čtvrtletní pozorování mi budou charakterizovat tržby na jaře, v létě, na podzim a v zimě, tj. 4 sezóny za rok. Budu tedy modelovat lineárním trend s konstantní sezónností. Tvar modelu je uveden v teoretické části 2.34.

Metodou nejmenších čtverců získám odhady parametrů uvedené v tabulce 6.1. V posledním sloupci tabulky jsou uvedeny normované odhadnuté parametry vypočtené dle vzorce 2.36, hodnota normovaného parametru v první řádku byla získána součtem odhadnutého parametru $\hat{\beta}_0$ a $\bar{\alpha}$ vypočtené dle vzorce 2.37.

Mám-li odhadnuty neznámé parametry, mohu vypočítat vyrovnané hodnoty dle vzorce 2.39. Graf srovnávající skutečné a odhadnuté hodnoty je uveden na Obrázku 6.2.

Parametr	Odhad	Normování
β_0	$51,62$	61,51
β_1	$1,08$	1,08
α_1	0	$-9,89$
α_2	$2,86$	$-7,02$
α_3	$3,40$	$-6,49$
α_4	$33,29$	$23,40$

Tabulka 6.1: Odhady parametrů a normované parametry modelu pro vývoj tržeb maloobchodů



Obrázek 6.2: Srovnání skutečných tržeb a odhadnutých tržeb dle modelu

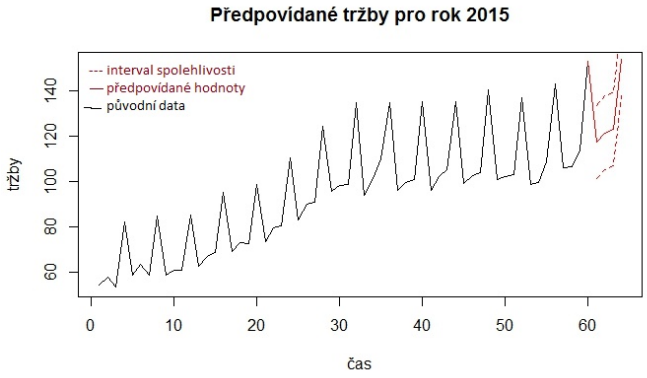
Na základě koeficientu determinace (dle vzorce 2.57 je $R^2 = 0,925$) mohu tvrdit, že model je vhodně zvolen. Mám-li odhadnuté parametry modelu a vhodně zvolený model, mohu přejít k předpovědím na rok 2015. Předpovědi na rok 2015 obdržím dle vzorce 2.40.

	odhad	skutečnost
1.čtvrtletí	$117,37$	$112,5$
2.čtvrtletí	$121,31$	$116,8$
3.čtvrtletí	$122,92$	$120,1$
4.čtvrtletí	$153,89$	$162,6$

Tabulka 6.2: Skutečné tržby vs. předpovídané tržby na rok 2015

Srovnání skutečných tržeb za rok 2015 s odhadnutými tržby mohu své předpovědi zhodnotit. Skutečné tržby za rok 2015 a odhadnuté tržby jsou uvedeny v tabulce 6.2. Střední čtvercová chyba předpovědi (MSEP) je $31,973$, MAEP je

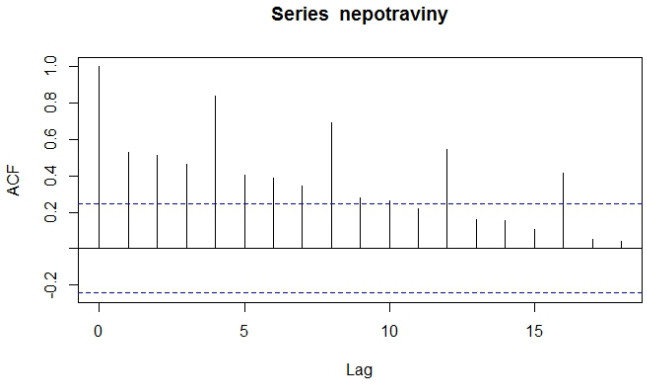
5, 23. Předpovědi s intervaly spolehlivosti jsou vykresleny na obrázku 6.3.



Obrázek 6.3: Předpovědí tržeb na rok 2015

6.2. Analýza vývoje tržeb maloobchodů pomocí přístupu Boxe a Jenkinse

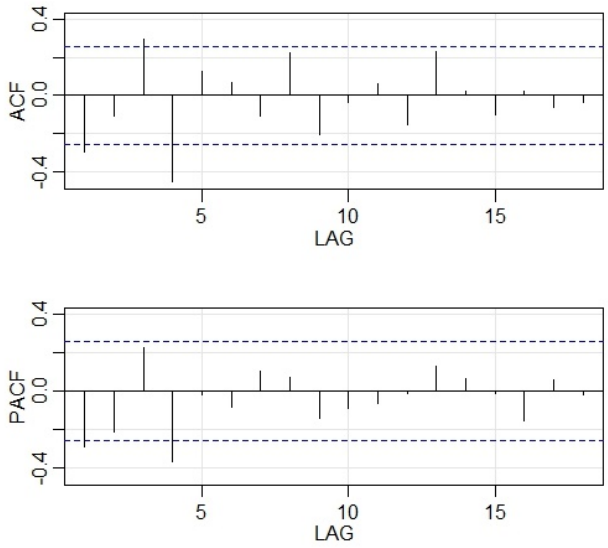
Podle obrázku 6.1 je zřejmé, že řada bude sezónní. To dokazuje i ACF vykreslena na obrázku 6.4, kde můžeme pozorovat opakující se periodu s délkou 4.



Obrázek 6.4: ACF a PACF původních dat

ACF navíc pomalu klesá, v dalším kroku budu tedy potřeba časovou řadu tržeb diferencovat. Provedu-li klasické diferencování 1. řádu dat, rozptyl původních

dat 605 klesne na 510,4. Rozptyl sezónně diferencované řady klesne na 21,99 a kombinací klasického a sezónního diferencování rozptyl řady klesne na 15,72. Porovnám-li výše uvedené odhadované rozptyly, nejmenší odhadovaný rozptyl dostanu pomocí kombinace klasického a sezónního diferencování. Kombinace sezónního a klasického diferencování je nejvhodnější. ACF a PACF po kombinovaném diferencování je uvedena na obrázku 6.5. ACF i PACF mají značně významnou 4 hodnotu. To svědčí buď ve prospěch sezónního parametru modelu AR(1) nebo MA(1).

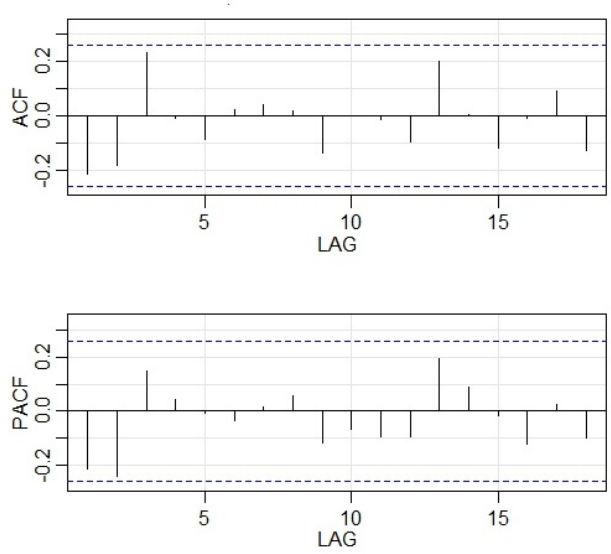


Obrázek 6.5: ACF a PACF po sezónním diferencováním

SARIMA$(0,1,0)(1,1,0)_4$

S přidáním sezónního parametru AR(1) do modelu je ACF a PACF vykreslena na obrázku 6.6. Odhadovaná hodnota sezónního parametru AR(1) je -0,44. Hodnoty ACF a PACF nemají již žádné významné hodnoty. Diagnostika modelu SARIMA$(0,1,0)(1,1,0)_4$ uvedená na obrázku 6.9 poukazuje na náhodnost reziduí. Dle autokorelační funkce reziduí, Q-Q plotu a Ljungova-Boxova testu tvoří rezidua v modelu bílý šum.

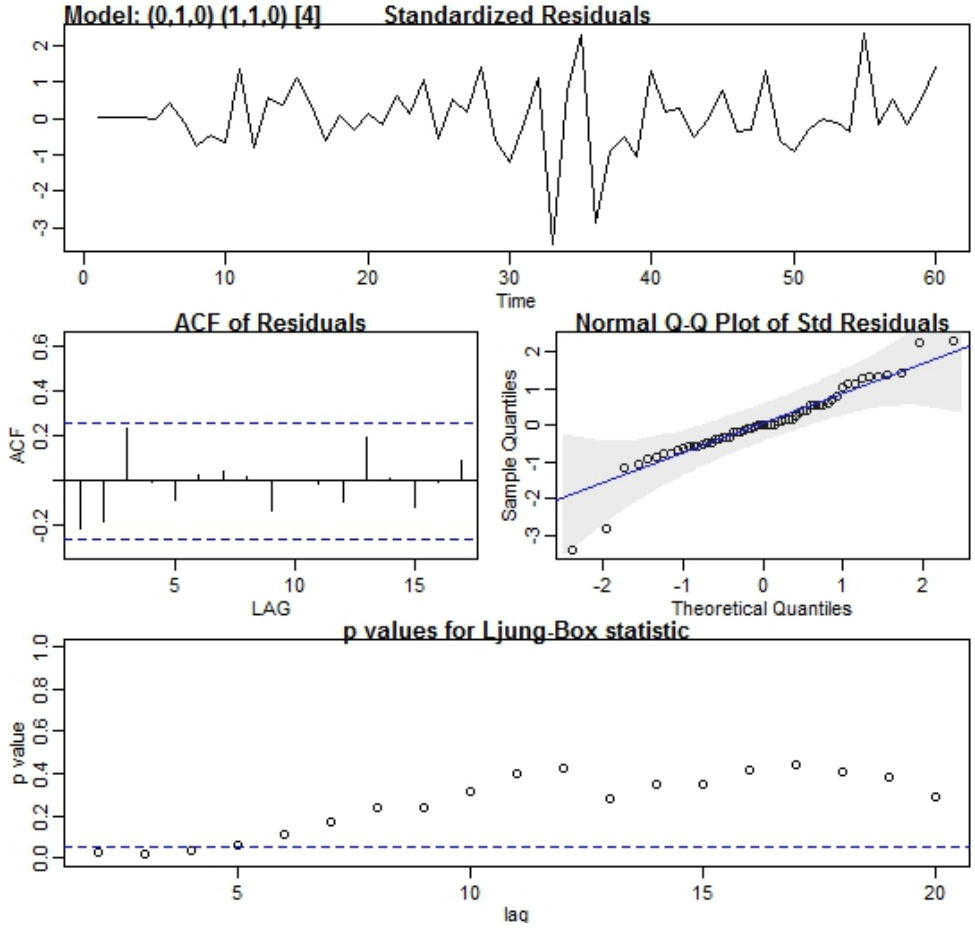
SARIMA$(0,1,0)(0,1,1)_4$



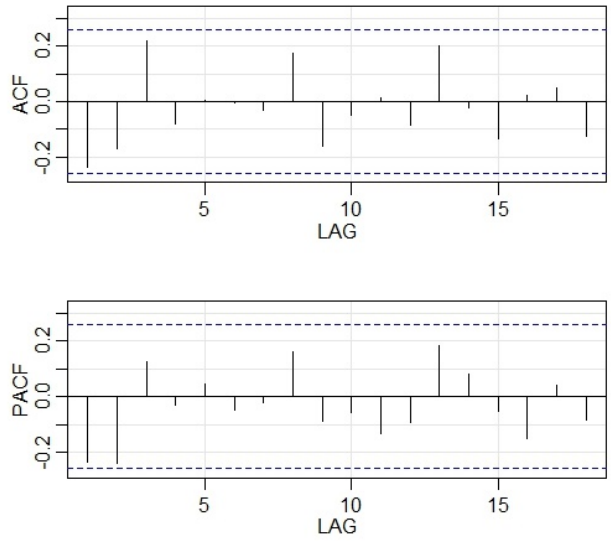
Obrázek 6.6: ACF a PACF modelu $SARIMA(0,1,0)(1,1,0)_4$

S přidáním sezónního parametru MA(1) do modelu je ACF a PACF vykreslena na obrázku 6.8. Odhadovaná hodnota sezónního parametru MA(1) je -0,39. Hodnoty ACF a PACF nemají již žádné významné hodnoty. Rezidua v modelu dle autokorelační funkce reziduí (2.8), Q-Q plotu a Ljungova-Boxova testu (2.8) tvoří bílý šum. Celková diagnostika modelu dle výše uvedených testů je vykreslena na obrázku 6.9. Model SARIMA$(0,1,0)(1,1,0)_4$ můžu tedy považovat za vhodně zvolený.

Graf s předpověďmi na rok 2015 je uveden na obrázku 6.10. Předpovědi pro rok 2015 dle modelu SARIMA$(0,1,0)(1,1,0)_4$ a modelu SARIMA$(0,1,0)(0,1,1)_4$ jsou dle obrázku 6.10 téměř totožné. Hodnoty předpovědí společně se skutečnými hodnotami jsou uvedeny v tabulce **??**. Pro modelování předpovědí pro rok 2015 se hodí oba modely.



Obrázek 6.7: Diagnostika odhadnutého modelu SARIMA$(0, 1, 0)(1, 1, 0)_4$



Obrázek 6.8: ACF a PACF pro model $SARIMA(0,1,0)(0,1,1)_4$

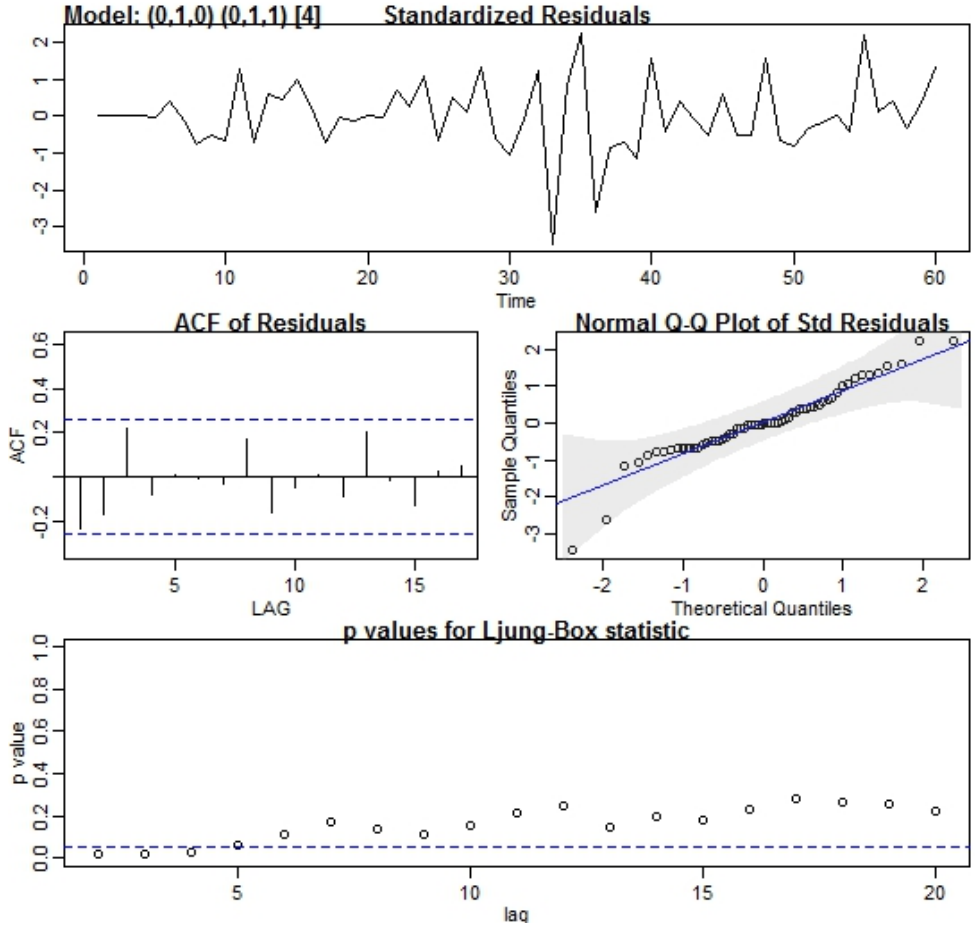
	Skutečnost	SARIMA$(0,1,0)(1,1,0)_4$	SARIMA$(0,1,0)(0,1,1)_4$
1.čtvrtletí	112, 5	115, 2	115,2
2.čtvrtletí	116, 8	116	116,3
3.čtvrtletí	120, 1	124, 1	123
4.čtvrtletí	162, 6	161	160

Tabulka 6.3: Skutečné tržby vs. předpovídané tržby modelem SARIMA$(0,1,0)(1,1,0)_4$ a modelem SARIMA$(0,1,0)(0,1,1)_4$ na rok 2015

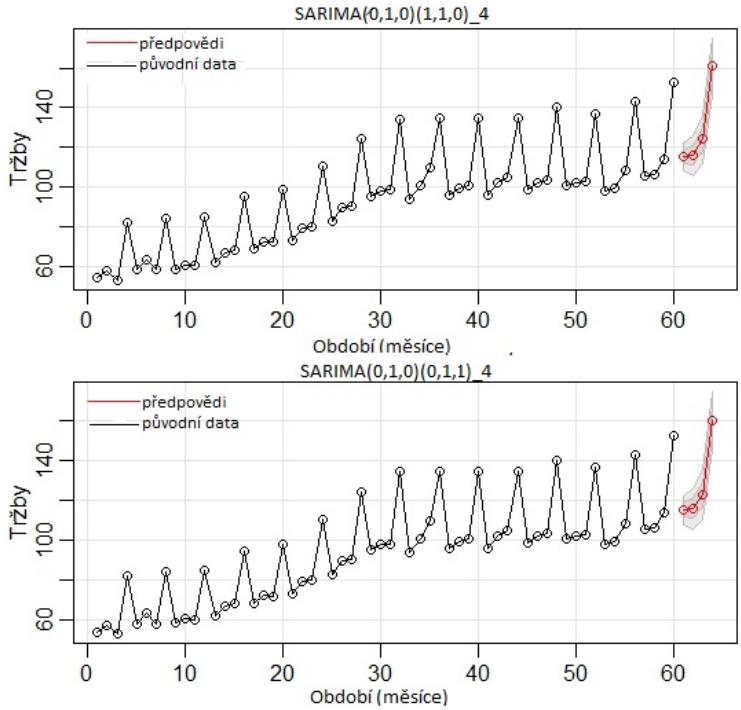
6.3. Srovnání a Závěr

Model dekompozičního přístupu a SARIMA modely budu srovnávat pomocí vybraných charakteristik pro model a pro předpovědi. Srovnání modelů je uvedeno v tabulce 6.3.

SARIMA modely si pomocí diferencování velmi dobře poradili s rostoucím trendem tržeb. Podle RSČ i R^2 nám lépe popisují původní data než dekompoziční přístup. Pomocí modelů SARIMA jsem také získala přesnější odhady předpovědi pro rok 2015 (dle MSEP (2.62) a MAEP (2.63)). Pro modelování vývoje tržeb maloobchodů a pro předpovědi tržeb se lépe hodí SARIMA modely.



Obrázek 6.9: Diagnostika odhadnutého modelu SARIMA$(0,1,0)(0,1,1)_4$



Obrázek 6.10: Předpovědi modelu SARIMA$(0,1,0)(1,1,0)_4$ (nahoře) a modelu SARIMA$(0,1,0)(0,1,1)_4$ (dole) pro rok 2015

Kapitola 7

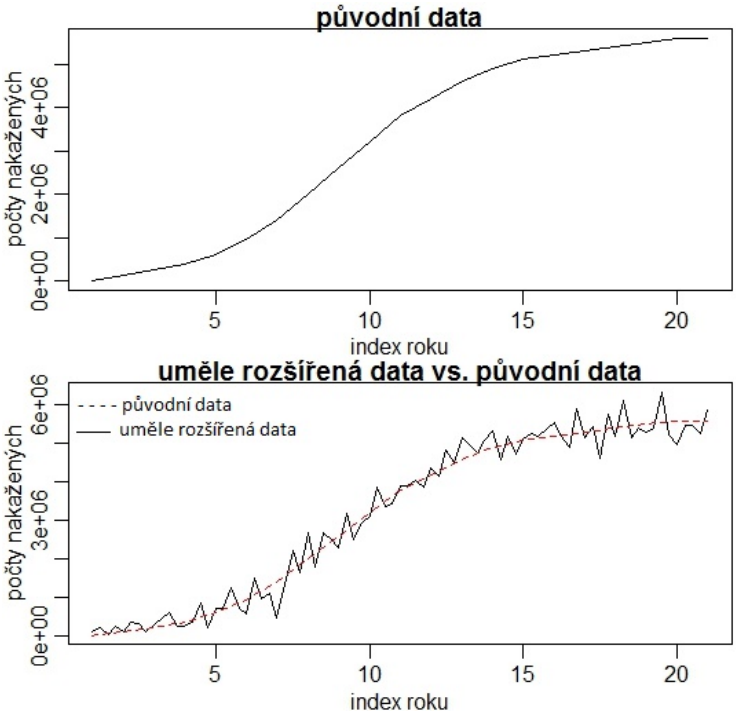
Analýza vývoje počtu nakažených lidí virem HIV v Jihoafrické republice

Pro modelování vývoje počtu nakažených lidí s HIV v Jihoafrické republice jsem ze zdroje [?] získala roční počty nakažených lidí virem HIV v období 1981-2009. Z důvodu nedostupnosti dat v roce 1982-1989 budu mít k dispozici pouze 21 pozorování. Data jsou vykreslena na obrázku 7.1

Pro modelování pomocí přístupu Boxe a Jenkinse je tento počet pozorování nedostačující. Proto původní data rozšířím pomocí lineární interpolace a následně je mírně "rozšumím"přičtením vygenerováných náhodných čísel pocházejících s normálního rozdělení s nulovou střední hodnotou a vhodně zvoleným rozptylem σ^2. Rozšířená data jsou vykreslena na obrázku 7.1. Pro modelování budu využívat uměle rozšířená data.

7.1. Modelování vývoje počtu nakažených lidí virem HIV pomocí dekompozičního přístupu

Tvar vývoje počtu nakažených lidí virem HIV (dle obrázku 7.1 má tvar S-křivky. To je charakteristické pro logistický trend.



Obrázek 7.1: Vývoj počtu nakažených lidí virem HIV v letech 1981-2009

Dle vzorce 2.31 a 2.30 získám modifikované odhadované hodnoty parametrů $\hat{k*}$ a $\hat{\beta}_0^*$ modelu 2.27. Následnou úpravou dle vzorce 2.32 a 2.29 získám odhadované hodnoty parametrů k, β_0 a β_1 uvedené společně s modifikovanými odhady v tabulce 7.1. Dosazením odhadnutých parametrů do vzorce 2.33 vypočtu odhadované hodnoty $\hat{y}_t$. Srovnání původních dat a odhadovaných hodnot ukazuje obrázek

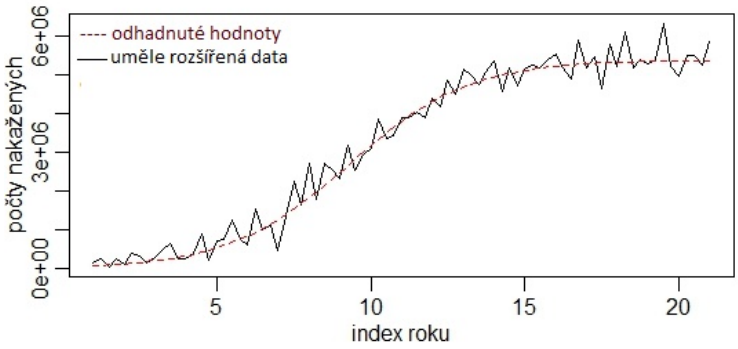
Parametr	Modifikovaný odhad	Odhad
β_0	$0,0007$	$3992,1$
β_1	$0,4007$	$0,4$
k	$1,8569E-O7$	5385226

Tabulka 7.1: Odhady parametrů logistického trendu

7.2. Odhadované hodnoty jsou shora omezené hodnotou parametru $k = 5385226$. Koeficientem dereminace (2.57) prověřím vhodnost zvoleného modelu. Hodnota koeficientu determinace je

$$R^2 = 1 - \frac{4062495898607}{2232107798690180} = 0,99,$$

kde ve zlomku v čitateli je uvedena hodnota RSČ vypočtena dle vzorce 2.56 a v jmenovateli hodnota celkové variability daného modelu vypočtena dle vzorce 2.58.

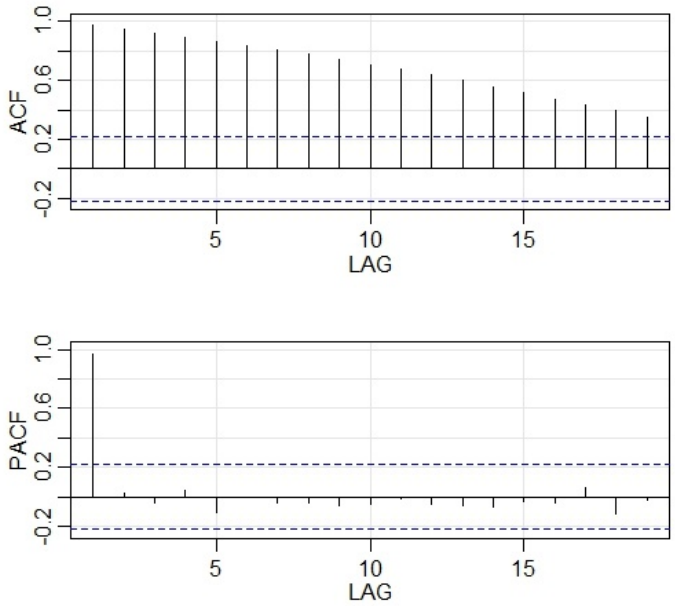


Obrázek 7.2: Srovnání původních dat a odhadnutých hodnot modelu s logistickým trendem

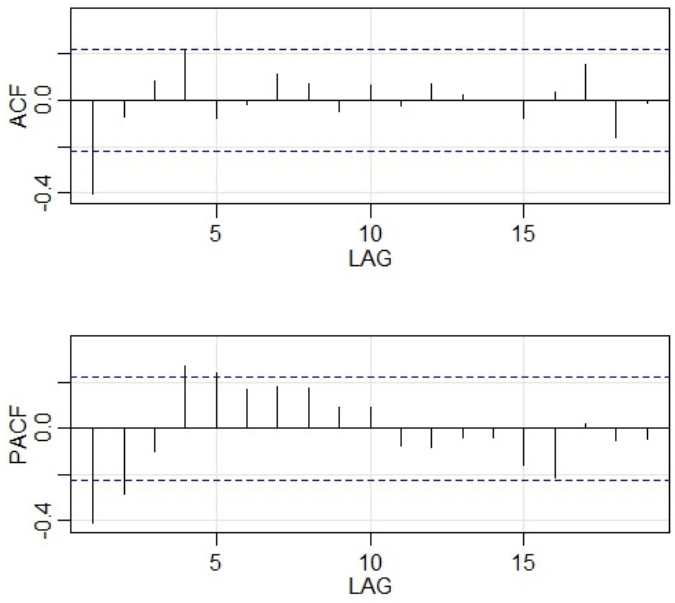
7.2. Modelování vývoje počtu nakažených lidí virem HIV pomocí přístupu Boxe a Jenkinse

Podle obrázku 7.1je vidět, že data nemají ustálené chování v čase. To značí porušení stacionarity. ACF vykreslena na obrázku 7.3 pomalu klesá. Pomocí diferencování se budu snažit data stacionarizovat. Nejmenší odhadovaný rozptyl získám pomocí klasického diferencování 1. řádu. Rozptyl původních dat (4.3e+12) pomocí diferencování 1. řádu klesl na 5.44e+10.

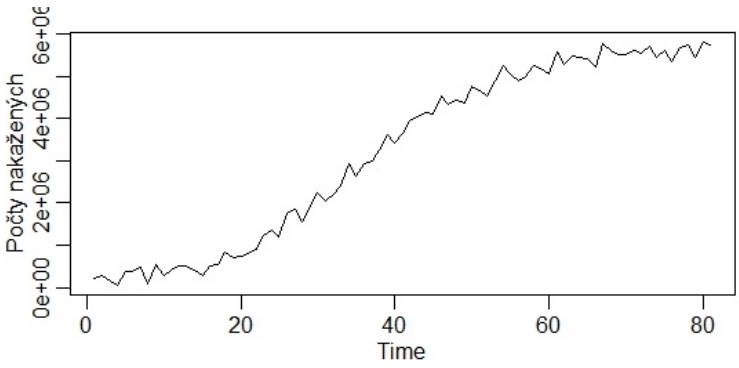
Příslušná ACF a PACF pro diferencovaná data je vykreslena na obrázku 7.4. ACF má významou první hodnotu, to napovídá o modelu MA(1). PACF má významné první dvě hodnoty, to napovídá o modelu AR(2). Popřípadě o smíšeném modelu ARMA(1,1). Nicméně vykreslím-li si diferencovaná data, zjistím, že diferencování 1. řádu neodstranilo nestacionaritu. Diferencovaná data dle obrázku 7.5 nemají stále ustálené chování v čase. Ani opětovným diferencováním nedosáhnu stacionarity. Z důvodu porušení stacionarity nemohu data modelovat pomocí přístupu Boxe a Jenkinse.



Obrázek 7.3: ACF a PACF uměle rozšířených dat



Obrázek 7.4: ACF a PACF pro diferencovaná data



Obrázek 7.5: Vývoj počtu nakažených (diferencovaná data)

Závěr

Cílem mé práce bylo nastudovat danou problematiku, najít vhodná data a složitější výpočty zpracovat v softwaru R. Teorie časových řad je rozebrána v prvních třech kapitolách, teoretické poznatky jsou poté demonstrovány na třech zvolených datových sadách uvedených v kapitole 5-7. Čtvrtá kapitola se věnuje zpracováním dat v softwaru R, kde byla naprogramována ACF, která by lépe vyhovovala mým požadavkům.

První časová řada popsaná v kapitole 5 popisuje časový vývoj teploty v ČR od roku 1961. Tato časová řada má lineárně sezónní vývoj. V modelování se ukazuje, že Boxův a Jenkinsův přístup nám dává srovnatelné výsledky jako dekompoziční přístup.

Druhá časová řada je rozebrána v kapitole 6 Vývoj tržeb maloobchodů. Zde se ukazuje, že Boxovy a Jenkinsovy modely si umí poradit v některých případech pomocí diferencování s rostoucími časovými řady. V tomto konkrétním případě jsme dokonce lepší výsledky obdrželi právě pomocí této metody.

V kapitole 7 je analyzována poslední zvolená časová řada popisující vývoj počtu nakažených virem HIV v Jihoafrické republice. Tato časová řada má tvar S-křivky. Taková časová řada je vhodná pro modelování logistického trendu pomocí dekompozičního přístupu. Zde se ukazuje, že pro modelování časových řad s logistickým trendem, nelze použít z důvodu porušení předpokladu stacionarity Boxův a Jenkinsův přístup. Stacionarity nedosáhneme ani diferencováním časové řady.

Závěrem mohu říct, že nebylo jednoduché najít data, v kterých by dekom-

poziční přístup vycházel lépe. V drtivé většině případů modelování dat jednoznačně lépe vycházel Boxův a Jenkinsův model.

Literatura

[1] Cipra, T.: Analýza časových řad s aplikacemi v ekonomii. 1. vydání. SNTL, 1986

[2] Arlt, J., Arltová, M.: Ekonomické časové řady. Professional Publishing, 2009

[3] Shumway,R., Stoffer, D.: Time series analysis and its applications with R examples. 3.vydání, Springer, 2011

[4] Arlt, J., Arltová, M.: Analýza ekonomických časových řad s příklady. Praha, 2002 http://nb.vse.cz/ arltova/vyuka/crsbir02.pdf

[5] HRON, Karel a Pavla KUNDEROVÁ. Základy počtu pravděpodobnosti a metod matematické statistiky. 1. vyd. Olomouc: Univerzita Palackého v Olomouci, 2013. ISBN 978-80-244-3396-7

[6] wikipedia [online]. 2011, [cit. 2016-03-30]. dostupné z: https://en.wikipedia.org/wiki/Ljung

[7] [online]. [cit. 2016-04-04]. dostupné z: https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc44a.htm.

[8] Přednášky katedry matematiky Jihočeské univerzity v Českých Budějovicích [online]. 2009, [cit. 2017-04-19]. dostupné z: http://www.pf.jcu.cz/stru/katedry/m/petraskova/crek-prednaska$_4$.*pdf*.