UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VYBRANÉ METÓDY PREDIKCIE ČASOVÝCH RADOV

DIPLOMOVÁ PRÁCA

2018 Bc. Ema LÖFFLEROVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VYBRANÉ METÓDY PREDIKCIE ČASOVÝCH RADOV

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	9.1.9. Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Soňa Kilianová, PhD.

Bratislava 2018 **Bc. Ema LÖFFLEROVÁ**

54036039

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Ema Löfflerová
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Vybrané metódy predikcie časových radov
Selected methods for time series prediction

Anotácia: Záujmom predstaviteľov mnohých inštitúcií v praxi je vedieť predikovať časové rady náhodných javov. K príkladom patria napríklad dáta rôznych ekonomických ukazovateľov, finančných hodnôt či prírodných javov. Existuje väčšie množstvo metód predikcie časových radov, založených na nástrojoch rôznych oblastí matematiky, ako napríklad štatistika či optimalizácia. V tejto práci budeme pozornosť venovať niektorým klasickým a moderným metódam a porovnáme ich efektívnosť.

Vedúci: Mgr. Soňa Kilianová, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, DrSc.

Dátum zadania: 26.01.2017

Dátum schválenia: 27.01.2017 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou by som sa chcela poďakovať svojej vedúcej diplomovej práce Mgr. Soni Kilianovej, PhD., za ochotu, pomoc a podnetné pripomienky pri písaní práce. Rovnako ďakujem mojej rodine a priateľom za trpezlivosť a podporu.

Abstrakt

LÖFFLEROVÁ, Ema: Vybrané metódy predikcie časových radov [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Soňa Kilianová, PhD., Bratislava, 2018, 73s.

V tejto diplomovej práci sa zaoberáme metódami, ktoré slúžia na predikciu časových radov. Ako prvú metódu si predstavíme analýzu singulárneho spektra, ktorá je založená na rozklade pôvodného časového radu na separovateľné zložky ako trend a sezónnosť. Ďalšou metódou budú ARMA modely a ako poslednú si uvedieme regresiu pomocou metódy oporných bodov. Ukážeme si použitie týchto metód, porovnáme ich a pozrieme sa, pre aké dáta dosahujú jednotlivé metódy najlepšie výsledky.

Kľúčové slová: časový rad, analýza singulárneho spektra, ARMA model, metóda oporných bodov

Abstract

LÖFFLEROVÁ, Ema: Selected methods for time series prediction [Master thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Soňa Kilianová, PhD., Bratislava, 2018, 73p.

In this thesis we deal with methods for time series prediction. Firstly we introduce singular spectrum analysis, which is based on decomposition of original time series into trend and seasonal component. The next method is ARMA models and as the last one we introduce regression using support vector machine. In the last section we apply these methods on our data, compare them and see, which method is suitable for the given data.

Key words: time series, singular spectrum analysis, ARMA model, support vector machine

Obsah

Úvod

Časový rad predstavuje postupnosť hodnôt, pozorovaní, zoradených chronologicky podľa času. Ich analyzovaním sa snažíme zistiť typické črty vývoja daných hodnôt, či už z dôvodu klasifikácie, detekcie nepravidelného vývoja alebo predikcie. V tejto práci sa zameriavame na porovnanie metód predikcie časových radov, ktoré sú založené na rôznych princípoch. K najzákladnejším metódam na popis ich štruktúry patrí ARMA model, ktorý sa skladá z dvoch častí - autoregresie (AR) a moving average (MA). Medzi prvé články, v ktorých boli autoregresné modely použité na dáta, patria [32] od G. U. Yule a [30] od G. Walkera, z rokov 1927 a 1931. V roku 1938 vyšla kniha [31] H. Wolda, kde bol predstavený a formalizovaný ARMA model pre stacionárne časové rady. G. E. P. Box a G. M. Jenkins vydali v roku 1970 knihu [3], ktorá obsahovala celý proces modelovania časového radu, spôsob odhadovania parametrov a metódy predikcie.

Menej známa technika analýzy časových radov je Analýza singulárneho spektra. Táto metóda sa prvýkrát objavila v článkoch [4], [5] od D. Broomheada a G. Kinga v 1986. Spája časti klasickej analýzy časových radov, viacrozmernej geometrie a spracovania signálov. Nezávisle od nich vznikla v bývalom Sovietskom zväze tzv. „húsenicová" SSA, ktorá je opísaná v knihách [8], [11].

Metódy strojového učenia sa stávajú čím ďalej, tým viac populárne. Spájajú oblasti umelej inteligencie a matematickej štatistiky. Na rozdiel od ARMA procesov dokážu metódy strojového učenia zachytiť aj nelineárne závislosti medzi dátami. Do tejto kategórie patria metódy, ktoré sú schopné učiť sa a následne sa zlepšovať na základe svojich predošlých skúseností ako napríklad rozhodovacie stromy, zhlukovanie, metóda hlavných komponentov, metóda oporných bodov, neurónové siete a ďalšie. Metóda oporných bodov bola predstavená v dnešnej podobe V. Vapnikom v publikácii [28] v roku 1995. Pôvodne bola určená na optické rozpoznávanie znakov [2], [6] a objektov [22]. V poslednom období sa do popredia dostávajú hlavne neurónové siete. Prvú umelú neurónovú sieť nazývanú Percepton vytvoril F. Rosenblatt v roku 1957 [18], [19], ktorá slúžila na rozpoznávanie a klasifikáciu objektov. V 90-tych rokov bola v článku [20] predstavená backpropagation metóda slúžiaca na výpočet váh v umelých neurónových sieťach. Ich použitie na predikciu časových radov môžeme vidieť v článkoch [1], [33].

V tejto práci si predstavíme 3 metódy - analýzu singularného spektra, ARMA modely a

metódu oporných bodov. Rozhodli sme vybrať tieto metódy kvôli ich odlišnostiam. ARMA modely predstavujú klasický prístup analýzy a predikcie časových radov. Zo strojového učenia sme sa rozhodli vybrať metódy oporného bodu, pri ktorých by nízky počet dát v trénovacej sade nemal prekážať. Analýza singulárneho spektra zas ponúka odlišný pohľad na predikciu, kde časový rad rozkladáme na pomocné časové rady a predikujeme ich hodnoty.

Naším hlavným cieľom a taktiež aj prínosom tejto diplomovej práce je podrobne spracovať teóriu k jednotlivým metódam, následne ich aplikovať na rôzne príklady a zistiť, ktoré metódy sú najpresnejšie pre ktoré typy dát. Členenie práce bude spôsobom striedania kapitol, v ktorých predstavíme a vysvetlíme metódy, s kapitolami, v ktorých aplikujeme dané metódy na dáta. Rozhodli sme sa zvoliť takéto členenie kvôli veľkej odlišnosti jednotlivých metód a pre lepšie porozumenie ich použitia.

1 Analýza singulárneho spektra

Analýza singulárneho spektra (SSA, z angl. *Singular spectrum analysis*) je technika na analýzu časových radov, ktorá kombinuje časti klasickej analýzy časových radov, viacrozmernej štatistiky, viacrozmernej geometrie a spracovania signálov. SSA sa snaží rozložiť pôvodný rad sa súčet niekoľkých interpretovateľných zložiek ako napríklad trend, oscilačná zložka a šum. Na jej použitie nie je potrebný predpoklad stacionarity časového radu [10].

Počiatok SSA sa zvykne spájať s vydaním článkov [4], [5] od Broomheada a Kinga. V tejto práci sa budeme zaoberať tzv. „húsenicovou" SSA, ktorá vznikla v bývalom Sovietskom zväze nezávisle od Broomheada a Kinga. Medzi knihy venované tejto metóde patria [11] a [8].

V tejto kapitole si predstavíme algoritmus základnej SSA, prejdeme si jednotlivé kroky a vysvetlíme si ich na ilustračných príkladoch. Ako zdroj sme použili knihy [10] a [11], ktoré sú zamerané na metodológiu SSA a vysvetľujú viaceré teoretické problémy, s ktorými sa SSA musí vysporiadať.

1.1 Algoritmus SSA

Nech $\mathbb{X}_N$ je reálny, nenulový časový rad, kde $N > 2$:

$$\mathbb{X}_N = (x_1, x_2, ..., x_N).$$

Číslo L $(1 < L < N)$ budeme nazývať dĺžka okna a K si zadefinujeme ako $K = N - L + 1$. Základná SSA je algoritmus, ktorý sa skladá z dvoch fáz - dekompozície a rekonštrukcie. Pri vysvetľovaní jednotlivých krokov, z ktorých sa skladajú tieto fázy, budeme postupovať podľa [10].

Dekompozícia

1.krok Vkladanie

V prvom kroku vezmeme pôvodný časový rad a vytvoríme postupnosť posunutých vektorov veľkosti L, kde $K = N - L + 1$ predstavuje ich počet:

$$X_i = (x_i, x_{i+1}, ..., x_{i+L-1})^T \quad (1 \leq i \leq K).$$

Potom maticou trajektórií $\mathbf{X}$ časového radu $\mathbb{X}$ budeme nazývať maticu, ktorej stĺpce sú tieto posunuté vektory $X_i, 1 \leq i \leq K$:

$$\mathbf{X} = [X_1 : ... : X_K] = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_K \\ x_2 & x_3 & x_4 & \dots & x_{K+1} \\ x_3 & x_4 & x_5 & \dots & x_{K+@} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \dots & x_N \end{bmatrix}$$

.

Príklad 1.1. Majme exponenciálnu funkciu $f(x) = 2^x$. Vezmeme si 4 hodnoty časového radu pre $x = \{-\frac{1}{2}; 0; \frac{1}{2}; 1\}$, t.j. $\mathbb{X}_4 = \{-\sqrt{2}, 1, \sqrt{2}, 2\}$. Matica trajektórií bude potom vyzerať ako:

$$\mathbf{X} = \begin{bmatrix} -\sqrt{2} & 1 & \sqrt{2} \\ 1 & \sqrt{2} & 2 \end{bmatrix}$$

2.krok Singulárny rozklad - SVD

V tomto kroku rozložíme maticu $\mathbf{X}$ pomocou singulárneho rozkladu (SVD, z angl. *Singular value decomposition*). Hodnoty $\lambda_1 \geq ... \geq \lambda_L$ označujú vlastné čísla matice $\mathbf{X}\mathbf{X}^T$ a teda $\sqrt{\lambda_1} \geq ... \geq \sqrt{\lambda_L}$ sú singulárne čísla matice $\mathbf{X}$. Vektory $u_1, ..., u_L$ sú prináležiace vlastné vektory matice $\mathbf{X}\mathbf{X}^T$ a zároveň sú ľavé singulárne vektory matice $\mathbf{X}$.

Označme si hodnosť matice $\mathbf{X}$ ako $d = rank(\mathbf{X}) = \max\{i,\ kde\ \lambda_i > 0\}$ a $v_i = \mathbf{X^T} u_i \frac{1}{\lambda_i}$, kde $i = 1, ..., d$. Tieto vektory nazývame pravé singulárne vektory matice $\mathbf{X}$. Potom SVD matice trajektórií $\mathbf{X}$ vyzerá ako:

$$\mathbf{X} = \mathbf{X_1} + ... + \mathbf{X_d}, \tag{1.1}$$

kde $\mathbf{X_i} = \sqrt{\lambda_i} u_i v_i^T$. Matice $\mathbf{X_i}$ sú hodnosti 1. Trojicu $(\sqrt{\lambda_i}, u_i, v_i)$ budeme nazývať vlastná trojica.

Príklad 1.2. Pre maticu trajektórií z Príkladu 1.1 si ukážeme, ako vypočítať jej vlastné trojice. Pre vlastné hodnoty a vlastné čísla matice

$$\mathbf{X}\mathbf{X}^T = \begin{bmatrix} -\sqrt{2} & 1 & \sqrt{2} \\ 1 & \sqrt{2} & 2 \end{bmatrix} \begin{bmatrix} -\sqrt{2} & 1 \\ 1 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix} = \begin{bmatrix} 5 & 2\sqrt{2} \\ 2\sqrt{2} & 7 \end{bmatrix}$$

platí

$$\mathbf{X}\mathbf{X}^T u = \lambda u.$$

Z toho dostávame vlastné čísla matice $\mathbf{X}\mathbf{X}^T \lambda_1 = 3$ a $\lambda_2 = 9$, t.j. singulárne čísla matice $\mathbf{X}$ sú $\sigma_1 = \sqrt{3}$ a $\sigma_2 = 3$. K nim prislúchajúce vlastné vektory sú $u_1 = \left(\frac{-\sqrt{2}}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)^T$ a $u_2 = \left(\frac{1}{\sqrt{3}}, \frac{\sqrt{2}}{\sqrt{3}}\right)^T$. Pravé singulárne vektory sú

$$v_1 = \frac{1}{\lambda_1}\mathbf{X}^{\mathbf{T}} u_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} -\sqrt{2} & 1 \\ 1 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix} \begin{bmatrix} \frac{-\sqrt{2}}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$v_2 = \frac{1}{\lambda_2}\mathbf{X}^{\mathbf{T}} u_2 = \frac{1}{3} \begin{bmatrix} -\sqrt{2} & 1 \\ 1 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{3}} \\ \frac{\sqrt{2}}{\sqrt{3}}. \end{bmatrix}$$

Dostali sme 2 vlastné trojice

$$\left(\sqrt{3}, \begin{bmatrix} \frac{-\sqrt{2}}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right),$$

$$\left(3, \begin{bmatrix} \frac{1}{\sqrt{3}} \\ \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix}, \begin{bmatrix} 0 \\ \frac{1}{\sqrt{3}} \\ \frac{\sqrt{2}}{\sqrt{3}} \end{bmatrix} \right).$$

Rekonštrukcia

3.krok Zoskupovanie vlastných trojíc

V tejto časti rozdelíme množinu indexov $1, ..., d$ do m disjunktných podmonožín $I_1, ..., I_m$. Následne dostávame rozklad

$$\mathbf{X} = \mathbf{X}_{I_1} + ... + \mathbf{X}_{I_m}.$$

Proces výberu týchto podmnožín $I_1, ..., I_m$ sa nazýva zoskupovanie vlastných trojíc a slúži na separovanie jednotlivých zložiek časového radu.

4.krok Diagonálne priemerovanie

V tomto kroku transformujeme každú maticu $\mathbf{X}_{I_j}$ do nového časového radu s dĺžkou N.

Majme maticu Y s rozmermi $L \times K$. Označíme si $L^\star = \min(L, K)$ a $K^\star = \max(L, K)$. Nech $y^\star_{ij} = y_{ij}$, ak $L < K$ a $y^\star_{ij} = y_{ji}$, ak $L > K$. Diagonálnym priemerovaním transformujeme maticu $\mathbf{Y}$ do časového radu $(y_1, ..., y_N)$

$$y_k = \begin{cases} \frac{1}{k} \sum\limits_{m=1}^{k} y^*_{m,k-m+1} & \text{pre } 1 \leq k < L^*, \\ \frac{1}{L^*} \sum\limits_{m=1}^{L^*} y^*_{m,k-m+1} & \text{pre } L^* \leq k \leq K^*, \\ \frac{1}{N-k+1} \sum\limits_{m=k-K+1}^{N-K^*+1} y^*_{m,k-m+1} & \text{pre } K^* < k \leq N. \end{cases} \tag{1.2}$$

Tento proces predstavuje priemerovanie prvkov matice $\mathbf{Y}$ po antidiagonálach, teda pre $k = 1$ dostaneme $y_1 = y_1$, pre $k = 2$ budeme mať $y_2 = (y_{1,2} + y_{2,1})/2$ atď.

Tento postup aplikujeme na $\mathbf{X}_{I_k}$ a dostaneme rekonštruovaný časový rad $\tilde{\mathbf{X}}^{(k)} = (\tilde{x}^{(k)}_1, ..., \tilde{x}^{(k)}_N)$. Pôvodný časový rad je rozložený do sumy m rekonštruovaných časových radov

$$x_n = \sum_{k=1}^{m} \tilde{x}^{(k)}_n \quad n = 1, ..., N.$$

1.2 Separovateľnosť

Hlavným cieľom SSA je rozloženie pôvodného časového radu na súčet niekoľkých radov, tak aby každý rad reprezentoval určitú štruktúru, ako napríklad trend, periodicitu alebo šum. Tento rozklad je využiteľný iba v prípade, že jednotlivé zložky sú separovateľné.

1.2.1 Slabá a silná separovateľnosť

Uvažujme SVD singulárny rozklad matice trajektórií $\mathbf{X}$ pôvodného časového radu $\mathbb{X}$ pre fixnú dĺžku okna L. Predpokladajme, že rad $\mathbb{X}$ je súčtom radov $\mathbb{X}^{(1)}$ a $\mathbb{X}^{(2)}$, t.j. $\mathbb{X} = \mathbb{X}^{(1)} + \mathbb{X}^{(2)}$.

V tomto prípade separovateľnosť radov $\mathbb{X}^{(1)}$ a $\mathbb{X}^{(2)}$ znamená, že súčet (1.1) v SVD matice trajektórií $\mathbf{X}$ môžeme rozdeliť do 2 skupín tak, že jednotlivé sumy budú tvoriť matice trajektórií $\mathbf{X}^{(1)}$ a $\mathbf{X}^{(2)}$ radov $\mathbb{X}^{(1)}$ a $\mathbb{X}^{(2)}$. Keďže SVD nemusí byť jednoznačne určené kvôli násobnosti singulárnych hodnôt, budeme rozlišovať 2 typy separovateľnosti - *slabú* a *silnú*. Rad budeme nazývať slabo separovateľným, ak existuje SVD matice trajektórií $\mathbf{X}$ taký, že jednotlivé prvky rozkladu budeme vedieť rozdeliť do 2 skupín tak, aby prvky v skupinách dávali matice trajektórií $\mathbf{X}^{(1)}$ a $\mathbf{X}^{(2)}$. To je ekvivalentné s podmienkou, že riadky a stĺpce matice trajektórií $\mathbf{X}^{(1)}$ a $\mathbf{X}^{(2)}$ sú ortogonálne. Separovateľnosť budeme nazývať silnou, ak takéto rozdelenie vieme spraviť pre každý SVD matice trajektórií.

1.2.2 Charakteristika separovateľnosti

Označme si $L^\star = \min(L, K)$ a $K^\star = \max(L, K)$. Váhy (1.3) predstavujú výskyt prvku x_i v matici trajektórií $\mathbf{X}$:

$$w_i = \begin{cases} i & \text{pre } 0 \le i < L^*, \\ L^* & \text{pre } L^* \le i \le K^*, \\ N - i + 1 & \text{pre } K^* < i \le N. \end{cases} \tag{1.3}$$

Zadefinujeme si súčin dvoch radov $\mathbb{X}^{(1)}$ a $\mathbb{X}^{(2)}$ dĺžky N ako

$$\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}\right)_w \stackrel{\text{def}}{=} \sum_{i=1}^{N} w_i x_i^{(1)} x_i^{(2)}. \tag{1.4}$$

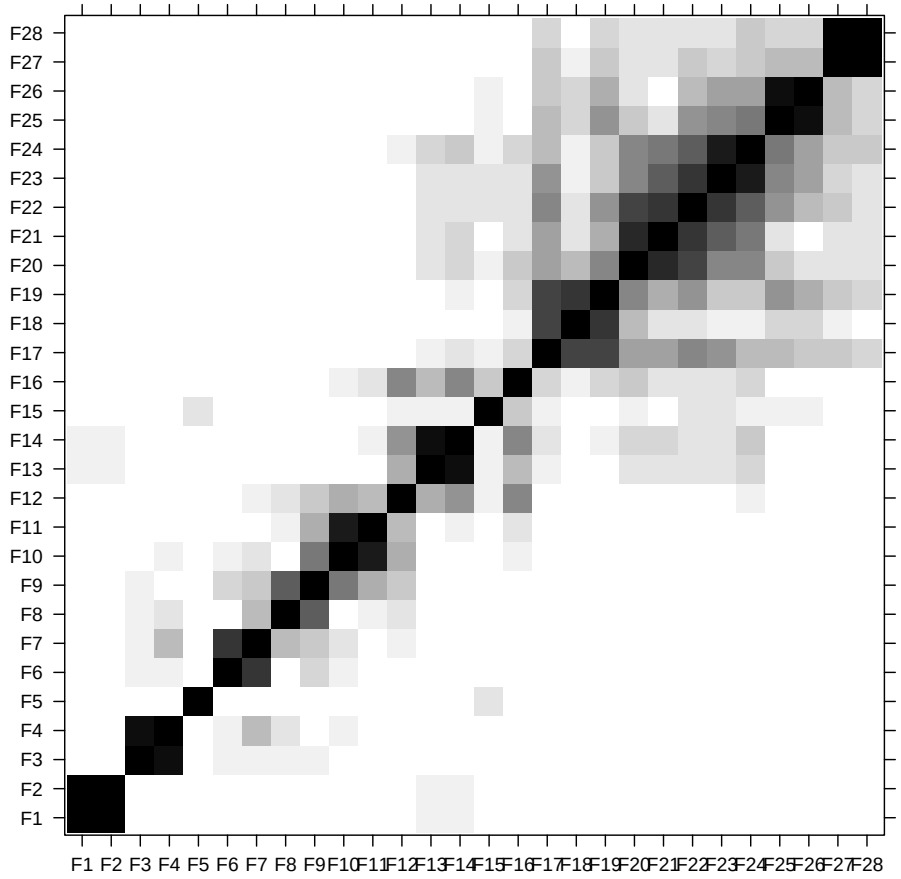
Rady $\mathbb{X}^{(1)}$ a $\mathbb{X}^{(2)}$ budeme nazývať *w-ortogonálne*, ak $\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}\right)_w = 0$.

Zavedieme si tzv. *w-koreláciu* na meranie približnej separovateľnosti dvoch radov $\mathbb{X}^{(1)}$ a $\mathbb{X}^{(2)}$

$$\rho^{(w)}\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}\right) \stackrel{\text{def}}{=} \frac{\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}\right)_w}{\|\mathbb{X}^{(1)}\|_w \|\mathbb{X}^{(2)}\|_w}, \tag{1.5}$$

kde $\|\mathbb{X}^{(1)}\|_w = \sqrt{(\mathbb{X}^{(1)}, \mathbb{X}^{(1)})_w}$. Ak $\rho^{(w)}\left(\mathbb{X}^{(1)}, \mathbb{X}^{(2)}\right) \simeq 0$, tak budeme hovoriť, že rady $\mathbb{X}^{(1)}$ a $\mathbb{X}^{(2)}$ sú približne separovateľné.

Na obrázku 1.1 je zobrazená matica w-korelácií 28 časových radov. Môžeme si všimnúť, že pri prvých dvoch časových radoch je w-korelácia znázornená čiernou farbou, t.j. má hodnotu 1, a teda tieto dva časové rady nie sú separovateľné od seba navzájom. Naopak vidíme, že w-korelácie týchto dvoch časových radov s ostatnými sú takmer nulové.



Obr. 1.1: Matica w-korelácií, biela - nulová w-korelácia, čierna - jednotková w-korelácia

1.3 Predikcia

V SSA predikciách sú predikčné modely popísané pomocou *lineárnych rekurentných vzťahov* (LRR). Predstavíme si definíciu LRR a ich súvis s časovými radmi, ktoré sme prebrali z [10].

1.3.1 Lineárne rekurentné vzťahy

Definícia 1.3. [10] Časový rad $\mathbb{S}_N = \{s_i\}_{i=1}^N$ sa riadi lineárnymi rekurentnými vzťahmi (LRR z angl. linear recurent relations), ak exisujú $a_1, a_2, ..., a_t$ také, že

$$s_{i+t} = \sum_{k=1}^{t} a_k s_{i+t-k}, \quad 1 \leq i \leq N-t, a_t \neq 0, t < N. \tag{1.6}$$

Číslo t sa nazýva stupeň LRR a $a_1, ..., a_t$ sú koeficienty LRR.

Definícia 1.4. [10] Polynóm $P_t(\mu) = \mu^t - \sum_{k=1}^{t} a_k \mu^{t-k}$ nazývame *charakteristický polynóm* LRR (1.6).

Nech časový rad $S_\infty = (s_1, ..., s_n, ...)$ spĺňa LRR (1.6) pre $a_t \neq 0$ a $i \geq 1$. Uvažujme charakteristický polynóm LRR a označme si rôzne (komplexné) korene $\mu_1, ..., \mu_p$, kde $1 \leq p \leq t$. Všetky tieto korene sú nenulové, keďže $a_t \neq 0$. Násobnosť koreňa μ_m označíme k_m, kde $1 \leq m \leq p$ a $k_1 + ... + k_p = t$. Nasledujúce tvrdenie poskytuje všeobecný tvar radu spĺňajúceho LRR.

Veta 1.5. [10] Časový rad $S_\infty = (s_1, ..., s_n, ...)$ spĺňa LRR (1.6) pre všetky $i \geq 0$ vtedy a len vtedy, ak

$$s_n = \sum_{m=1}^{p} \left(\sum_{j=0}^{k_m - 1} c_{mj} n^j \right) \mu_m^n, \tag{1.7}$$

kde komplexné koeficienty c_{mj} závisia od prvých t bodov $s_1, ..., s_t$.

1.3.2 Formalizácia problému predikcie

Uvažujme rad $\mathbb{X}_N = \mathbb{X}_N^{(1)} + \mathbb{X}_N^{(2)}$, v ktorom chceme predikovať $\mathbb{X}_N^{(1)}$. Ak je časový rad hodnosti $r < L$, potom generuje nejaký L-trajekčný podpriestor dimenzie r. Tento podpriestor odráža štruktúru $\mathbb{X}_N^{(1)}$, a preto ho môžeme brať ako bázu pre predikciu. Predikcia v rámci tohto podpriestoru znamená pokračovanie L-posunutých vektorov predikovaného

radu tak, aby ležali v alebo veľmi blízko vybraného podpriestoru z $\mathbb{R}^L$ [10].

Vstupy do predikčného algoritmu sú:

(a) časový rad $\mathbb{X}_N = (x_1, ..., x_N), N > 2$,

(b) dĺžka okna $L, 1 < L < N$,

(c) lineárny priestor $\mathcal{L}_r \subset \mathbb{R}^L$. Predpokladáme, že $e_L \notin \mathcal{L}_r$, kde $e_L = (0, 0, ..., 0, 1)^T \in \mathbb{R}^L$; inak povedané, $\mathcal{L}_r$ nie je vertikálny priestor,

(d) číslo M predstavuje počet predikovaných hodnôt.

Pre lepšiu orientáciu si zavedieme nasledovné značenie:

(a) $\mathbf{X} = [X_1 : ... : X_K]$ (pre $K = N - L + 1$) je matica trajektórií radu $\mathbb{X}_N$,

(b) $P_1, ..., P_r$ je ortonormálna báza $\mathcal{L}_r$,

(c) $\widehat{\mathbf{X}} \stackrel{\text{def}}{=} [\widehat{X}_1, ..., \widehat{X}_K] = \sum_{i=1}^{r} P_i P_i^T \mathbf{X}$. Vektor $\widehat{X}_i$ je ortogornálnou projekciou X_i do priestoru $\mathcal{L}_r$,

(d) $\widetilde{\mathbf{X}} = \mathcal{H}\widetilde{\mathbf{X}} = [\widetilde{X}_1 : ... : \widetilde{X}_K]$ je výsledok hankelizácie matice $\widehat{\mathbf{X}}$. Matica $\widetilde{\mathbf{X}}$ je matica trajektórií nejakého radu $\widetilde{\mathbb{X}}_N = (\tilde{x}_1, ..., \tilde{x}_N)$,

(e) pre akýkoľvek vektor $Y \in \mathbb{R}^L$ budeme označovať $\bar{Y} \in \mathbb{R}^{L-1}$ vektor tvorený poslednými $L - 1$ zložkami vektora Y a naopak $\underline{Y} \in \mathbb{R}^{L-1}$ bude vektor pozostávajúci z prvých $L - 1$ zložiek Y,

(f) zavedieme si $\nu^2 = \pi_1^2 + ... + \pi_r^2$, kde π_i označuje poslednú zložku vektora $P_i (i = 1, ..., r)$. Pretože ν^2 je štvorec kosínu uhla medzi vektorom e_L a lineárnym priestorom $\mathcal{L}_r$, budeme ho nazývať *koeficient vertikálnosti* $\mathcal{L}_r$. Keďže $e_L \notin \mathcal{L}_r$, tak $\nu^2 < 1$.

Uvedieme si tvrdenie z [12], ktorého špeciálny prípad využijeme pri definovaní algoritmu predikcie.

Zavedieme si pojem obmedzenie vektora $X = (x_1, ..., x_n)^T \in \mathbb{R}^n$ na množinu indexov $\mathcal{S}$, čo predstavuje vektor $X|_{\mathcal{S}} = (x_{i_1}, ..., x_{i_{|\mathcal{S}|}})^T \in \mathcal{R}^{|\mathcal{S}|}$, kde $\mathcal{S} = \{i_1, ..., i_{|\mathcal{S}|}\}$. Obmedzenie

matice na množinu indexov je matica, pozostávajúca z obmedzení stĺpcov na túto množinu. V nasledujúcom tvrdení si ukážeme vzťah na výpočet obmedzenia vektora $X|_{\mathcal{P}}$, kde vektor $X \in \mathcal{L}_r$, pomocou ortonormálnej bázy $P_1, ..., P_r$ a $X|_{\mathcal{I}\setminus\mathcal{P}}$.

Tvrdenie 1.6. [12] Nech $\mathcal{P}$ je usporiadaná množina indexov a $R = [P_1 : ... : P_r]$. Nech matica $I_| - R|_{\mathcal{P}}(R|_{\mathcal{P}})^T$ je regulárna matica. Potom pre každý vektor $X \in \mathcal{L}_r$ je nasledovný vzťah pre $X|_{\mathcal{P}}$ platný:

$$X|_{\mathcal{P}} = (I_| - R|_{\mathcal{P}}(R|_{\mathcal{P}})^T)^{-1} R|_{\mathcal{P}}(R|_{\mathcal{I}\setminus\mathcal{P}})^T X|_{\mathcal{I}\setminus\mathcal{P}}. \tag{1.8}$$

Dôkaz. Pre jednoduchosť značenia, nech $\mathcal{P} = \{1, ..., |\mathcal{P}|\}$. Označíme si $X_1 = X|_{\mathcal{P}}$, $X_2 = X|_{\mathcal{I}\setminus\mathcal{P}}, R_1 = R|_{\mathcal{P}}, R_2 = R|_{\mathcal{I}\setminus\mathcal{P}}$. Keďže $P_1, ..., P_r$ je ortonormálna báza $\mathcal{L}_r$, tak $RR^T X = X$ pre $X \in \mathcal{L}_r$. $RR^T =$ si môžeme zapísať podľa nášho zavedeného značenia ako:

$$RR^T = \begin{bmatrix} R_1 R_1^T & R_1 R_2^T \\ R_2 R_1^T & R_2 R_2^T \end{bmatrix}.$$

Pre vektor X_1 máme

$$X_1 = R_1 R_1^T X_2 + R_1 R_2^T X_2.$$

V originálnom značení dostávame rovnicu

$$X|_{\mathcal{P}} = (I_| - R|_{\mathcal{P}}(R|_{\mathcal{P}})^T)^{-1} R|_{\mathcal{P}}(R|_{\mathcal{I}\setminus\mathcal{P}})^T X|_{\mathcal{I}\setminus\mathcal{P}},$$

čo je vzťah, ktorý sme chceli dokázať. □

Tvrdenie 1.6 má priamy súvis s predikciou pomocou SSA. Keď v 3.kroku SSA zoskupujeme vlastné trojice, tak pri diagonálnom priemerovaní a následnej rekonštrukcii časového radu pracujeme ďalej samostatne s jednotlivými rekonštruovanými radmi. Pre rad tvorený vlastnými trojicami s indexami z množiny I_j dostávame priestor $\mathcal{L}_r$. To znamená, že pre špeciálny prípad Tvrdenia 1.6, kde $\mathcal{P} = \{L\}$ dostávame vzťah ako vypočítať posledný prvok ľubovoľného vektora $Y \in \mathcal{L}_r$. Inak povedané, ako predikovať novú hodnotu časového radu pomocou $L - 1$ predchádzajúcich, tak aby výsledný časový rad, resp. vektor patril do priestoru $\mathcal{L}_r$.

Tvrdenie 1.7. [10] Posledná zložka y_L ľubovoľného vektora $Y = (y_1, ..., y_L)^T \in \mathcal{L}_r$ je lineárnou kombináciou prvých zložiek $y_1, ..., y_{L-1}$:

$$y_L = a_1 y_{L-1} + a_2 y_{L-2} + ... + a_{L-1} y_1,$$

kde vektor $R = (a_{L-1}, ..., a_1)^T$ môžeme vyjadriť ako

$$R = \frac{1}{1-\nu^2} \sum_{i=1}^{r} \pi_i \underline{P_i} \tag{1.9}$$

a nezávisí od výberu bázy $P_1, ..., P_r$ v lineárnom priestore $\mathcal{L}_r$.

1.3.3 Rekurentná predikcia

Pomocou Tvrdenia 1.7 vieme vypočítať jednu nasledujúcu predikovanú hodnotu a teda algoritmus na rekurentné predikovanie (R-predikcia) je sformulovaný nasledovne [10]:
1. Časový rad $\mathbb{Y}_{N+M} = (y_1, ..., y_{N+M})$ je definovaný ako:

$$y_i = \begin{cases} \tilde{x}_i & \text{pre } i = 1, ..., N, \\ \sum\limits_{j=1}^{L-1} a_j y_{i-j} & \text{pre } i = N+1, ..., N+M. \end{cases} \tag{1.10}$$

2. Hodnoty $y_{N+1}, ..., y_{N+M}$ tvoria M členov rekurentnej predikcie.

R-predikcia je teda počítaná priamo pomocou LRR s koeficientami $\{a_j, j = 1, ..., L-1\}$ z Tvrdenia 1.7.

2 Príklady použitia SSA

V tejto kapitole si ukážeme použitie SSA na dátach s rôznou štruktúrou. Jednotlivé príklady môžeme zatriediť do 3 kategórií, t.j. dáta s trendom, s periodickou zložkou a dáta bez trendu a periodickej zložky. Prejdeme si kroky SSA rozkladu a taktiež samotnú rekurentnú predikciu, ktorej aplikácia v softvéri R je opísaná v [13]. Dáta sme čerpali z verejne prístupných internetových databáz.

Kvalitu predikcie jednotlivých metód budeme merať pomocou chyby $NRMSE$ (z angl. normalized root mean square error)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad NRMSE = \frac{RMSE}{y_{max} - y_{min}},$$

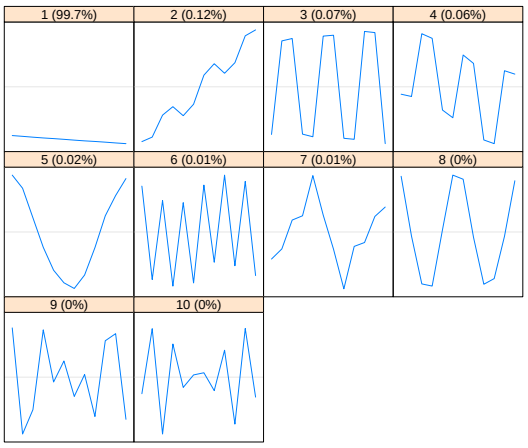
kde $\hat{y}_i$ sú predikované a y_i skutočné hodnoty.

2.1 HDP Slovenska

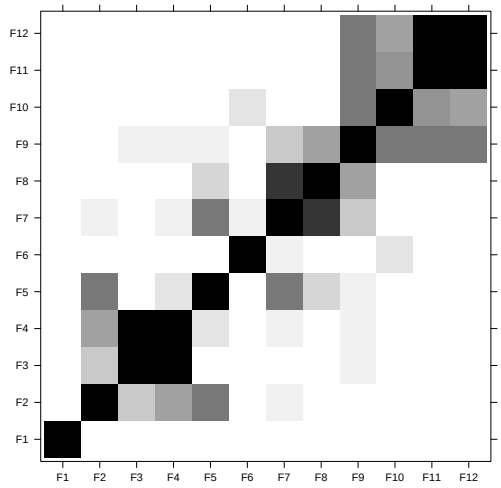
Ako prvý príklad sme si zvolili mesačné hodnoty HDP Slovenska, ktoré sme získali z makroekonomickej databázy NBS [14]. Jedná sa o sezónne neočistené dáta v miliónoch Eur od 1Q 1995 po 4Q 2014, t.j. $N = 80$. Na Obrázku 2.1 vidíme priebeh HDP Slovenska. Z grafu si môžeme všimnúť určitú periodicitu, ale taktiež rastúci trend.

Obr. 2.1: Kvartálne hodnoty HDP Slovenska v miliónoch EUR v rokoch 1995 - 2014 [14]

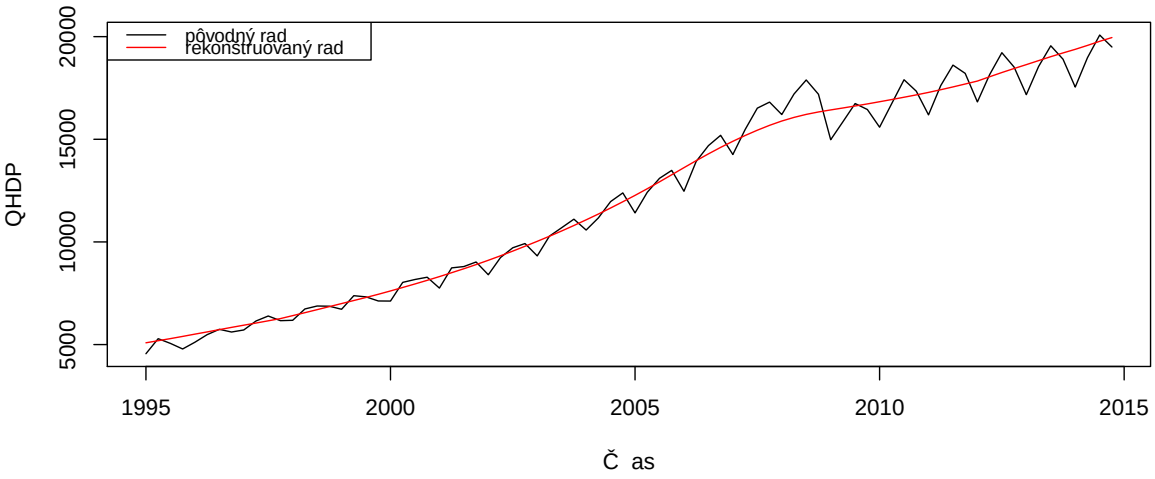
Našou úlohou je správne zvoliť parameter L, tak aby sa dali jednotlivé zložky časového radu odseparovať. Zároveň vieme, že by to mal byť násobok čísla 4, keďže v prípade kvartálnych dát číslo 4 predstavuje prirodzenú periódu a preto by sme chceli, aby bola zachytená aj v posunutých časových radoch.



(a) Vlastné vektory z SVD, $L = 12$



(b) w-korelačná matica, $L = 12$

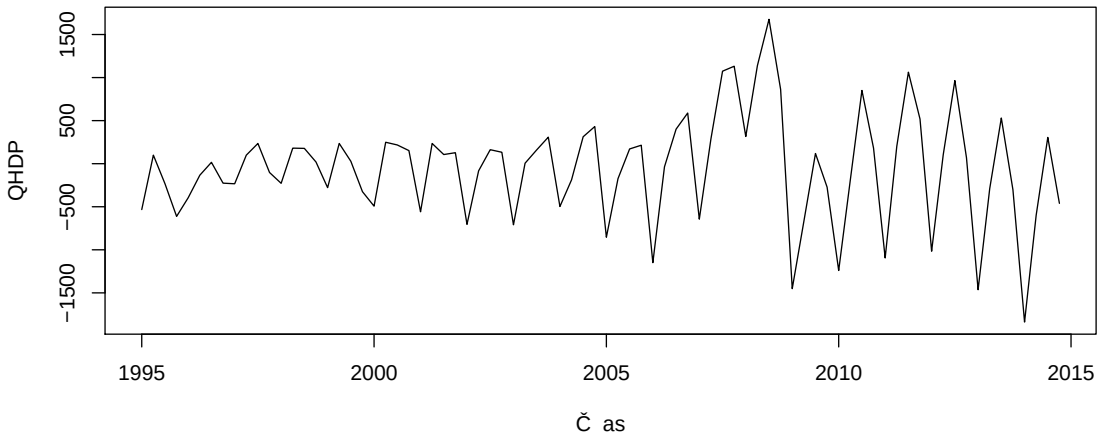


(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

Obr. 2.2: Rekonštrukcia časového radu pomocou prvej vlastnej trojice pre $L = 12$

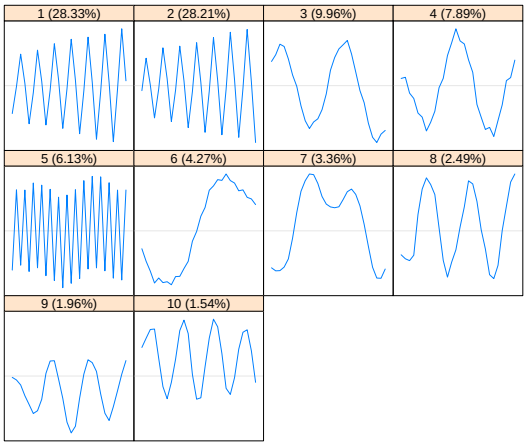
Zvolili sme si $L = 12$. Na Obrázku 2.2a sú zobrazené časové rady generované jednotlivými vlastnými vektormi z SVD (2.krok SSA). Prvý vlastný vektor, ako jediný reprezentujúci rastúci trend, prispieva najväčšou váhou do nášho pôvodného časového radu. Napravo v Obrázku 2.2b vidíme, že jeho w-korelácie s ostatnými čiastkovými časovými

radmi sú prakticky nulové. Preto sme sa rozhodli vybrať prvú vlastnú trojicu ako samostatnú skupinu (3.krok SSA) a dostávame prvú množinu indexov $I_1 = \{1\}$. Časový rad rekonštruovaný pomocou prvej vlastnej trojice (4.krok SSA) je zobrazený na Obrázku 2.2c červenou farbou, zatiaľ čo pôvodný časový rad je čiernou farbou.

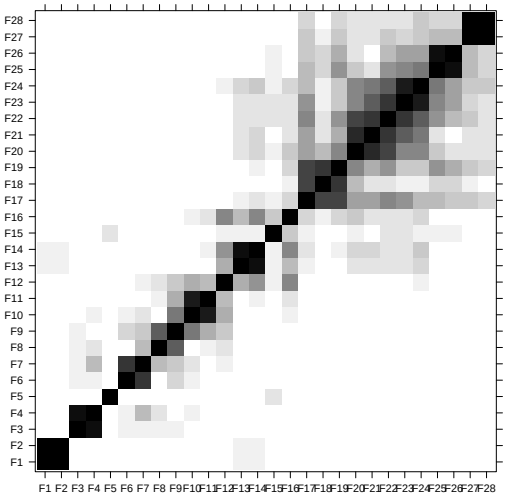


Obr. 2.3: Pôvodný časový rad bez rastúceho trendu

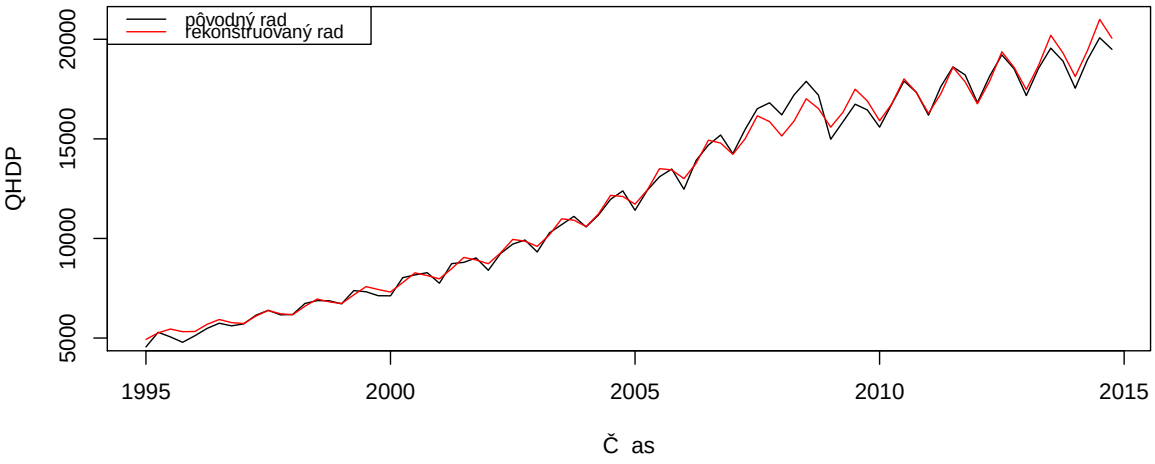
Po vybratí trendu z pôvodného časového radu ďalej pracujeme už len s časovým radom, ktorý je na Obrázku 2.3. Tentokrát sme sa rozhodli zvoliť veľkosť okna $L = 28$. Na Obrázku 2.4a vidíme, že prvé 2 vlastné vektory reprezentujú rovnakú periódu a spolu tvoria takmer 60% nášho ostávajúceho časového radu. Rovnako aj vo w-korelačnej matici (Obr. 2.4b) vidíme, že ich w-korelácie so sebou navzájom sú takmer 1 a s ostatnými časovými radmi až na 13. a 14. sú nulové. Prvá a druhá vlastná trojica budú reprezentovať periodickú zložku.



(a) Vlastné vektory z SVD, $L = 28$



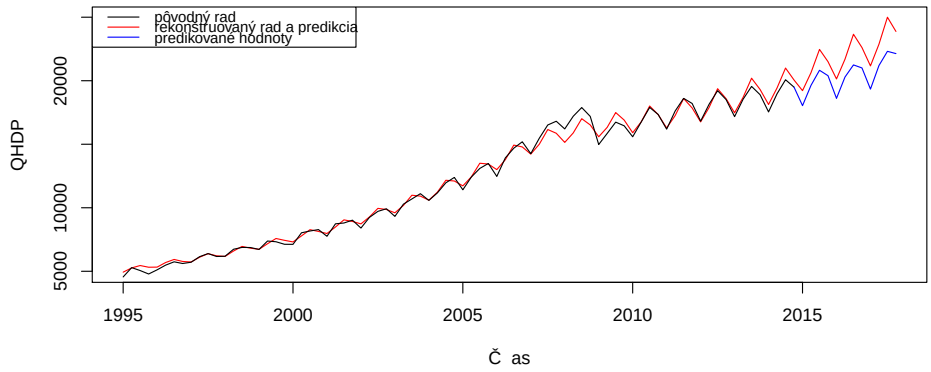
(b) w-korelačná matica, $L = 28$



(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

Obr. 2.4: Rekonštrukcia časového radu pomocou 3 vlastných trojíc

Po sčítaní časových radov predstavujúcich trend a periodickú zložku dostávame časový rad vyznačený červenou farbou na Obrázku 2.4c, ktorého nasledujúce hodnoty budeme predikovať. Rozdelenie pôvodného časového radu na sumu časových radov, kde každý reprezentuje určitú štruktúru, je dôležité pri predikcii. Ak sme vyberali tieto časové rady tak, aby boli od seba čo najviac separovateľné, tak predikcie jednotlivých časových radov budú presnejšie.



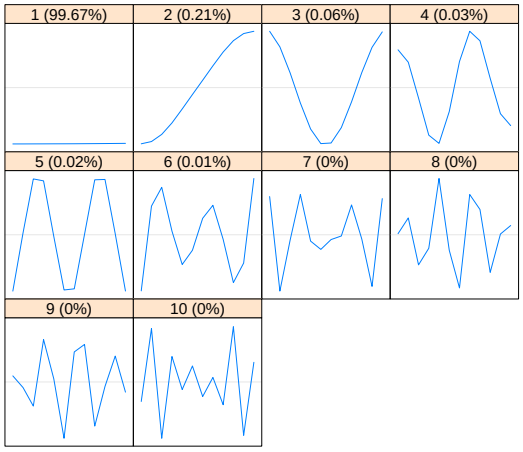
Obr. 2.5: Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená) a jeho predikované hodnoty pre roky 2015 - 2017, reálne hodnoty pre roky 2015 - 2017 (modrá)

Obrázok 2.5 zobrazuje pôvodný časový rad čiernou farbou, reálne hodnoty HDP pre roky 2015 - 2017 modrou farbou a červenou rekonštruovaný časový rad s jeho predikovanými hodnotami na roky 2016 - 2017. Priemerná chyba pri 12 predikovaných hodnotách nám vyšla 8%. Chyba metódy pri predikcách nám vyšla $NRMSE = 0,4016775$, zatiaľ čo pri pôvodnom časovom rade bola len $0,02512148$.

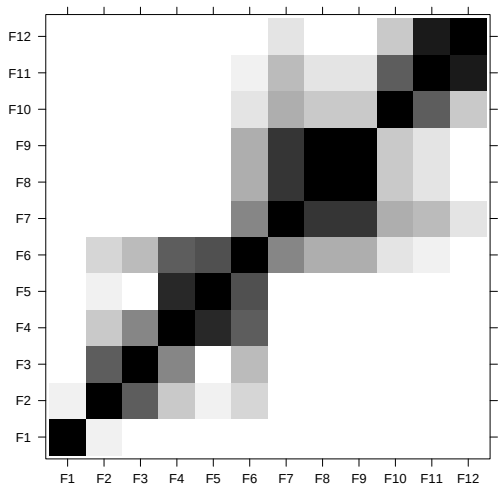
2.2 Miera nezamestnanosti

Obr. 2.6: Mesačné hodnoty miery nezamestnanosti Slovenska v rokoch 1997 - 2016 [14]

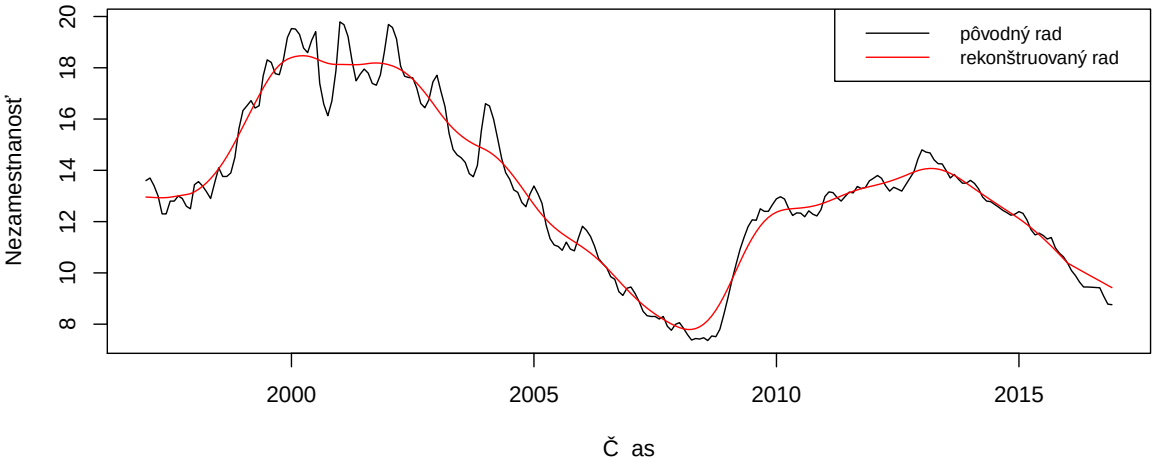
V tomto príklade používame mesačné dáta evidovanej miery nezamestnanosti Slovenska pochádzajúce z Úradu práce, sociálnych vecí a rodiny, ktoré sú dostupné na [14]. Hodnoty sú vyjadrené v % za obdobie od januára 1997 do decembra 2016. Počet dát je $N = 240$. Na obrázku 2.6 na rozdiel od predchádzajúceho príkladu, v tomto grafe nevidíme žiaden trend alebo periodicitu.



(a) Vlastné vektory z SVD, $L = 12$



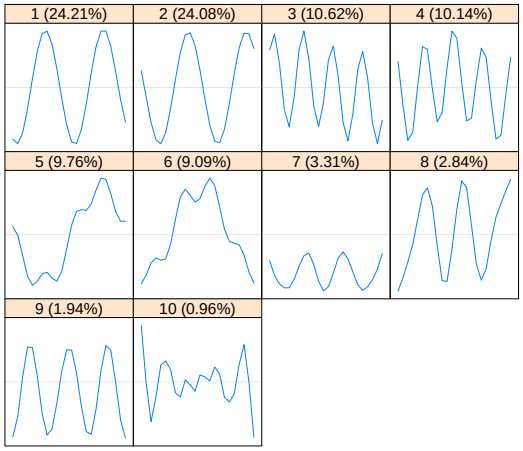
(b) w-korelačná matica, $L = 12$



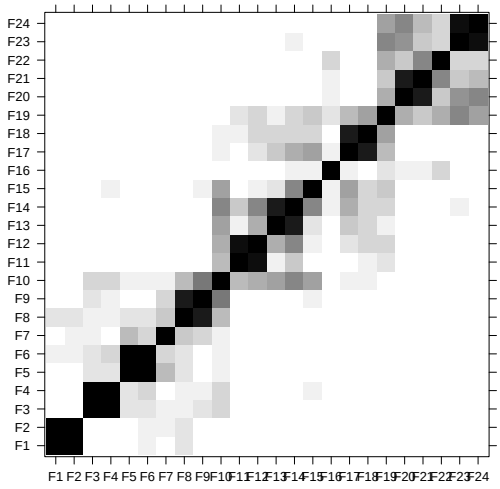
(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

Obr. 2.7: Rekonštrukcia časového radu pomocou prvej vlastnej trojice pre $L = 12$

Rozhodli sme sa zvoliť dĺžku okna $L = 12$, keďže pracujeme s mesačnými dátami. Pre takto zvolený parameter vidíme na Obrázku 2.7b, že časový rad generovaný prvou vlastnou trojicou má nulovú w-koreláciu so všetkými zvyšnými časovými radmi okrem druhého. Zároveň predstavuje viac ako 99% pôvodného časového radu a preto si prvú vlastnú trojicu vyberieme ako prvú množinu indexov. Rekonštruovaný časový rad pomocou jednej vlastnej trojice vidíme na obrázku 2.7c.



(a) Vlastné vektory z SVD, $L = 24$



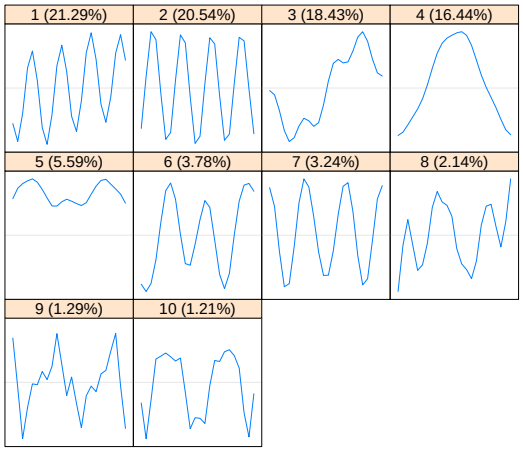
(b) w-korelačná matica, $L = 24$

(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

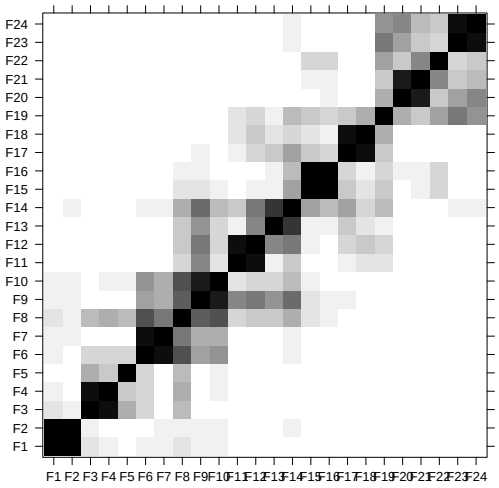
Obr. 2.8: Rekonštrukcia časového radu pomocou 3 vlastných trojíc

V ďalšom kroku sme zvolili parameter $L = 24$. Príslušná w-korelačná matica sa aj s vlastnými vektormi nachádza na Obrázkoch 2.8a a 2.8b. Tentokrát prvé 2 časové rady majú jednotkovú w-koreláciu medzi sebou a spolu reprezentujú až 48% zvyšného časového radu. Na Obrázku 2.8c vidíme časový rad rekonštruovaný pomocou troch vlastných trojíc.

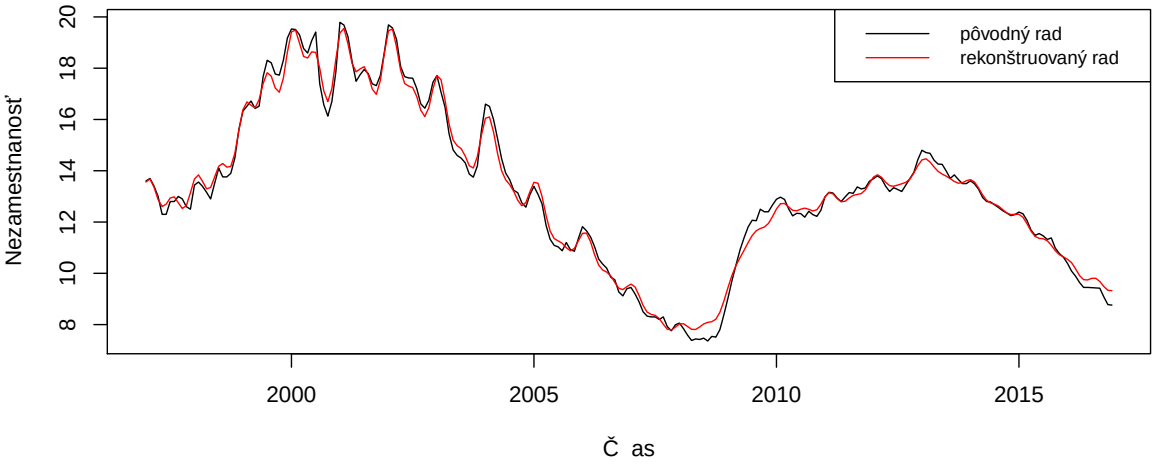
Rozhodli sme sa pozrieť ešte raz na rozklad ostávajúceho časového radu. Dĺžku okna sme nastavili na $L = 24$. Rovnako ako v predchádzajúcom rozklade máme silnú w-koreláciu medzi prvou a druhou vlastnou trojicou (Obrázok 2.9b), ktoré spolu tvoria 41% zvyšného časového radu.



(a) Vlastné vektory z SVD, $L = 24$



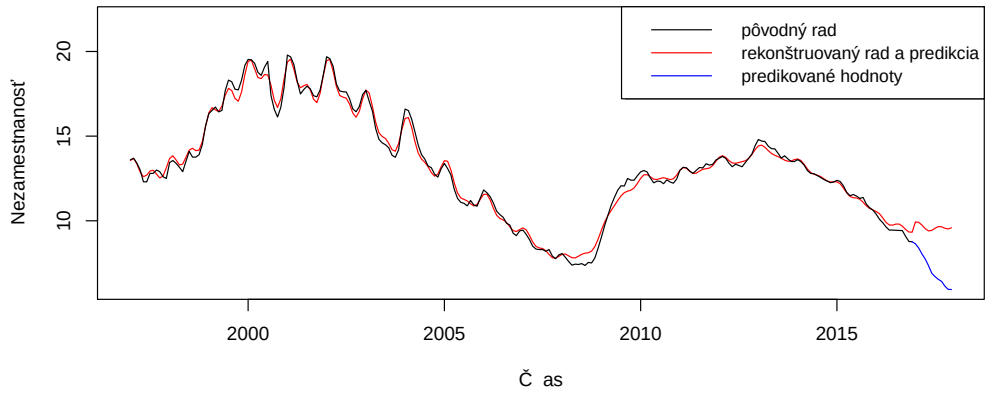
(b) w-korelačná matica, $L = 24$



(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

Obr. 2.9: Rekonštrukcia časového radu pomocou 5 vlastných trojíc

Rekonštruovaný časový rad sa nachádza na Obrázku 2.9c červenou farbou a pôvodný časový rad čiernou. Na obrázku 2.10 máme modrou vyznačené skutočné hodnoty miery nezamestnanosti a červenou naše predikcie. Predikovali sme 12 hodnôt, t.j. celý rok 2017. Chyba metódy NRMSE pri pôvodnom časovom rade vyšla $0,02449586$ a pri predikcii $0,9958386$. Skutočná miera nezamestnanosti pokračovala v roku 2017 v klesajúcom trende a dostala sa na historické minumum $5,94\%$. Keďže najväčšie zastúpenie v rekonštruovanom časovom rade mala prvá vlastná trojica predstavujúca takmer konštatný priebeh, aj predikcia sa udržiava približne na rovnakej hodnote a teda sa nezhoduje s realitou.

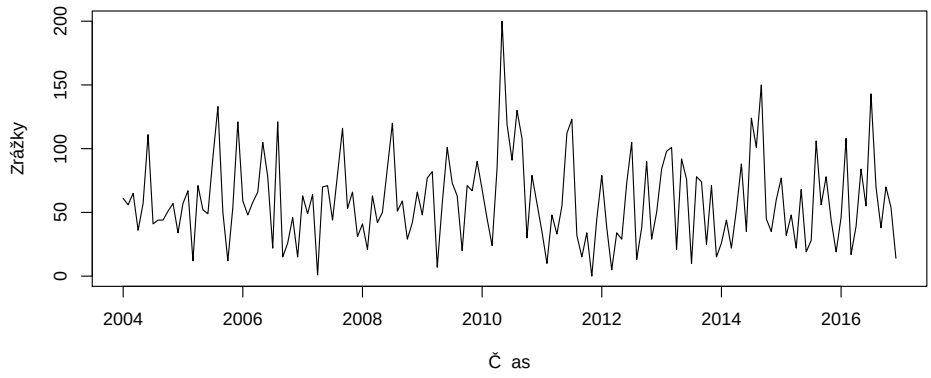


Obr. 2.10: Reálne hodnoty (čierna), predikované hodnoty (modrá), rekonštruovaný rad a predikcia (červená)

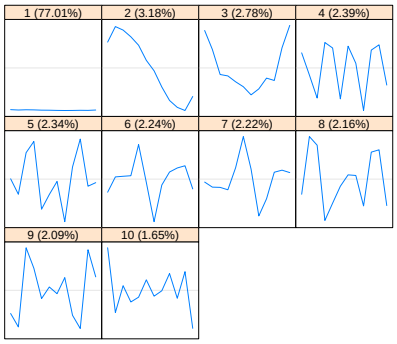
2.3 Priemerný úhrn zrážok

Ako ďalší príklad sme zvolili dáta predstavujúce priemerný mesačný úhrn zrážok v [mm] na západnom Slovensku od januára 2004 do decembra 2016. Dáta pochádzajú z [17] a boli spracované P. Faškom zo SHMÚ. Počet hodnôt je 156 a ich priebeh môžeme sledovať na Obrázku 2.11. Dáta nevykazujú nejaký výrazný trend, držia sa okolo priemernej hodnoty $59,25$ mm. Taktiež nevidno žiadnu sezónnosť, cyklus, ktorý by sa opakoval.

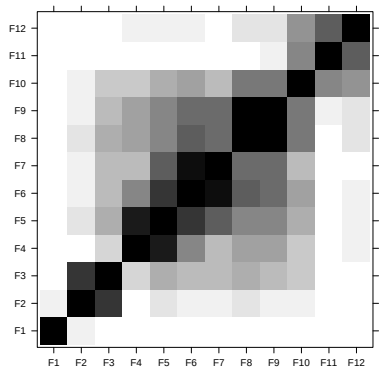
Parameter L sme si zvolili 12, keďže máme mesačné dáta a teda hodnota 12 predstavuje prirodzenú periódu. Na Obrázku 2.12a vidíme, že prvý vlastný vektor reprezentuje mierne klesajúci trend a reprezentuje vyše 70% pôvodného časového radu. W-korelácia (Obr. 2.12b) časového radu tvoreného prvou vlastnou trojicou s ostatnými časovými radmi je takmer nulová a preto prvú vlastnú trojicu vyberieme ako samostatnú skupinu.



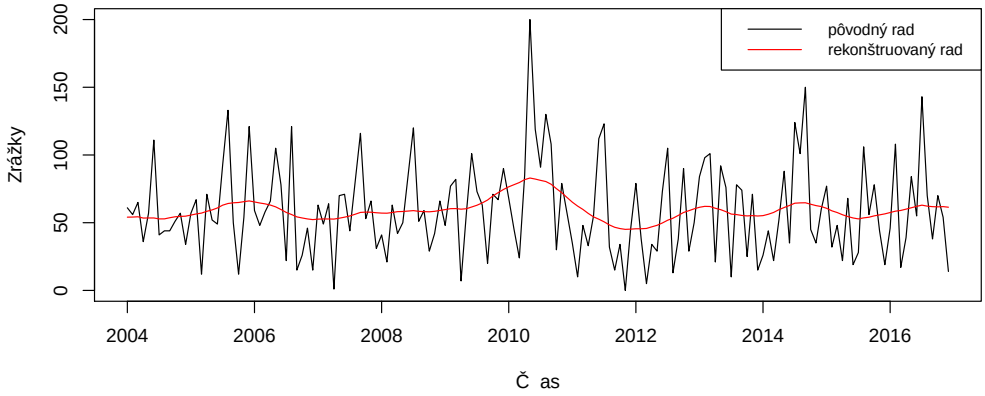
Obr. 2.11: Priemerný mesačný úhrn zrážok na západnom Slovensku v mm od roku 2004 do roku 2016 [17]



(a) Vlastné vektory z SVD, $L = 12$



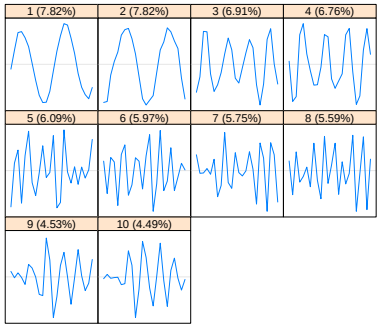
(b) w-korelačná matica, $L = 12$



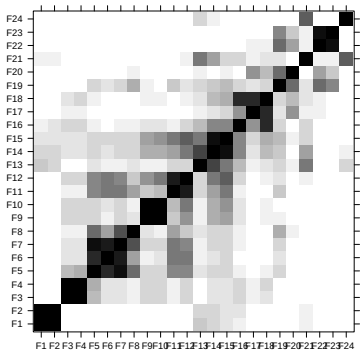
(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

Obr. 2.12: Rekonštrukcia časového radu pomocou prvej vlastnej trojice pre $L = 12$

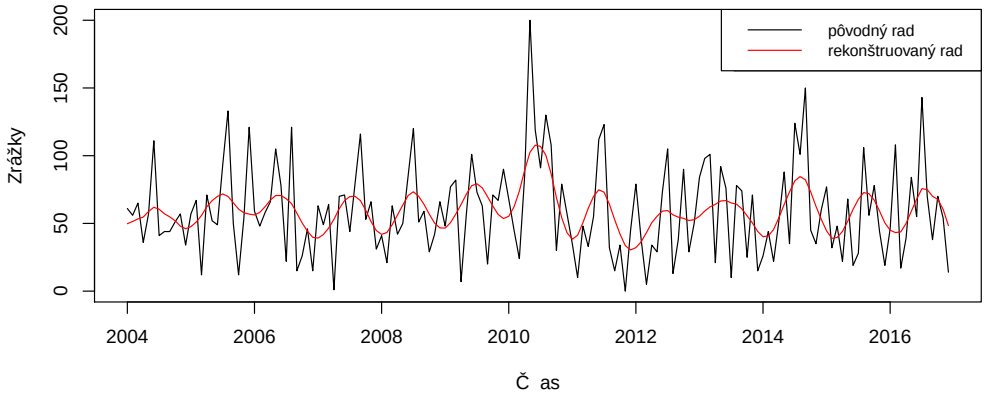
Časový rad rekonštruovaný pomocou prvej vlastnej trojice sa nachádza na Obrázku 2.12c červenou farbou. Ďalej budeme pokračovať s $L = 24$, kde príslušné grafy môžeme vidieť na Obrázkoch 2.13a a 2.13b. Prvé dve vlastné trojice predstavujú určitú cyklickú zložku a keďže majú aj silnú w-koreláciu so sebou navzájom a veľmi slabú s ostatnými, tak si ich vyberáme ako druhú skupinu vlastných trojíc.



(a) Vlastné vektory z SVD, $L = 24$



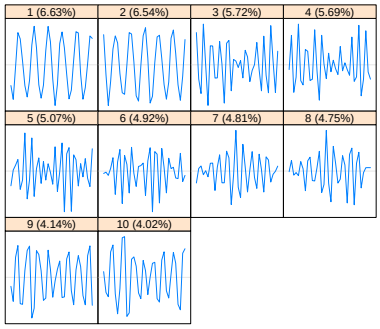
(b) w-korelačná matica, $L = 24$



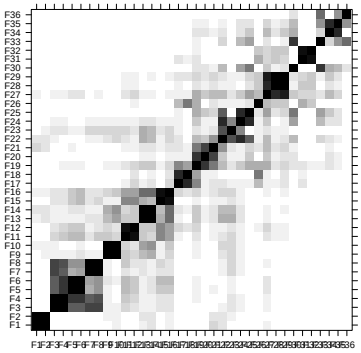
(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

Obr. 2.13: Rekonštrukcia časového radu pomocou 3 vlastných trojíc

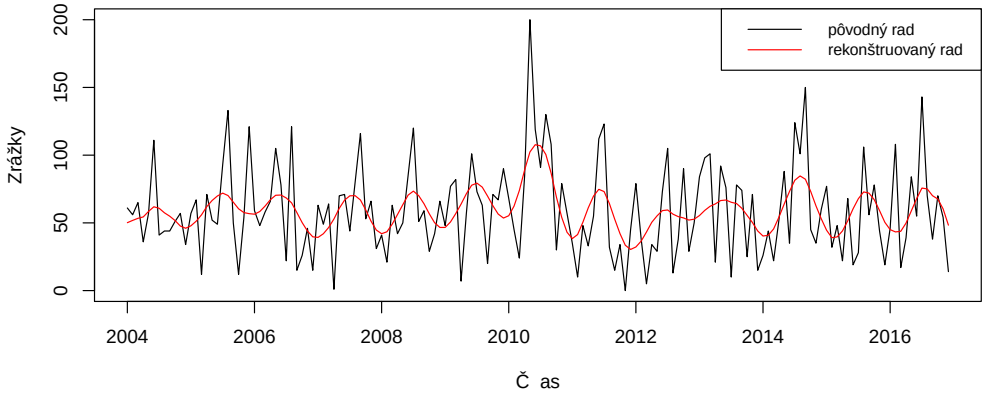
Na Obrázku 2.13c je časový rad zrekonštruovaný z dvoch skupín vlastných trojíc. Môžeme si všimnúť, že tentokrát nám už vstúpila aj sezónna zložka. Pokračujeme v rozkladaní nášho zvyškového časového radu, tentokrát pre $L = 36$. Vlastné vektory a w-korelačná matica sú zobrazené na Obrázkoch 2.14a a 2.14b, kde sa znovu rozhodneme vybrať prvé 2 vlastné vektory. Časový rad rekonštruovaný z 5 vlastných trojíc sa nachádza na Obrázku 2.14c.



(a) Vlastné vektory z SVD, $L = 36$



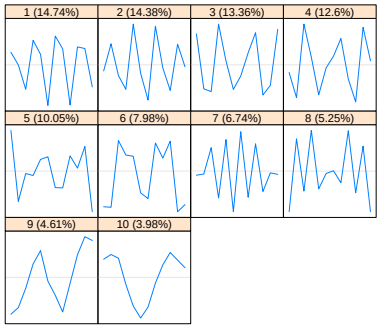
(b) w-korelačná matica, $L = 36$



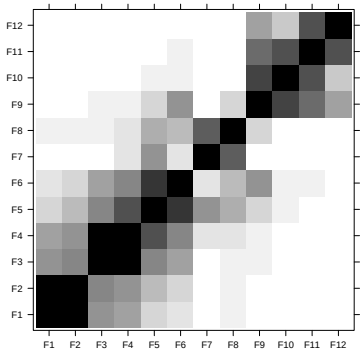
(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

Obr. 2.14: Rekonštrukcia časového radu pomocou 5 vlastných trojíc

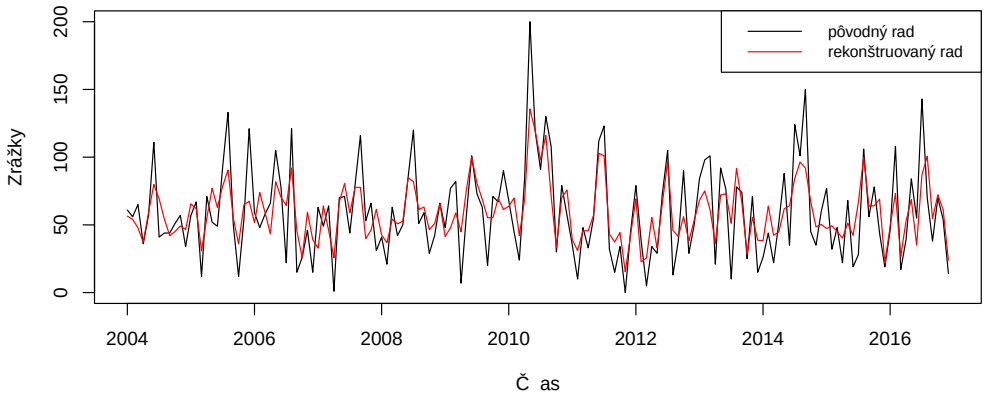
Pozrieme sa ešte posledný krát na ostávajúci časový rad a rozložíme ho pre dĺžku okna $L = 12$. Tentokrát w-korelácia nie je výrazne nízka medzi prvými dvomi vlastnými trojicami a zvyškom (Obr. 2.15b). Vyberieme prvé dve vlastné trojice a dostávame rekonštruovaný časový rad pomocou 7 vlastných trojíc na Obrázku 2.15c.



(a) Vlastné vektory z SVD, $L = 12$



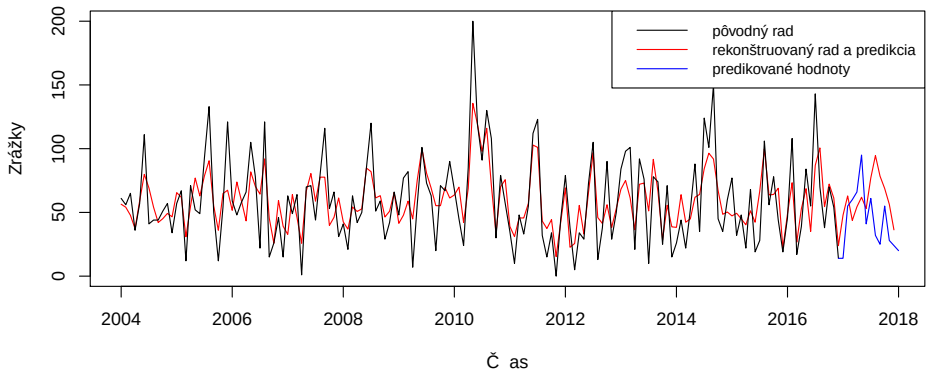
(b) w-korelačná matica, $L = 12$



(c) Pôvodný časový rad (čierna), rekonštruovaný časový rad (červená)

Obr. 2.15: Rekonštrukcia časového radu pomocou 7 vlastných trojíc

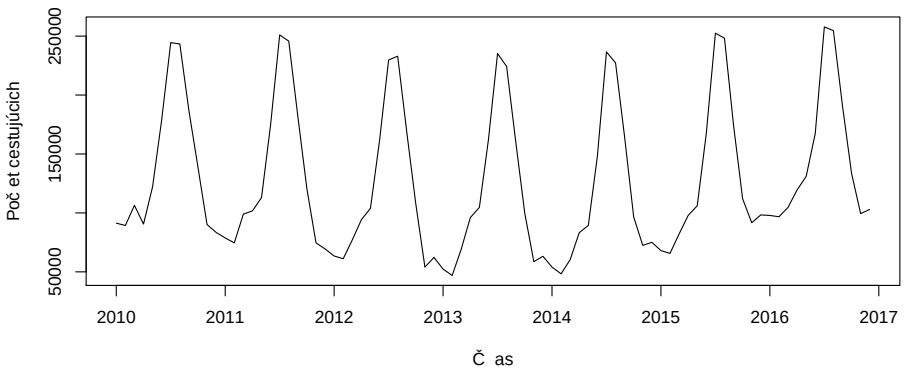
Predikované dáta na rok 2017 môžeme vidieť na Obrázku 2.16. Červenou farbou je zobrazený rekonštruovaný rad s predikovanými hodnotami. Modrou farbou sú skutočné mesačné úhrny zrážok. Vidíme, že v tomto prípade si SSA neporadila, kedže sa tu nenachádzal žiaden zjavný cyklus a chyba predikcie vyšla $0,3742457$. Medzi najupršanejšie mesiace patrí august a september s priemerným úhrnom zrážok približne 78 mm. V predikcii dosahujú zrážky za mesiac august najväčšiu hodnotu, hoci v roku 2017 bol najupršanejší mesiac máj.



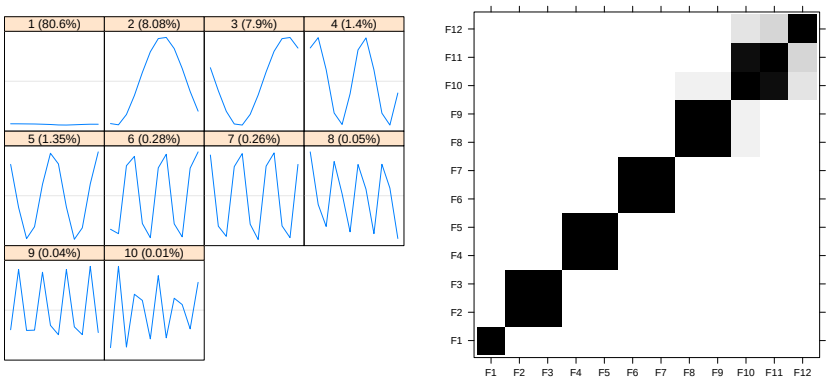
Obr. 2.16: Pôvodný časový rad (čierna), rekonštruovaný rad a predikcia (červená), skutočné predikované hodnoty (modrá)

2.4 Počet cestujúcich na letisku

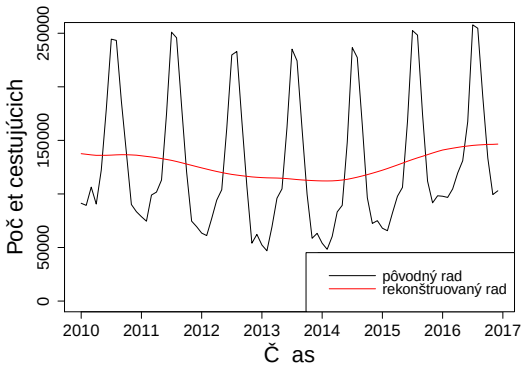
Ako posledný príklad sme vybrali mesačné dáta počtu cestujúcich na Letisku M. R. Štefánika v Bratislave. Tieto údaje sú dostupné na stránke [21]. Analyzovať budeme dáta od januára 2010 do decembra 2016, t.j. $N = 84$ a potom následne budeme predikovať počty cestujúcich pre rok 2017. Na Obrázku 2.17 vidíme, že tieto dáta majú veľmi silnú sezónnu zložku, jednotlivé roky majú rovnaký priebeh, ale trend sa tu nevyskytuje. V roku 2012 si môžeme všimnúť mierny pokles v počte cestujúcich, ktorý bol spôsobený výpadkom ponuky Českých aerolinií a rovnako aj redukciou kapacity ponúkanej spoločnosťou Ryanair [29].



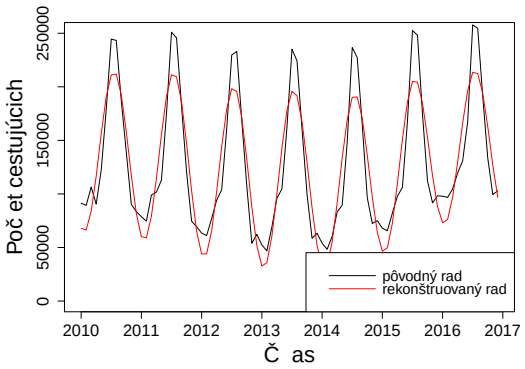
Obr. 2.17: Počty cestujúcich na bratislavskom letisku od roku 2010 do roku 2016 [21]



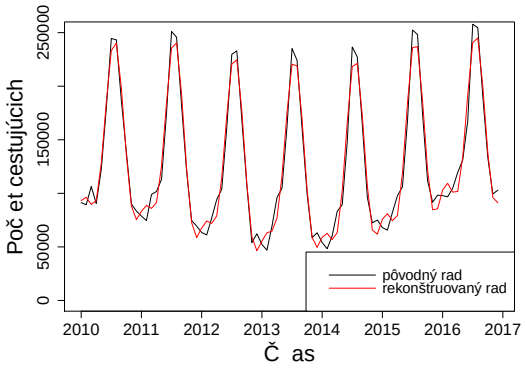
Obr. 2.18: Vľavo - vlastné vektory z SVD, vpravo - w-korelačná matica, $L = 12$



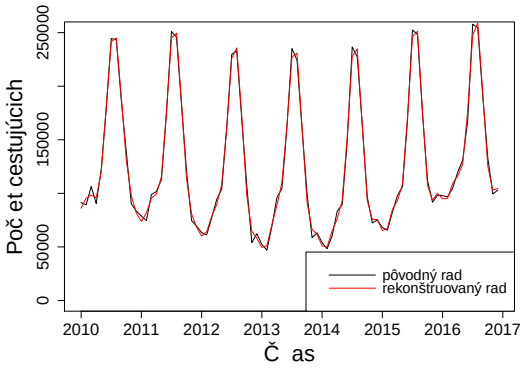
(a) Časový rad rekonštruovaný pomocou 1 vlastnej trojice



(b) Časový rad rekonštruovaný pomocou 3 vlastných trojíc



(c) Časový rad rekonštruovaný pomocou 5 vlastných trojíc

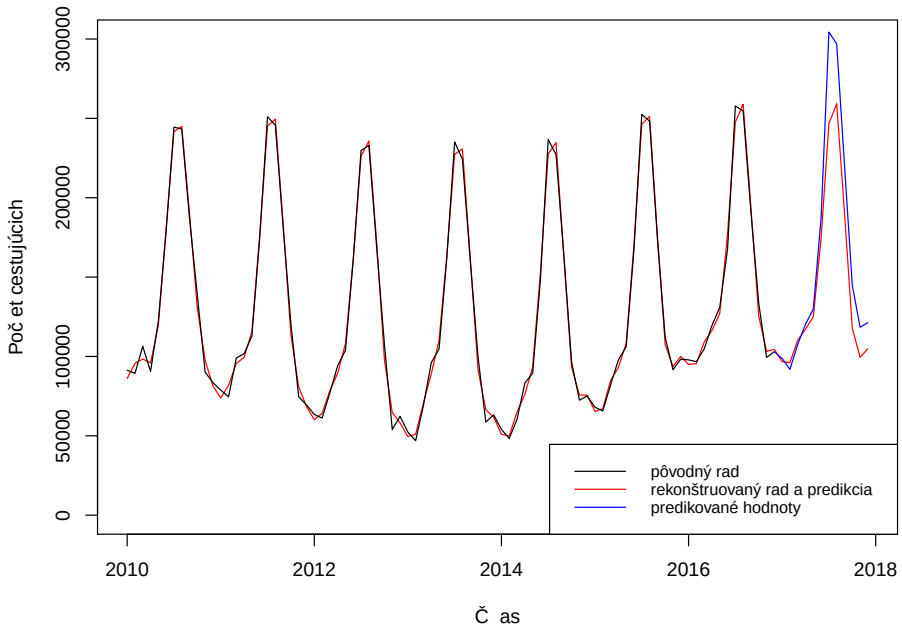


(d) Časový rad rekonštruovaný pomocou 7 vlastných trojíc

Obr. 2.19: Rekonštrukcia časového radu

Pre $L = 12$ dostávame vlastné vektory a maticu w-korelácií, ktoré sú na Obrázku 2.18. Okrem prvého časového radu sú zvyšné časové rady vždy po dvojiciach nekorelované. V takomto prípade nemusíme robiť viac rozkladov a môžeme si jednotlivé skupiny vlastných trojíc vybrať už z prvého rozkladu.

Postupne na Obrázkoch 2.19(a)-2.19(d) je vykreslený časový rad rekonštruovaný postupne 1 až 7 vlastnými trojicami. Pri poslednej rekonštrukcii je chyba metódy $0,02315252$. Na Obrázku 2.20 vidíme skutočné počty cestujúcich za rok 2017 oproti predikovaným. Chyba predikcie bola $0,1155622$, čo je spôsobené hlavne výrazným nárastom cestujúcich v letných mesiacoch, ktorý sa nedal predikovať z historických dát. V roku 2017 sa počet destinácií zvýšil z predošlých 30 na 41, kde spoločnosť Wizz Air pridala pravidelné linky do Bosny a Hercegoviny, Bulharska, Poľska či Rumunska a zároveň sa zvýšili frekvencie niektorých liniek.



Obr. 2.20: Pôvodný časový rad (čierna), rekonštruovaný rad a predikcia (červená), skutočné predikované hodnoty (modrá)

3 ARIMA model

Autoregressive integrated moving average modely vychadzajú z knihy [3]. Skladajú sa z 3 častí - autoregresívneho procesu (AR), integrovaného procesu (I) a procesu kĺzavých priemerov (MA). Pri AR modeli predpokladáme, že hodnota časového radu závisí lineárne od predchádzajúcich hodnôt, MA model zas zahŕňa modelovanie rezíduí pomocou predchádzajúcich chýb. V tejto kapitole si vysvetlíme jednotlivé časti, ktoré tvoria ARIMA modely a predpoklady, ktoré musia byť splnené. Predstavíme si základné pojmy, ktoré súvisia s týmito procesmi.

Stacionárny proces je taký proces, ktorého dáta oscilujú okolo určitej rovnovážnej hodnoty. Musí spĺňať nasledujúce 2 podmienky [25]:

- $E(x_t) = \mu \quad \forall t$
- $cov(x_t, x_s) = \gamma(|t - s|) \quad \forall t, s.$

Biely šum $\{u_t\}$ je definovaný vlastnosťami [25]:

- $E(u_t) = 0 \quad \forall t$
- $Var(u_t) = \sigma^2 \quad \forall t$
- $cov(u_t, u_s) = 0 \quad \forall t \neq s.$

Biely šum spĺňa podmienky stacionárneho procesu a v modeloch bude reprezentovať rezíduá.

Waldova reprezentácia je tvar, v ktorom sa dá zapísať každý stacionárny proces [31]. Je tvorená sumou bieleho šumu a konštanty [25]:

$$x_t = \mu + \sum_{j=0}^{\inf} \psi_j u_{t-j}. \tag{3.1}$$

Autokorelačná funkcia (ACF) stacionárneho procesu vyjadruje závislosť medzi x_t a hodnotami $x_{t+1}, x_{t+2}, ...$ Definovaná je ako [25]:

$$\rho = cor(x_t, x_{t+\tau}) = \frac{cov(x_t, x_{t+\tau})}{\sqrt{var(x_t)var(x_{t+\tau})}}. \tag{3.2}$$

Keď si korelácie a kovariancie označíme

$$\rho(\tau) = cor(x_t, x_{t+\tau} \quad \text{pre } \tau = 1, 2, 3, ...$$
$$\gamma(\tau) = cov(x_t, x_{t+\tau}) \quad \text{pre } \tau = 1, 2, 3, ...,$$

tak vzťah pre ACF môžeme prepísať na

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)},$$

pričom platí $\rho(0) = 1$ a $\rho(\tau) = \rho(-\tau)$. Pri výpočtoch budeme používať asymptoticky nevychýlený odhad ACF, tzv. výberovú ACF.

Parciálna autokorelačná funkcia (PACF) je využívaná pri určovaní radu AR procesov. Predstavuje koreláciu medzi x_t a x_{t-k} očistenú o lineárnu závislosť $x_{t-1}, x_{t-2}, ..., x_{t-k+1}$. Označovať ju budeme $\{\phi_k k\}_{k=1}^{\inf}$ a jej hodnoty získame vypočítaním koeficientov z [25]:

$$x_t = \phi_k 1 x_{t-1} + \phi_k 2 x_{t-2} + ... + \phi_k k x_{t-k} + u_t.$$

Invertovateľnosťou budeme nazývať vlastnosť, keď sa proces bude dať zapísať v tvare [25]:

$$x_t = \hat{\mu} + u_t + \psi_1 x_{t-1} + \psi_2 x_{t-2} + ...$$

3.1 Autoregresný proces

$AR(p)$ predstavuje autoregresný proces radu p definovaný ako:

$$x_t = \delta + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + ... + \alpha_p x_{t-p} + u_t,$$

kde $\delta, \alpha_1, ..., \alpha_p$ sú konštanty a u_t biely šum. Po prepise tohto vzťahu pomocou operátora posunu L, kde $x_t = L x_{t-1}$, dostávame

$$(1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p) x_t = \delta + u_t,$$

z čoho vyplýva podmienka stacionarity, že korene $(1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p)$ musia byť v absolútnej hodnote väčšie ako 1. Autoregresný proces je vždy invertovateľný a hodnoty PACF $\phi_{kk} = 0$, pre $k > p$ [25].

3.2 Proces kĺzavých priemerov

$MA(q)$ označuje procesy kĺzavých priemerov (moving average) radu q. Hodnotu časového radu vyjadríme pomocou bielych šumov $u_t, u_{t-1}, ..., u_{t-q}$

$$x_t = \mu + u_t - \beta_1 u_{t-1} - ... - \beta_q u_{t-q}$$

Toto vyjadrenie prepíšeme pomocou operátora posunu L na:

$$x_t = \mu + (1 - \beta_1 L - ... - \beta_q L^q) u_t$$

a následne dostávame podmienku invertovateľnosti, ktorá hovorí, že korene polynómu $(1-\beta_1 L-...-\beta_q L^q)$ musia byť v absolútnej hodnote väčšie ako 1. Hodnoty ACF $\rho(k) = 0$ pre $k > q$ [25].

3.3 ARMA procesy

$ARMA(p, q)$ proces vzniká spojením autoregresného a moving average procesu. Vyjadrujeme ho vzťahom

$$x_t = \delta + \alpha_1 x_{t-1} + ... + \alpha_p x_{t-p} + u_t - \beta_1 u_{t-1} - ... - \beta_q u_{t-q}. \tag{3.3}$$

Prepísaním pomocou operátora posunu máme

$$(1 - \alpha_1 L - \alpha_2 L^2 - ... - \alpha_p L^p) x_t = \delta + (1 - \beta_1 L - ... - \beta_q L^q) u_t. \tag{3.4}$$

Podmienky stacionarity a invertovateľnosti ostávajú a teda, korene polynómov $(1-\alpha_1 L-\alpha_2 L^2-...-\alpha_p L^p)$ a $(1-\beta_1 L-...-\beta_q L^q)$ sa musia nachádzať mimo jednotkového kruhu. Pri ARMA procesoch sa hodnoty ACF a PACF nevynulujú [25].

3.4 ARIMA procesy

ARIMA modely sa líšia od ARMA modelov pridaním integrovaného procesu. V prípade, že proces nie je stacionárny, zvykneme sa pozerať na jeho diferencie. Ak d-tymi diferenciami časového radu dostaneme biely šum, tak hovoríme, že časový rad je integrovaný proces rádu d [25], t.j.

$$(1 - L)^d x_t = \epsilon_t, \tag{3.5}$$

kde ϵ_t je biely šum. To znamená, že $ARIMA(p, d, q)$ predstavuje aplikovanie ARMA(p, q) na integrovaný proces rádu d.

3.5 SARIMA modely

SARIMA modely označujú sezónne ARIMA modely. Zapisujeme ich v tvare SARIMA $(p, d, q) \times (P, D, Q)_S$, kde p, d, q sú parametre z ARIMA modelu, P je počet sezónnych AR členov, D - počet sezónnych diferencovaní, Q počet sezónnych MA členov a S je perióda predstavujúca sezónnosť dát. Formálne môžeme tento model zapísať ako [25]:

$$A(L^S)\alpha(L)(x_t - \mu) = B(L^S)\beta(L)w_t,$$

kde nesezónne zložky sú:

$$AR : \alpha(L) = 1 - \alpha_1 L - ... - \alpha_p L^p$$
$$MA : \beta(L) = 1 + \beta_1 L + ... + \beta_q L^q$$

a sezónne zložky sú:

$$sezónneAR : A(L^S) = 1 - A_1 L^S - ... - A_P L^{PS}$$
$$sezónneMA : B(L^S) = 1 + B_1 L^S + ... + B_Q L^{QS}$$

4 Príklady použitia ARIMA modelov

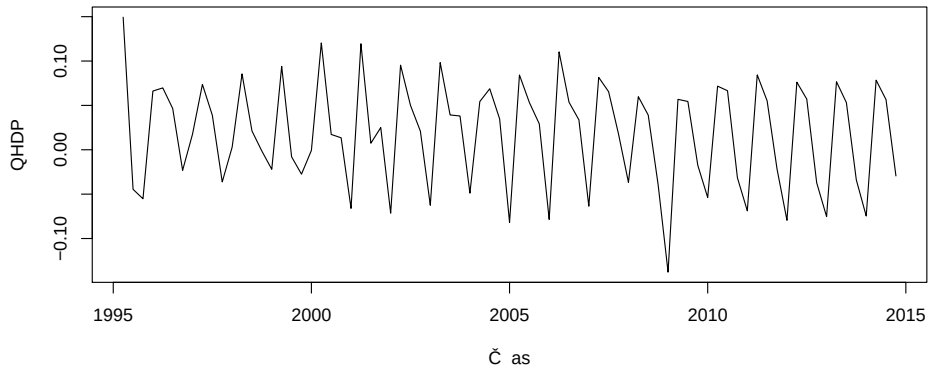
V predchádzajúcej kapitole sme si predstavili ARIMA modely, ktoré v tejto kapitole použijeme na príklady z Kapitoly 2. Analýza časových radov pomocou ARIMA modelov a nasledná jej aplikácia v R je opísaná v [7] a taktiež na stránke [25]. Budeme používať knižnice `astsa` a `urca` v softvéri R.

4.1 HDP Slovenska

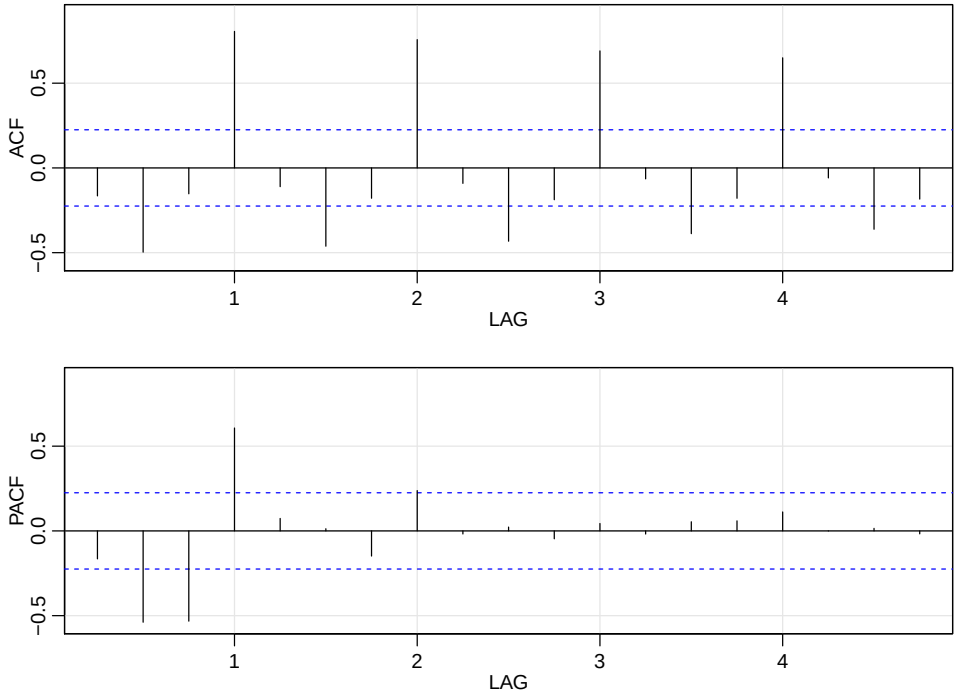
Dáta pre kvartálne hodnoty HDP Slovenska sme si už predstavili v podkapitole 2.1. Ich graficky znázornený priebeh môžeme vidieť na obrázku 4.5. Pre zredukovanie disperzie budeme naďalej pracovať s logaritmami pôvodných hodnôt a navyše si vypočítame diferencie hodnôt, aby sme odstránili rastúci trend (Obr. 4.2).

Obr. 4.1: Kvartálne hodnoty HDP Slovenska v miliónoch EUR v rokoch 1995 - 2015

Najprv sa pozrieme na ACF a PACF pomocou funkcie `acf2()`, ktoré sú zobrazené na Obrázku 4.3. Pri výberovej autokorelačnej funkcii vidíme, že pre posuny zodpovedajúce jednému, dvom, trom rokom t.j. $lag = 4, 8, 12$, máme hodnoty signifikantne väčšie od nuly. Prezrádza nám to, že v našich dátach sa vyskytuje sezónnosť, čo vyplýva aj z toho, že sme použili kvartálne hodnoty. Namiesto obyčajných diferencií použijeme aj sezónne diferencie zodpovedajúce jednému roku ako $x_t - x_{t-4}$.

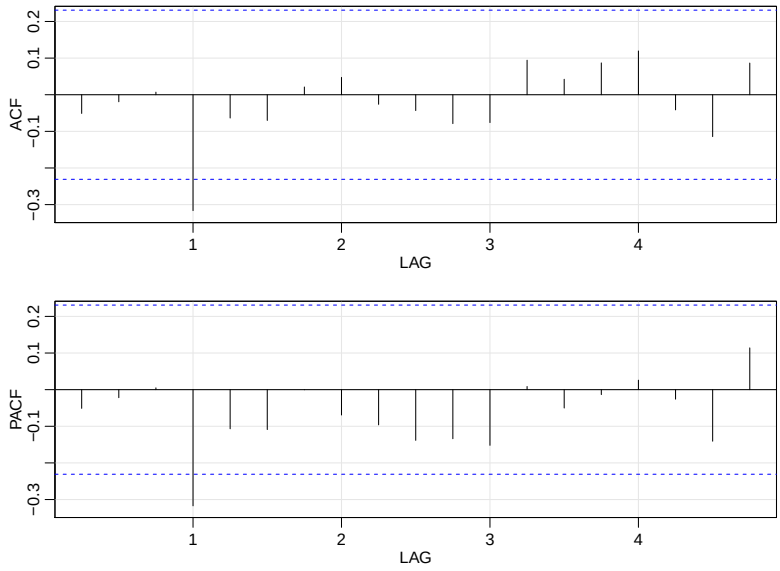


Obr. 4.2: Diferencie logaritmov kvartálnych hodnôt HDP Slovenska

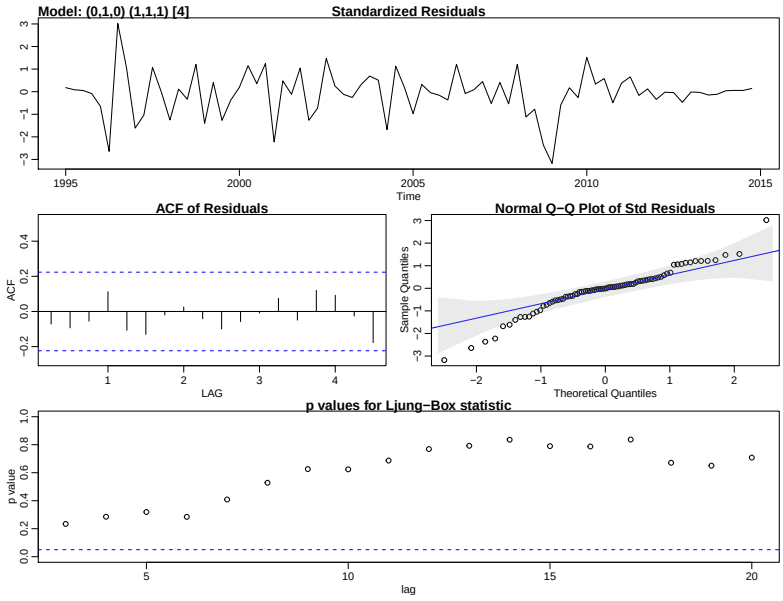


Obr. 4.3: Výberová ACF a PACF pre dáta HDP Slovenska

Výberovú ACF pre tieto dáta vidíme na Obrázku 4.4. Časový rad si tentokrát upravíme pomocou klasických diferencií, aby sme sa zbavili trendu. Novú výberovú ACF môžeme vidieť na Obrázku **??**. Keďže hodnota pre lag predstavujúci 1 rok je štatisticky významná pri ACF aj PACF, pretože sa nachádza nad modrou čiarou predstavujúcu p-hodnotu, vyberáme si sezónny ARMA(1,1) proces.



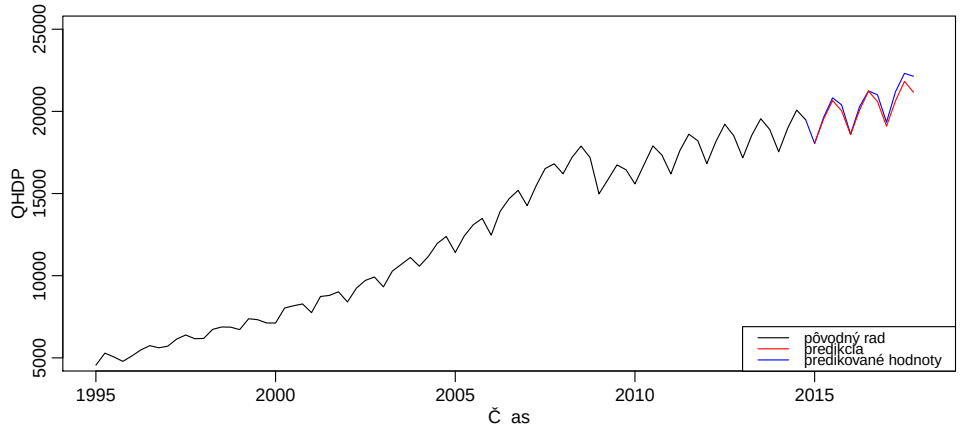
Obr. 4.4: Výberová ACF po sezónnom a klasickom diferencovaní



Obr. 4.5: Model SARIMA $(0,1,0) \times (1,1,1)_{S=4}$

Na overenie, že časový rad je po diferencovaní bez jednotkového koreňa, sme použili funkciu `ur.df()`, ktorá nám potvrdila, že zamietame hypotézu o jednotkovom koreni. Pre naše dáta sme sa rozhodli použiť $SARIMA\ (0,1,0) \times (1,1,1)_S$, kde $S = 4$, ktorá je v programe R pod názvom `sarima()`.

Pre nami zvolený model sa nachádza na Obrázku 4.5 zhrnutie, či sú rezíduá vyhovujúce. Hodnoty výberovej ACF nevybiehajú z intervalu spoľahlivosti a taktiež Ljung-Boxova štatistika sa nachádza nad p-hodnotou. Predikcie na najbližšie 3 roky vytvoríme použitím funkcie `sarima.for()`.



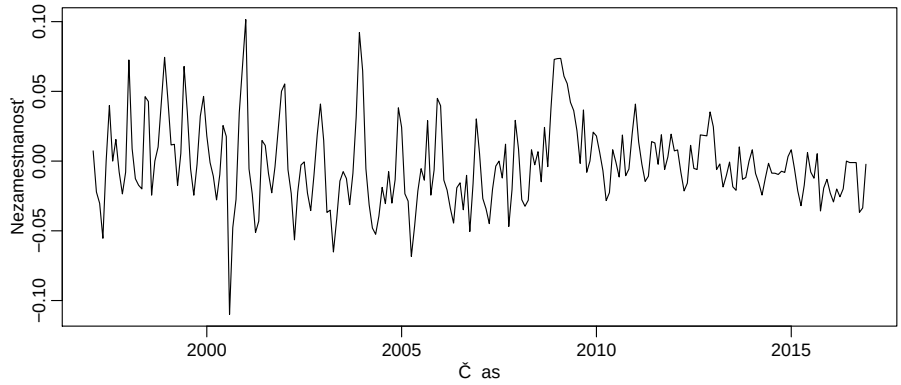
Obr. 4.6: Skutočné hodnoty HDP (čierna a modrá) v porovnaní s predikciami (červená)

Obrázok 4.6 ukazuje porovnanie skutočných hodnôt HDP, zobrazené čiernou a modrou farbou, s hodnotami predikcie zobrazenými červenou farbou. Vidíme, že sa modelu podarilo skopírovať periodický priebeh a zachovať rastúci trend. Chyba predikcie je v tomto prípade $0,09551474$.

4.2 Miera nezamestnanosti

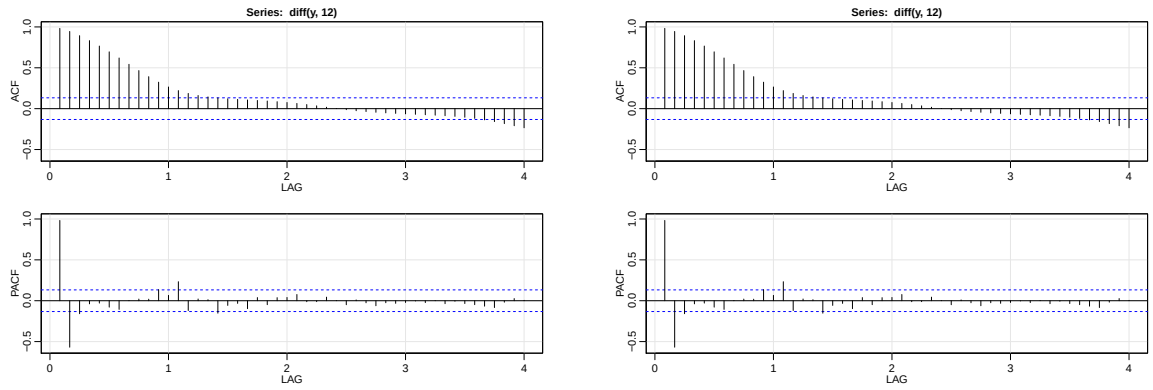
Obr. 4.7: Mesačné hodnoty miery nezamestnanosti Slovenska v rokoch 1997 - 2016

Mesačné hodnoty miery nezamestnanosti Slovenska z podkapitoly 2.2 môžeme vidieť na Obrázku 4.7. Rovnako ako v predchádzajúcom príklade budeme pracovať s logaritmami týchto hodnôt, aby sme znížili disperziu našich dát.



Obr. 4.8: Mesačné hodnoty miery nezamestnanosti Slovenska po zlogaritmovaní a zdiferencovaní

Miera nezamestanosti vykazovala určitý trend, preto sme dáta zdiferencovali (Obrázok 4.8). Výberová ACF na Obrázku 4.9a dáva vysoké hodnoty pre lagy reprezentujúce roky. Upravíme si diferencie, aby predstavovali 12-mesačné rozdiely, t.j. $x_t - x_{t-12}$.

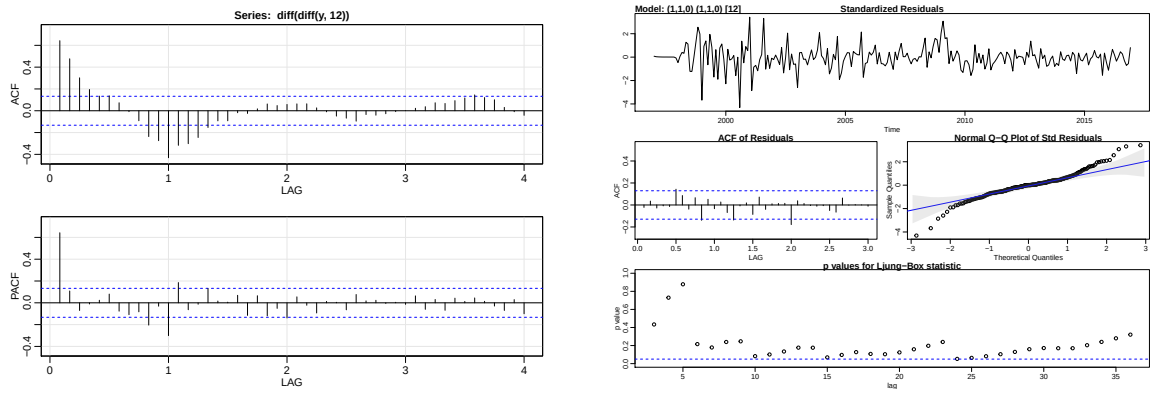


(a) **(b)**

Obr. 4.9: (a) Výberová ACF a PACF pre mieru nezamestnanosti,
(b) Výberová ACF a PACF pre mieru nezamestnanosti po sezónnom diferencovaní

Na Obrázku 4.9b dosahuje výberová ACF vysoké hodnoty, hoci pre vyššie lagy pomaly klesá až pod hranicu signifikantnosti. Naznačuje nám to, že dáta treba znova differencovať, hoci tentoraz klasicky, nie sezónne.

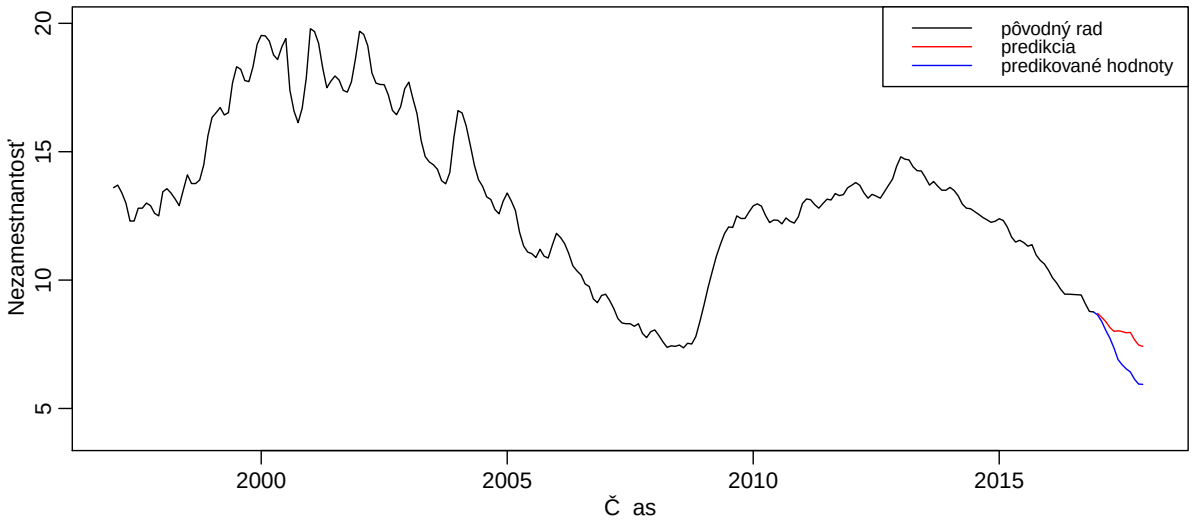
Výberová ACF pre klasicky aj sezónne diferencované dáta sa vynuluje (Obr. 4.10a). Na druhej strane výberová PACF nadobúda výrazne vyššie hodnoty pre $lag = 1$ a $lag = 12$, a preto zvolíme AR(1) a sezónny AR(1) proces. Na Obrázku 4.10b vidíme, že reziduá modelu sú v poriadku.



(a) **(b)**

Obr. 4.10: (a) Výberová ACF po sezónnom diferencovaní a klasickom diferencovaní, (b) model SARIMA $(1,1,0) \times (1,1,0)_{S=12}$

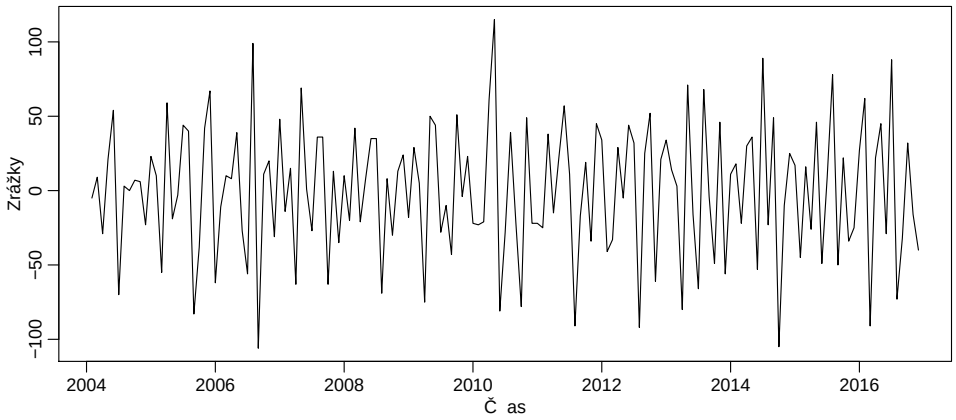
Predikované hodnoty znázornené červenou farbou na Obrázku 4.11 dodržiavajú klesajúci trend posledných rokov, hoci pokles skutočnej miery nezamestnanosti bol výraznejší, a preto chyba predikcie je $0,4128969$.



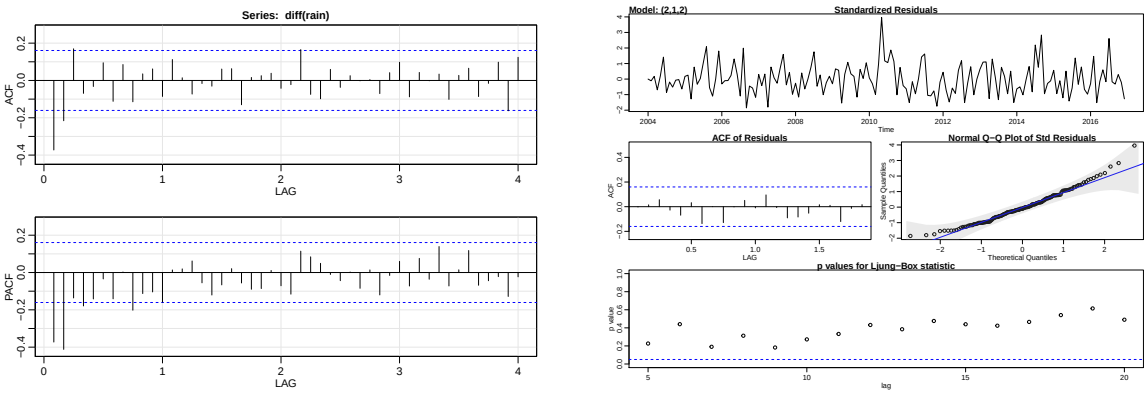
Obr. 4.11: Pôvodný časový rad (čierna), predikcia (červená), skutočné hodnoty (modrá)

4.3 Priemerný úhrn zrážok

Na Obrázku 4.12 sú zobrazené diferencie priemerného mesačného úhrnu zražok. Výberová ACF aj PACF na Obrázku 4.13a dosahujú vyššie hodnoty pre $lag = 2$ a následne sa obe vynulujú. Zvolíme autoregresný proces rádu 2 a taktiež MA(2).



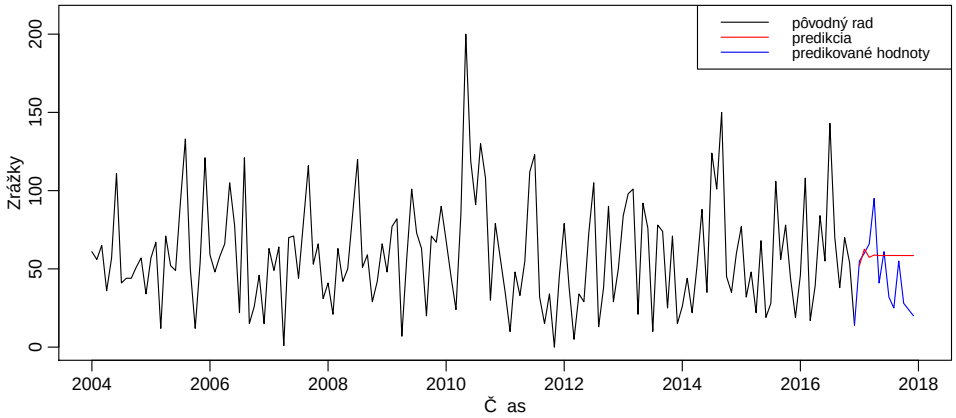
Obr. 4.12: Diferencie priemerného mesačného úhrnu zrážok na západnom Slovensku



(a) **(b)**

Obr. 4.13: (a) Výberová ACF pre diferencované dáta, (b) model ARIMA $(2, 1, 2)$

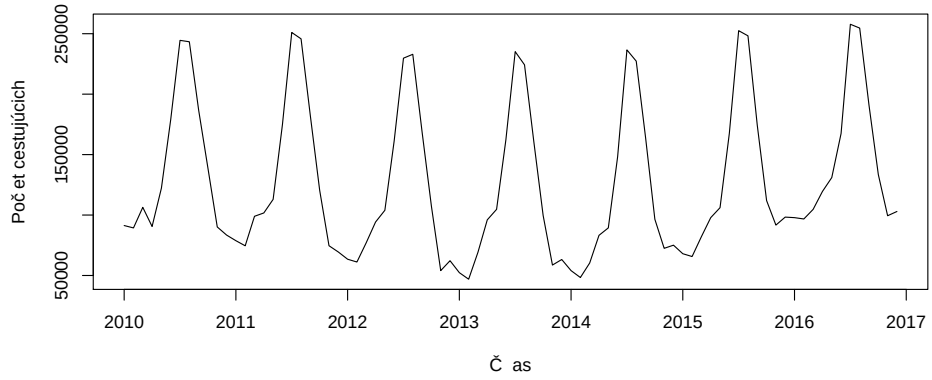
Reziduá pre zvolený model $ARIMA(2, 1, 2)$ sú v poriadku (Obr. 4.13b). Predikované hodnoty môžeme vidieť červenou farbou na Obrázku 4.14. Predikcia vyšla výrazne odlišne oproti skutočnosti s chybou predikcie $NRMSE = 0,32567$. Oproti SSA predikcií sme získali síce menšiu chybu, ale na rozdiel od SSA nám vyšli predikované hodnoty blízko priemeru časového radu, ktoré vôbec nepripomínajú priebeh pôvodného časového radu.



Obr. 4.14: Pôvodný časový rad (čierna), predikcia (červená), skutočné hodnoty (modrá)

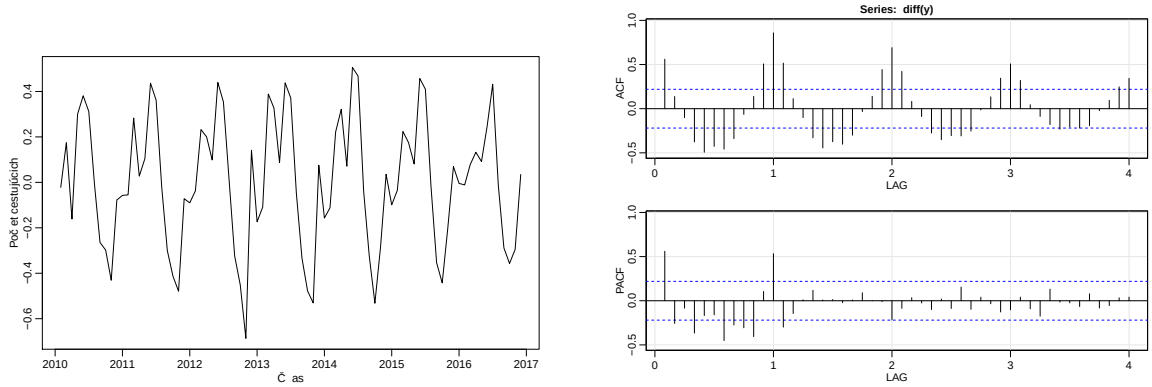
4.4 Počet cestujúcich na letisku

V tomto príklade budeme pracovať s dátami z podkapitoly 2.4, konkrétne s ich logaritmami, aby sme zredukovali ich disperziu. Na Obrázku 4.15 je zobrazený priebeh dát a na Obrázku 4.16a sú diferencie zlogaritmovaného pôvodného časového radu.



Obr. 4.15: Počet cestujúcich na letisku v Bratislave

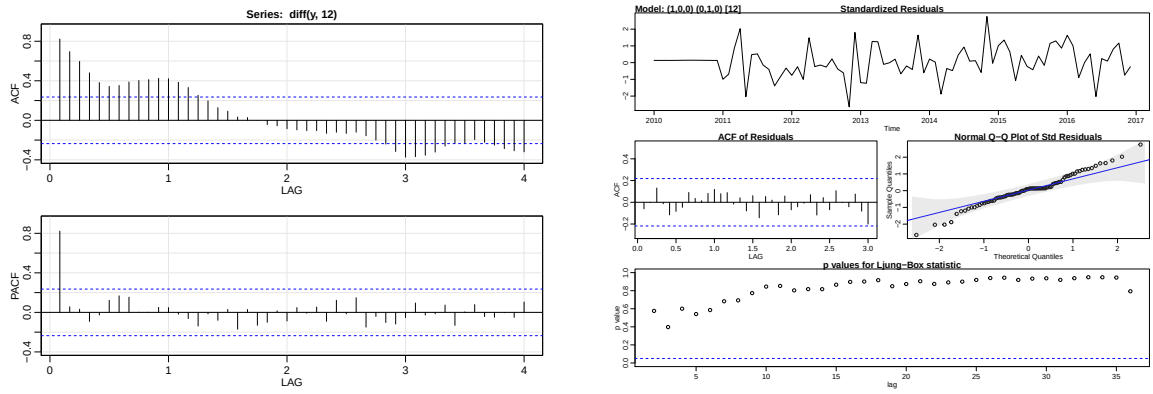
Výberová ACF (Obr. 4.16b) vyznačuje vyššie hodnoty pre ročné lagy, a preto namiesto klasických diferencií použijeme sezónne diferencie pre $lag = 12$. Pre dáta s ročnými diferenciami môžeme vidieť na Obrázku 4.17a, že výberová ACF postupne klesá až k nule. Na druhej strane výberová PACF má pre $lag = 1$ výrazne vysokú hodnotu, zatiaľ čo pre nasledujúce lagy klesá pod hranicu signifikantnosti. Rozhodneme sa pre proces $AR(1)$. Pri testovaní hypotézy o jednotkovom koreni dostávame hodnotu testovacej štatistiky nižšiu



(a) **(b)**

Obr. 4.16: (a) Diferencie zlogaritmovaných hodnôt počtu cestujúcich, (b) výberová ACF pre diferencované dáta

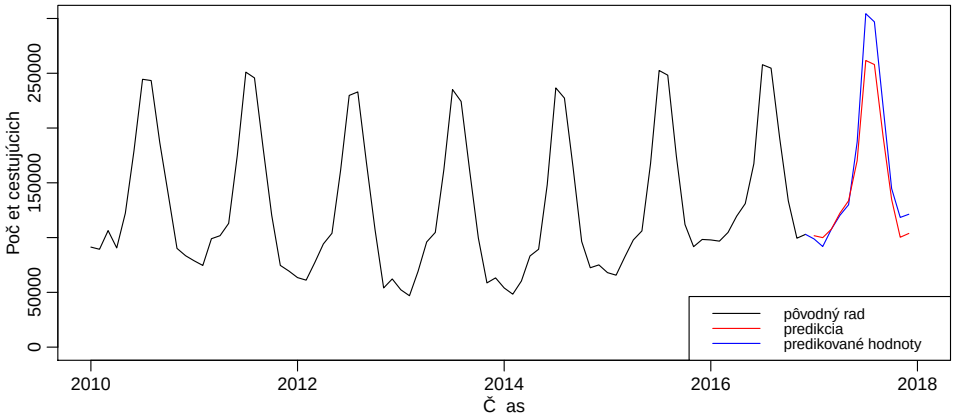
ako danú kritickú hodnotu, a teda hypotézu zamietame.



(a) **(b)**

Obr. 4.17: (a) Výberové ACF a PACF, (b) model SARIMA $(1,0,0) \times (0,1,0)_{S=12}$

Po použití modelu $SARIMA\ (1,0,0) \times (0,1,0)_{S=12}$ vidíme na Obrázku 4.17b, že výberová ACF rezíduí nepresahuje hranice signifikantnosti a taktiež všetky p-hodnoty Ljung-Boxovej štatistiky sa nachadzajú nad hranicou 5%. Vypočítané predikcie môžeme vidieť na Obrázku 4.18. Takisto ako v podkapitole 2.4 sa nám nepodarilo predikovať taký výrazný nárast cestujúcich v letných mesiacoch, chyba predikcie vyšla $0,09813038$.



Obr. 4.18: Pôvodný časový rad (čierna), predikcia (červená), skutočné hodnoty (modrá)

5 Metóda oporných bodov

Ako tretiu metódu si predstavíme metódu oporných bodov (SVM z angl. *Support vector machine*), ktorej dnešnú podobu poznáme hlavne z publikácie V. Vapnika z roku 1995 [28]. Medzi jej prvé aplikácie patrilo optické rozpoznávanie znakov [2], [6] a rozpoznávanie objektov [22], [23]. V tejto kapitole si vysvetlíme základnú myšlienku SVM regresie, pričom budeme vychádzať z [23].

5.1 Základná myšlienka

Majme trénovacie dáta

$$\{(x_1, y_1), ..., (x_t, y_t)\} \subset \mathcal{X} \times \mathbb{R}.$$

Cieľom je nájsť funkciu $f(x)$, ktorej odchýlka od skutočných hodnôt y_i bude maximálne ϵ pre všetky trénovacie dáta, ide o tzv. $\epsilon - regresiu$. Pre lineárnu funkciu $f(x)$ tvaru

$$f(x) = \langle w, x \rangle + b, \quad \text{kde } w \in \mathcal{X}, b \in \mathbb{R}, \tag{5.1}$$

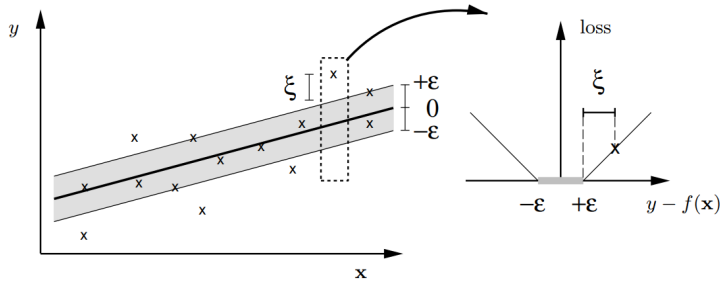
môžeme náš problém formulovať ako úlohu konvexnej optimalizácie

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 \\ s.t. \quad & y_i - \langle w, x_i \rangle - b \leq \epsilon \\ & \langle w, x_i \rangle + b - y_i \leq \epsilon. \end{aligned} \tag{5.2}$$

Keďže takto definovaná úloha nemusí byť prípustná, zavedieme si doplnkové premenné ξ_i, ξ_i^*, ktoré nám dajú určitú voľnosť pri hľadaní funkcie f. Na Obrázku 24 vidíme grafické zobrazenie SVM regresie. Chyby, ktoré sú menšie ako ϵ, nás nezaujímajú, ale pokiaľ prekročia túto hranicu, budú vstupovať do minimalizačnej funkcie. Pridaním premenných ξ_i, ξ_i^* dostávame nasledujúcu formuláciu [28]:

$$\begin{aligned} \min \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{t}(\xi_i + \xi_i^*) \\ s.t. \quad & y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i \\ & \langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, \end{aligned} \tag{5.3}$$

kde $C > 0$ je penalizačná konštanta vyjadrujúca, nakoľko sme ochotní tolerovať chyby väčšie ako ϵ.



Obr. 5.1: SVM regresia [24]

Namiesto úlohy (5.3) budeme pri riešení pracovať s úlohou k nej duálnou, čo sa ukáže ako výhodné pri nelineárnom prípade SVM. Pre úlohu nelineárneho programovania v tvare (5.4) má duálna úloha tvar (5.5).

$$\min\left\{f_0(x)|f_i(x) \leq 0, i = 1, ..., t\right\} \tag{5.4}$$

$$\max\left\{L(x,u)|\nabla_x L(x,u) = 0, u \geq 0\right\} \tag{5.5}$$

$$L(x,y) = f_0(x) + \sum_{i=1}^{t} y_i f_i(x), \quad x \in \mathbb{R}^n, Y = \mathbb{R}^t_+ \tag{5.6}$$

Pre úlohu (5.3) má Lagrangeova funkcia (5.6) tvar

$$\begin{aligned} L := \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{t}(\xi_i + \xi_i^*) - \sum_{i=1}^{t}\alpha_i(\epsilon + \xi_i - y_i + \langle w, x_i\rangle + b) \\ - \sum_{i=1}^{t}\alpha_i^*(\epsilon + \xi_i^* + y_i - \langle w, x_i\rangle - b) - \sum_{i=1}^{t}(\eta_i\xi_i + \eta_i^*\xi_i^*). \end{aligned} \tag{5.7}$$

Parciálne derivácie Lagrangeovej funkcie (5.7) podľa primárnych premenných b, w, ξ_i, ξ_i^*,

$i = 1, ..., t$ sú

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{t}(\alpha_i^* + \alpha_i) = 0 \tag{5.8}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{t}(\alpha_i - \alpha_i^*)x_i = 0 \tag{5.9}$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i, \quad \forall i = 1, ..., t \tag{5.10}$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* = 0, \quad \forall i = 1, ..., t, \tag{5.11}$$

kde $\eta_i, \eta_i^*, \xi_i, \xi_i^*$ sú Lagrangeove multiplikátory. Z rovníc (5.10), (5.11) a (5.11) vyjadríme premenné w, ξ_i, ξ_i^*, dosadíme ich do Lagrangeovej funkcie a dostaneme tvar (5.12).

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i,j=1}^{t}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle + \epsilon\sum_{i=1}^{t}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{t} y_i(\alpha_i + \alpha_i^*) \\ & s.t. \quad \sum_{i=1}^{t}(\alpha_i - \alpha_i^*) = 0 \\ & \alpha_i, \alpha_i^* \in [0, C], \end{aligned} \tag{5.12}$$

5.2 Nelineárne SVM

V predchádzajúcej časti sme si ukázali SVM regresiu pre prípad, že funkcia f je lineárna. Pri rozšírení pre nelineárne funkcie budeme pracovať s hodnotami $\phi(x_i)$ funkcie $\phi : \mathcal{X} \to \mathcal{F}$ namiesto x_i. Ideou je previesť lineárne neseparovateľné vstupy do vyššej dimenzie, kde ich separovateľnosť už bude možná. V prípade primárnej úlohy (5.3) by sme potrebovali zistiť explicitný tvar funkcie ϕ. Tu sa ukazuje výhoda riešenia duálnej úlohy, v ktorej nám vystupuje x_i iba v rámci skalárneho súčinu $\langle x_i, x_j\rangle$ (5.12), ktorý nahradíme $\langle\phi(x_i), \phi(x_j)\rangle$. Vďaka tomu nepotrebujeme zistiť presný tvar funkcie ϕ, ale bude nám stačiť jej skalárny súčin $k(x, x') := \langle\phi(x), \phi(x')\rangle$. Funkcia $k(x, x')$ sa nazýva kernel funkcia a medzi jej najpoužívanejšie tvary patria:

lineárna	$\langle x, x'\rangle$
polynomiálna	$\gamma(\langle x, x'\rangle)^d$
Gaussovská (radiálna)	$\exp(-\gamma\|x - x'\|_2^2)$
sigmoid	$\tanh(\gamma\langle x, x'\rangle + c_0)$

Úlohu (5.12) môžeme prepísať na tvar (5.13)

$$\begin{aligned} \min \quad & \frac{1}{2}\sum_{i,j=1}^{t}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(x_i,x_j)+\epsilon\sum_{i=1}^{t}(\alpha_i+\alpha_i^*)-\sum_{i=1}^{t}y_i(\alpha_i+\alpha_i^*) \\ & s.t. \quad \sum_{i=1}^{t}(\alpha_i-\alpha_i^*)=0 \\ & \alpha_i,\alpha_i^*\in[0,C], \end{aligned} \tag{5.13}$$

Na riešenie úlohy (5.13) sa používa algoritmus typu SMO (z angl. sequential minimal optimization) [9], ktorý je implementovaný aj v softvéri R v balíku `e1071`. V prípade riešenia duálnej úlohy je výhodou jednoduchší výpočet predikcie pre vstupné dáta. Rovnicu (5.1) môžeme pomocou (5.10) a kernel funkcie prepísať na

$$f(x)=\sum_{i=1}^{t}(\alpha_i-\alpha_i^*)k(x_i,x),$$

kde α_i,α_i^* sú výstupy z (5.13), x_i sú vstupné dáta trénovacej sady a x sú dáta, pre ktoré robíme predikciu.

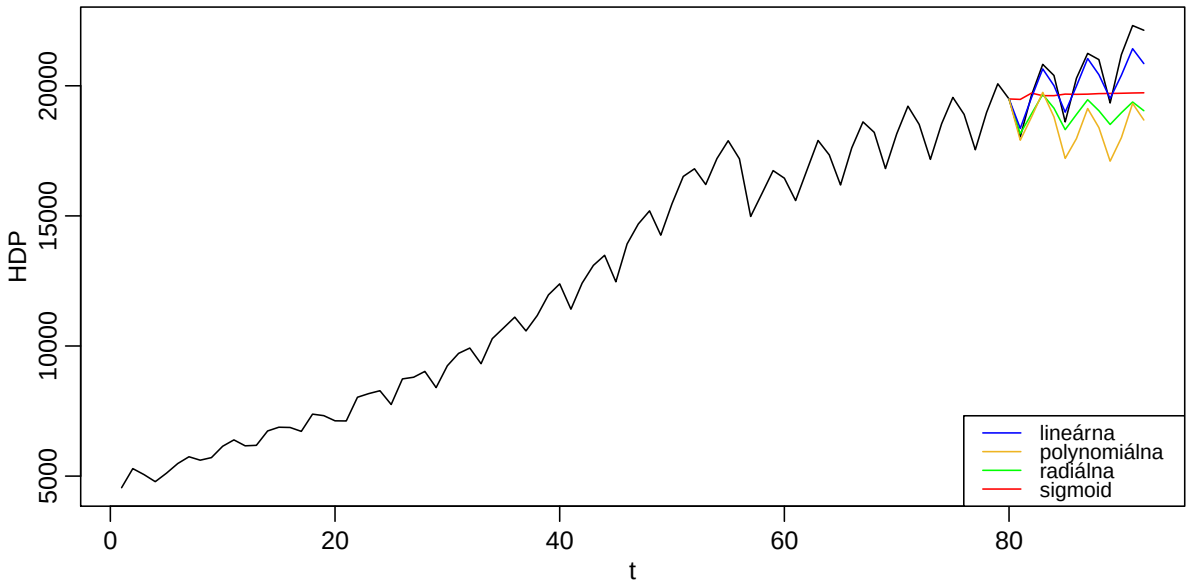
6 Príklady použitia SVM

Algoritmus SVM vyžaduje, aby boli dáta v tvare vstupov x_i a výstupov y_i. Ako výstupy budeme používať hodnoty daného časového radu a vstupy budú vždy predchádzajúce hodnoty. Dostávame vzťah $x_t = f(x_{t-1}, ..., x_{t-q})$. Parameter q bude predstavovať počet použitých predchádzajúcich hodnôt, tzv. posun. Dáta si rozdelíme do 2 skupín - na trénovacie a testovacie dáta v pomere $3 : 1$. Funkcia `tune()` nám vypočíta optimálne hodnoty parametrov vstupujúcich do algoritmu, kde najprv počíta optimálne hodnoty pre trénovaciu sadu a následne vypočíta chybu pri použití týchto parametrov na testovacej sade. Túto funkciu spustíme pre rôzne hodnoty q a model vyberieme na základe najnižšej chyby. Tento postup zopakujeme pre všetky typy jadrových funkcií. Na aplikáciu SVM budeme používať knižnicu `e1071` v softvéri R.

6.1 HDP Slovenska

Pre kvartálne hodnoty HDP sme posun vyberali z hodnôt 2 až 8, kde 8 predstavuje 2-násobok prirodzeného cyklu. V Tabuľke 6.1 vidíme voľby parametrov, posun, s ktorým sme pracovali a chybu predikcie pre jednotlivé typy jadrových funkcií. Najlepšie výsledky sme dosahovali so sigmoid jadrovou funkciou, kde sme ako vstupy použili hodnoty do času $t - 3$. Na Obrázku 6.1 vidíme predikcie pomocou jednotlivých kernel funkcií. Predikcia pomocou sigmoid kernel funkcie (červená) je najkonzervatívnejšia, udržuje sa približne stále na rovnakej hodnote a jediná nedodržiava typický priebeh kvartálneho HDP.

kernel	ϵ	C	γ	c_0	d	posun	chyba predikcie
lineárna	0,1	2	-	-	-	8	0,1347675
polynomiálna	0,1	8	0,1	0,5	2	8	0,5190495
radiálna	0,1	8	0,2	-	-	5	0,3349916
sigmoid	0,2	1	0,1	0,1	-	3	0,04798223

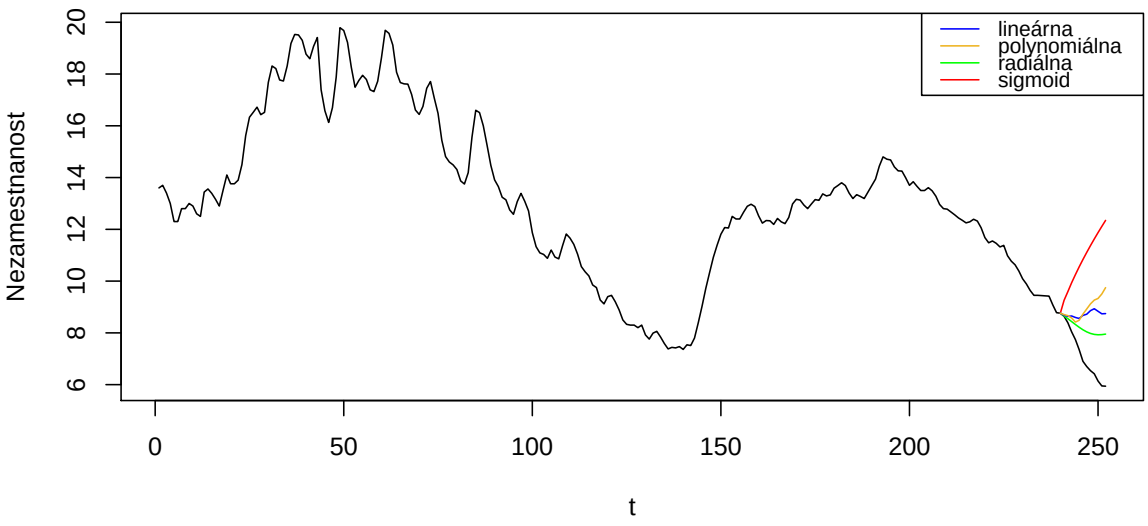


Obr. 6.1: Reálne hodnoty HDP (čierna) s predikciami pomocou lineárnej (modrá), polynomiálnej (žltá), radiálnej (zelená) a sigmoid (červená) kernel funkcie

6.2 Miera nezamestnanosti

Mesačnú mieru nezamestnanosti Slovenska sme modelovali pre vstupné dáta predstavujúce posun od 2 do 24 hodnôt. Hodnoty optimálnych parametrov, posun s ktorým sme pracovali a chyby predikcie sa nachádzajú v Tabuľke 6.2. V tomto prípade dostávame najlepšie výsledky pre radiálnu kernel funkciu s chybou predikcie $0,4794341$. Priebeh predikcií môžeme vidieť na Obrázku 6.2. Radiálna kernel funkcia (zelená) ako jediná predikovala pokles miery nezamestnanosti. Tentokrát vyšla predikcia pomocou sigmoid kernelu funkcie (červená) výrazne najhoršie zo všetkých metód, kde chyba predikcie dosiahla hodnotu $1,58324$.

kernel	ϵ	C	γ	c_0	d	posun	chyba predikcie
lineárna	0,1	4	-	-	-	24	0,7128078
polynomiálna	0,1	8	0,1	1	2	15	0,8448795
radiálna	0,1	16	0,3	-	-	3	0,4794341
sigmoid	0,3	1	0,1	0,1	-	2	1,58324

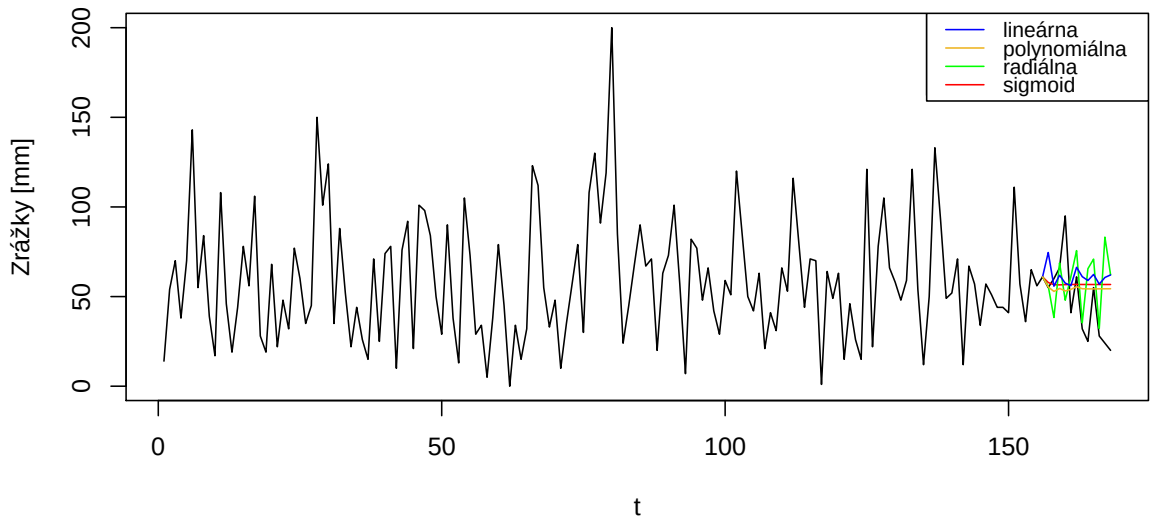


Obr. 6.2: Reálne hodnoty miery nezamestnanosti (čierna) s predikciami pomocou lineárnej (modrá), polynomiálnej (žltá), radiálnej (zelená) a sigmoid (červená) kernel funkcie.

6.3 Priemerný úhrn zrážok

Pre mesačné dáta priemerného úhrnu zrážok sme uvažovali hodnoty posunu od 2 do 24. V Tabuľke 6.3 sa nachádzajú optimálne hodnoty parametrov. Vo všetkých prípadoch vyšli chyby predikcie veľmi podobné. Na Obrázku 6.3 vidíme, že predikcie pomocou sigmoid (červená) a polynomiálnej (žltá) kernel funkcie sú veľmi konzervatívne a dosahujú najnižšie chyby predikcie, obe približne $0,3$. Pri použití lineárnej (modrá) a radiálnej (zelená) kernel funkcie majú predikované hodnoty väčšiu disperziu, hoci sa stále pohybujú okolo priemeru.

kernel	ϵ	C	γ	c_0	d	posun	chyba predikcie
lineárna	1	1	-	-	-	9	0,3461508
polynomiálna	0,5	1	0,1	0,1	4	11	0,3051453
radiálna	0,1	2	0,2	-	-	12	0,3945647
sigmoid	0,9	1	0,1	0,2	-	3	0,3153998

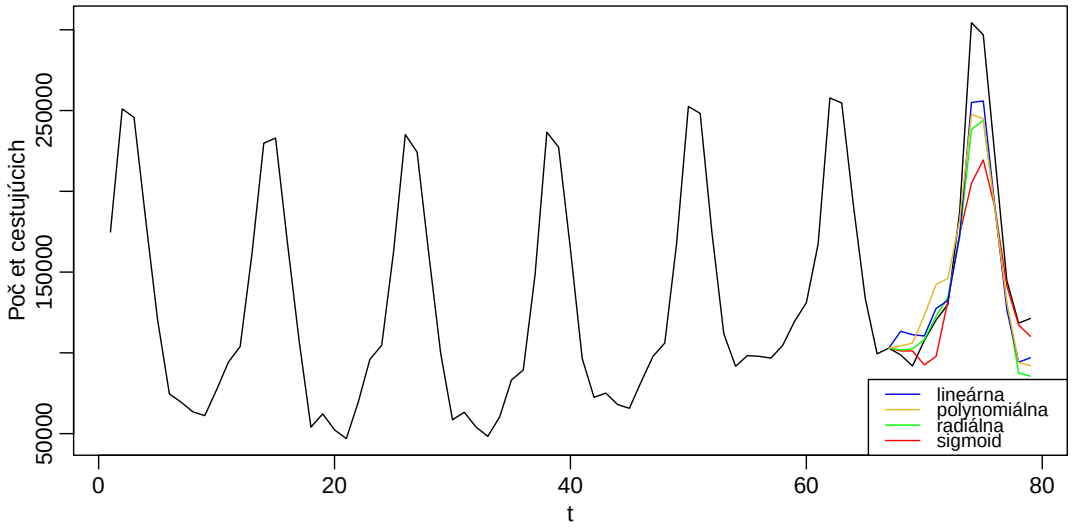


Obr. 6.3: Reálne hodnoty priemerného úhrnu zrážok (čierna) s predikciami pomocou lineárnej (modrá), polynomiálnej (žltá), radiálnej (zelená) a sigmoid (červená) kernel funkcie

6.4 Počet cestujúcich na letisku

Posun q sme pre mesačné dáta počtu cestujúcich vyberali z hodnôt 2 až 24. Optimálne hodnoty parametrov môžeme vidieť v Tabuľke 6.4. Najnižšie hodnoty predikcie sme dostali pri použití lineárnej kernel funkcie a to $0,1172723$. Priebeh predikcií vidíme na Obrázku 6.4. Rozdiely medzi jednotlivými kernel funkciami sú minimálne, rovnako to vidieť aj na chybách predikcií. V letných mesiacoch sa až na sigmoid kernel takmer zhodujú. Jediný výraznejší rozdiel nastal začiatkom roka, kde regresia pri použití sigmoid kernel funkcie ako jediná nepredikovala zvyčajný vyšší počet cestujúcich.

kernel	ϵ	C	γ	c_0	d	posun	chyba predikcie
lineárna	0,1	1	-	-	-	15	0,1172723
polynomiálna	0,1	1	0,1	1	2	17	0,1347131
radiálna	0,1	4	0,1	-	-	16	0,1425461
sigmoid	0,8	1	0,1	0,1	-	15	0,1822829



Obr. 6.4: Reálne počty cestujúcich (čierna) s predikciami pomocou lineárnej (modrá), polynomiálnej (žltá), radiálnej (zelená) a sigmoid (červená) kernel funkcie

7 Porovnanie metód

V kapitolách 2, 4 a 6 sme si ukázali použitie 3 metód predikcie časových radov na 4 príkladoch. Jednotlivé sady mali rozličný charakter a najlepšie výsledky dosahovali pre rôzne metódy. V tejto kapitole spravíme porovnanie, či je možné na základe zaradenia dát do jednej zo štyroch kategórií určiť, ktorá metóda predikcie je najvhodnejšia na použitie.

HDP	chyba predikcie
SSA	0,4016775
ARIMA	0,09551474
SVM - lineárna	0,1347675
SVM - polynomiálna	0,5190495
SVM - radiálna	0,3349916
SVM - sigmoid	**0,04798223**

Nezamestnanosť	chyba predikcie
SSA	0,9958386
ARIMA	**0,4128969**
SVM - lineárna	0,7128078
SVM - polynomiálna	0,8448795
SVM - radiálna	0,4794341
SVM - sigmoid	1,58324

Tabuľka 7.1: Porovnanie NRMSE pre dáta HDP Slovenska a miery nezamestnanosti Slovenska

Kvartálne dáta HDP Slovenska reprezentovali kategóriu s výrazným trendom a periodickou zložkou. V Tabuľke 7.1 vidíme, že najlepšie výsledky dosiahla metóda SVM s polynomiálnou kernel funkciou. Druhý príklad, miera nezamestnanosti, kde sme pracovali s mesačnými dátami, nevykazoval výraznú periodicitu a celý priebeh dát spočíval v miernom nelineárnom trende. Najnižšiu hodnotu chyby predikcie NRMSE sme dostali pri použití ARIMA modelu. V príklade s úhrnom zrážok sa dáta pohybovali okolo priemernej hodnoty a nebola pozorovateľná výrazná sezónnosť. Najnižšiu chybu sme získali pri po-

Zrážky	chyba predikcie
SSA	0,3742457
ARIMA	0,32567
SVM - lineárna	0,3461508
SVM - polynomiálna	**0,3051453**
SVM - radiálna	0,3945647
SVM - sigmoid	0,3153998

Letisko	chyba predikcie
SSA	0,1155622
ARIMA	**0,09813038**
SVM - lineárna	0,1172723
SVM - polynomiálna	0,1347131
SVM - radiálna	0,1425461
SVM - sigmoid	0,1822829

Tabuľka 7.2: Porovnanie NRMSE pre dáta úhrnu zrážok a počtu cetujúcich na letisku

užití SVM regresie s polynomiálnym kernelom. Posledná sada dát, počet cestujúcich na bratislavskom letisku, bola príkladom periodicity bez výrazného rastúceho či klesajúceho trendu. Rovnako ako v druhom príklade sme dosiahli najnižšiu hodnotu NRMSE pri použití ARIMA modelu.

V každej kategórii si predstavíme 3 sady dát, pre ktoré spravíme predikcie. Pozrieme sa, či dosiahneme najlepšie výsledky pre rovnaké metódy ako v príkladoch z Kapitoly 2. V kategórii HDP, dáta s trendom a sezónnou zložkou, máme dáta priemerných mesačných miezd na Slovensku v rokoch 2008 až 2016 pochádzajúce z [14], mesačné dáta pôrodnosti na Slovensku pre roky 1993 až 2016 z [27] a kvartálne dáta systémových dávok nemocenského poistenia na Slovensku za obdobie rokov 1995 až 2014 z [14].

	mzdy	pôrodnosť	dávky
SSA	0,2688858	0,3852906	0,515798
ARIMA	**0,1690502**	0,1697844	**0,4907857**
SVR - lineárna	0,3787617	**0,1594707**	0,8060034
SVR - polynomiálna	0,5585707	0,2004942	0,6479373
SVR - radiálna	1,043093	0,2011318	0,6755024
SVR - sigmoid	1,663597	0,5524541	2,593362

Tabuľka 7.3: Porovnanie NRMSE pre dáta z kategórie HDP

Vo všetkých troch uvedených príkladov dosiahla SVM regresia za použitia sigmoid kernel fukncie najhoršie výsledky (Tab. 7.3). V dvoch prípadoch sme najnižšie hodnoty chyby predikcie dostali pri použití ARIMA modelu.

Druhou kategóriou sú dáta bez sezónnosti s miernym trendom ako sme mali v príklade nezamestnanosti. Ako prvé sme predikovali hodnoty mesačných dát ceny ropy, ktoré sme mali k dispozícii za obdobie rokov 1993 až 2016. Ďalej sme pracovali s kvartálnymi dátami priemerných cien nehnuteľností za m^2 na Slovensku za roky 2002 až 2015. Obe sady pochádzali z [14]. Ako posledná bola sada mesačných dát cien potravinárskej pšenice z [27]. Údaje sme mali pre roky 2010 - 2016.

	ropa	nehnuteľnosti	pšenica
SSA	0,9048073	0,2623794	0,4712774
ARIMA	0,5999012	**0,153308**	0,7391462
SVR - lineárna	0,4773018	0,5059938	**0,160925**
SVR - polynomiálna	0,3905076	0,4419517	0,4815422
SVR - radiálna	0,7034383	0,7205617	1,029932
SVR - sigmoid	**0,2859937**	0,6872102	0,4060127

Tabuľka 7.4: Porovnanie NRMSE pre dáta z kategórie nezamestnanosť

Pri predikcii miery nezamestnanosti sme dosiahli najlepšie výsledky pri ARIMA modeloch. V Tabuľke 7.4 vidíme, že len jednom príklade získala ARIMA metóda najnižšiu hodnotu chyby NRMSE. V dvoch prípadoch bola metóda oporných bodov najlepšia. Predikcia pomocou radiálneho kernelu dosiahla v každom prípade najhoršie alebo druhé najhoršie výsledky.

Priemerný úhrn zrážok charakterizuje tretiu kategóriu dát bez sezónnoti a pohybujúcej sa okolo priemernej hodnoty. Vybrali sme si mesačné dáta počtu usmrtených osôb na cestách v dopravných nehodách na Slovensku. Dáta máme od roku 2010 do roku 2016 a pochádzajú z [16]. Druhá sada sú mesačné dáta dovozu elektriny na Slovensko v GWh z [27]. Posledné sú mesačné dáta registrovaných automobilov na Slovensku za roky 2008 až 2016 z [14].

	usmrteni	elektrina	automobily
SSA	**0,3097751**	0,3852906	**0,2395952**
ARIMA	0,3208192	**0,3425063**	0,3088445
SVR - lineárna	0,3445158	0,4758367	0,3124442
SVR - polynomiálna	0,333809	0,38353	3,269554
SVR - radiálna	0,3282796	0,3944728	0,3516725
SVR - sigmoid	0,4060127	0,8919862	0,5433425

Tabuľka 7.5: Porovnanie NRMSE pre dáta z kategórie zrážky

Predikcia metódou SVM za použitia polynomiálnej kernel funkcie nedosiahla v ani jednom prípade najnižšiu chybu predikcie (Tab. 7.6), hoci v prvých dvoch príkladoch jej výsledky neboli výrazne horšie oproti ostatným metódam. Najlepšie výsledky sme dosahli pri metódach SSA a ARIMA.

Poslednú kategóriu reprezentujú dáta cestujúcich na letisku, sezónnosť bez rastúceho alebo klesajúceho trendu. Patria sem dáta mesačných počtov dopravných nehôd za Slovensku na roky 2011 - 2016 pochádzajúce z [16]. Druhou sadou boli mesačné dáta priemerných teplôt v Moste pri Bratislave za obdobie 2010 až 2016 [15]. Ako posledné sme použili mesačné dáta počtu sobášov v Lotyšsku za roky 2009 až 2015 [26].

	nehody	teploty	sobáše
SSA	0,2207064	0,07710743	0,1019485
ARIMA	**0,07354533**	0,0827738	**0,06121206**
SVR - lineárna	0,223652	0,08106167	0,06881735
SVR - polynomiálna	0,2207867	**0,07330694**	0,1068443
SVR - radiálna	0,1731442	0,08436764	0,08330493
SVR - sigmoid	0,2859937	0,08373107	0,1565601

Tabuľka 7.6: Porovnanie NRMSE pre dáta z kategórie letisko

Môžeme si všimnúť, že v prípade sady dát teploty bol minimálny rozdiel medzi jednotlivými metódami (Tab 7.6). V dvoch prípadoch dosiahla dosiahol ARIMA model najnižšiu chybu predikcie.

Len v troch prípadoch dosiahla najnižšiu chybu predikcie rovnaká metóda ako v príkladoch predstavených v predchádzajúcich kapitolách. Predikcie pomocou ARIMA modelov dosiahli v šiestich prípadoch najnižšie hodnoty chyby predikcie. Len v prípade cien pšenice dosiahli výrazne horšie výsledky, a teda celkovo bola táto metóda najspoľahlivejšia. Metóda SSA v dvoch príkladoch dosiahla najnižšiu chybu predikcie. Najhoršie výsledky dosahovala pri dátach bez sezónnej zložky. Pri predikciách pomocou tejto metódy je najdôležitejším krokom vybrať pri rozklade časové rady, ktoré sú od seba separovateľné. V príkladoch s výraznou sezónnou zložkou sme videli, že vo w-korelačnej matici dosahovali časové rady nulové korelácie medzi sebou a tým pádom separovateľnosť bola dodržaná.

Nevýhodou SSA a ARIMA modelu je, že je potrebné vybrať parametre na základe analýzy priebehu časového radu. Na rozdiel od týchto metód, pri SVM môžeme urobiť predikiciu bez nejakej predošlej analýzy časového radu. Na druhej strane, pri použití metódy opor-ných bodov je potrebné riešiť optimalizačnú úlohu kvadratického programovania a tým pádom je táto metóda výpočtovo a teda aj časovo najnáročnejšia.

Záver

Cieľom tejto diplomovej práce bolo predstaviť a porovnať rôzne metódy predikcie časových radov. Na porovnanie sme vybrali tri metódy - analýzu singulárneho spektra, ARIMA modely a regresiu použitím metódy oporných bodov.

Na výsledky predikcií sa môžeme pozrieť z dvoch perspektív. Prvou je, aké chyby predikcie dosiahla daná metóda pri rôznych typoch dát. Metóda SSA dosahovala najlepšie výsledky pri dátach, v ktorých bola sezónna zložka bez výrazného trendu. Najväčšie problémy mala s dátami typu nezamestnanosť, ktoré nemali sezónnu zložku. V týchto príkladoch sme mali pri rozklade časového radu problém so separovateľnosťou, a preto boli pri výsledkoch predikcie výraznejšie chyby. Predikcie za použitia ARIMA modelov vychádzali globálne najpresnejšie. Z použitých metód nemali nikdy najväčšiu chybu. Nevýhodou SSA a ARIMA modelov je, že je potrebné každý jeden príklad osobne analyzovať a rozhodovať o výbere parametrov. Na druhej strane, predikciu pomocou SVM regresie je možné zautomatizovať a nie je potrebné vstupovať do procesu predikcie. Časovo bola táto metóda najnáročnejšia, z dôvodu, že sme v každom príklade prechádzali rôzne možnosti výberu počtu predchádzajúcich hodnôt použitých na predikciu. Tento problém by sa dal obísť, ak by sme vždy počet použitých hodnôt vyberali na základe frekvencie dát, hoci v tomto prípade by chyby predikcie mohli byť vyššie.

Cieľom tejto práce bolo zistiť, či je možné určiť, ktorá metóda dosiahne najnižšiu chybu predikcie len na základe toho, či ide o dáta s alebo bez sezónnej zložky alebo trendu. Na základe chýb predikcií sme zistili, že nie je možné určiť, ktorá metóda dosiahne najlepšie výsledky pri danom type dát. V rámci jednotlivých kategórií bolo možné vylúčiť použitie niektorých metód, ale presne určiť, pomocou ktorej metódy získame najnižšiu chybu predikcie, nebolo možné.

Zoznam použitej literatúry

[1] Azoff, E. M.: *Neural Network Time Series Forecasting of Financial Markets*, John Wiley & Sons, New York, USA, 1994

[2] Boser, B.E., Guyon, I.M., Vapnik, V.N.: *Atraining algorithm for optimal margin classifiers*, Proceedings of the Annual Conference on Computational Learning Theory (1992), Pittsburgh, 144–152

[3] Box, G.E.P, Jenkins, G.M.: *Time Series Analysis: Forecasting and Control*, San Francisco, CA: Holden-Day, 1970

[4] Broomhead, D., King, G.: *Extracting qualitative dynamics from experimental data*, Physica D 20 (1986), 217-236

[5] Broomhead, D., King, G.: *On the Qualitative Analysis of Experimental Dynamical Systems*, 1986, v knihe: Nonlinear Phenomena and Chaos, Publisher: Adam Hilger, Bristol, 113-144

[6] Cortes C., Vapnik, V.: *Support vector networks*, Machine Learning 20 (1995), 273–297

[7] Cowpertwait, P. S. P., Metcalfe, A. V.: *Introductory Time Series with R*, Springer, Dordrecht, 2009

[8] Danilov, D., Zhigljavsky, A.: *Main components of time series: the "Caterpillar" method*, St. Petersburg Press, St. Petersburg, 1997

[9] Fan, R. E., Chen, P. H., Lin, C. J.: *Working set selection using second order information for training support vector machines*, Journal of machine learning research 6 (2005), 1889-1918

[10] Golyandina, N., Zhigljavsky, A.: *Singular Spectrum Analysis for Time Series*, Springer, 2013

[11] Golyandina, N., Nekrutkin, V., Zhigljavsky, A.: *Analysis of Time Series Structure: SSA and related techniques*, Chapman & Hall/CRC, New York - London, 2001

[12] Golyandina, N., Osipov, E.: *The "Caterpillar"-SSA method for analysis of time series with missing values*, Journal of Statistical Planning and Inference 137 (2007), 2642 – 2653

[13] Golyandina, N., Korobeynikov, A.: *Basic Singular Spectrum Analysis and forecasting with R*, Computational Statistics & Data Analysis, Volume 71 (2014), 934-954

[14] Makroekonomická databáza NBS, dostupné na internete (11.12.2017): https://www.nbs.sk/sk/menova-politika/makroekonomickadatabaza/makroekonomicke-ukazovatele-graf

[15] Meteorologická stanica Most pri Bratislave, internetová stránka, dostupné na internete (15.4.2018): http://meteocentrum.sk/most/wxtempsummary.php?

[16] Ministerstvo vnútra SR, internetová stránka, dostupné na internete (16.4.2018): https://www.minv.sk/?dopravna-nehodovost-podla-mesiacov

[17] Lapin, M., Oddelenie meteorológie a klimatológie, FMFI UK, Bratislava, webová stránka (2.2.2018): http://www.dmc.fmph.uniba.sk/public_html/main8.html

[18] Rosenblatt, F.: *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review 65 (1958), 386-408

[19] Rosenblatt, F.: *Principles of Neurodynamics*, Spartan Books, Washington D.C., 1962

[20] Rumelhart, D., Hinton, G., Williams, R.: *Learning representations by back-propagating errors*, Nature 323 (1986), 533–536

[21] Slovak Aviation, webová stránka (1.3.2018): https://slovakaviation.sk/airport-info/bratislava/statistiky

[22] Schölkopf, B.: *Support Vector Learning*, Oldenbourg Verlag, München, 1997

[23] Schölkopf, B., Smola, A. J.: *A Tutorial on Support Vector Regression*, Statistics and Computing 14 (2004), 199–222, dostupné na internete (11.3.2018): https://alex.smola.org/papers/2004/SmoSch04.pdf

[24] Schölkopf, B., Smola, A. J.: *Learning with Kernels*, MIT Press, 2002

[25] Stehlíková B., študijné materiály k predmetu Časové rady, dostupné na internete (4.3.2018):
http://www.iam.fmph.uniba.sk/institute/stehlikova/cr16.html

[26] Štatistický úrad Lotyšska, databáza, dostupné na internete (20.4.2018):
http://data.csb.gov.lv/pxweb/en/Sociala/Sociala___isterm___iedz

[27] Štatistický úrad SR, databáza DATAcube, dostupné na internete (17.4.2018):
http://datacube.statistics.sk

[28] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer, New York, 1995

[29] Výročná správa bratislavského letiska za rok 2012, dostupné na internete (11.3.2018):
https://www.bts.aero/downloads/rocne-spravy/rocna-sprava-2012.pdf

[30] Walker, G.: *On Periodicity in Series of Related Terms*, Proceedings of the Royal Society of London, Ser. A, Volume 131 (1931), 518–532

[31] Wold, H.: *A study in the analysis of stationary time series*, Almqvist&Wiksell, Stockholm, 1938

[32] Yule, G. U.: *On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers* Philosophical Transactions, Ser. A, Volume 226 (1927), 267-298

[33] Zhang, G.P.: *Time series forecasting using a hybrid ARIMA and neural network model*, Neurocomputing 50 (2003), 159–175

Príloha A

Kód k metóde SSA v jazyku R:

```
1
2  library(Rssa)
3
4  priklad <- read.table("porodnost.txt", header = TRUE)
5  priklad <- ts(priklad, frequency = 12, start = 1993)
6
7  ######################################
8
9  p1 <- ssa(priklad[,1], L = 12)
10 summary(p1)
11
12 plot(p1)
13 plot(p1, type = "vectors")
14 plot(p1, type = "paired")
15 plot(wcor(p1), main = "␣")
16
17 res1       <- reconstruct(p1, groups = list(1))
18 trend      <- res1$F1
19 res.trend <- residuals(res1)
20
21 plot(res1, add.residuals = FALSE, plot.type = "single",
22      col = c("black", "red"))
23
24 ######################################
25
26 p2 <- ssa(res.trend, L=24)
27 plot(p2)
28 plot(p2, type = "vectors")
29 plot(p2, type = "paired")
30 plot(wcor(p2))
31
32 res2 <- reconstruct(p2, groups = list(1:2))
33 seasonality <- res2$F1
34
35 plot(priklad, lwd = 1)
36 lines(trend+seasonality, col = "red", lwd = 1)
37
38 res3 <- reconstruct(p2, groups = list(3:4))
39 seasonality2 <- res3$F1
40
41 plot(priklad)
42 lines(trend+seasonality+seasonality2, col = "red")
43
44 ######################################
45
46 # chyba NRMSE
47 REKONSTRUCT <- trend+seasonality+seasonality2
48 rmse        <- sqrt(sum((REKONSTRUCT-priklad[,1])^2)/length(priklad))
49 (nrmse      <- rmse/(max(priklad)-min(priklad)))
50
51 ######################################
52
53 # R-predikcia
54 n_predikcia <- 12
55 n_rad       <- dim(priklad)[1]
56 rtrend      <- rforecast(p1, groups = list(1),
57                          len = n_predikcia, only.new = FALSE)
58 rseason1    <- rforecast(p2, groups = list(1:2),
59                          len = n_predikcia, only.new = FALSE)
60 rseason2    <- rforecast(p2, groups = list(3:4),
61                          len = n_predikcia, only.new = FALSE)
```

```
rpredikt     <- rtrend + rseason1 + rseason2
prikladP     <- read.table("porodnostP.txt", header = TRUE)
prikladP     <- ts(prikladP, frequency = 12, start = c(2016, 12))

# vykreslenie casoveho radu s predikciou
plot(rpredikt, col = "red")
lines(priklad, col = "black")
lines(prikladP, col = "blue")

##########################################

# chyba predikcie
rmseP     <- sqrt(sum((rpredikt[(n_rad+1):(n_rad+12)]
                       -prikladP[-1,1])^2)/length(prikladP[-1,1]))
(nrmseP   <- rmseP/(max(prikladP[-1,1])-min(prikladP[-1,1])))
```

Listing 1: Kód SSA

Kód k metóde ARIMA v jazyku R:

```
library(astsa)
library(urca)

priklad <- read.table("usmrteni.txt", header = TRUE)
priklad <- ts(priklad, frequency = 12, start = 2010)

# vykreslenie casoveho radu
plot(priklad)
plot(log(priklad))
plot(diff(log(priklad)))

#####################################

prikladN <- log(priklad)
plot(prikladN)
acf2(prikladN)

plot(diff(priklad, 12))
acf2(diff(priklad, 12))

# testovanie hypotezy o jednotkovom koreni
# nulova hypoteza H0: je jednotkovy koren, L=1
mean(diff(prikladN, 12))
dftest <- ur.df(diff(prikladN, 12), lags = 12, type='none',
                selectlags="BIC")
summary(dftest)

# zostavenie modelu a predikcia
model0 <- sarima(prikladN, 0, 0, 0, 1, 1, 1, 12)
predikcie <- sarima.for(prikladN, 12, 0, 0, 0, 1, 1, 1, 12)

prikladP <- read.table("usmrteniP.txt", header=TRUE)
prikladP <- ts(priklad2, frequency = 12, start = c(2016,12))
plot(prikladP)

plot(priklad, col = "black", xlim = c(2010, 2018))
lines(prikladP, col = "blue")
lines( exp(predikcie$pred), col = "red")

# chyba NRMSE
rmseP    <- sqrt(sum((exp(predikcie$pred)-priklad2[-1,1])^2)
                 /length(priklad2[-1,1]))
nrmseP   <- rmseP/(max(priklad2[-1,1])-min(priklad2[-1,1]))
```

Listing 2: Kód ARIMA

Kód k metóde SVM v jazyku R:

```
1
2  library(e1071)
3
4  priklad   <- read.table("usmrteni.txt", header = TRUE)
5  priklad   <- priklad[,1]
6  n         <- length(priklad)
7  freq      <- 12
8
9  prikladP  <- read.table("usmrteniP.txt", header = TRUE)
10 prikladP  <- prikladP[-1,1]
11
12 Start.time <- Sys.time()
13 chyba1 <- NA; chyba2 <- NA; chyba3 <- NA; chyba4 <- NA
14 parametre1 <- NA; parametre2 <- NA; parametre3 <- NA
15 parametre4 <- NA
16
17 for (j in 2:(2*freq))
18 {
19   priklad_input <- priklad[j:(n-1)]
20   for (i in 2:j)
21   {
22     priklad_input <-
23       as.data.frame(cbind(priklad_input, priklad[((j+1)-i):(n-i)]))
24   }
25   priklad_input   <-
26       as.data.frame(cbind(priklad_input, priklad[(j+1):n]))
27
28   # rozdelenie dat na trenovaciu a testovaciu sadu
29   dlzka <- floor(dim(priklad_input)[1]*0.75)
30   priklad_train <- priklad_input[1:dlzka,]
31   priklad_test <- priklad_input[(dlzka+1):(n-j),]
32
33   #odhad parametrov
34   priklad_tune1 <- tune(svm, train.x = priklad_train[,1:j],
35                         train.y = priklad_train[,j+1],
36                         validation.x = priklad_test[,1:j],
37                         validation.y = priklad_test[,j+1],
38                         kernel = "linear",
39                         ranges = list(epsilon=seq(0.1,1,0.1),
40                           cost=2^(0:4)),
41                         type = "eps-regression")
42
43   priklad_tune2 <- tune(svm, train.x = priklad_train[,1:j],
44                         train.y = priklad_train[,j+1],
45                         validation.x = priklad_test[,1:j],
46                         validation.y = priklad_test[,j+1],
47                         kernel = "polynomial",
48                         ranges = list(epsilon=seq(0.1,1,0.1),
49                           cost=2^(0:4), gamma = seq(0,0.4,0.1),
50                           degree = 2:5, coef0 = seq(0.1,1,0.1)),
51                         type = "eps-regression")
52
53   priklad_tune3 <- tune(svm, train.x = priklad_train[,1:j],
54                         train.y = priklad_train[,j+1],
55                         validation.x = priklad_test[,1:j],
56                         validation.y = priklad_test[,j+1],
57                         kernel = "radial",
58                         ranges = list(epsilon=seq(0.1,1,0.1),
59                           cost=2^(0:4), gamma = seq(0,0.4,0.1)),
60                         type = "eps-regression")
```

```
  priklad_tune4 <- tune(svm, train.x = priklad_train[,1:j],
                        train.y = priklad_train[,j+1],
                        validation.x = priklad_test[,1:j],
                        validation.y = priklad_test[,j+1],
                        kernel = "sigmoid",
                        ranges = list(epsilon=seq(0.1,1,0.1),
                          cost=2^(0:4), gamma = seq(0,0.4,0.1),
                          coef0 = seq(0.1,0.5,0.1)),
                        type = "eps-regression")

  parametre1[((j-2)*2+1):((j-1)*2)] <- priklad_tune1$best.parameters
  chyba1[j-1] <- priklad_tune1$best.performance

  parametre2[((j-2)*5+1):((j-1)*5)] <- priklad_tune2$best.parameters
  chyba2[j-1] <- priklad_tune2$best.performance

  parametre3[((j-2)*3+1):((j-1)*3)] <- priklad_tune3$best.parameters
  chyba3[j-1] <- priklad_tune3$best.performance

  parametre4[((j-2)*4+1):((j-1)*4)] <- priklad_tune4$best.parameters
  chyba4[j-1] <- priklad_tune4$best.performance
}
End.time <- Sys.time()
(cas <- End.time - Start.time)

#########################################################

#LINEARNA KERNEL FUNKCIA

# vyberame hodnoty parametrov, s ktorymi sme dosiahli
# najnizsiu chybu predikcie na testovacej sade
j <- which(chyba1 == min(chyba1)) + 1
para1 <- parametre1[((j-2)*2+1):((j-1)*2)]

# vytvorenie matice vstupnych dat
priklad_input <- priklad[j:(n-1)]
for (i in 2:j)
{
  priklad_input <-
    as.data.frame(cbind(priklad_input, priklad[((j+1)-i):(n-i)]))
}
priklad_input   <-
    as.data.frame(cbind(priklad_input, priklad[(j+1):n]))

for (i in 1:j)
{
  colnames(priklad_input)[i] <- paste0("x", i)
}
colnames(priklad_input)[j+1] <- "y"

dlzka <- floor(dim(priklad_input)[1]*0.75)
priklad_train <- priklad_input[1:dlzka,]
priklad_test <- priklad_input[(dlzka+1):(n-j),]

# vytvorenie modelu
model1 <- svm(priklad_input$y ~ ., data = priklad_input,
              kernel = "linear", type = "eps-regression",
              epsilon = para1[1], cost = para1[2])
```

```
# predikovanie hodnot
priklad_aktualne <- priklad_input
for (i in 1:12)
{
  n1 <- dim(priklad_aktualne)[1]
  priklad_aktualne <-
          rbind(priklad_aktualne, c(priklad_aktualne[n1,j+1],
          priklad_aktualne[n1,1:(j-1)],0, use.names = FALSE))
  predikcia1 <- predict(model1, priklad_aktualne[n1+1,1:j])
  priklad_aktualne[n1+1,j+1] <- predikcia1
}

n <- dim(priklad_aktualne)[1]
priklad_predikcia1 <- priklad_aktualne$y[n:(n+freq)]

# chyba predikcie NRMSE
rmseP1     <- sqrt(sum((priklad_predikcia1[-1]-prikladP)^2)
                   /length(prikladP))
(nmrseP1  <- rmseP1/(max(prikladP) - min(prikladP)))
```

Listing 3: Kód SVR