

Roman PAVELKA
Štatistický úrad SR

ANALÝZY ČASOVÝCH ŘAD POMOCÍ PROGRAMOVACÍHO JAZYKA R

ANALYSES OF TIME SERIES WITH THE R PROGRAMMING LANGUAGE

1. ÚVOD

Posloupnost hodnot sledovaného ekonomického, fyzikálního či jiného ukazatele, které jsou uspořádány v čase, se nazývá časová řada [6]. S fenoménem časové řady se dříve nebo později setká každý, kdo je postaven před problém analyzovat určitý ekonomický jev, ať už jde o makroekonomické ukazatele – např. vývoj agregátů tvorby a užití hrubého domácího produktu, inflace, nezaměstnanosti apod. – nebo o některé dílčí ukazatele, jako je vývoj kurzů cizích měn, peněžní zásoby, cen akcií na kapitálovém trhu či průmyslové nebo stavební produkce.

Analýza časových řad je jednou z nejvýznamnějších aplikací kvantitativních metod v ekonomii. Kvantitativní metody a techniky používané v analýzách časových řad zpravidla patří k velmi komplexním a výpočetně velmi náročným. Z tohoto důvodu si příslušná statistická analýza vyžaduje zapojení výkonné výpočetní techniky včetně nezbytného specializovaného statistického softwaru. Ke statistickým analýzám časových řad existuje množství statistických programů. Většina specializovaných statistických programů (např. Stata, EViews, SPSS, apod.) nabízí srovnatelné analytické metody a techniky, které umožňují požadovanou analýzu a modelování časových řad. Jejich společnou nevýhodou je však skutečnost, že jejich pořízení a aktualizace je ekonomicky velmi nákladné.

Jedním z volně dostupných (a tedy je jeho použití zcela zdarma) specializovaných statistických programů je programovací jazyk R. Uvedený programovací jazyk se vyvinul v komplexní systém na manipulaci s daty, jejich statistickou analýzu a zpracování i grafickou reprezentaci. Programovací jazyk R vychází z jazyka S, který byl vyvinut v Bell Laboratories (předtím AT&T, nyní Lucent Technologies) jako komerční softwarový nástroj pro analýzy a vyhodnocování dat. Programovací jazyk R však představuje volně dostupný softwarový nástroj, který je šířen a používán v rámci tzv. open-source¹ projektů nadace *Free Software Foundation*². Snadná dostupnost a volné používání programovacího jazyka R umožňuje jeho neustálý rozvoj, který je realizován zejména v podobě mnoha doplňkových programových balíčků (*add-on packages*) s knihovnamí a funkcemi na různé typy specializovaných analýz. Proto programovací jazyk R získává stále větší význam nejen v komerční a akademické sféře, resp. ve vědeckovýzkumných pracovištích, ale nabývá na důležitosti i při

¹ *Otevřený software (anglicky open-source software nebo open software, zkratka OSS) je počítačový software s otevřeným zdrojovým kódem. Otevřenost zde znamená jak technickou dostupnost kódu, tak legální dostupnost – licenci software, která umožňuje, při dodržení jistých podmínek, uživatelům zdrojový kód využívat, například prohlížet a upravovat.*

² *Nadace Free Software Foundation, česky Nadace pro svobodný software, byla založena v roce 1985 s cílem podporovat práva uživatelů počítačů používat, studovat, kopírovat, modifikovat a redistribuovat počítačové programy.*

produkci výstupů státní statistiky v národních statistických úřadech. V současné době se programovací jazyk R používá v produkčních procesech národních statistických úřadů Rakouska, Itálie, Holandska, Rumunska a částečně i Slovenské republiky a jiných států EU i v samotném Eurostatu.

2. POUŽITÉ ČASOVÉ ŘADY

Pro účely tohoto příspěvku byly vybrány ekonomické časové řady, které popisují nejdůležitější ekonomické ukazatele produkované státní statistikou. Jedná se o časovou řadu hrubého domácího produktu měřeného v běžných cenách v mld. EUR, která byla zjišťována se čtvrtletní periodicitou od 1. čtvrtletí 1995 do 2. čtvrtletí roku 2016 [8]. Další sledovanou řadou je časová řada meziměsíční míry inflace (v jednotkách %), jejíž hodnoty byly pozorovány s periodicitou měsíční, a to od ledna roku 2002 do srpna roku 2016 [9]. Jako poslední časová řada použitá v rámci tohoto příspěvku byla vybrána časová řada indexů spotřebitelských cen podle Klasifikace individuální spotřeby podle účelu (COICOP) zjišťovaná měsíčně ve stejném časovém období jako předchozí časová řada [10]. Vybrané časové řady jsou volně dostupné na WWW stránkách Statistického úřadu SR v databázi SlovStat³. Výběr časových řad byl proveden tak, aby časové řady byly dostatečně dlouhé a také významné z ekonomického pohledu. Délka jednotlivých časových řad je určena dostupností údajů v databáze SlovStat.

3. ZÁKLADNÍ METODY ANALÝZY ČASOVÝCH ŘAD POMOCÍ R

3.1. Import časové řady do programovacího jazyku R a její vizualizace

Aby bylo možné časové řady analyzovat v programovacím jazyku R, je nutné data zkoumaných časových řad do tohoto systému importovat. Import dat do R může být proveden různými způsoby – například R nabízí programové prostředky pro import ASCII a binárních dat, dat z jiných aplikací nebo import dat přímo z databází pomocí ODBC připojení. Způsobem přípravy dat, který lze považovat jako optimální, je vytvoření textového souboru, který obsahuje výlučně samotná data časových řad strukturovaně uložených do pojmenovaných sloupců. Data z každého sloupce jsou oddělena středníkem, desetinná místa odděluje čárka. Každý řádek souboru reprezentuje jedno pozorování časové řady v příslušném časovém období. Jakmile importovaná data dosáhnou tuto podobu, celý takto vytvořený soubor se uloží ve formě textového souboru s rozšířením CSV⁴.

Po konverzi dat do textového formátu CSV data časových řad mohou být vloženy do programovacího jazyka R. Příkazem

³ Základní informace o databázi: databáze SLOVSTAT obsahuje časové řady ukazatelů hospodářského a sociálně-ekonomického vývoje za Slovenskou republiku. Údaje z různých statistických okruhů se prezentují ve formě tabulek v měsíčních, čtvrtletních nebo ročních časových řadách.

⁴ CSV (Comma-separated values, hodnoty oddělené čárkami) je jednoduchý souborový formát určený pro výměnu tabulkových dat. Soubor ve formátu CSV sestává z řádků, ve kterých jsou jednotlivé položky odděleny znakem čárka (.). Hodnoty položek mohou být uzavřeny do uvozovek ("), což umožňuje, aby text položky obsahoval čárku. Pokud text položky obsahuje uvozovky, jsou tyto zdvojeny. Jelikož se v některých jazycích včetně češtiny čárka používá v číslech jako oddělovač desetinných míst, existují varianty, které používají jiný znak pro oddělování položek než čárku, nejčastěji středník, případně tabulátor (taková varianta se pak někdy označuje jako TSV, Tab-separated values). Variantu se středníkem (ale stále pod názvem CSV) používá např. Microsoft Excel v české verzi Microsoft Windows (řídí se oddělovačem zadaným v místním a jazykovém nastavení). Díky jednoduchosti, nenáročnosti a čitelnosti i bez specializovaného software se tento formát používá pro výměnu informací mezi různými systémy.

```
HDP <- read.csv("D:/pom/HDP/dataTab_col.csv", header=T, dec="," , sep=";")
```

se časová řada obsahující informace o čtvrtletních hodnotách HDP načte do programovacího jazyka R. Použitý příkaz jazyka R obsahuje mimo potřebné cesty k importovanému souboru ("D:/pom/HDP/dataTab_col.csv") také informaci o tom, jestli importovaný soubor obsahuje i názvy datových sloupců a čím jsou odděleny jednotlivé údaje na 1 řádku. Proměnná nazývaná HDP obsahuje ukazatel na datové pole (v programovacím jazyku R se nazývá *dataframe*). Časové řady se ukládají jako datové matice, k nimž je možné přistupovat pomocí hranatých závorek ([]). Řádky a sloupce datových matic jsou dostupné pomocí příkazů *název_dat_matice[řádek, sloupec]*.

Programové prostředí jazyka R obsahuje z pohledu časových řad struktury pro reprezentaci a analýzu dat časových řad. Fundamentální strukturami [7] pro časové řady jsou programové třídy nazývané *ts* a *zoo*. Objekty třídy *ts* jsou buď numerické vektory v případě jednorozměrných řad, nebo numerická matice pro vícerozměrné řady. Tyto objekty slouží především pro uchování rovnoměrně rozložené časové řady (pomocí číselných časových razítek). Proto se velmi dobře hodí pro roční, měsíční, čtvrtletní data, atd. Atribut třídy *ts* nazývaný *tsp* (*time series properties*) reprezentuje základní vlastnosti časové řady. Jedná se o vektor o délce 3 prvků, který obsahuje počátku časové řady, ukončení časové řady (v časových jednotkách) a její periodicitu (frekvenci). Zobecněním objektů typu *ts* jsou objekty typu *zoo*, které umožňují pracovat i s časovými řadami neperiodickými, resp. umožňují práci i s časovými řadami s chybějícími pozorováními.

Příkazem programovacího jazyka R

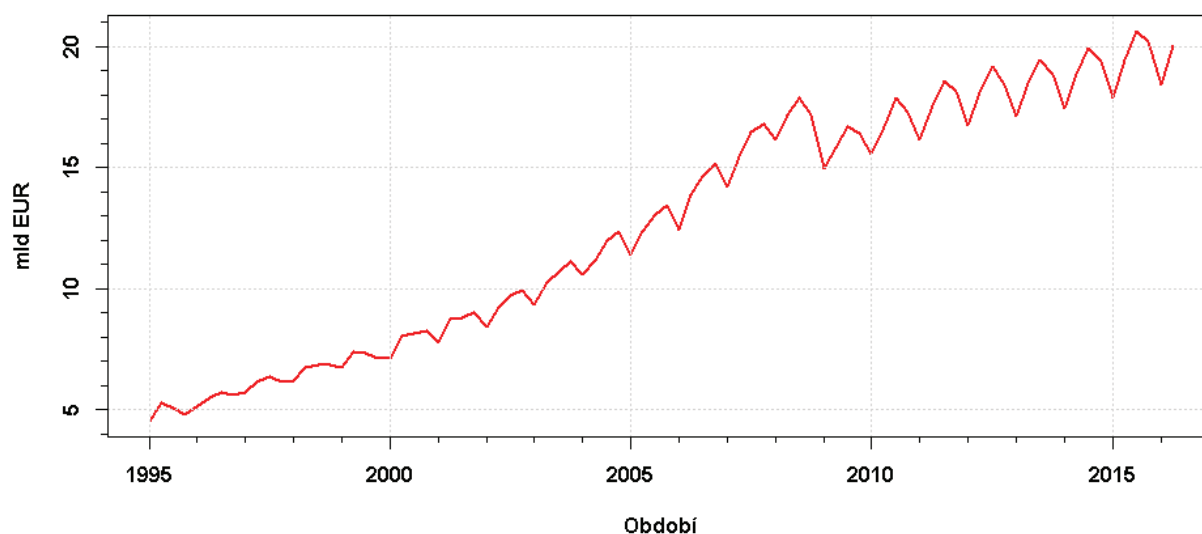
```
HDPTs<-ts(HDP[,2], frequency = 4, start=c(1995,1))
```

se vytvoří v prostředí programu R datová struktura jednorozměrné časové řady, která obsahuje pozorované hodnoty HDP v běžných cenách v miliardách EUR, která začíná v 1. čtvrtletí roku 1995 se čtvrtletní frekvencí pozorování.

Prvotní informace pro analýzu časových řad získáme ze spojnicových grafů. Jejich princip spočívá v zakreslení jednotlivých hodnot časové řady do souřadných os, na kterých jsou vyznačeny příslušné stupnice. Například zakreslení průběhu časové řady HDP v běžných cenách miliard EUR bylo realizováno pomocí příkazu

```
plot(HDPTs, type="l", lwd=2, col="red", xlab="Čtvrtletí", ylab="mld EUR", main="HDP v b. c. v mld EUR")
```

Příkaz k vykreslení grafu *plot()* dovoluje mnoho volitelných parametrů, jeden z parametrů je *type="l"*, který nastavuje typ grafu na spojnicový. Výsledkem uvedeného grafického příkazu je graf průběhu časové řady „HDP v b. c. v mld. EUR“, který je uveden na obrázku 1.

Obrázek č. 1: Vizuální průběh časové řady „HDP v b. c. v mld. EUR“

Zdroj: vlastní konstrukce

Podobným způsobem se zobrazí vizuální průběh časové řady roční míry inflace a časové řada indexů spotřebitelských cen (CPI) podle Klasifikace individuální spotřeby podle účelu (COICOP) zjišťovaných měsíčně.

Do programového prostředí jazyka R lze importovat časové řady nejen textového formátu (s koncovkou .CSV), ale také formátu SAS, strukturovaných formátů, jako je DBF, resp. DB, apod.

3.2. Analýza časové řady pomocí rozkladu na její složky

Klíčovým konceptem v tradiční analýze časových řad je rozklad časové řady Y_t na složku trendu T_t , sezónní složku S_t a složku reziduální e_t .

Nejběžnější metodou pro získání trendu časové řady je použití lineární filtrace [2] na dané časové řadě. Trend T_t časové řady Y_t lze vyjádřit ve tvaru

$$T_t = \sum_{i=-\infty}^{\infty} \lambda_i y_{t+i}, \quad (1)$$

kde λ_i jsou koeficienty filtrace a y_t jsou pozorované hodnoty časové řady v čase t .

K nejpoužívanější třídě lineárních filtrů náležejí tzv. klouzavé průměry se stejnými váhami (koeficienty):

$$T_t = \frac{1}{2a+1} \sum_{i=-a}^a y_{t+i}. \quad (2)$$

Platí-li pro trend časové řady vztah (2), jsou filtrované hodnoty časových řad v periodě φ reprezentované klouzavými průměry $x_{\varphi-a}, \dots, x_{\varphi}, \dots, x_{\varphi+a}$. Potom koeficienty lineární filtrace nabývají hodnot $\left\{ \frac{1}{2a+1}, \dots, \frac{1}{2a+1} \right\}$.

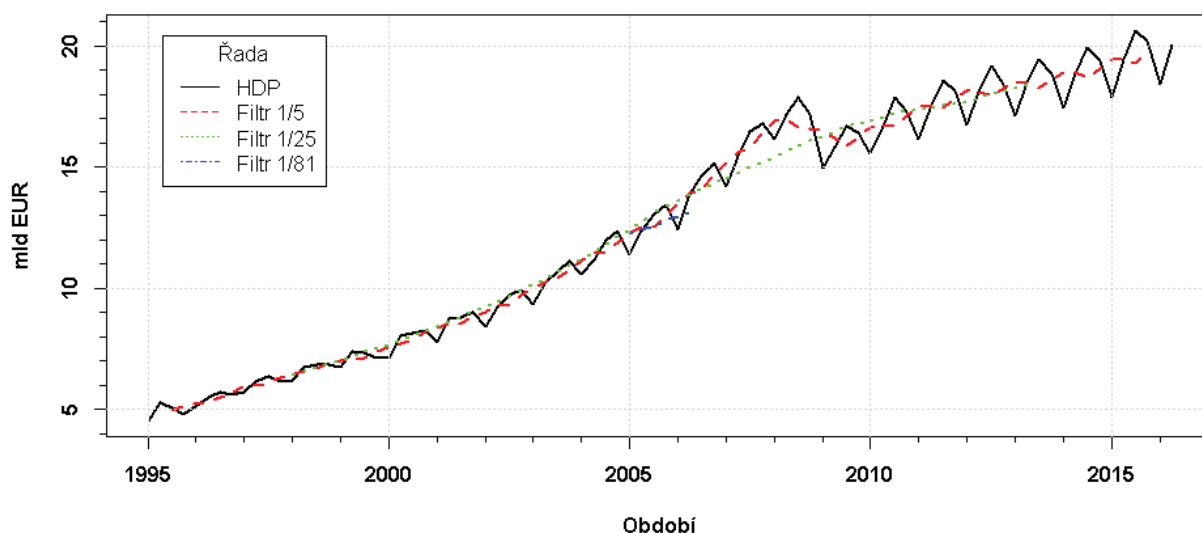
Použije-li se metoda klouzavých průměrů s hodnotami koeficientu $a = 2, 12$ a 40 na hledání trendu časové řady HDP v běžných cenách miliard EUR, vzniknou následující lineární filtry:

$$a = 2 : \lambda_i = \left\{ \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5} \right\}. \quad (3)$$

$$a = 12 : \lambda_i = \underbrace{\left\{ \frac{1}{25}, \dots, \frac{1}{25} \right\}}_{25 \text{ x}} \quad (4)$$

$$a = 40 : \lambda_i = \underbrace{\left\{ \frac{1}{81}, \dots, \frac{1}{81} \right\}}_{81 \text{ x}} \quad (5)$$

Obrázek č. 2: Časová řada „HDP v b. c. v mld. EUR“ a odhad trendu pro vybrané filtry



Zdroj: vlastní konstrukce

Filtrování časových řad v programovacím jazyku R je zajištěno pomocí příkazu `filter()`. Sekvence příkazů programovacího jazyka R, které realizují lineární filtraci časové řady HDP v běžných cenách v mld. EUR podle vztahů (3 - 5), je následující:

```
plot(HDPts,type="l", xlab="", ylab="", main="")
HDP.T1 <- filter(HDPts,rep(1/5,5))
HDP.T2 <- filter(HDPts,rep(1/25,25))
HDP.T3 <- filter(HDPts,rep(1/81,81))
lines(HDP.T1, col="red", lty=2, lwd=2)
lines(HDP.T2, col="green", lty=3, lwd=2)
lines(HDP.T3, col="blue", lty=4, lwd=2)
```

Výsledek lineární filtrace časové řady HDP v běžných cenách v mld. EUR je znázorněn na obrázku 2.

Další možností pro odhad trendu časové řady Y_t je dekompozice pomocí funkce `decompose` časové řady Y_t na složku trendu T_t , sezónní složku S_t . Z diferencí $Y_t - T_t$ je odhadnuta složka reziduální e_t . Funkce nejprve determinuje trendovou složku pomocí klouzavých průměrů, kterou vyděluje z časové řady. Potom je počítána sezónní složka průměrováním pozorovaných hodnot pro každou časovou jednotku během celého sledovaného období. Nakonec je odstraněním složky trendové a sezónní determinována reziduální komponenta časové řady. Funkce `decompose` poskytuje multiplikativní (6) nebo aditivní (7) model dekompozice časové řady, tj.

$$Y_t = T_t * S_t * e_t, \quad (6)$$

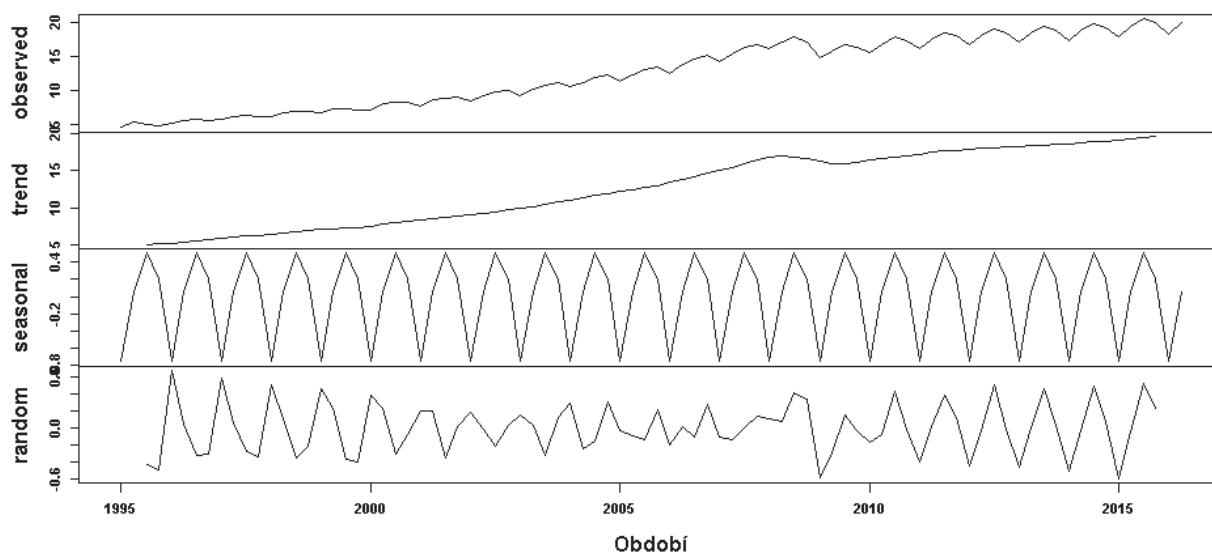
$$Y_t = T_t + S_t + e_t, \quad (7)$$

Symbol S_t představuje označení pro sezónní komponentu časové řady Y_t , e_t reprezentuje reziduální složku řady Y_t . Trendová složka Y_t časové řady je označena jako T_t . Funkce *decompose* se používá ve tvaru

```
compYAdd <- decompose(HDPts , type = "additive")
plot(compYAdd)
```

Vytvořený datový objekt *compYAdd* je maticová struktura jednorozměrné časové řady Y_t , která obsahuje pozorované hodnoty HDP v běžných cenách v miliardách EUR, řadu hodnot reprezentující trend T_t časové řady Y_t , hodnoty její sezónní složky S_t a řadu hodnot reziduální složky e_t . Aditivní model podle (7) časové řady HDP v běžných cenách v mld. EUR je znázorněn na obrázku 3.

Obrázek č. 3: Časová řada „HDP v b. c. v mld. EUR“ a její aditivní dekompozice

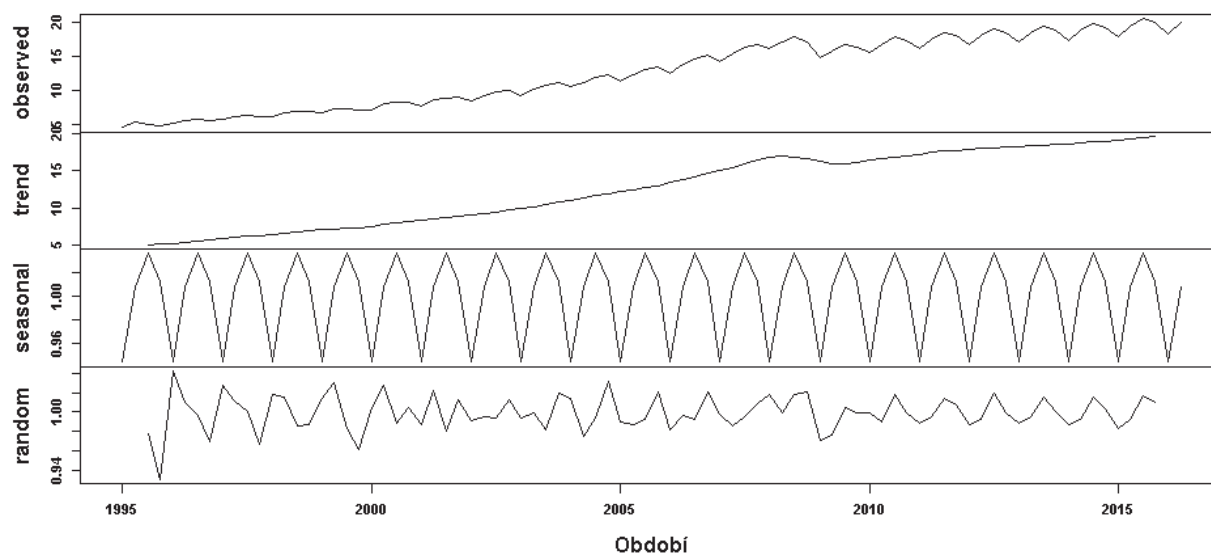


Zdroj: vlastní konstrukce

Stejnou funkcí lze provést také i multiplikativní dekompozici datové řady, tj.

```
compYMul <- decompose(HDPts , type = "multiplicative")
plot(compYMul)
```

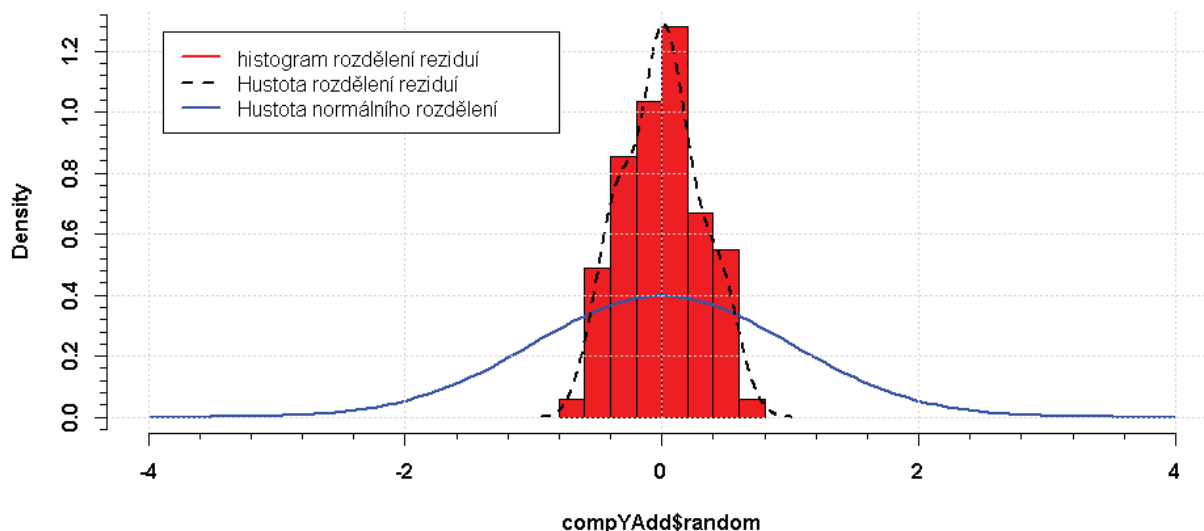
Multiplikativní model podle (6) časové řady HDP v běžných cenách v mld. EUR je znázorněn na obrázku 4.

Obrázek č. 4: Časová řada „HDP v b. c. v mld. EUR“ a její multiplikativní dekompozice

Zdroj: vlastní konstrukce

Často je užitečné posoudit vlastnosti časových řad z hlediska jejich pravděpodobnostního rozdělení. Tento přístup umožňuje na jedné straně porovnávat charakter časových řad, hledat jejich obdobné a rozdílné rysy a činit předběžné závěry o jejich závislosti, na druhé straně vytváří podmínky pro jejich modelování. V této souvislosti je třeba zdůraznit, že při modelování ekonomických časových řad se často vychází z dekompozičního principu, tj. jejich rozkladu na jednotlivé složky – trendovou, cyklickou, sezónní a reziduální. Kritériem kvalitně provedeného rozkladu, tedy oddělení systematické části od části nesystematické, je posouzení vlastností odhadu reziduální složky, tato složka by měla mít nesystematický charakter [2].

Pravděpodobnostní vlastnosti časové řady mohou být posuzovány například pomocí histogramu reziduální složky ve srovnání s hustotou normálního rozdělení.

Obrázek č. 5: Rozdělení reziduální složky časové řady „HDP v b. c. v mld. EUR“

Zdroj: vlastní konstrukce

Ke zjištění pravděpodobnostních vlastností reziduální složky e_t časové řady HDP v běžných cenách v miliardách EUR po aditivní dekompozici (podle (7)) byl zkonstruován histogram a jádrová hustota rozdělení reziduální složky ve srovnání s hustotou normálního rozdělení, což zachycuje obrázek 5. K vytvoření tohoto grafu byly použity následující příkazy:

```
hist(compYAdd$random, xlim=c(-4,4), prob=T, col="red")
mu<-mean(compYAdd$random)
sigma<-sd(compYAdd$random)
x<-seq(-4,4,length=100)
y<-dnorm(x, mu, sigma)
lines(x,y,lwd=2,col="blue")
lines(density(compYAdd$random), lwd=2, col="black", lty=2)
```

Normalitu rozdělení reziduální složky e_t časové řady Y_t lze posuzovat i podle jiných grafů – například pomocí pravděpodobnostních nebo kvantilových grafů, případně je možné reziduální složku testovat pomocí různých statistických testů.

3.3. Regresní analýza časové řady

Pro dekompozici časové řady Y_t lze využít i neparametrickou metodu lokálně váženou polynomičnou regresi. Jedná se o váženou metodu nejmenších čtverců, kde váhová funkce dává největší váhu bodům datovým bodům nejbližším bodu odhadu a nejnižší váhu datovým bodům nejvzdálenějším. Příkladem váhové funkce může být:

$$w(t) = (1 - |t|)^3 \cdot I[|t| < 1], \quad (8)$$

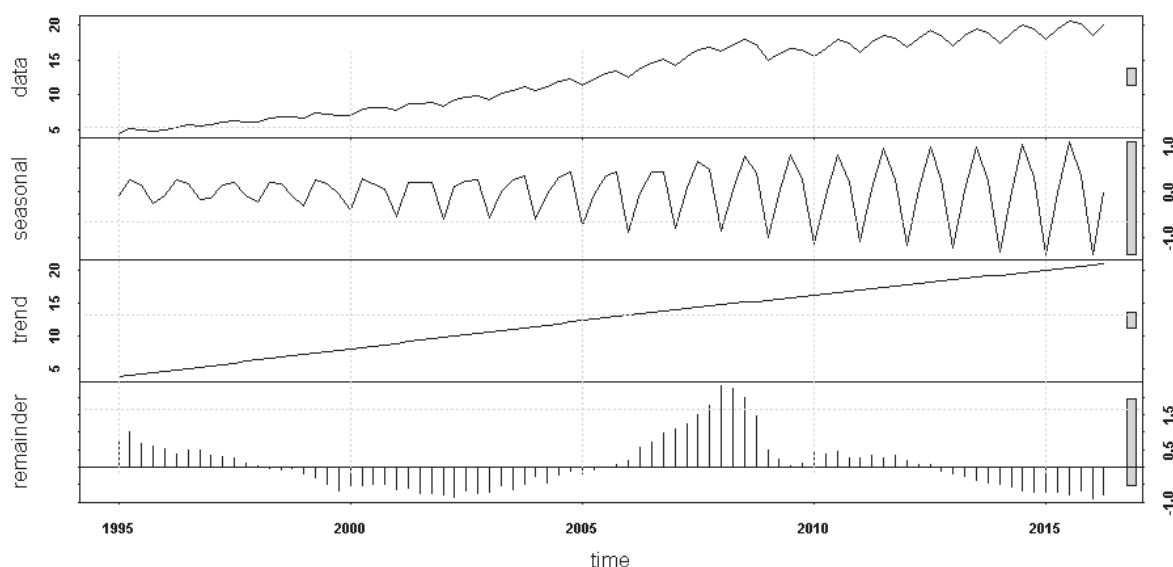
kde $w(t)$ je označením pro váhovou funkci.

V programovacím systému R lze provést dekompozici pomocí neparametrické regrese prostřednictvím příkazu *stl*, v jehož rámci lze zadat odpovídající frekvenci pozorování i délku analyzované časové řady:

```
decompY <- stl(HDPts, s.window = frequency(HDPts), t.window = length(HDPts))
```

Po provedené dekompozici je časová řada i její složky zobrazeny na obrázku 6.

Obrázek č. 6: Modelování časové řady „HDP v b. c. v mld. EUR“

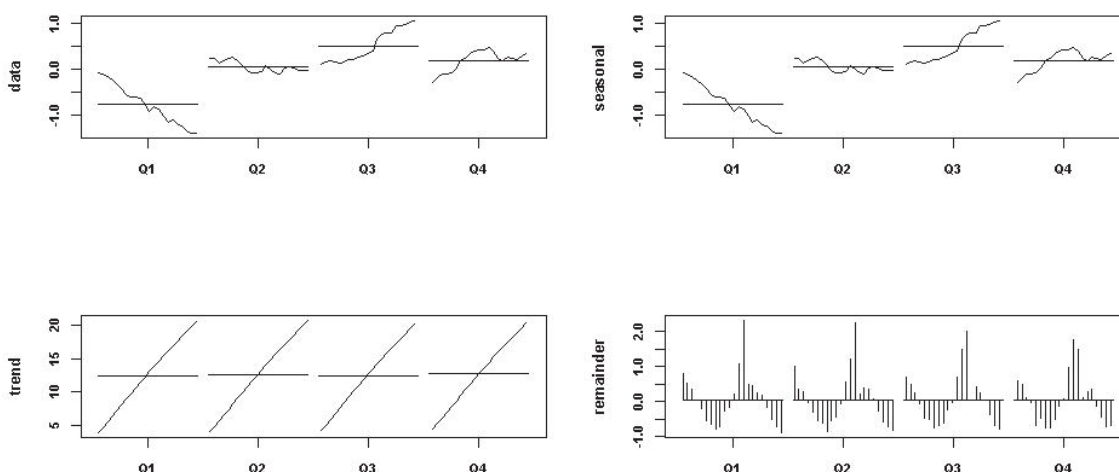


Zdroj: vlastní konstrukce

Příkazem `summary(decompY)` lze získat všechny souhrnné hodnoty odhadnutých složek časové řady Y_t . K jednotlivým složkám časové řady lze přistupovat pomocí vytvořené datové struktury `decompY`, v níž jsou tyto složky uloženy.

Programovací jazyk R umožňuje také analýzu časové řady po částech. Například lze analyzovat časovou řadu včetně dekompozice v rámci jednotlivých čtvrtletí. Toto přibližuje obrázek 7, kde je proveden rozbor chování časové řady pro jednotlivá čtvrtletí.

Obrázek č. 7: Modelování časové řady „HDP v b. c. v mld. EUR“ v jednotlivých čtvrtletích



Zdroj: vlastní konstrukce

Číselné charakteristiky a jednotlivé složky časové řady pro každé čtvrtletí, které jsou graficky reprezentovány na obrázku 7, lze získat pomocí vytvořeného datového objektu *decompY*.

K regresní analýze časových řad nabízí programovací jazyk R funkce *lsfit* především pro regrese metodou nejmenších čtverců a *lm* pro obecnější metody regrese. Větší možnosti pro regresní analýzu, zejména pokud jde o testování významnosti odhadovaných koeficientů, nabízí funkce *lm*. Proto potřebná regresní analýza pro účely analýzy časových řad by měla být pomocí příkazu *lm*.

3.4. Exponenciální vyrovnání a predikce časové řady

Přirozeným odhadem pro předpovídání další hodnoty dané časové řady Y_t v čase $t = \tau$ je vážená suma minulých pozorování, tj.

$$\hat{y}_{(t=\tau)}(1) = \lambda_0 \cdot y_\tau + \lambda_1 \cdot y_{\tau-1} + \dots, \quad (9)$$

kde $\hat{y}_{(t=\tau)}(1)$ značí hodnotu časové řady Y_t predikovanou v čase $t = \tau$, λ_i představují koeficienty váhy pro každé pozorování, $i = 1, \dots, n$.

Pro predikci pozorování časové řady se zdá rozumné, že nedávná pozorování mají větší vliv než pozorování v minulosti. Z tohoto důvodu je jednou z možností použití geometrických vah, které jsou dány tvarem:

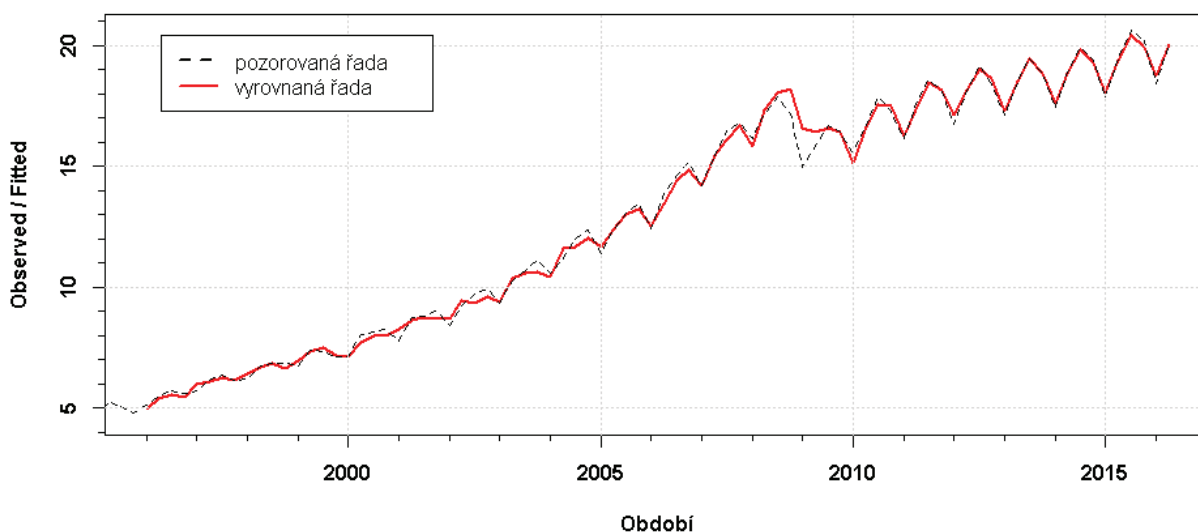
$$\lambda_i = \alpha(1 - \alpha)^i; \text{ pro } 0 < \alpha < 1, i = 1, \dots, n. \quad (10)$$

A po dosazení do vztahu (9) pro predikované pozorování dostaneme výraz

$$\hat{y}_{(t=\tau)}(1) = \alpha \cdot y_\tau + \alpha \cdot (1 - \alpha)^1 y_{\tau-1} + \alpha \cdot (1 - \alpha)^2 y_{\tau-2} + \dots, \quad (11)$$

Exponenciální vyrovnávání ve své základní formě (11) by mělo být použito pouze pro časové řady bez systematického trendu anebo sezónních komponent. Za účelem úspěšného použití exponenciálního vyrovnávání i pro časové řady obsahující jak trendovou složku, tak i složku sezónní, obsahuje programovací jazyk R zobecněnou verzi tohoto vyrovnávání, tzv. Holtovo-Wintersovo vyrovnávání. V tomto případě jsou tři parametry vyhlazování, a to α pro úroveň, β pro zachycení trendu a γ pro sezónní variace časových řad.

Obrázek č. 8: Exponenciální vyrovnávání časové řady „HDP v b. c. v mld. EUR“



Zdroj: vlastní konstrukce

3.5. Modelování časové řady pomocí Boxovy-Jenkinsovy metodologie

Boxova-Jenkinsova metodologie bere v úvahu při konstrukci modelu časové řady reziduální složku, která může být tvořena korelovanými (závislými) náhodnými veličinami. Boxova-Jenkinsova metodologie tedy nejen může zpracovávat časové řady s navzájem závislými pozorováními, ale dokonce těžiště jejich postupů spočívá právě ve zkoumání těchto závislostí neboli tzv. korelační analýze. Kombinují se autoregresivní modely $AR(p)$ (p je řád AR procesu) s modely klouzavých průměrů reziduální složky $MA(q)$ (q je řád MA procesu).

V případě nestacionární časové řady se provádí stacionarizace např. diferencováním a zjišťuje se řád diferencování s parametrem d . Výsledný model se potom označuje jako $ARIMA(p,d,q)$, v případě sezónních vlivů $SARIMA$ modely.

Informaci o síle lineární závislosti mezi veličinami y_t a y_{t-k} poskytuje autokorelační funkce definovaná výrazem

$$\rho(t, t-k) = \frac{E[(y_t - \mu_t)(y_{t-k} - \mu_{t-k})]}{\sqrt{E(y_t - \mu_t)} \cdot \sqrt{E(y_{t-k} - \mu_{t-k})}}, \quad (12)$$

kde y_t je pozorování časové řady v čase t , μ_t je střední hodnota časové řady Y_t .

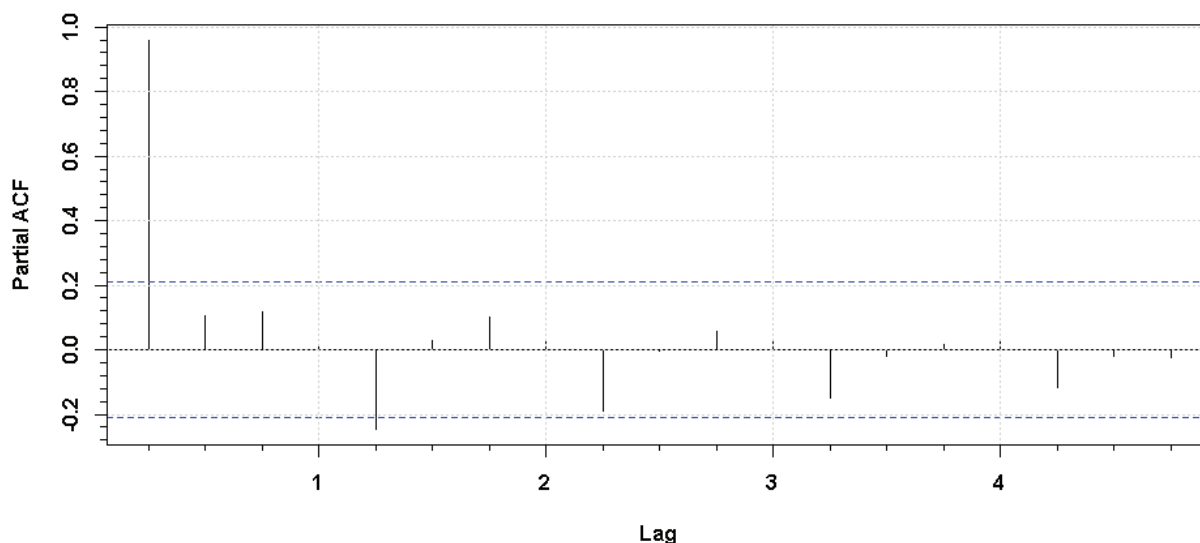
Informaci o síle lineární závislosti mezi veličinami y_t a y_{t-k} , která je očištěna o vliv veličin ležících mezi nimi, poskytuje výběrová parciální autokorelační funkce definovaná rekurentním vztahem [2]

$$f_{kk} = \frac{r_k - \sum_{j=1}^{k-1} f_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} f_{k-1,j} r_j}, \quad \text{kde } f_{kj} = f_{k-1,j} - f_{kk} f_{k-1,k-j}, \quad \text{a } j = 1, 2, \dots, k-1. \quad (13)$$

Výběrovou parciální autokorelační funkci časové řady HDP v běžných cenách v miliardách EUR lze v programovacím jazyku R získat pomocí příkazu `pacf(HDPts, main="HDP v b. c. v mld EUR")`

Výsledkem vykonání příkazu je graf výběrové parciální autokorelační funkce časové řady, z níž je patrná sezónní korelace mezi pozorování posunutými o čtyři čtvrtletí. Graf výběrové parciální autokorelační funkce časové řady je uveden na obrázku 9.

Obrázek č. 9: Výběrová parciální autokorelační funkce řady „HDP v b. c. v mld. EUR“



Zdroj: vlastní konstrukce

Identifikace modelu časové řady je provedena na základě výběrové parciální autokorelační funkce. Na základě [1] je vhodné zvolit jako model časové řady založený na Boxově-Jenkinsově metodologii ve tvaru $ARIMA(1, 0, 1)$, tedy řád AR procesu $p = 1$, řád MA procesu $q = 1$ a řád diferencování je $d = 0$. Ve skutečnosti se jedná o model $ARMA(1, 1)$, protože se diferencování neuplatňuje.

Odhady parametrů modelu časové řady v programovacím jazyku R se provádí pomocí příkazu *arima*, kde jako parametry příkazu vstupují modelovaná časová řada a parametry odhadovaného modelu, tj.

```
HDPts_arima<-arima(HDPts,order=c(1,0,1))
```

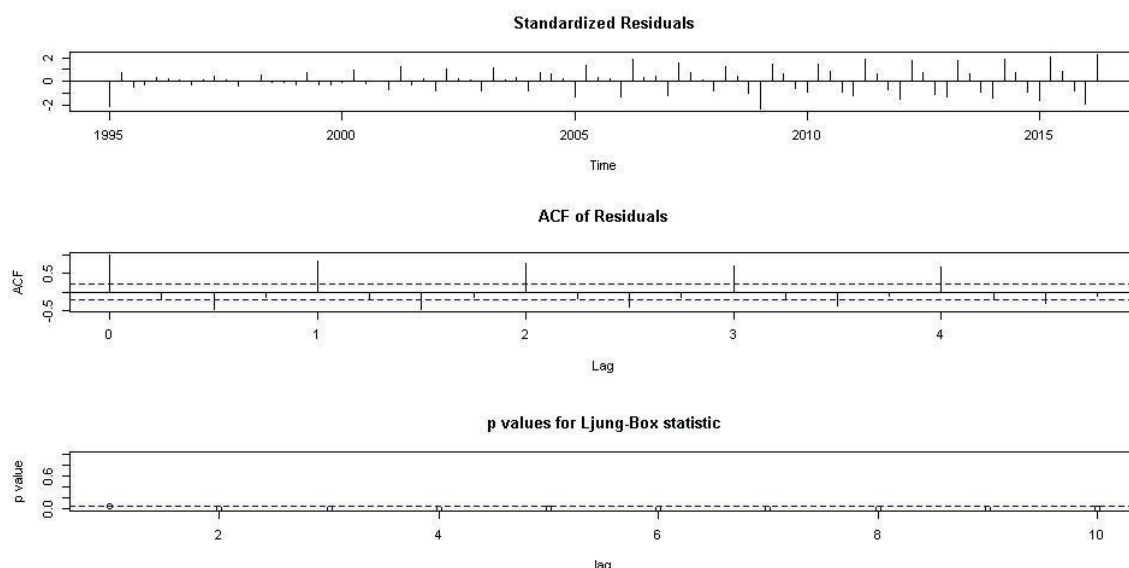
Po provedení odhadu parametrů navrženého modelu $ARMA(1, 1)$ vzniká datová struktura v prostředí programovacího jazyka R s odhadnutými koeficienty, reziduálními hodnotami i hodnotami k diagnostice modelu (například Akaikeho informační kritérium a další).

Následuje diagnostika odhadnutého modelu, která spočívá zejména v analýze reziduí. K tomu má programovací jazyk R funkci *tsdiag*, která zajistí celkovou diagnostiku navrženého modelu, tj.

```
tsdiag(HDPts_arima)
```

Výsledky diagnostiky modelu $ARMA(1, 1)$ navrženého jako modelu pro časovou řadu HDP v běžných cenách v mld. EUR jsou naznačeny na obrázku 10. I model $ARMA(1, 1)$ potvrdil sezónní korelaci mezi pozorováními posunutými o čtyři čtvrtletí.

Obrázek č. 10: Výsledky testování modelu $ARMA(1, 1)$ řady „HDP v b. c. v mld. EUR“



Zdroj: vlastní konstrukce

Jakmile je model identifikován a byly odhadnuty jeho parametry, je dalším cílem předpovědět budoucí hodnoty časové řady. Budoucí hodnoty časové řady (predikce) umožňuje programovací jazyk R modelovat z odhadnutého modelu pomocí funkce

```
HDP.pred<-predict(HDPts_arima,n.ahead=20).
```

Uvedený příkaz dovoluje na základě odhadnutého ARMA(1, 1) modelu časové řady HDP v běžných cenách v mld. EUR predikovat následujících 20 pozorování modelované časové řady.

3.6. Spektrální analýza časové řady

Časová řada (po odstranění trendové složky) může být při spektrální analýze reprezentována jako součet sinusovek a kosinusovek od nižších po vyšší frekvenci o různých amplitudách.

Frekvence ω časové řady je definována jako počet pozorování během 1 cyklu. Časová řada, kdy jsou jednotlivé hodnoty pozorovány měsíčně, je frekvence $\omega = 12$, u čtvrtletních pozorování je frekvence $\omega = 4$. Obecně nemusí být rozložení frekvencí u časové řady známé. Relativní hustotu každé frekvence ve spektru odvozené z analyzované časové řady lze odhadovat periodogramem $I(\omega)$. Periodogram $I(\omega)$ je definován ve tvaru [2]

$$I(\omega) = \frac{1}{2\pi n} \left[\left(\sum_{t=1}^n y_t \cos(\omega t) \right)^2 + \left(\sum_{t=1}^n y_t \sin(\omega t) \right)^2 \right], \quad (14)$$

kde n je počet pozorování časové řady, y_t jsou pozorované hodnoty časové řady v čase t a ω reprezentuje frekvenci. Periodogram umožňuje identifikovat hodnoty frekvencí ω , které jsou v analyzované řadě významné.

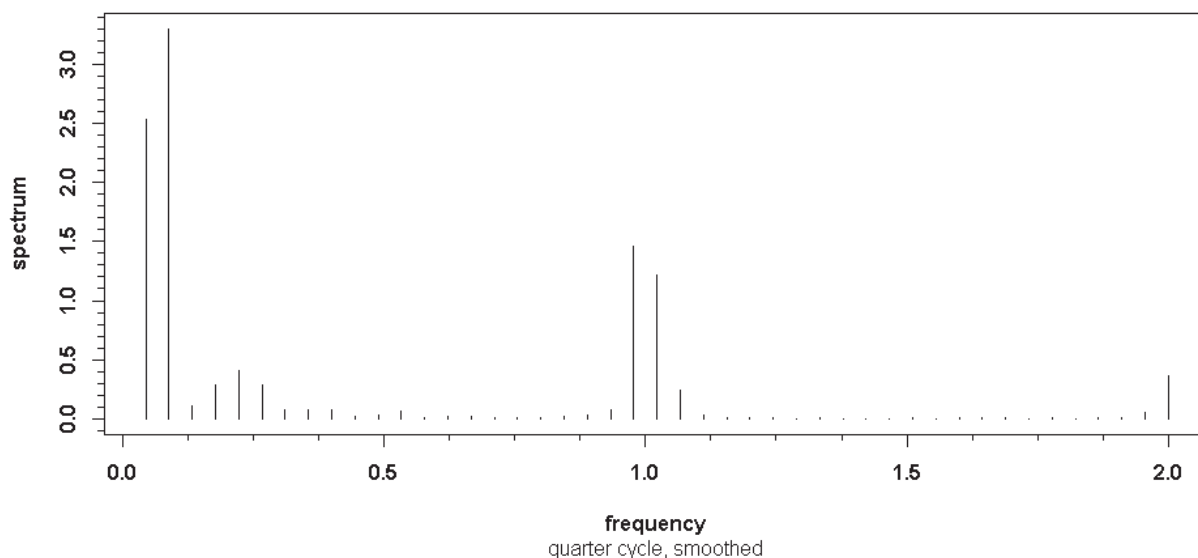
V případě čtvrtletní časové řady HDP v běžných cenách v mld. EUR je možné předpokládat, že frekvence s nejvyšší amplitudou bude 4.

Periodogram analyzované časové řady se v prostředí programovacího jazyka R vytvoří pomocí příkazu

```
spectrum(HDPts, log = "no", main = "Periodogram řady HDP v b. c. v mld EUR", sub = "quarter cycle, smoothed", type = "h").
```

Odhadnutý periodogram vybrané sledované řady je znázorněn na obrázku 11.

Obrázek č. 11: Periodogram časové řady „HDP v b. c. v mld. EUR“



Zdroj: vlastní konstrukce

4. SHRNUÍ A ZÁVĚR

Základní funkcionality programového prostředí R je dodávána s množstvím funkcí užitečných pro časové řady, zejména v doplňkovém programovém balíčku (add-on package) statistiky. Toto je doplněno dalšími doplňkovými programovými balíčky (add-on packages), které jsou vybaveny dalšími funkcionalitami pro oblast analýzy časových řad. Jedná se jak o klasické metody, tak i o velmi pokročilé techniky [5]. Doplňkové programové balíčky pro analytické zkoumání a modelování časových řad neobsahují pouze funkcionality pro samotné časové řady, ale i funkcionality z hlediska postavení časových řad v ekonometrii a financích a do značné míry se svými nástroji pro časové řady překrývají. Z tohoto pohledu lze zhruba strukturovat doplňkové programové balíčky do následujících témat, které se dotýkají problematiky časových řad:

- Základní funkcionality pro manipulaci s napozorovanými údaji ve tvaru časových řad představují zejména doplňkové programové balíčky *Time Series Analysis* [5].
- Pokročilé funkcionality programového prostředí jazyka R pro analytické metody a předvídání časových řad nabízejí doplňkové programové balíčky *Econometrics* [3].
- Metody a techniky analýzy časových řad z oblasti finanční a pojistné matematiky nabízí skupina doplňkových programových balíčků *Empirical Finance* [4].

Článek nabízí – aspoň pro některé doplňkové programové balíčky, které jsou nejdůležitější z pohledu používaných metod a technik na poli časových řad – stručný přehled o možnostech programu R a předložil pár jednoduchých a názorných příkladů praxe v používání programu R. Cílem článku bylo především upozornit na obrovský potenciál a velmi vhodné možnosti používání programu R při modelování časových řad.

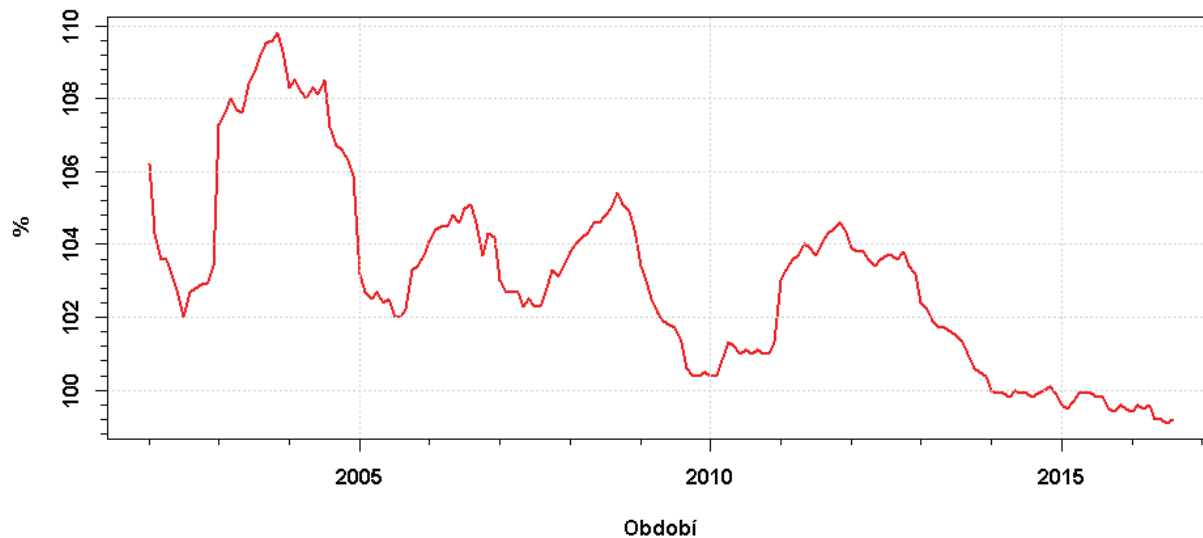
Využití specializovaných programových balíčků v rámci programovacího prostředí R na metody analýzy a prognózování časových řad dovoluje zvládat problémy, které jsou řešitelné ostatními komerčními statistickými programy. Na rozdíl od jiných specializovaných statistických programů lze programovací prostředí R získat a používat zcela zadarmo. Neustálý vývoj nových programových balíčků, na jejichž vzniku se zpravidla podílejí významní odborníci i zkušení praktici v této oblasti, zajišťuje možnost přístupu k moderním metodám a technikám analýzy časových řad a jejich předpovídání celé komunitě uživatelů programu R.

Výše provedený přehled metod a technik, které lze uplatnit na poli modelování časových řad, představuje pouhý výsek možností, které poskytuje programové prostředí programu R. Využitím tohoto programového vybavení, a to zejména doplňkových programových balíčků, lze získat velmi efektivní a sofistikované nástroje (a to zcela zdarma) k řešení i těch nejsložitějších odborných problémů a otázek v oblasti časových řad. Jedná se především o odhad parametrů klasického modelu časové řady, dekompozici časové řady, otázky Boxovy-Jenkinsovy metodologie, nestacionaritu časových řad, regresní metody i strukturální modely a další odborné otázky v této oblasti. Vícerozměrné časové řady, jejich diagnostika, spektrální analýza, modely GARCH a další lze poměrně jednoduchým způsobem řešit pomocí zabudovaných funkcionalit programového prostředí jazyka R.

V příloze jsou uvedeny ukázky analýzy ostatních vybraných ekonomických časových řad z databáze SlovStat.

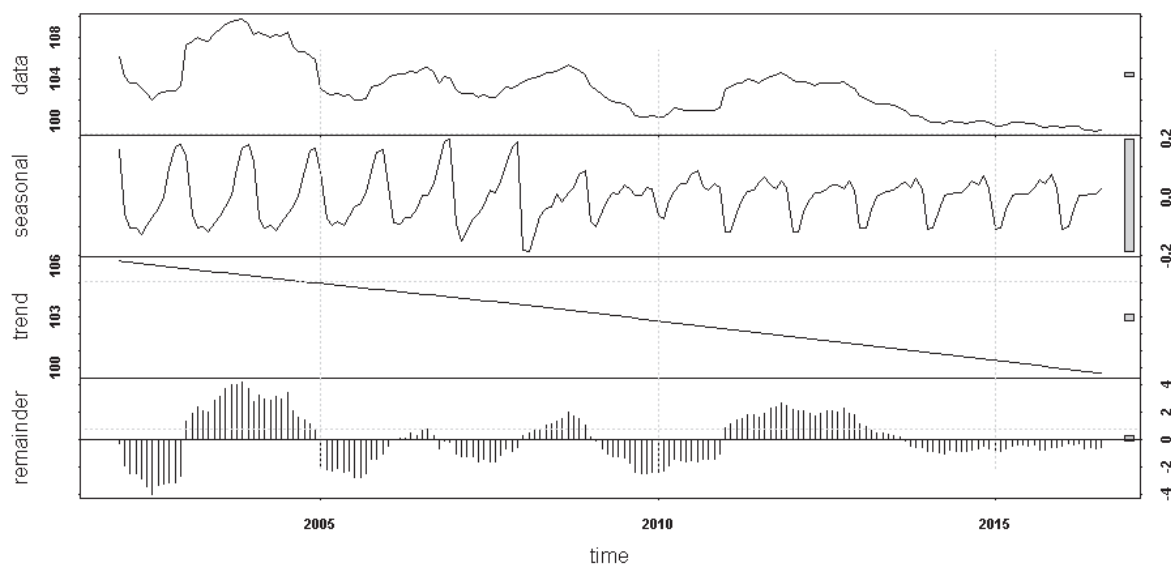
PŘÍLOHA

Obrázek č. 1: Vizuální průběh řady „Indexy spotřebitelských cen podle COICOP“

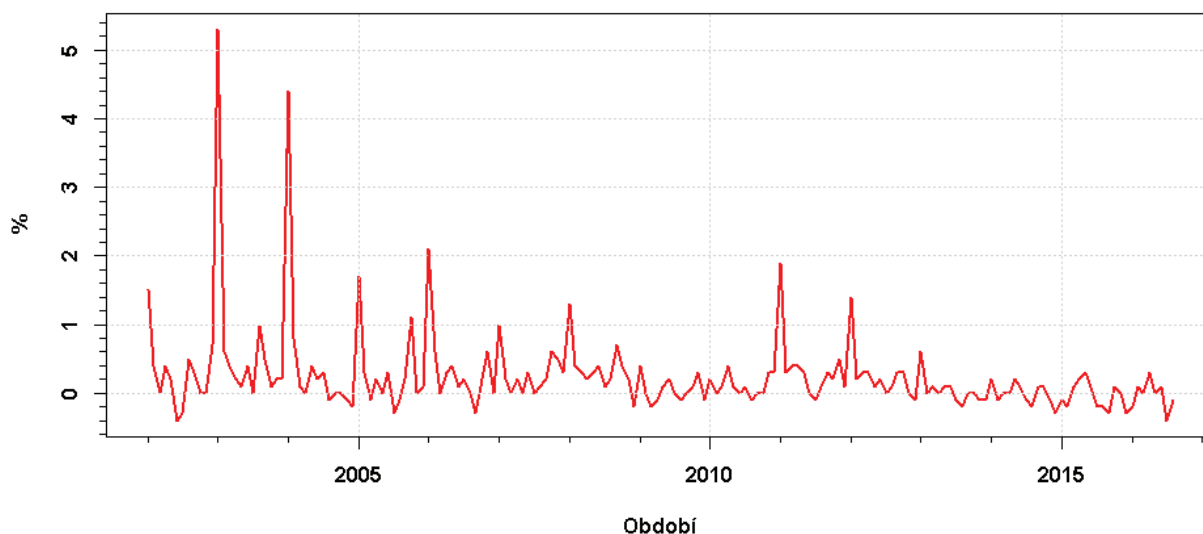
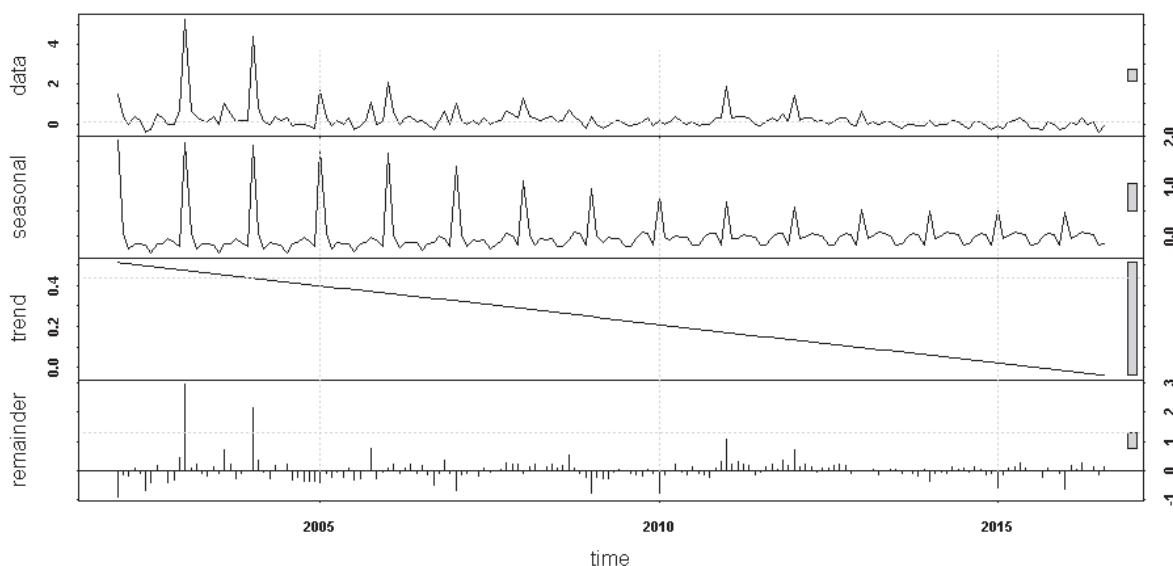


Zdroj: vlastní konstrukce

Obrázek č. 2: Modelování časové řady „Indexy spotřebitelských cen podle COICOP“



Zdroj: vlastní konstrukce

Obrázek č. 3: Vizuální průběh časové řady „Meziměsíční inflace v %“**Zdroj: vlastní konstrukce****Obrázek č. 4: Modelování časové řady „Meziměsíční inflace v %“****Zdroj: vlastní konstrukce**

LITERATURA

- [1] ARLT, J. – ARLTOVÁ, M. – RUBLÍKOVÁ, E.: Analýza ekonomických časových řad s příklady. VŠE Praha, 2004. 148 s. ISBN 80-245-0777-3.
- [2] BROCKWELL, P., J. – DAVIS, R., A.: Introduction to Time Series and Forecasting. Second Edition. New York: Springer-Verlag New York, Inc., 2002. 450 p. ISBN 0-387-95351-5.
- [3] CRAN Task View: Econometrics [on-line]. 23. 04. 2016 [cit. 09. 05. 2016]. WWW: <<https://cran.rstudio.com/web/views/Econometrics.html>>.
- [4] CRAN Task View: Empirical Finance [on-line]. 14. 04. 2016 [cit. 09. 05. 2016]. WWW: <<https://cran.rstudio.com/web/views/Finance.html>>.
- [5] CRAN Task View: Time Series Analysis [online]. 22. 01. 2016 [cit. 09.05.2016]. WWW: <<https://cran.rstudio.com/web/views/TimeSeries.html>>.
- [6] CYHELSKÝ, L. – SOUČEK, E.: Základy statistiky. Praha: VŠFS, o. p. s., 2009. 164 s. ISBN 978-80-7408-013-5.

- [7] KLEIBER, CH. – ZEILEIS, A.: Applied Econometrics with R. New York: Springer Science+BusinessMedia, LLC, 2008. 222 p. ISBN: 978-0-387-77318-6 (eBook).
- [8] SlovSTAT – HDP na obyvateľa v stálych cenách prechádzajúceho roka (s. c. p. r.) (tis. EUR) – metodika ESA 2010, podľa štvrtročných NÚ. WWW: <http://www.statistics.sk/pls/elisw/objekt.send?uic=1854&m_sso=3&m_so=81&ic=188>.
- [9] SlovSTAT – Inflácia meraná CPI (národná metodika) – celková, jadrová, čistá (2002M01 - 2016M03). WWW: <http://www.statistics.sk/pls/elisw/objekt.send?uic=1486&m_sso=3&m_so=16&ic=46>.
- [10] SlovSTAT – Inflácia meraná CPI (národná metodika) – celková, jadrová, čistá (2002M01 - 2016M03). WWW: <http://www.statistics.sk/pls/elisw/objekt.send?uic=1486&m_sso=3&m_so=16&ic=46>.

Ing. Roman Pavelka, PhD., pracuje v odbore metód štatistických zisťovaní Štatistického úradu SR.