UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Hodnocení ustálenosti náhodného procesu a odhad jeho asymptotické střední hodnoty

Katedra matematické analýzy a aplikací matematiky
Vedoucí práce: **doc. Mgr. Ondřej Vencálek, Ph.D.**
Vypracoval(a): **Bc. Matyáš Kovařík**
Studijní program: N0541A170026 Aplikovaná matematika
Studijní obor: Aplikovaná matematika
Forma studia: prezenční
Rok odevzdání: 2024

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Matyáš Kovařík

Název práce: Hodnocení ustálenosti náhodného procesu a odhad jeho asymptotické střední hodnoty

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. Mgr. Ondřej Vencálek, Ph.D.

Rok obhajoby práce: 2024

Abstrakt: Diplomová práce se zabývá analýzou časových řad, které v sobě obsahují výraznou periodickou složku. Cílem je periodicitu v datech odhalit, data vyhladit a rozhodnout, zda jsou časové řady ustálené či nikoliv. V případě ustálené řady je důležité odhadnout také její asymptotickou střední hodnotu. Analýza je provedena pomocí metod jako je rychlá Fourierova transformace, metoda klouzavých průměrů či s využitím modifikovaného exponenciálního trendu.

Klíčová slova: Náhodný proces, asymptotická střední hodnota, časové řady, dekompozice, trend, periodická složka, klouzavé průměry, Fourierova transformace, modifikovaný exponenciální trend

Počet stran: 71

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Matyáš Kovařík

Title: Assessment of the stability of a random process and estimation of its asymptotic mean

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2024

Abstract: The thesis deals with analysis of time series which contain significant periodic components. The aim is to detetct the periodicity, smooth the data and decide whether the time series are stable or not. In the case of a stable series, it is also important to estimate its asymptotic mean. The analysis is performed using methods such as the fast Fourier transform, the moving average method or using a modified exponential trend.

Key words: Random process, asymptotic mean, time series, decomposition, trend, periodic component, moving average, Fourier transform, modified exponential trend

Number of pages: 71

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracoval samostatně pod vedením pana doc. Mgr. Ondřeje Vencálka, Ph.D. a všechny použité zdroje jsem uvedl v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Poděkování

Rád bych poděkoval vedoucímu mé diplomové práce, panu doc. Mgr. Ondřeji Vencálkovi, Ph.D., za odborné vedení mé diplomové práce a cenné rady a připomínky, které mi pomohly dotáhnout práci do zdárného konce.

Úvod

Časové řady představují důležitý nástroj v analýze dat, který umožňuje zkoumat vývoj hodnot určité proměnné v čase. Jejich analýza a modelování nachází široké uplatnění v různých oblastech, ať už jde o ekonomii, biologii, meteorologii či například průmysl. Právě průmyslová data budou v rámci této diplomové práce použita a analyzována. Data pocházejí ze společnosti Sigma, která se zabývá výrobou čerpadel, a vznikla na základě simulací účinnosti různých druhů čerpadel. V práci bude analyzována účinnost čtyř různých čerpadel, každé z nich při třech různě velkých průtocích vody. Celkově tedy máme k dispozici dvanáct časových řad, na kterých budeme analýzu provádět.

Ve všech dvanácti časových řadách hraje významnou roli periodická složka. Naším cílem v rámci diplomové práce bude vypořádat se s periodicitou, vyhladit data a zejména prohlásit, zda je časová řada ustálená, či nikoliv. V případě ustálených časových řad pak bude důležité odhadnout asymptotickou střední hodnotu náhodného procesu. Důležitost rozhodnutí o ustálenosti procesu a asymptotické střední hodnotě spočívá v tom, že čerpadla, jejichž účinnost analyzujeme, ve skutečnosti neexistují, data jsou simulovaná a až na základě jejich analýzy padne rozhodnutí, který typ čerpadel je vhodné vyrábět. Důkladná analýza tedy může firmě pomoci ušetřit značné finanční prostředky, které by jinak vynaložila na výrobu nevhodného typu čerpadla.

Práce je členěna do dvou hlavních kapitol. Obsahem první kapitoly jsou základní teoretické poznatky nutné k praktické analýze. Popíšeme zde dekompoziční přístup k modelování časových řad a zaměříme se primárně na trendovou a periodickou složku. V rámci periodické složky představíme Fourierovu analýzu, která nám pomůže v detekci nejvýznamnějších period v da-

tech. Popíšeme také metodu klouzavých průměrů, která slouží k vyhlazování dat a tím k redukci šumu a periodicity. Ve druhé kapitole pak provedeme praktickou analýzu časových řad ze společnosti Sigma s využitím zmíněných teoretických poznatků a s důrazem na hlavní cíl práce, tedy zhodnocení ustálenosti řad a odhad jejich asymptotické střední hodnoty.

1. Časové řady z pohledu teorie

V analýze časových řad, kterými se tato diplomová práce zabývá, využijeme různé přístupy a metody z teorie časových řad. První kapitola diplomové práce tedy slouží k osvětlení problematiky z teoretického hlediska. K vypracování této celé kapitoly byly využity zdroje [1], [2], [3], [4], [5], [6], [7], [8], [9] a [10]. Konkrétněji pak v kapitolách 1.1 a 1.2 vycházíme zejména z [1] a [2], ale využijeme také publikace [3], [4] a [5]. Pro vypracování kapitoly 1.3 byly použity zdroje [1], [2], [3] a [4]. Kapitola 1.4 pak vychází zejména z [6], ale čerpá také z [9] a [10] a částečně i [7] a [8].

1.1. Časové řady a jejich dekompozice

Pojmem *časová řada* rozumíme soubor dat, která jsou uspořádána chronologicky dle času, kdy byla pozorována. Jedná se tedy o posloupnost náhodných veličin Y_t, kde t označuje čas pozorování. V praxi uvažujeme konečnou množinu časových indexů $t = 1, \ldots, n$.

Dekompozice časových řad spočívá v rozkladu časové řady na jednotlivé složky, což usnadňuje identifikaci a modelování různých aspektů dat. Těmito složkami standardně rozumíme trend T_t, sezónní složku S_t, cyklickou složku C_t a náhodnou (reziduální) složku ϵ_t. Alternativou k dekompozičnímu přístupu je například Box-Jenkinsova metodologie.

Dekompoziční model může být dle tvaru rozkladu dvojího typu, a to *aditivní* a *multiplikativní*. Více využívaná je aditivní dekompozice, která je tvaru:

$$y_t = T_t + S_t + C_t + \epsilon_t. \tag{1.1}$$

Multiplikativní dekompozice pak vypadá následovně:

$$y_t = T_t \times S_t \times C_t \times \epsilon_t. \tag{1.2}$$

Je zřejmé, že z multiplikativního tvaru (1.2) můžeme snadno pomocí logaritmické transformace získat aditivní tvar (1.1), proto si zde vystačíme s aditivním přístupem.

Trendová složka T_t popisuje a pomáhá vysvětlit dlouhodobé chování daného ukazatele v čase. Rozlišujeme rostoucí, klesající a konstantní trend. V závislosti na charakteru dat lze trend popisovat různě komplexními modely. Na některé z nich se zaměříme v další kapitole věnované speciálně trendu.

Sezónnost S_t lze chápat jako pravidelně se opakující odchylku od dlouhodobého trendu. V kontextu ekonomických časových řad platí, že délka periody je nejvýše jeden rok. Typická je například týdenní či měsíční periodicita.

Cyklická složka C_t je, podobně jako sezónnost, fluktuací kolem trendu. Zásadní rozdíl ovšem je v délce periody, která je u cyklu delší než jeden rok. V případě cyklu navíc jde o fluktuace dlouhodobé a nepravidelné s neznámou periodou a často i různou amplitudou. Často se o cyklické složce hovoří v souvislosti s hospodářským cyklem.

Náhodná složka ϵ_t je taková část časové řady, kterou nelze vysvětlit pomocí trendu ani sezónní či cyklické složky. Jedná se o náhodné fluktuace vzniklé například náhodnou událostí či nepřesným měřením. V praxi předpokládáme nulovou střední hodnotu náhodné chyby, tj.:

$$E(\epsilon_t) = 0, \qquad t = 1, \ldots, n. \tag{1.3}$$

Dalším standardním předpokladem je homoskedasticita náhodné chyby, tedy

v čase konstantní rozptyl:

$$var(\epsilon_t) = \sigma^2, \qquad t = 1, \ldots, n. \tag{1.4}$$

V neposlední řadě požadujeme nezávislost a tím také nekorelovanost náhodných chyb, tedy:

$$cov(\epsilon_i, \epsilon_j) = cor(\epsilon_i, \epsilon_j) = 0, \qquad i, j = 1, \ldots, n, \qquad i \neq j. \tag{1.5}$$

Jsou-li splněny vlastnosti (1.3), (1.4) a (1.5), říkáme, že řada ϵ_t tvoří tzv. *bílý šum*.

Není nutností, aby v dané časové řadě existovaly všechny zmíněné složky. Zejména cyklická část je sporná a využívá se primárně v časových řadách ekonomického charakteru. Někdy se cyklická složka zahrnuje pod trend jako jeho část. Také můžeme považovat sezónní a cyklickou složku za periodicitu časové řady, tedy $P_t = S_t + C_t$. Periodicitě bude věnována jedna z dalších kapitol diplomové práce.

1.2. Trendová složka

Trendová složka, jak již bylo zmíněno, slouží k popisu dlouhodobých tendencí v datech. V závislosti na charakteristice dat můžeme popisovat trend pomocí různých *trendových funkcí*. Příkladem těchto funkcí může být konstantní, lineární či kvadratický trend, ale také například exponenciální či modifikovaný exponenciální trend. Podívejme se na předpisy těchto funkcí.

Konstantní trend je nejjednodušší formou trendu. Jak již naznačuje samotný název, jeho předpisem je konstanta, tedy:

$$T_t = \beta_0, \qquad t = 1, \ldots, n, \tag{1.6}$$

kde β_0 je neznámý parametr. Jeho odhad provádíme jednoduše pomocí aritmetického průměru pozorování, tj. $\hat{\beta}_0 = \sum_{t=1}^{n} Y_t$.

Lineární trend je díky své univerzálnosti nejvyužívanějším typem trendu. Má podobu přímky, což umožňuje získat základní informace o analyzované časové řadě. Můžeme jej vyjádřit ve tvaru:

$$T_t = \beta_0 + \beta_1 t, \qquad t = 1, \ldots, n, \tag{1.7}$$

kde β_0 a β_1 jsou neznámé parametry. Jejich odhad je rovněž poměrně jednoduchý, neboť funkce je lineární v parametrech a parametry díky tomu lze odhadnout pomocí metody nejmenších čtverců.

V případě, kdy lineární trendová funkce nedokáže dostatečně vystihnout povahu dat, můžeme využívat také polynomiální trendové funkce s vyšším stupněm polynomu. Z nich je nejčastěji využíván *kvadratický trend*, který můžeme vyjádřit jako:

$$T_t = \beta_0 + \beta_1 t + \beta_2 t^2, \qquad t = 1, \ldots, n, \tag{1.8}$$

kde β_0, β_1 a β_2 jsou neznámé parametry. Odhadujeme je opět pomocí metody nejmenších čtverců, protože i tato funkce je lineární v parametrech. Při rozhodování se mezi použitím lineárního a kvadratického trendu se dá využít například test platnosti podmodelu, kde se testuje nulová hypotéza $H_0 : \beta_2 = 0$ oproti alternativě $H_A : \beta_2 \neq 0$.

Exponenciální a *modifikovaný exponenciální trend* jsou alternativou k polynomiálním typům trendu. Jejich využití je zejména tehdy, když se růst či pokles hodnot sledované veličiny zrychluje či zpomaluje s časem. Na základě toho rozlišujeme exponenciální růst a exponenciální pokles. Exponenciální

trendovou funkci můžeme zapsat ve tvaru:

$$T_t = \alpha\beta^t, \qquad t = 1, \ldots, n, \tag{1.9}$$

kde α a $\beta > 0$ jsou neznámé parametry. Na rozdíl od polynomiálních typů trendu zde nemůžeme pro odhad parametrů použít metodu nejmenších čtverců, protože funkce není lineární v parametrech. Proto pro jejich odhady využíváme například tzv. metodu linearizující transformace či metodu vybraných bodů. V praxi je mnohdy vhodnější využít modifikovaný exponenciální trend, kterému je věnována následující část této kapitoly.

1.2.1. Modifikovaný exponenciální trend

Modifikovaný exponenciální trend vychází svou podobou z exponenciálního trendu, konkrétněji je jeho zobecněním. Změnou je zde posun funkce o určitou konstantu, což může být v aplikacích značnou výhodou. Tento trend má následující podobu:

$$T_t = \gamma + \alpha\beta^t, \qquad t = 1, \ldots, n, \tag{1.10}$$

kde γ, α a $\beta > 0$ jsou neznámé parametry. Ze stejných důvodů jako u exponenciálního trendu nemůžeme ani zde využít k odhadu parametrů metodu nejmenších čtverců. Parametry tedy odhadujeme jinými metodami, například metodou částečných (postupných) součtů nebo metodou vybraných bodů.

Před osvětlením samotného principu těchto metod je vhodné interpretovat parametry. Parametr γ lze chápat jako asymptotu, tedy platí:

$$\lim_{t\to\infty} \gamma + \alpha\beta^t = \lim_{t\to\infty} T_t = \gamma, \qquad \beta \in (0, 1). \tag{1.11}$$

Parametr α můžeme interpretovat jako $\alpha = \gamma - T_t(0)$. Parametr β pak vyjadřuje průměrný podíl $\frac{T_{t+1}-\gamma}{T_t-\gamma} = \beta$.

Metoda částečných součtů spočívá v rozdělení časové řady na tři stejně dlouhé úseky délky m. V případě, kdy celkový počet pozorování n není dělitelný třemi, tedy neplatí $n = 3m$, tak vynecháme jedno, nebo dvě pozorování ze začátku řady a pokračujeme v postupu s takto upravenou řadou. Pro každý úsek spočítáme příslušný částečný součet následovně:

$$S_1 = \sum_{t=1}^{m} y_t, \qquad S_2 = \sum_{t=m+1}^{2m} y_t, \qquad S_3 = \sum_{t=2m+1}^{3m} y_t. \tag{1.12}$$

Nyní položíme empirické součty (1.12) do rovnosti s jejich teoretickými protějšky:

$$\begin{aligned} S_1 &= \sum_{t=1}^{m} T_t = \sum_{t=1}^{m} \gamma + \alpha\beta^t = m\gamma + \alpha \sum_{t=1}^{m} \beta^t, \\ S_2 &= \sum_{t=m+1}^{2m} T_t = \sum_{t=m+1}^{2m} \gamma + \alpha\beta^t = m\gamma + \alpha \sum_{t=m+1}^{2m} \beta^t, \\ S_3 &= \sum_{t=2m+1}^{3m} T_t = \sum_{t=2m+1}^{3m} \gamma + \alpha\beta^t = m\gamma + \alpha \sum_{t=2m+1}^{3m} \beta^t. \end{aligned} \tag{1.13}$$

Pro další úpravu součtů využijeme větu o součtu geometrické řady. Z ní plyne platnost vzorce pro součet prvních m členů geometrické řady:

$$\sum_{t=i}^{i+(m-1)} \beta^t = \beta^i \frac{\beta^m - 1}{\beta - 1}. \tag{1.14}$$

S využitím (1.14) lze soustavu (1.13) přepsat do následující soustavy tří rov-

nic o třech neznámých α, β, γ:

$$\begin{aligned} S_1 &= m\gamma + \alpha\frac{\beta(\beta^m - 1)}{\beta - 1}, \\ S_2 &= m\gamma + \alpha\frac{\beta^{m+1}(\beta^m - 1)}{\beta - 1}, \\ S_3 &= m\gamma + \alpha\frac{\beta^{2m+1}(\beta^m - 1)}{\beta - 1}. \end{aligned} \tag{1.15}$$

Řešením této soustavy rovnic dostáváme postupně odhady parametrů:

$$\begin{aligned} \hat{\beta} &= \left(\frac{S_3 - S_2}{S_2 - S_1}\right)^{\frac{1}{m}}, \\ \hat{\alpha} &= \frac{\hat{\beta} - 1}{\hat{\beta}(\hat{\beta}^m - 1)^2}(S_2 - S_1), \\ \hat{\gamma} &= \frac{1}{m}\left[S_1 - \hat{\alpha}\frac{\hat{\beta}(\hat{\beta}^m - 1)}{\hat{\beta} - 1}\right], \end{aligned} \tag{1.16}$$

kde $\hat{\beta}$ je odhadem β, $\hat{\alpha}$ je odhadem α a $\hat{\gamma}$ je odhadem γ.

Je dobré si uvědomit, že pro odhad $\hat{\beta}$ potřebujeme mít stejná znaménka v čitateli a jmenovateli. Znamená to tedy, že pro tuto metodu musí platit $S_1 > S_2 > S_3$, nebo $S_1 < S_2 < S_3$. Formálně můžeme například v programu R uvažovat v odhadu $\hat{\beta}$ absolutní hodnotu, ovšem nesplnění zmíněných nerovností naznačuje, že pro daná data není použití modifikovaného exponenciálního trendu příliš vhodné.

Alternativou k metodě postupných součtů je například *metoda vybraných bodů*. Její princip spočívá v tom, že zvolíme z časové řady tři body, a to v časech t (např. $t = 0$), $t + m$ a $t + 2m$. V těchto bodech položíme hodnotu

trendu rovnu pozorováním, tedy:

$$\begin{aligned} y_t =& \gamma + \alpha\beta^t, \\ y_{t+m} =& \gamma + \alpha\beta^{t+m}, \\ y_{t+2m} =& \gamma + \alpha\beta^{t+2m}, \end{aligned} \tag{1.17}$$

čímž vznikla soustava tří rovnic o třech neznámých. Jejím řešením dostáváme odhady parametrů:

$$\begin{aligned} \hat{\beta} =& \left(\frac{y_{t+2m} - y_{t+m}}{y_{t+m} - y_t}\right)^{\frac{1}{m}}, \\ \hat{\alpha} =& \frac{y_{t+m} - y_t}{\hat{\beta}^m - 1}, \\ \hat{\gamma} =& y_t - \hat{\alpha}\hat{\beta}^t. \end{aligned} \tag{1.18}$$

1.3. Metoda klouzavých průměrů

Metodu klouzavých průměrů řadíme mezi tzv. adaptivní přístupy k modelování trendu. Své využití tato metoda nalézá zejména ve chvílích, kdy je časová řada dlouhá a v průběhu času se mění charakter trendu. Nedá se tak dost dobře popsat trend jednou křivkou s neměnnými parametry, jako tomu bylo v kapitole 1.2. Smyslem metody je data vyhladit a zbavit je krátkodobé náhodné fluktuace a šumu. Díky tomu je pak snadnější identifikace dlouhodobého trendu.

Metoda klouzavých průměrů pracuje s trendem lokálně. Princip spočívá ve volbě délky okna, což je malá část dat určité délky, a následném zprůměrování hodnot v tomto okně. Následně se okno posouvá o jedno pozorování dále a postup je opakován. V závislosti na způsobu průměrování dat můžeme rozlišovat prosté, vážené nebo centrované klouzavé průměry. Alternativou ke klouzavým průměrům je například exponenciální vyrovnávání.

1.3.1. Prosté klouzavé průměry

Prosté klouzavé průměry jsou nejjednodušším typem klouzavých průměrů. Pro jejich konstrukci budeme předpokládat lichou délku okna (též délka klouzavé části), tj. délka $p = 2m + 1$, kde $m = 1, 2, 3, \ldots$. Označme čas uprostřed okna jako s a všechny časy v rámci daného okna jako $\{s + \tau\}$, kde

$$\tau = \{-m, \ldots, -1, 0, 1, \ldots, m\}. \tag{1.19}$$

Dané okno tedy obsahuje pozorování v časech:

$$\{s - m, \ldots, s - 1, s, s + 1, \ldots, s + m\}. \tag{1.20}$$

Dále předpokládejme, že na každé klouzavé části je definován lokálně lineární trend. Trendová funkce pak bude mít předpis:

$$T_{s+\tau} = \beta_0(s) + \beta_1(s)\tau, \tag{1.21}$$

kde $\beta_0(s)$ a $\beta_1(s)$ jsou neznámé parametry v daném časovém okně.

Parametr $\beta_0(s)$ lze interpretovat jako hodnotu trendu uprostřed okna a parametr $\beta_1(s)$ jako směrnici či změnu trendu za jednotku času. Tyto parametry budeme v každém časovém okamžiku (v každém okně) odhadovat pomocí metody nejmenších čtverců. Úkolem je vyřešit následující minimalizační úlohu:

$$\min f(\beta_0, \beta_1) = \sum_{\tau=-m}^{m} (y_{s+\tau} - \beta_0 - \beta_1\tau)^2. \tag{1.22}$$

Minimalizace výrazu (1.22) dosáhneme parciální derivací podle parametrů β_0

a β_1 a následným položením těchto derivací rovno nule:

$$\begin{aligned}\frac{\partial f}{\partial \beta_0} &= 2\sum_{\tau=-m}^{m}(y_{s+\tau}-\beta_0-\beta_1\tau)(-1)=0,\\ \frac{\partial f}{\partial \beta_1} &= 2\sum_{\tau=-m}^{m}(y_{s+\tau}-\beta_0-\beta_1\tau)(-\tau)=0.\end{aligned} \tag{1.23}$$

Tyto rovnice lze zjednodušit do podoby:

$$\begin{aligned}\sum_{\tau=-m}^{m}(y_{s+\tau}-\beta_0-\beta_1\tau)&=0,\\ \sum_{\tau=-m}^{m}(y_{s+\tau}-\beta_0-\beta_1\tau)\tau&=0.\end{aligned} \tag{1.24}$$

Vzniklá soustava (1.24) dvou rovnic o dvou neznámých se nazývá normální soustava. Jejím řešením získáme odhady parametrů β_0 a β_1. V rámci metody klouzavých průměrů ovšem odhad $\hat{\beta}_1$ nepotřebujeme, protože principem metody je nahrazení příslušného okna jedním číslem, a to průměrem. Stačí se tedy zaměřit na odhad $\hat{\beta}_0$. V prvním z výrazů (1.24) provedeme úpravu převedením β_0 a $\beta_1\tau$ na druhou stranu rovnosti:

$$\sum_{\tau=-m}^{m} y_{s+\tau} = \beta_0 + \beta_1 \sum_{\tau=-m}^{m} \tau. \tag{1.25}$$

Vzhledem k liché volbě délky okna $p = 2m + 1$ je však zajištěno následující:

$$\sum_{\tau=-m}^{m} \tau = 0. \tag{1.26}$$

Díky tomu můžeme z výrazu (1.25) získat odhad $\hat{\beta}_0$ následovně:

$$\hat{\beta}_0 = \frac{1}{2m+1}\sum_{\tau=-m}^{m} y_{s+\tau} = \bar{y}_s. \tag{1.27}$$

Výraz (1.27) nazýváme *prostý klouzavý průměr*. Tento průměr počítáme postupně pro všechna okna, díky čemuž získáme vyhlazená data.

Na závěr je třeba zmínit nevýhodu, kterou tato metoda má. Vzhledem ke způsobu konstrukce klouzavých průměrů totiž bude vyhlazená řada kratší než řada původní. Časová řada délky n bude mít po vyhlazení pomocí klouzavých průměrů s okny velikosti $p = 2m+1$ délku $n-(p-1)$. Konkrétněji pak bude ve vyhlazené řadě chybět prvních $\frac{p-1}{2}$ a posledních $\frac{p-1}{2}$ pozorování.

1.3.2. Vážené klouzavé průměry

Někdy nám nemusí stačit vyhlazování dat pomocí prostých klouzavých průměrů a můžeme chtít namísto lokálně lineárního trendu využívat lokálně kvadratický nebo obecně lokálně polynomiální trend. V takovém případě přicházejí na řadu *vážené klouzavé průměry*. Jejich základní myšlenka spočívá v přiřazení různých vah různým pozorováním v daném okně. Největší váhu bude mít pozorování ve středu okna, nejmenší pak naopak pozorování na okrajích okna. Budeme opět jako v kapitole 1.3.1 předpokládat lichou délku okna $p = 2m + 1$ se středem s. Předpokládejme nyní lokálně kvadratický trend:

$$T_{s+\tau} = \beta_0(s) + \beta_1(s)\tau + \beta_2(s)\tau^2. \tag{1.28}$$

Parametry β_0, β_1, β_2 jsou neznámé parametry. Jejich odhad lze získat opět pomocí metody nejmenších čtverců, neboť trendová funkce je lineární v parametrech. I zde nás zajímá primárně parametr β_0, parametry β_1 a β_2 k výpočtu klouzavých průměrů nepotřebujeme. Budeme řešit minimalizační úlohu:

$$\min f(\beta_0, \beta_1, \beta_2) = \sum_{\tau=-m}^{m} (y_{s+\tau} - \beta_0 - \beta_1\tau - \beta_2\tau^2)^2. \tag{1.29}$$

Výraz parciálně zderivujeme podle jednotlivých proměnných a položíme ro-

ven nule:

$$\begin{aligned}
\frac{\partial f}{\partial \beta_0} &= 2 \sum_{\tau=-m}^{m} (y_{s+\tau} - \beta_0 - \beta_1 \tau - \beta_2 \tau^2)(-1) = 0, \\
\frac{\partial f}{\partial \beta_1} &= 2 \sum_{\tau=-m}^{m} (y_{s+\tau} - \beta_0 - \beta_1 \tau - \beta_2 \tau^2)(-\tau) = 0, \\
\frac{\partial f}{\partial \beta_2} &= 2 \sum_{\tau=-m}^{m} (y_{s+\tau} - \beta_0 - \beta_1 \tau - \beta_2 \tau^2)(-\tau^2) = 0.
\end{aligned} \tag{1.30}$$

Po úpravě získáme normální rovnice ve tvaru:

$$\begin{aligned}
\sum_{\tau=-m}^{m} y_{s+\tau} &= p\beta_0 + \beta_1 \sum_{\tau=-m}^{m} \tau + \beta_2 \sum_{\tau=-m}^{m} \tau^2, \\
\sum_{\tau=-m}^{m} \tau y_{s+\tau} &= \beta_0 \sum_{\tau=-m}^{m} \tau + \beta_1 \sum_{\tau=-m}^{m} \tau^2 + \beta_2 \sum_{\tau=-m}^{m} \tau^3, \\
\sum_{\tau=-m}^{m} \tau^2 y_{s+\tau} &= \beta_0 \sum_{\tau=-m}^{m} \tau^2 + \beta_1 \sum_{\tau=-m}^{m} \tau^3 + \beta_2 \sum_{\tau=-m}^{m} \tau^4.
\end{aligned} \tag{1.31}$$

Nyní můžeme díky platnosti $\sum_{\tau=-m}^{m} \tau = 0$ a $\sum_{\tau=-m}^{m} \tau^3 = 0$ výrazy zjednodušit a omezit se na soustavu dvou rovnic o dvou neznámých, ze kterých budeme odhadovat parametr β_0:

$$\begin{aligned}
\sum_{\tau=-m}^{m} y_{s+\tau} &= p\beta_0 + \beta_2 \sum_{\tau=-m}^{m} \tau^2, \\
\sum_{\tau=-m}^{m} \tau^2 y_{s+\tau} &= \beta_0 \sum_{\tau=-m}^{m} \tau^2 + \beta_2 \sum_{\tau=-m}^{m} \tau^4.
\end{aligned} \tag{1.32}$$

Dle [1] platí:

$$\sum_{\tau=-m}^{m} \tau^2 = \frac{p(p^2-1)}{12}, \qquad \sum_{\tau=-m}^{m} \tau^4 = \frac{p(p^2-1)(3p^2-7)}{240}. \tag{1.33}$$

Díky tomu můžeme vyjádřit z první rovnice v (1.32) parametr β_2 pomocí

parametru β_0 následovně:

$$\beta_2 = \frac{12\left(\sum_{\tau=-m}^{m} y_{s+\tau} - \beta_0 p\right)}{p(p^2-1)}. \tag{1.34}$$

Pro odhad parametru β_0 nám tak stačí vyřešit druhou rovnici z (1.32) jakožto rovnici o jediné neznámé β_0:

$$\beta_0 = \frac{12\left(\sum_{\tau=-m}^{m} \tau^2 y_{s+\tau} - \frac{12\left(\sum_{\tau=-m}^{m} y_{s+\tau} - \beta_0 p\right)}{p(p^2-1)} \cdot \frac{p(p^2-1)(3p^2-7)}{240}\right)}{p(p^2-1)}. \tag{1.35}$$

To budeme postupně upravovat. Nejprve vykrácením $p(p^2-1)$ a roznásobením závorky v čitateli:

$$\beta_0 = \frac{12\sum_{\tau=-m}^{m} \tau^2 y_{s+\tau} - \frac{3}{5}\left(3p^2\sum_{\tau=-m}^{m} y_{s+\tau} - 3p^3\beta_0 - 7\sum_{\tau=-m}^{m} y_{s+\tau} + 7\beta_0 p\right)}{p(p^2-1)}.$$

Provedeme další roznásobení čitatele a převedeme část obsahující β_0 na levou stranu rovnosti:

$$\beta_0 - \frac{\beta_0\left(9p^2-21\right)}{5(p^2-1)} = \frac{12\sum_{\tau=-m}^{m} \tau^2 y_{s+\tau}}{p(p^2-1)} - \frac{\left(9p^2-21\right)\sum_{\tau=-m}^{m} y_{s+\tau}}{5p(p^2-1)}.$$

Vytkneme β_0 na levé straně rovnosti a osamostatníme jej:

$$\beta_0 = \frac{\frac{12\sum_{\tau=-m}^{m} \tau^2 y_{s+\tau}}{p(p^2-1)} - \frac{\left(9p^2-21\right)\sum_{\tau=-m}^{m} y_{s+\tau}}{5p(p^2-1)}}{\frac{5p^2-5-9p^2+21}{5(p^2-1)}}.$$

Nyní se zbavíme složeného zlomku pomocí roznásobení výrazu a vykrácení (p^2-1), čímž získáme tvar:

$$\beta_0 = \frac{60\sum_{\tau=-m}^{m} \tau^2 y_{s+\tau} - \left(9p^2-21\right)\sum_{\tau=-m}^{m} y_{s+\tau}}{p(-4p^2+16)}.$$

Závěrečnými úpravami dostaneme finální tvar pro odhad parametru β_0:

$$\hat{\beta}_0 = \sum_{\tau=-m}^{m} \frac{3(3p^2 - 7 - 20\tau^2)}{4p(p^2-4)} \cdot y_{s+\tau} = \bar{y}_s, \tag{1.36}$$

kde $p = 2m+1$. Vzorec (1.36) nazýváme *vážený klouzavý průměr* a hodnoty:

$$W_\tau = \frac{3(3p^2 - 7 - 20\tau^2)}{4p(p^2-4)}, \qquad \tau = -m, \ldots, -1, 0, 1, \ldots, m \tag{1.37}$$

jsou symetrické váhy splňující podmínku $\sum_{\tau=-m}^{m} W_\tau = 1$.

Vážený klouzavý průměr můžeme též zapsat za pomoci normovací konstanty G ve výpočetně vhodnějším tvaru:

$$\hat{\beta}_0 = \bar{y}_s = \frac{1}{G} \sum_{\tau=-m}^{m} w_\tau y_{s+\tau}, \tag{1.38}$$

kde $w_\tau = GW_\tau$ pro $\tau = -m, \ldots, -1, 0, 1, \ldots, m$.

Na základě vzorce (1.37) můžeme sestavit systém vah pro liché délky okna. Například pro $p = 5$, a tedy $m = 2$, budeme počítat pět vah, a to W_{-2}, W_{-1}, W_0, W_1 a W_2. Pro W_{-2} pak po dosazení do (1.37) dostáváme:

$$W_{-2} = \frac{3(3 \cdot 5^2 - 7 - 20 \cdot (-2)^2)}{4 \cdot 5(5^2-4)} = -\frac{36}{420} = -\frac{3}{35}. \tag{1.39}$$

Obdobně získáme $W_{-1} = \frac{12}{35}$, $W_0 = \frac{17}{35}$, $W_1 = \frac{12}{35}$ a $W_2 = -\frac{3}{35}$. Systém vah pro $p = 5$ pak lze zapisovat následujícím způsobem:

$$\frac{1}{35}\left[-3, 12, 17, 12, -3\right], \tag{1.40}$$

kde $G = 35$. Podobným způsobem můžeme konstruovat systém vah pro libovolné liché p.

1.3.3. Centrované klouzavé průměry

Centrované klouzavé průměry nám pomohou v situaci, kdy potřebujeme volit sudý rozsah okna, obecně $p = 2m$. Využití tohoto typu klouzavých průměrů v praxi se objevuje například u čtvrtletních dat (okno velikosti 4), měsíčních dat (12 pozorování v okně) či hodinových dat (24 pozorování). Oproti kapitolám 1.3.1 a 1.3.2 zde není střed okna celočíselný. Například pro čtvrtletní data s pozorováními v časech $(1, 2, 3, 4)$ by byl střed v čase $\frac{5}{2}$. Obecně bude střed okna v čase $m + \frac{1}{2}$.

Uvažujme pro jednoduchost prosté klouzavé průměry se sudou velikostí okna $p = 2m$. Můžeme určit vyrovnanou hodnotu v neceločíselném čase $m+\frac{1}{2}$ jako aritmetický průměr hodnot v prvním okně:

$$\hat{y}_{m+\frac{1}{2}} = \frac{y_1 + \cdots + y_p}{p}. \tag{1.41}$$

Nyní posuneme okno o jedno pozorování doprava a počítáme vyrovnanou hodnotu v čase $m + \frac{3}{2}$:

$$\hat{y}_{m+\frac{3}{2}} = \frac{y_2 + \cdots + y_{p+1}}{p}. \tag{1.42}$$

Vyrovnanou hodnotu v celočíselném čase $m+1$ pak získáme zprůměrováním (1.41) a (1.42):

$$\hat{y}_{m+1} = \frac{\hat{y}_{m+\frac{1}{2}} + \hat{y}_{m+\frac{3}{2}}}{2}, \tag{1.43}$$

což lze zapsat také přímo:

$$\hat{y}_{m+1} = \frac{1}{2p}\left(y_1 + 2y_2 + \cdots + 2y_p + y_{p+1}\right). \tag{1.44}$$

Výraz (1.44) nazýváme *centrovaný klouzavý průměr*. Krajní hodnoty okna

jsou zde brány s poloviční vahou oproti ostatním hodnotám.

Podobným způsobem můžeme konstruovat také vážený centrovaný průměr. Například pro výpočet čtyřčlenného váženého klouzavého průměru bychom použili následující systém vah:

$$\frac{1}{32}\left[-1, 8, 18, 8, -1\right] \tag{1.45}$$

a pro dvanáctičlenný vážený centrovaný klouzavý průměr pak systém vah:

$$\frac{1}{224}\left[-9, -8, 10, 24, 34, 40, 42, 40, 34, 24, 10, -8, -9\right]. \tag{1.46}$$

1.4. Periodicita a Fourierova analýza

V předchozích kapitolách již bylo zmíněno, že sezónnost a cykličnost časové řady dohromady tvoří *periodickou složku*. Periodicita se vyskytuje v datech různého typu, ať už jsou to data ekonomická, průmyslová či například biologická. Zkoumání periodicity hraje při analýze časových řad významnou roli, neboť nám umožňuje identifikovat opakující se vzorce v datech a pomáhá nám porozumět odchylkám od trendu. Detekce všech významných period v časové řadě je důležitá pro důkladnou analýzu a porozumění vlastnostem a chování časové řady. Jedním z nejpoužívanějších nástrojů k tomuto účelu je *Fourierova analýza*.

Fourierova analýza umožňuje rozložit složité signály nebo časové řady na jednodušší periodické složky, což nám poskytuje hlubší vhled do struktury dat. Je důležité sledovat významnost každé periodické složky v celkovém signálu. Mnohdy se v datech objevuje více periodických složek, ale ne všechny musejí být významné. V praxi je proto dobré věnovat práci též interpretaci jednotlivých period. Mnohdy se v datech různého typu objevuje například

denní či měsíční periodicita.

1.4.1. Fourierovy řady

Princip Fourierovy analýzy a Fourierových řad spočívá v tom, že jestliže je funkce $f(x)$ 2π-periodická a po částech hladká, tak ji můžeme zapsat jako nekonečnou kombinaci sinů a cosinů s postupně se zvyšující frekvencí. Fourierovy řady jsou základním stavebním kamenem pro pochopení a použití Fourierovy transformace, která se zabývá rozkladem nejen periodických, ale i neperiodických funkcí. *Fourierova řada* pro 2π-periodickou funkci $f(x)$ je daná následovně:

$$f(x) = \frac{A_0}{2} + \sum_{k=1}^{\infty}(A_k \cos(kx) + B_k \sin(kx)). \tag{1.47}$$

Koeficienty A_k, B_k jsou určeny jako:

$$\begin{aligned} A_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)\,\mathrm{d}x, \\ B_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)\,\mathrm{d}x. \end{aligned} \tag{1.48}$$

Jinou možností zápisu koeficientů A_k, B_k je zápis pomocí skalárního součinu. Skalární součin obecně komplexních funkcí $f(x)$ a $g(x)$ definovaných na intervalu $\langle a, b\rangle$ je dán předpisem:

$$\langle f(x), g(x)\rangle = \int_a^b f(x)\overline{g}(x)\,\mathrm{d}x, \tag{1.49}$$

kde $\overline{g}(x)$ značí komplexně sdružené číslo k $g(x)$.

Pro lepší pochopení skalárního součinu funkcí je vhodné podívat se na diskrétní případ. Budeme uvažovat datové vektory, obecně komplexních čísel, $\mathbf{f} = [f_1, \ldots, f_n]$ a $\mathbf{g} = [g_1, \ldots, g_n]$, vzniklé diskretizací funkcí $f(x)$ a $g(x)$.

Platí tedy $f_k = f(x_k)$ a $g_k = g(x_k)$ pro všechna $k = 1, \ldots, n$. Skalární součin těchto datových vektorů můžeme vyjádřit následovně:

$$\langle \mathbf{f}, \mathbf{g} \rangle = \mathbf{g}^T \mathbf{f} = \sum_{k=1}^{n} f_k \overline{g}_k = \sum_{k=1}^{n} f(x_k) \overline{g}(x_k). \tag{1.50}$$

Limitně pro $n \to \infty$ jsou skalární součiny (1.49) a (1.50) shodné. [6]

Předpis pro výpočet koeficientů A_k, B_k s využitím skalárního součinu je pak následující:

$$\begin{aligned} A_k &= \frac{1}{\| \cos(kx) \|^2} \langle f(x), \cos(kx) \rangle, \\ B_k &= \frac{1}{\| \sin(kx) \|^2} \langle f(x), \sin(kx) \rangle, \end{aligned} \tag{1.51}$$

kde $\| \cos(kx) \|^2 = \| \sin(kx) \|^2 = \pi$. Platnost této rovnosti vychází z platnosti vztahů:

$$\| \cos(kx) \|^2 = \int_{-\pi}^{\pi} \cos^2(kx)\, \mathrm{d}x = \int_{-\pi}^{\pi} 1 - \sin^2(kx)\, \mathrm{d}x = 2\pi - \int_{-\pi}^{\pi} \sin^2(kx)\, \mathrm{d}x$$

a dále

$$\int_{-\pi}^{\pi} \left[\cos^2(kx) - \sin^2(kx) \right] \, \mathrm{d}x = \int_{-\pi}^{\pi} \cos(2kx)\, \mathrm{d}x = 0.$$

Na základě toho lze říci, že

$$\int_{-\pi}^{\pi} \cos^2(kx)\, \mathrm{d}x = \int_{-\pi}^{\pi} \sin^2(kx)\, \mathrm{d}x$$

a tedy:

$$\int_{-\pi}^{\pi} \cos^2(kx)\, \mathrm{d}x = 2\pi - \int_{-\pi}^{\pi} \cos^2(kx)\, \mathrm{d}x.$$

Odtud je tedy zřejmé, že

$$\| \cos(kx) \|^2 = \| \sin(kx) \|^2 = \int_{-\pi}^{\pi} \cos^2(kx)\, \mathrm{d}x = \int_{-\pi}^{\pi} \sin^2(kx)\, \mathrm{d}x = \pi. \tag{1.52}$$

Namísto 2π-periodických funkcí budeme ale obecně uvažovat L-periodické funkce, které jsou lépe využitelné v praxi. Fourierova řada pro L-periodické funkce na intervalu $\langle 0, L)$ je daná jako:

$$f(x) = \frac{A_0}{2} + \sum_{k=1}^{\infty} \left(A_k \cos\left(\frac{2\pi kx}{L}\right) + B_k \sin\left(\frac{2\pi kx}{L}\right)\right). \tag{1.53}$$

Její koeficienty A_k, B_k pak získáme následovně:

$$\begin{aligned} A_k &= \frac{2}{L}\int_0^L f(x)\cos\left(\frac{2\pi kx}{L}\right)\,\mathrm{d}x, \\ B_k &= \frac{2}{L}\int_0^L f(x)\sin\left(\frac{2\pi kx}{L}\right)\,\mathrm{d}x. \end{aligned} \tag{1.54}$$

V rámci Fourierovy analýzy je výhodné pracovat s komplexními funkcemi. Budeme tedy uvažovat Fourierovu řadu pro komplexní 2π-periodickou funkci $f(x)$. Využijeme Eulerovu formuli:

$$\mathrm{e}^{ikx} = \cos(kx) + i\sin(kx) \tag{1.55}$$

a komplexní koeficienty:

$$c_k = \alpha_k + i\beta_k. \tag{1.56}$$

Fourierovou řadu v komplexní formě pak můžeme vyjádřit jako:

$$f(x) = \sum_{k=-\infty}^{\infty} c_k \mathrm{e}^{ikx} = \sum_{k=-\infty}^{\infty} (\alpha_k + i\beta_k)(\cos(kx) + i\sin(kx)). \tag{1.57}$$

Rovnost (1.57) lze dále upravovat:

$$\begin{aligned} f(x) = \sum_{k=-\infty}^{\infty} (\alpha_k \cos(kx) - \beta_k \sin(kx)) + i \sum_{k=-\infty}^{\infty} (\beta_k \cos(kx) + \alpha_k \sin(kx)) = \\ = (\alpha_0 + i\beta_0) + \sum_{k=1}^{\infty} [(\alpha_{-k} + \alpha_k) \cos(kx) + (\beta_{-k} - \beta_k) \sin(kx)] + \\ + i \sum_{k=1}^{\infty} [(\beta_{-k} + \beta_k) \cos(kx) - (\alpha_{-k} - \alpha_k) \sin(kx)] . \end{aligned} \tag{1.58}$$

Pokud by byla funkce $f(x)$ reálná, bude celá imaginární část nulová. Musí tedy platit $\beta_{-k} + \beta_k = 0$ a zároveň $\alpha_{-k} - \alpha_k = 0$, tedy $\beta_{-k} = -\beta_k$ a $\alpha_{-k} = \alpha_k$. Na základě toho jsou čísla c_{-k} a c_k pro reálnou funkci $f(x)$ komplexně sdružená.

Funkce $\phi_k = \mathrm{e}^{ikx}$ pro $k \in \mathbb{Z}$ tvoří bázi pro komplexní periodické funkce na intervalu $\langle 0, 2\pi)$. Tyto funkce jsou navíc ortogonální, což lze vidět, když rozepíšeme skalární součin funkcí ϕ_j a ϕ_k:

$$\langle \phi_j, \phi_k \rangle = \int_{-\pi}^{\pi} \mathrm{e}^{jx} \mathrm{e}^{-ikx} \,\mathrm{d}x = \int_{-\pi}^{\pi} \mathrm{e}^{(j-k)x} \,\mathrm{d}x = \left[\frac{\mathrm{e}^{i(j-k)x}}{i(j-k)} \right]_{-\pi}^{\pi} = \begin{cases} 0 & \text{pro } j \neq k \\ 2\pi & \text{pro } j = k. \end{cases}$$

1.4.2. Fourierova transformace

Dalším důležitým pojmem v rámci Fourierovy analýzy je *Fourierova transformace*. Vychází z Fourierových řad a v podstatě se jedná a jejich limitní verzi. Fourierovy řady jsou pro L-periodické funkce definované na intervalu $\langle -L, L)$. V případě Fourierovy transformace rozšíříme definiční obor funkce na interval $(-\infty, \infty)$. Fourierova transformace je na rozdíl od Fourierových řad definovaná pro neperiodické funkce. Fourierova řada na intervalu $\langle -L, L)$

má následující podobu:

$$f(x) = \frac{A_0}{2} + \sum_{k=1}^{\infty} \left(A_k \cos\left(\frac{\pi k x}{L}\right) + B_k \sin\left(\frac{\pi k x}{L}\right)\right) = \sum_{k=-\infty}^{\infty} c_k \mathrm{e}^{ik\pi x/L}, \tag{1.59}$$

přičemž koeficienty c_k získáme jako:

$$c_k = \frac{1}{2L}\langle f(x), \psi_k\rangle = \frac{1}{2L}\int_{-L}^{L} f(x)\mathrm{e}^{-ik\pi x/L}\,\mathrm{d}x. \tag{1.60}$$

Označme $\omega = \frac{k\pi}{L}$ a dále $\Delta\omega = \frac{\pi}{L}$. Odtud plyne $L = \frac{\pi}{\Delta\omega}$. Nyní vezmeme limitu $L \to \infty$ a tedy $\Delta\omega \to 0$. Získáme následující předpis:

$$f(x) = \lim_{\Delta\omega\to 0} \sum_{k=-\infty}^{\infty} \frac{\Delta\omega}{2\pi} \int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(\xi)\mathrm{e}^{-ik\Delta\omega\xi}\,\mathrm{d}\xi \mathrm{e}^{ik\Delta\omega x}, \tag{1.61}$$

přičemž část $\int_{-\pi/\Delta\omega}^{\pi/\Delta\omega} f(\xi)\mathrm{e}^{-ik\Delta\omega\xi}\,\mathrm{d}\xi$ můžeme také vyjádřit jako skalární součin $\langle f(x), \psi_(x)\rangle$. Provedeme-li nyní limitní operaci, získáme po úpravách následující Fourierovskou transformační dvojici $f(x)$ a $\hat{f}(\omega)$:

$$\hat{f}(\omega) = F(f(x)) = \int_{-\infty}^{\infty} f(x)\mathrm{e}^{-i\omega x}\,\mathrm{d}x, \tag{1.62}$$

$$f(x) = F^{-1}(\hat{f}(\omega)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)\mathrm{e}^{i\omega x}\,\mathrm{d}\omega. \tag{1.63}$$

Výraz (1.62) nazýváme *Fourierova transformace* funkce f a výraz (1.63) nazýváme *inverzní Fourierova transformace*.

1.4.3. Diskrétní a rychlá Fourierova transformace

Dosud jsme uvažovali Fourierovu řadu a Fourierovu transformaci pro spojité funkce $f(x)$. V praxi ale máme k dispozici většinou data diskrétního typu. Proto je nutné aproximovat Fourierovu transformaci pro diskrétní da-

tové vektory. K tomuto účelu se využívá *diskrétní Fourierova transformace* označovaná krátce jako DFT. Tuto zkratku budeme dále používat. V případě DFT se v podstatě jedná o diskrétní verzi Fourierových řad pro vektor dat $\mathbf{f} = [f_0, f_1, \ldots, f_{n-1}]^T$.

DFT je velmi užitečným nástrojem pro analýzu periodických i neperiodických signálů. Pomůže nám při numerické aproximaci i výpočtech, ovšem zásadní nevýhodou této metody je složitost výpočtu. Provedení DFT totiž vyžaduje řádově n^2 operací, což je pro velká n silně nepraktické. V praxi je proto využíván algoritmus označovaný jako *rychlá Fourierova transformace* (Fast Fourier Transform - FFT). Jeho provedení vyžaduje pouze $n \log(n)$ operací.

Diskrétní Fourierova transformace vycházející z (1.62) je daná jako:

$$\hat{f}_k = \sum_{j=0}^{n-1} f_j \mathrm{e}^{-i2\pi jk/n}. \tag{1.64}$$

Podobně můžeme vyjádřit také inverzní diskrétní Fourierovu transformaci:

$$f_k = \frac{1}{n} \sum_{j=0}^{n-1} \hat{f}_j \mathrm{e}^{i2\pi jk/n}. \tag{1.65}$$

DFT tedy můžeme vnímat jako lineární operátor zobrazující body z $\mathbf{f}$ do frekvenční domény $\hat{\mathbf{f}}$:

$$\{f_0, f_1, \ldots, f_{n-1}\} \overset{\text{DFT}}{\Longrightarrow} \left\{\hat{f}_0, \hat{f}_1, \ldots, \hat{f}_{n-1}\right\}. \tag{1.66}$$

Označme nyní $\omega_n = \mathrm{e}^{-2\pi i/n}$. Pro daný počet bodů n můžeme DFT počítat

pomocí maticového násobení:

$$\begin{pmatrix} \hat{f}_0 \\ \hat{f}_1 \\ \hat{f}_2 \\ \vdots \\ \hat{f}_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \cdots & \omega_n^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)^2} \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{n-1} \end{pmatrix}. \tag{1.67}$$

Výstupní vektor $\hat{\mathbf{f}}$ obsahuje Fourierovy koeficienty, které lze interpretovat jako čísla, která nám říkají, jak moc je daná frekvence zastoupena ve vstupním vektoru $\mathbf{f}$. Matice

$$\mathbf{F} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \cdots & \omega_n^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)^2} \end{pmatrix} \tag{1.68}$$

je tzv. unitární Vandermondeova matice. Tato matice je symetrická a obsahuje komplexní hodnoty. Každý její řádek i sloupec je cosinová funkce se zvyšující se frekvencí. Prvky matice $\mathbf{F}$ obsahují informace o frekvencích a amplitudách přítomných ve vstupním signálu $\mathbf{f}$. Každý prvek matice $\mathbf{F}$ reprezentuje konkrétní harmonickou frekvenci a její příspěvek k celkovému spektru signálu.

Algoritmus FFT umožňuje efektivní výpočet DFT pomocí snížení počtu potřebných operací. Základní myšlenka FFT spočívá v tom, že DFT může být implementována mnohem efektivněji, pokud počet datových bodů n je mocninou čísla 2. Uvažujme případ, kde $n = 1024 = 2^{10}$. Matici DFT $\mathbf{F}_{1024}$

pak můžeme rozepsat následujícím způsobem:

$$\hat{\mathbf{f}} = \mathbf{F}_{1024}\mathbf{f} = \begin{pmatrix} \mathbf{I}_{512} & -\mathbf{D}_{512} \\ \mathbf{I}_{512} & -\mathbf{D}_{512} \end{pmatrix} \begin{pmatrix} \mathbf{F}_{512} & 0 \\ 0 & \mathbf{F}_{512} \end{pmatrix} \begin{pmatrix} \mathbf{f}_{sud} \\ \mathbf{f}_{lich} \end{pmatrix}, \tag{1.69}$$

kde $\mathbf{f}_{sud}$ představuje prvky vektoru $\mathbf{f}$ se sudými indexy a $\mathbf{f}_{lich}$ jsou prvky s lichými indexy. Matice $\mathbf{I}_{512}$ je jednotková matice typu 512×512 a matice $\mathbf{D}_{512}$ je daná následovně:

$$\mathbf{D}_{512} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \omega & 0 & \cdots & 0 \\ 0 & 0 & \omega^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \omega^{511} \end{pmatrix}. \tag{1.70}$$

Výraz (1.69) lze získat s využitím (1.64) a (1.67). V případě, kdy $n = 2^p$ můžeme zopakovat celý proces a matici $\mathbf{F}_{512}$ lze vyjádřit za pomoci matice $\mathbf{F}_{256}$. Tímto způsobem můžeme pokračovat dále:

$$\mathbf{F}_{256} \to \mathbf{F}_{128} \to \mathbf{F}_{64} \to \mathbf{F}_{32} \to \mathbf{F}_{16} \to \mathbf{F}_8 \to \mathbf{F}_4 \to \mathbf{F}_2. \tag{1.71}$$

V případě kdy n není mocninou čísla 2, doplní se vektor $\mathbf{f}$ nulami, dokud nebude $n = 2^p$. Algoritmus FFT zahrnuje efektivní střídání sudých a lichých indexů podvektorů $\mathbf{f}$ a výpočet rozložený do několika menších DFT výpočtů typu 2×2. Odtud je zřejmé výrazné zrychlení výpočtu z n^2 operací na $n \log n$ operací. V praxi tak je využíván převážně algoritmus FFT, který je vbudován mimo jiné i do programu *R*.

2. Praktická analýza časových řad

Kapitola 1 byla zaměřena na teoretický základ některých metod využívaných v oblasti analýzy časových řad. V této kapitole budou metody naopak prakticky použity na reálná data získaná od průmyslové firmy Sigma. Samotná analýza je provedena v prostředí softwaru *R*. Zdroje využité pro tuto analýzu a sepsání této kapitoly jsou následující: [5], [6], [7], [8], [11], [12], [13] a [14]. Výrazným způsobem byla využita zejména nápověda vbudovaná do programu *R* a tedy zdroj [11].

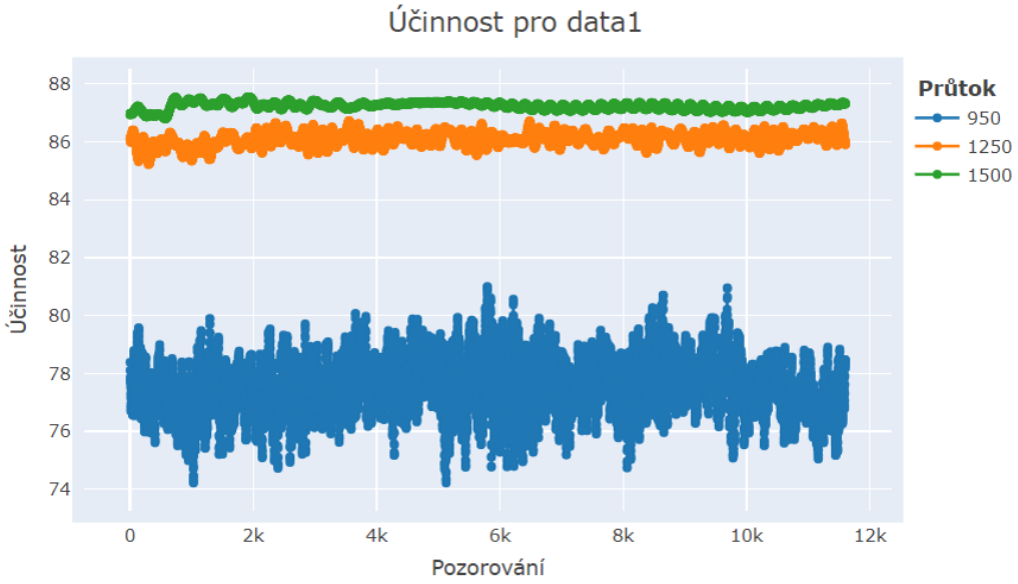
2.1. Datová sada a její vizualizace

Data, která budeme dále analyzovat, vznikla simulacemi ve společnosti Sigma. Data zobrazují odsimulované účinnosti různých čerpadel (v procentech) při různě velkém průtoku vody (v litrech za minutu). Jednotlivá čerpadla se liší například tvarem, velikostí a počtem lopatek. Tato konkrétní čerpadla reálně neexistují, jedná se o jejich návrhy a modely, ze kterých jen ty nejvhodnější budou nakonec vyráběny a reálně testovány. Vzhledem k tomu, že se jedná o velká a drahá průmyslová čerpadla, je nutná předchozí důkladná analýza simulovaných dat, aby bylo minimalizováno riziko zbytečně vynaložených financí na výrobu nevhodného typu čerpadla. Část analýzy provedeme a ukážeme v této diplomové práci.

Analyzovaná budou data pro čtyři různá čerpadla (ozn. data1, data2, data3 a data4). Každé z nich pak při třech různě silných průtocích vody, a to 950 $l \cdot min^{-1}$ (ozn. písmenem a, dále budou využity zkratky jako dat1a), 1250 $l \cdot min^{-1}$ (ozn. pomocí písmena b, např. dat3b) a 1500 $l \cdot min^{-1}$ (ozn. písmenem c). Celková délka každé z datových sad přesahovala 13000 pozorování, ovšem prvních přibližně 2000 pozorování bylo vlivem simulací extrémně

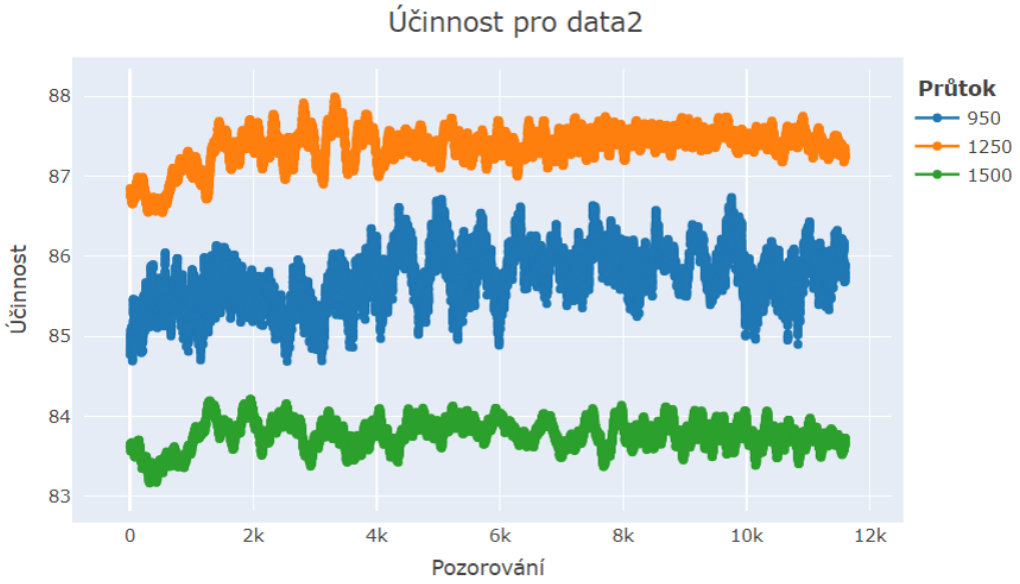
rozkolísaných a nestabilních s velkou amplitudou, proto prvních 2000 pozorování vynecháme. Také zkrátíme některá data řádově o jednotky pozorování z konce časové řady, abychom pracovali se stejně dlouhými časovými řadami a usnadnila se tak interpretace. Výsledná délka všech uvažovaných časových řad, se kterými budeme pracovat, pak dosahuje 11605 pozorování.

Podívejme se nejprve na všechna data vizuálně. Využijeme k tomu vykreslování grafů v programu *R* pomocí knihovny *plotly*. Tento způsob vykreslování grafů byl zvolen z důvodu, že se jedná o poměrně velký datový soubor, kde by mohlo být užitečné přibližovat si konkrétní části časové řady. To nám interaktivní prostředí plotly dovoluje. Vizualizace účinností pro první datovou sadu je následující:



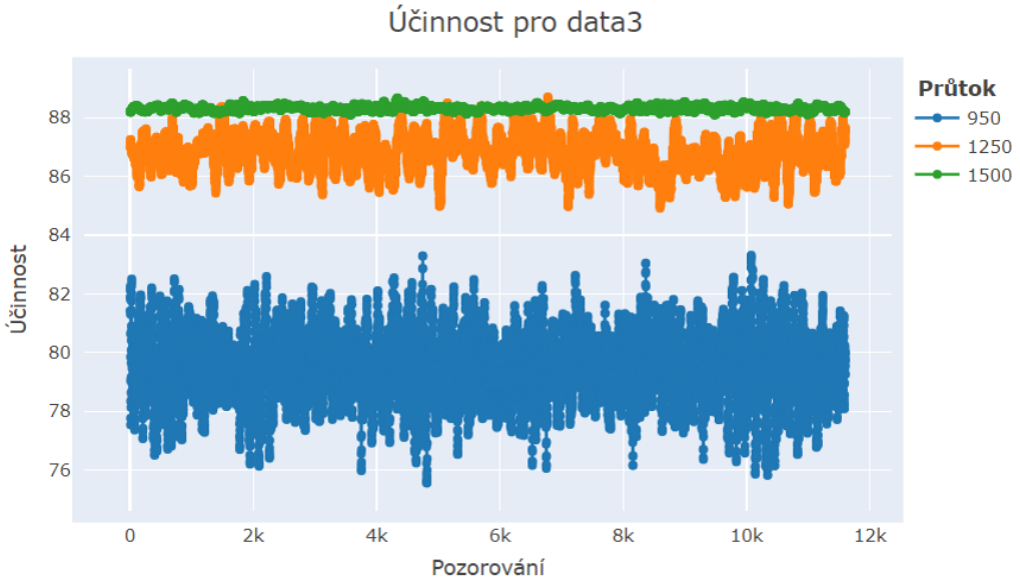
Obrázek 1: Vizualizace účinnosti prvního čerpadla

Z obrázku 1 můžeme vyčíst několik věcí. Je zřejmé, že při nejmenším průtoku vody (950) má čerpadlo nejmenší střední hodnotu účinnosti a zároveň největší rozpětí hodnot (rozdíl maximální a minimální hodnoty). Naopak při největším průtoku vody (1500) dosahuje účinnost nejvyšších hodnot a zároveň nejmenšího rozpětí dat. Podívejme se stejným způsobem graficky na další datový soubor:

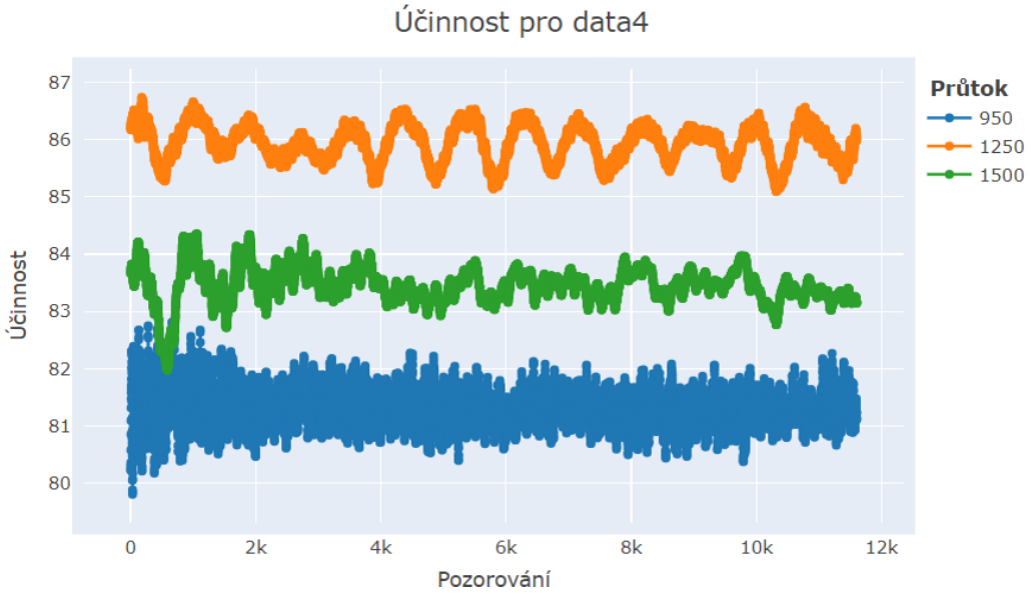


Obrázek 2: Vizualizace účinnosti druhého čerpadla

I zde je viditelně největší rozpětí hodnot u nejslabšího průtoku vody a nejmenší rozpětí má nejsilnější průtok. Ten však v tomto případě dosahuje nejmenší účinnosti. Nejvyšší střední hodnotu má průtok 1250. Podívejme se ještě na dvě zbývající datové sady:



Obrázek 3: Vizualizace účinnosti třetího čerpadla



Obrázek 4: Vizualizace účinnosti čtvrtého čerpadla

Obrázky 3 a 4 mají podobné vlastnosti jako obrázky 1 a 2. Konkrétní charakteristiky (zaokrouhlené na dvě desetinná místa) získané v programu *R* pomocí příkazů `summary(data)` můžeme zobrazit v tabulce:

Data	Průtok	Stř. hodnota	Minimum	Maximum	Rozpětí
data1	950	77.53	74.21	81.00	6.79
	1250	86.10	85.21	86.73	1.52
	1500	87.24	86.80	87.55	0.75
data2	950	85.73	84.69	86.74	2.05
	1250	87.36	86.55	87.99	1.44
	1500	83.78	83.17	84.21	1.04
data3	950	79.45	75.56	83.31	7.75
	1250	86.85	84.91	88.70	3.79
	1500	88.32	88.08	88.67	0.59
data4	950	81.30	79.81	82.81	3.00
	1250	85.94	85.09	86.73	1.64
	1500	83.45	81.97	84.35	2.38

Tabulka 1: Základní charakteristiky vypočtených hodnot účinnosti čerpadel

Můžeme pozorovat, že ve všech případech dosahují největšího rozpětí dat časové řady s nejmenším průtokem. S výjimkou čtvrté datové sady pak nejmenší rozpětí mají řady s nejsilnějším průtokem. Co se týče středních hodnot, tam situace není jednoznačná. Nejvyšší účinnosti dosahují buď průtoky 1250, nebo 1500. V případě první a třetí časové řady dosahují čerpadla při průtoku 950 výrazně horší účinnosti než při silnějších průtocích. U druhé a čtvrté řady rozdíl není tak markantní.

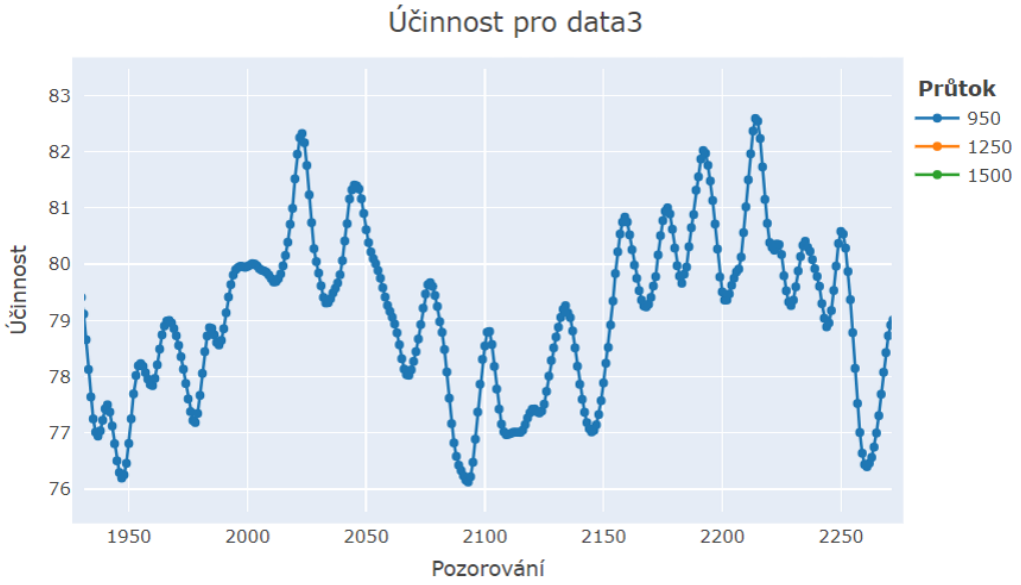
2.2. Periodicita v datech

Dalším specifikem těchto časových řad je periodicita. Ta se vyskytuje u všech dvanácti datových souborů, které máme k dispozici. Na periodickou složku je vhodné podívat se nejprve z grafického hlediska a následně vyzkou-

šet na data aplikovat Fourierovu analýzu, zejména pak rychlou Fourierovu transformaci popsanou v části 1.4.3.

2.2.1. Grafický pohled na periodicitu

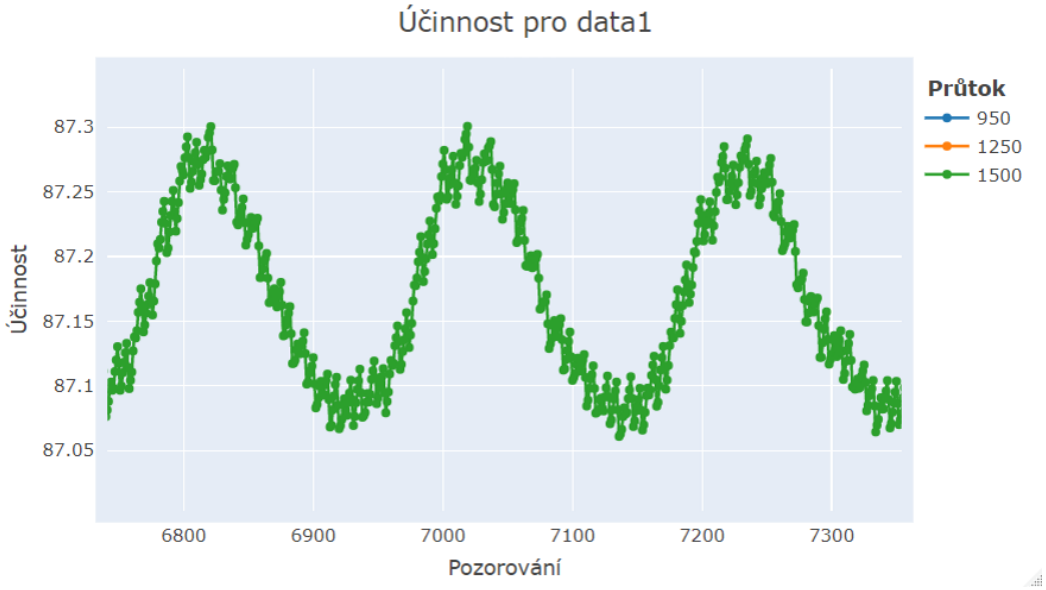
Kromě výše uvedeného je v datech patrná periodicita. Někde, jako třeba v sadě data4 při průtoku 1250, jsou periodické vzory jasně viditelné. V některých případech to ovšem z výše vložených obrázků zřejmé není. Například při průtoku 950 v sadě data3 je vzhledem k množství zhuštěných dat velmi složité odhalit periodicitu. Ovšem stačí data v prostředí plotly přiblížit a periody můžeme pozorovat poměrně snadno:



Obrázek 5: Přiblížení dat3a

Na obrázku 5 vidíme periodicitu dvojího typu. Dvakrát se zde vyskytuje perioda dlouhá přibližně 160 pozorování. Kromě ní můžeme na obrázku pozorovat také asi patnáctkrát zopakovanou periodu o délce kolem 25 pozorování. Podobným způsobem lze odhalit periodičnost u všech dvanácti časových řad, které máme k dispozici. Hezkou ilustrací periodicity je například i časová řada

ze sady data1 při průtoku 1500, kde po přiblížení můžeme pozorovat velmi zřetelné periody délky lehce přes 200 pozorování:



Obrázek 6: Přiblížení dat1c

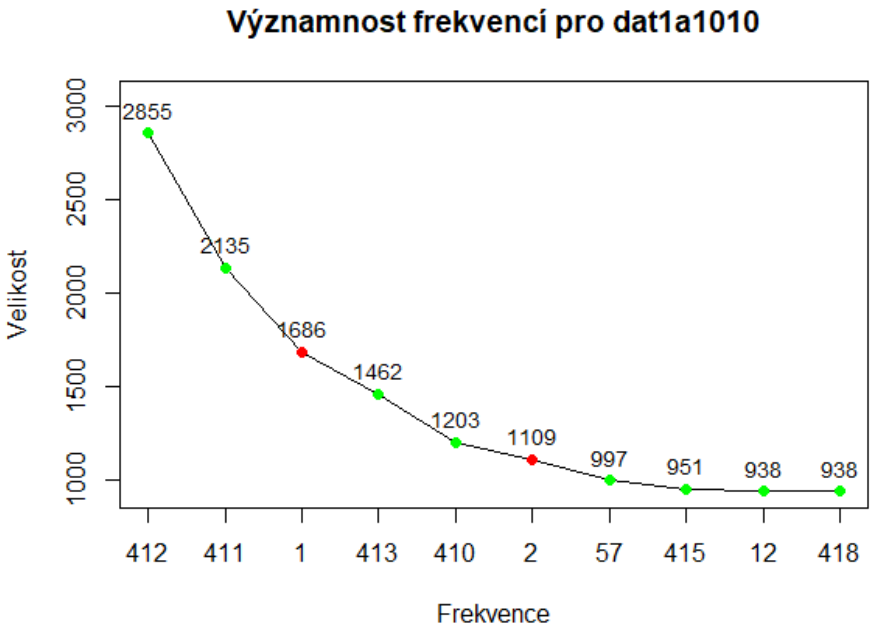
2.2.2. Využití FFT

Díky vizualizaci můžeme získat přibližnou představu o periodicitě v datech, ovšem pro důkladnější analýzu je třeba sáhnout po některé ze statistických metod. My zde využijeme, v teoretické části již vysvětlenou, Fourierovu analýzu, konkrétněji fast Fourier transform.

Cílem je detekce nejvýznamnějších frekvencí či period a určení, jak moc významné jsou. Budou-li mezi nejvýznamnějšími periodami i periody velmi dlouhé, může to naznačovat neustálenost časové řady. Kratší periody pak můžeme vyhlazovat například pomocí metody klouzavých průměrů.

V softwaru *R* je přímo vbudovaná funkce `fft()` provádějící rychlou Fourierovu transformaci. Ta je také základním stavebním kamenem, pro účel diplomové práce vytvořené, funkce `fourier()`. Pomocí ní provedeme pro

vybraná data FFT, díky čemuž detekujeme nejvýznamnější frekvence. Poznamenejme, že pojem *frekvence* je zde i dále v textu použit ve významu počtu opakování periodického děje za jednotku času, přičemž za jednotku času budeme vždy brát aktuální celkovou délku dané řady. Výsledky funkce `fourier()` můžeme prezentovat graficky s využitím další vytvořené funkce pojmenované `vyznamnost_frekvenci()`. Deset nejvýznamnějších frekvencí pro dat1a, tedy první datový soubor při průtoku 950, vykreslíme následovně:

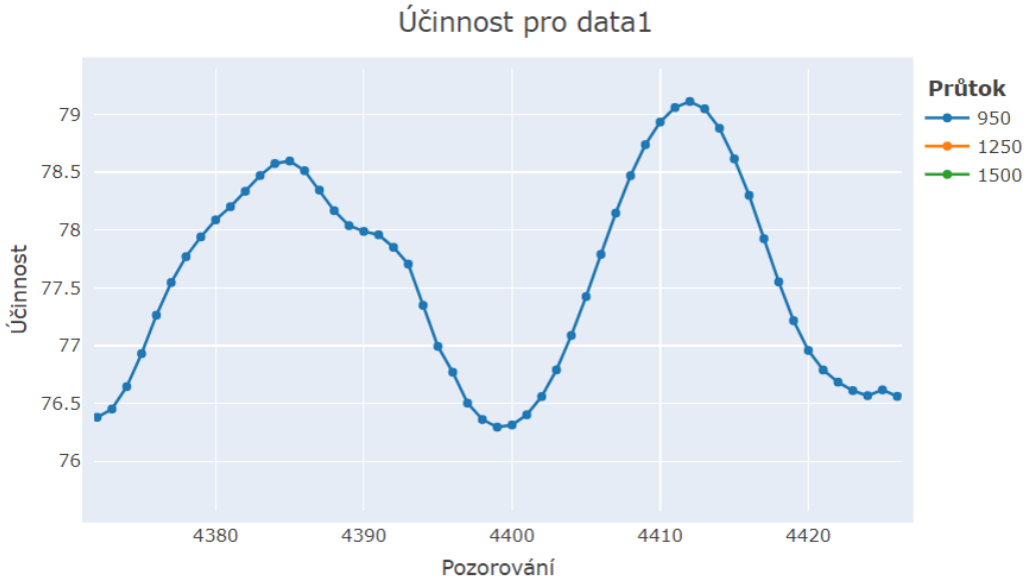


Obrázek 7: Významnost frekvencí pro dat1a

Obrázek 7 má na ose x vypsaných deset nejvýznamnějších frekvencí na základě FFT. Osa y pak zobrazuje (v absolutní hodnotě a zaokrouhlenou na jednotky) velikost příspěvku (amplitudu) dané frekvence k celkovému signálu. Pro každou z nejvýznamnějších frekvencí je v grafu velikost příspěvku znázorněna bodem s číselným popiskem. Body jsou zobrazeny zelenou barvou pro frekvence > 10 a červenou pro frekvence ≤ 10. Označení dat1a1010

z obrázku 7 tedy znamená vykreslení deseti nejvýznamnějších frekvencí pro datovou sadu 1 při nejmenším průtoku a červeném vyznačení frekvencí ≤ 10

Vidíme, že nejvýznamnější frekvencí je poměrně výrazně frekvence 412, což odpovídá pro délku časové řady $n = 11605$ periodě o délce $\frac{11605}{412} = 28.17$ pozorování. Nejvýznamnější částí periodické složky jsou tedy periody obsahující asi 28 pozorování. Tento vzor se v datech zopakuje 412krát. Že se jedná o nejvýraznější periodický vzor v datech potvrzuje i fakt, že druhá, čtvrtá, pátá, osmá i desátá nejvýznamnější frekvence se pohybuje mezi 410 a 418 opakování v datech, což opět dává délku periody pohybující se kolem 28 pozorování. O pravdivosti výskytu takto dlouhé periody se můžeme přesvědčit i vizuálně po přiblížení grafu:

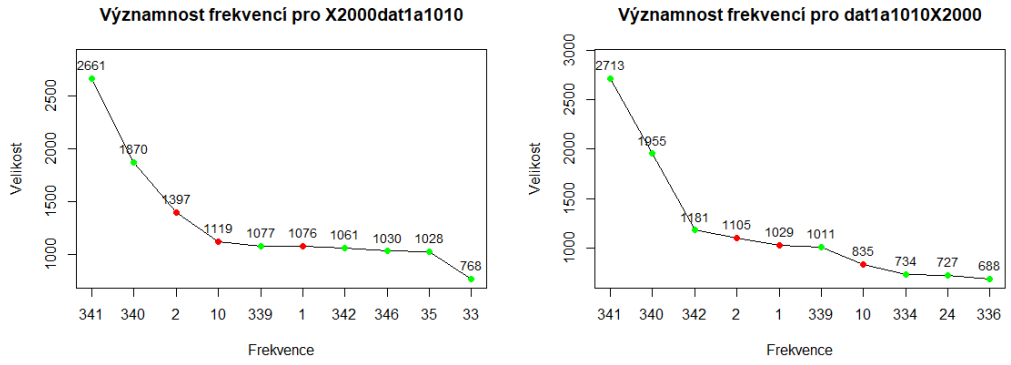


Obrázek 8: Přiblížení dat1a

Pozastavit se musíme u třetí nejvýznamnější frekvence, což je 1. Celá časová řada by tedy dle tohoto výsledku byla jednou periodou. Interpretovat to můžeme tak, že třetí nejvýznamnější frekvence tvrdí, že více než periodicita hraje v datech roli nějaký dlouhodobý trend. V kontextu hodnocení

ustálenosti procesu to značí nestabilitu a značnou komplikaci pro predikce.

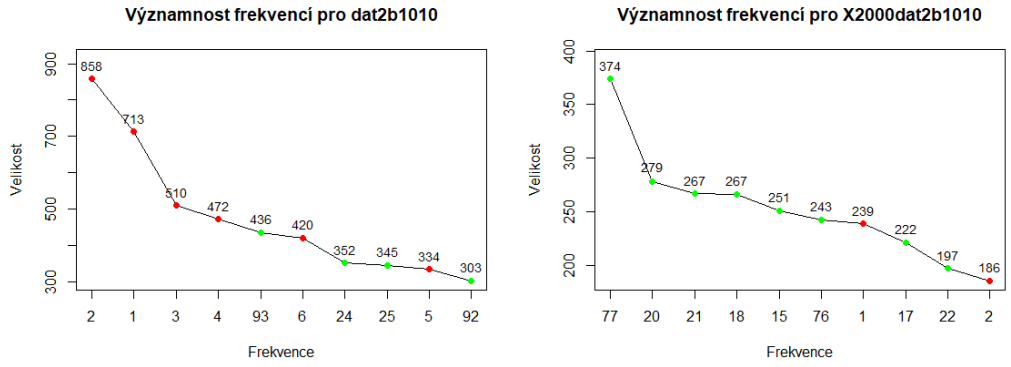
V některých případech může být výhodné zkrácení časové řady. Zkrácením zepředu můžeme eliminovat vliv vyšší rozkolísanosti z důvodu simulací. Zkrácení zezadu pak může být zajímavé pro srovnání s řadou zkrácenou zepředu, případně pak pro predikce. Významnosti frekvencí ve zkrácených řadách jsou následující:



Obrázek 9: Významnost frekvencí pro zkrácené dat1a

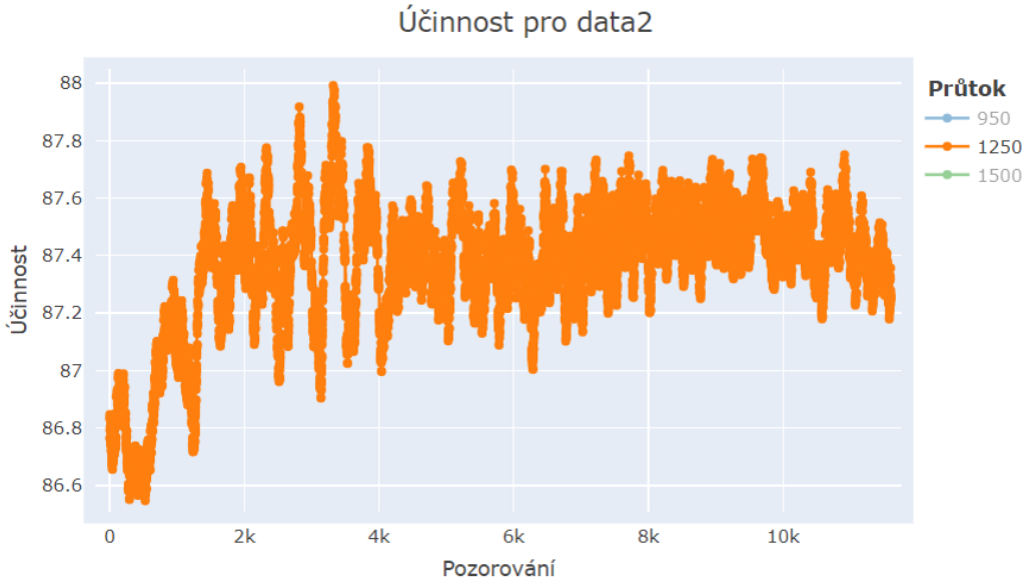
Levý graf na obrázku 9 znázorňuje významnosti frekvencí po zkrácení řady o 2000 pozorování zepředu, pravá část pracovala se řadou zkrácenou o 2000 pozorování z konce. V obou grafech můžeme vysledovat podobnosti. Nejvýznamnější je v obou případech frekvence 341. Vzhledem k tomu, že nyní $n = 9605$, tak délka nejvýznamnější periody je $\frac{9605}{341} = 28.17$, což plně odpovídá výsledkům pro nezkrácenou časovou řadu. Interpretace pro další významné frekvence je obdobná. Za povšimnutí stojí, že poměrově ku nejvýznamnější frekvenci se v obou případech snížila významnost frekvence 1.

Podobným způsobem je možné pokračovat pro všechny datové sady. Zajímavé jsou například výsledky pro dat2b, tedy druhou datovou sadu při průtoku 1250. Nejvýznamnější frekvence pro celou časovou řadu a pro časovou řadu po zkrácení o prvních 2000 pozorování můžeme vykreslit následovně:



Obrázek 10: Významnost frekvencí pro dat2b

Vidíme zde obrovský rozdíl ve významnosti frekvencí. Zatímco u nezkrácené časové řady jasně dominují malé frekvence (tedy velmi dlouhé periody), tak u zkrácené je vliv nízkých frekvencí potlačen. V kratší řadě je nejvýznamnější frekvence 77, tedy $\frac{9605}{77} = 124.74$ pozorování v rámci periody, což odpovídá páté nejvýznamnější frekvenci v delší řadě, a to 93, protože $\frac{11605}{93} = 124.78$.



Obrázek 11: Vizualizace dat2b

Důvod, proč při zkrácení časové řady dojde k utlumení významu dlouhých period je zřejmý z prostého pohledu na danou časovou řadu. Z obrázku 11 je totiž patrné, že funkce na prvních 2000 pozorováních prudce roste a až poté se relativně ustaluje kolem hodnoty 87.4 %. Pokud tedy bereme všechna pozorování, řada je neustálená. Vynecháme-li ale prvních 2000 pozorování, vykazují data značně ustálenější chování.

Podobným způsobem můžeme analyzovat všechny datové sady, které máme k dispozici. Z úsporných důvodů ale nebudeme vykreslovat další grafy. Místo toho můžeme výsledky pro všechny datové soubory zahrnout do tabulek. Čtenář si případně může grafy nechat vykreslit v příloze diplomové práce pojmenované jako *spousteni_funkci.R*. Podívejme se nyní na tabulku pro sadu data1:

Data	Frek.:	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1a	F:	412	411	1	413	410	2	57	415	12	418
	V:	2855	2135	1686	1462	1203	1109	997	951	938	938
x1a	F:	341	340	2	10	339	1	342	346	35	33
	V:	2661	1870	1397	1119	1077	1076	1061	1030	1028	768
1b	F:	76	80	2	1	16	13	83	3	63	70
	V:	693	429	411	409	408	382	381	361	353	349
x1b	F:	63	64	56	68	58	52	13	43	39	51
	V:	660	376	313	307	302	301	290	262	259	251
1c	F:	1	57	4	3	6	5	7	55	11	10
	V:	477	346	263	246	207	205	201	197	184	165
x1c	F:	1	47	3	46	26	25	27	24	50	1601
	V:	383	290	253	198	74	72	62	62	52	51

Tabulka 2: Nejvýznamnější frekvence (F) a jejich významnost (V) pro data1

Tabulka 2 zobrazuje deset nejvýznamnějších frekvencí a velikosti amplitud daných frekvencí (zaokrouhlených na jednotky) pro sadu data1. Pod zkratkou „1a“ jsou myšlena data1 při průtoku 950. Zkratka „x1a“ pak značí řadu data1 při průtoku 950 po zkrácení dat o prvních 2000 pozorování. Podobným

způsobem jsou značena data pro větší průtoky, jen s použitím písmen „b“ a „c“. Zkratka „F:“ označuje frekvenci, „V:“ pak značí velikost příspěvku dané frekvence k celkovému signálu. Červenou barvou jsou označeny frekvence ≤ 10, které mohou signalizovat neustálenost řady.

Na základě hodnot frekvencí lze vysledovat, že datové soubory vykazují vyšší sklon k ustálenosti po odečtení prvních 2000 pozorování. Jednoznačně patrné to je v případě řady 1b, kde ve zkrácené řadě zmizely všechny nízké frekvence a tedy dlouhé periody. Řada 1c se jeví jako nejméně ustálená, byť z tabulky 1 víme, že má nejmenší rozpětí dat. Pro sadu data2 je situace následující:

Data	Frek.:	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
2a	F:	1	20	19	17	5	10	16	4	6	289
	V:	1519	921	529	495	458	451	403	399	390	373
x2a	F:	1	16	14	2	4	3	18	235	6	244
	V:	1003	866	522	446	438	421	415	349	336	321
2b	F:	2	1	3	4	93	6	24	25	5	92
	V:	858	713	510	472	436	420	352	345	334	303
x2b	F:	77	20	21	18	15	76	1	17	22	2
	V:	374	279	267	267	251	243	239	222	197	186
2c	F:	3	17	1	4	100	98	19	29	5	6
	V:	546	522	435	421	355	314	310	274	251	226
x2c	F:	14	81	83	24	3	16	13	15	1	28
	V:	392	311	271	254	252	209	207	199	198	176

Tabulka 3: Nejvýznamnější frekvence (F) a jejich významnost (V) pro data2

Z tabulky 3 můžeme opět vyčíst několik zajímavých informací. Při nejnižším průtoku je pro data2 nejvýznamnější frekvence 1, podle amplitudy docela výrazně, což jasně nasvědčuje neustálenosti řady. V případě řad 2b a 2c je velmi zřetelný vliv prvních 2000 pozorování, protože po zkrácení řady můžeme pozorovat značné ustálení. Důvodem je výrazný růst hodnot účin-

nosti během prvních 2000 pozorování. Výsledky FFT pro sadu data3 ukazuje následující tabulka:

Data	Frek.:	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
3a	F:	69	413	414	82	599	64	605	72	602	609
	V:	2013	1871	1307	1237	1231	1225	1223	1184	1167	1132
x3a	F:	342	57	48	68	495	54	500	489	337	53
	V:	2267	1469	1104	1102	1057	971	952	951	950	948
3b	F:	61	55	2	44	72	66	8	49	1	41
	V:	1104	992	963	910	902	869	862	858	851	832
x3b	F:	36	2	48	41	7	1	51	8	57	34
	V:	1136	981	947	907	878	843	775	770	767	735
3c	F:	107	5	102	6	104	30	110	645	105	41
	V:	182	175	166	160	130	110	107	101	96	88
x3c	F:	4	89	84	5	87	86	31	25	88	534
	V:	158	146	132	129	110	107	101	101	96	82

Tabulka 4: Nejvýznamnější frekvence (F) a jejich významnost (V) pro data3

Sada data3 vykazuje známky ustáleného chování pro nejmenší průtok. Při středním průtoku situace není tak jasná, významnou roli hraje frekvence 2 a mezi desítkou nejvýznamnější frekvencí se objevuje i jednička. U časové řady s největším průtokem je pak významná frekvence 5. Vizuálně je tato frekvence v datech vidět špatně, jasnější posouzení ustálenosti by mohlo být po vyhlazení časové řady.

Při pohledu na tabulku 5 ukazující významnosti frekvencí pro sadu data4 můžeme jasně vidět, že při nejslabším průtoku vykazuje účinnost známky ustálenosti, neboť významné jsou pouze vysoké frekvence. Při středním průtoku má jednoznačně nejvyšší vliv frekvence 13, což jen potvrzuje vizuální představu z obrázku 4. Mezi deseti nejvýznamnějšími se objevují i nižší frekvence, ovšem jejich významnost je dle amplitudy poměrně nízká. Nejvyšší průtok je z hlediska interpretace složitější. Významná je frekvence 2 i 3 což naznačuje neustálenost. Nejdůležitější je ale frekvence 13 (respektive 11 po

zkrácení řady) a frekvence 1 se zde nevyskytuje. Možná nám i zde pomůže vyhlazení časové řady.

Data	Frek.:	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
4a	F:	308	305	307	310	53	381	931	1238	1239	76
	V:	1084	626	514	371	359	345	339	338	332	331
x4a	F:	255	253	316	771	257	44	55	1026	254	258
	V:	654	487	418	415	338	311	248	236	235	230
4b	F:	13	12	11	14	3	5	20	25	16	18
	V:	1908	906	660	571	372	352	348	302	290	264
x4b	F:	10	11	12	9	13	7	21	4	17	1
	V:	1312	1111	516	458	445	313	295	283	258	226
4c	F:	13	2	16	8	12	54	108	109	14	3
	V:	1330	732	664	609	592	559	515	455	436	392
x4c	F:	11	7	3	45	2	90	89	13	8	25
	V:	696	615	553	527	446	332	307	260	210	202

Tabulka 5: Nejvýznamnější frekvence (F) a jejich významnost (V) pro data4

2.3. Vyhlazení dat pomocí klouzavých průměrů

Jak již bylo naznačeno, při další analýze nám může pomoci vyhlazení dat. Pokud se povede dostatečně odfiltrovat šum a periodickou strukturu dat, může být snadnější rozhodnout o ustálenosti procesu. V datech, se kterými pracujeme, se periodická složka i šum vyskytují ve velké míře. Pro vyhlazení těchto časových řad využijeme metodu klouzavých průměrů, která byla teoreticky popsána v kapitole 1.3.

Pro implementaci klouzavých průměrů v programu *R* je využívána knihovna zoo a funkce `rollmean()`. Myšlenka vyhlazení dat pomocí klouzavých průměrů v tomto případě spočívá ve využití nejvýznamnějších frekvencí získaných pomocí rychlé Fourierovy transformace pro výpočet délky klouzavého

okna. Délku okna vypočítáme jako $k = \frac{n}{frekvence}$, kde n je délka vyhlazovaných dat a $frekvence$ je daná frekvence získaná pomocí FFT.

Vyhlazování dat chceme samozřejmě co nejvíce automatizovat, proto pro účely této diplomové práce vznikla funkce `fourier_vyhlazeni()`, která v sobě spouští další speciálně vytvořené funkce. Konkrétně již dříve zmíněnou funkci `fourier()`, dále funkci `klouzave_prumery()`, která má za úkol pomocí funkce `rollmean()` vyhladit postupně data, také funkci `vykresleni()` pro vizualizaci výsledků a v neposlední řadě `filtr_frekvenci()`.

U poslední zmíněné funkce se na chvíli zastavme. Myšlenka jejího vytvoření vychází z pohledu na tabulky deseti nejvýznamnějších frekvencí pro různá data. Podíváme-li se například v tabulce 2 na nejvýznamnější frekvence pro řadu 1a, uvidíme, že zde mezi deseti nejvýznamnějšími frekvencemi najdeme hned šest frekvencí pohybujících se mezi hodnotami 410 a 418. Jak již bylo zmíněno, tyto frekvence odpovídají přibližně 28 pozorováním v rámci klouzavého okna. Nemá tedy smysl vyhlazovat data za pomoci všech těchto frekvencí, bude nám stačit jedna z nich. Podobně tak je tomu i u dalších řad.

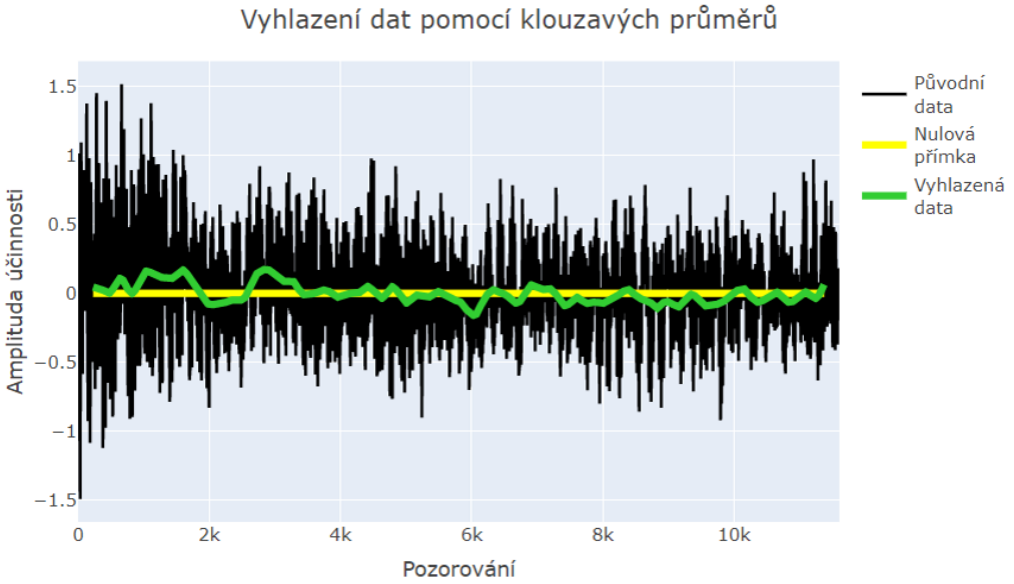
Je zbytečné využít opakovaně na danou řadu vyhlazování pomocí okna stejné či podobné délky. Vyhlazení tím nevylepšíme, naopak bychom se mohli připravit o část informace, neboť každým vyhlazováním se časová řada zkracuje. Velmi citelné by to bylo například u řady 4b, kde v tabulce 5 můžeme vidět, že čtyři nejdůležitější frekvence jsou 13, 12, 11 a 14. Vzhledem k tomu, že to znamená klouzavá okna o délkách takřka 1000 pozorování, tak by při využití všech těchto frekvencí byla ztráta informace vlivem chybějících hodnot výrazná. Z důvodu velké ztrátovosti informace pro nízké frekvence budeme k vyhlazování pomocí klouzavých průměrů využívat jen frekvence > 10.

Funkce `filtr_frekvenci()` je nastavená tak, aby ze zadaných frekvencí vybrala a zachovala jen ty, které se od sebe výrazněji liší. Konkrétně vždy

zachová první nejdůležitější frekvenci a každá další příchozí frekvence se musí lišit alespoň o 10 % od všech předchozích zachovaných frekvencí, jinak je vyřazena a k vyhlazení nebude využita. Tím zajistíme dostatečné a zároveň o informaci ne příliš ochuzující vyhlazení.

2.3.1. Ustálené časové řady

Po vyhlazení jednotlivých časových řad můžeme pozorovat různé typy výsledků. Prvním typem jsou řady, po jejichž vyhlazení můžeme prohlásit, že jsou ustálené. Například pro dat4a, tedy nejmenší průtok ze čtvrté datové sady, vypadá vyhlazení po odečtení střední hodnoty účinnosti následovně:



Obrázek 12: Vyhlazení dat4a pomocí klouzavých průměrů

Obrázek 12 zobrazuje vyhlazení časové řady 4a pomocí metody klouzavých průměrů. Černou barvou jsou reprezentována původní data, zelenou jsou vykreslena vyhlazená data. Obojí je vykresleno po odečtení průměru původních dat. Žlutá přímka, symbolizující nulovou konstantu, je do grafu přidána čistě pro jednodušší grafickou představu o kvalitě vyhlazení. Funkce

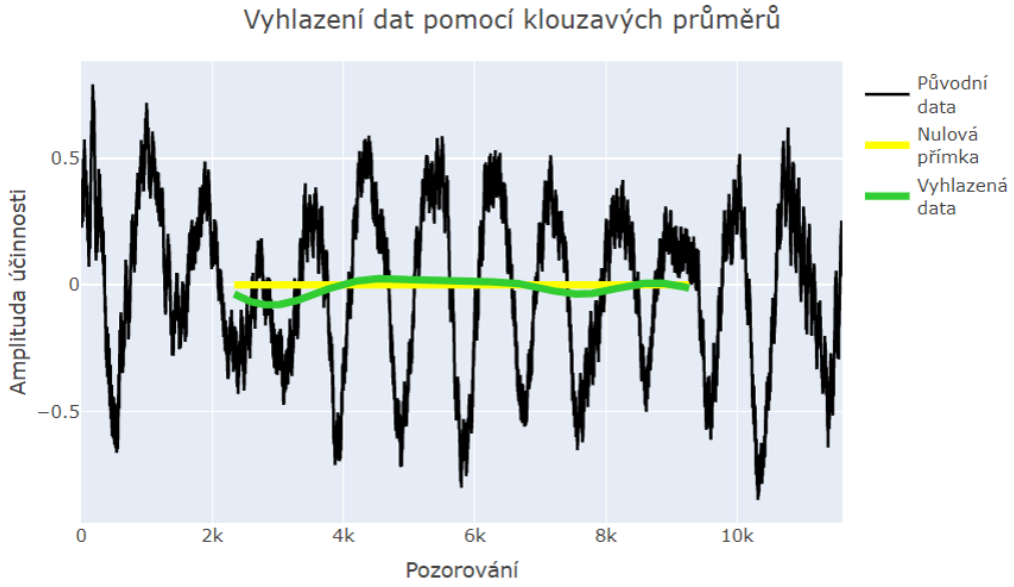
`fourier_vyhlazeni()` společně s algoritmem `filtr_frekvenci()` zde zachovala frekvence 308, 53, 381, 931, 1238 a 76. Pomocí nich vypočítala délky klouzavého okna a metodou klouzavých průměrů pro tyto délky okna vyhladila řadu do podoby zelené křivky v grafu.

Na základě tohoto obrázku opravdu můžeme prohlásit, že řada 4a je ustálená. Konkrétně u této řady se nám jen potvrdila již vizuální představa z vykreslení samotné řady a také fakt, že mezi deseti nejvýznamnějšími frekvencemi se neobjevovala žádná nízká frekvence, která by mohla signalizovat neustálenost. Důležitým ukazatelem ustálenosti je kromě grafického posouzení vyhlazené křivky také srovnání rozpětí hodnot pro původní časovou řadu a rozpětí ve vyhlazené řadě. V tomto případě je hodnota původního rozpětí (zaokrouhlena na dvě desetinná místa) 3.00 a rozpětí vyhlazené řady pouze 0.33, což činí pouhých 11.15 % původního rozpětí. To opět potvrzuje ustálenost časové řady.

Můžeme tedy prohlásit, že časová řada 4a je ustálená, a také že asymptotická střední hodnota by měla být prakticky totožná se střední hodnotou nám známé časové řady 4a, respektive střední hodnotě jejího vyhlazení. Můžeme tedy vypočítat průměr původní časové řady, což činí (po zaokrouhlení na dvě desetinná místa) 81.30 % účinnosti a průměr vyhlazené řady, jehož hodnota je 81.29 % účinnosti. Přibližně takovou asymptotickou střední hodnotu bude tato řada mít.

Velmi podobně pak vypadá vyhlazení časové řady 3a. Ustálené jsou také řady 4c a 1a. Z úsporných důvodů je zde nebudeme vykreslovat a výsledky pouze shrneme pomocí tabulky na konci této kapitoly. Vyhlazení všech řad pak je možné vyzkoušet v příloze diplomové práce *spousteni_funkci.R*. Pro ilustraci se ale podívejme ještě na jeden, trošku specifický, příklad ustálené časové řady. Konkrétně jde o řadu 4b.

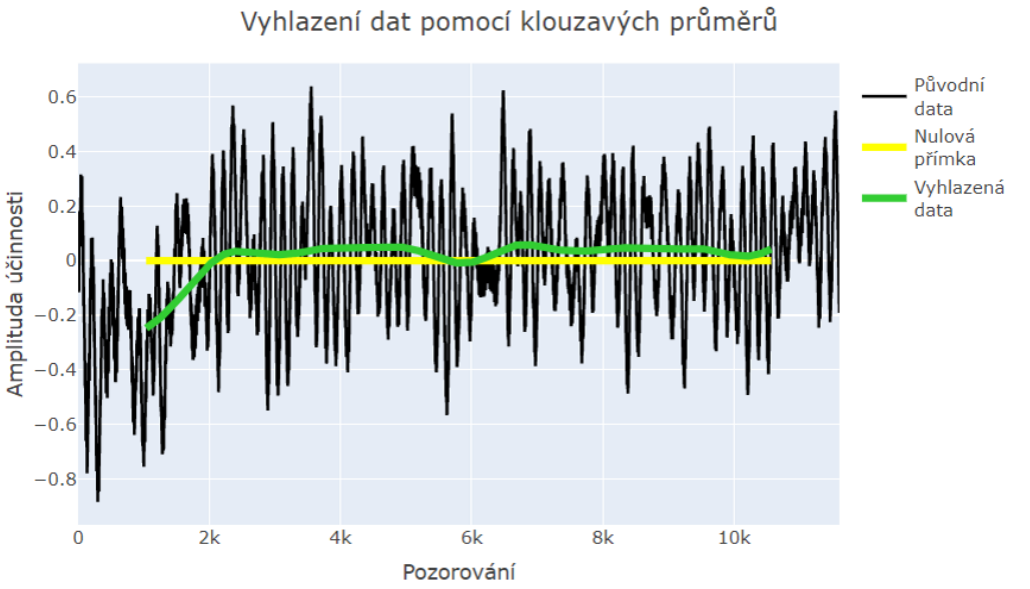
Řada 4b má na první pohled zřejmou velmi výraznou periodicitu. Ovšem vyhlazením pomocí frekvencí 13, 11, 20, 25, 16 a 18 jsme dosáhli vyhlazení, patrného z obrázku 13, takřka do podoby konstantní přímky. Přestože jsme ztratili vyhlazením poměrně velkou část dat, nebrání nám to v rozhodnutí o ustálenosti řady. Rozpětí dat se z 1.64 snížilo až na 0.10, což činí pouhých 6.37 % původního rozpětí. Vše tedy signalizuje ustálenost procesu. Asymptotická střední hodnota časové řady zde dosahuje asi 85.93 % účinnosti.



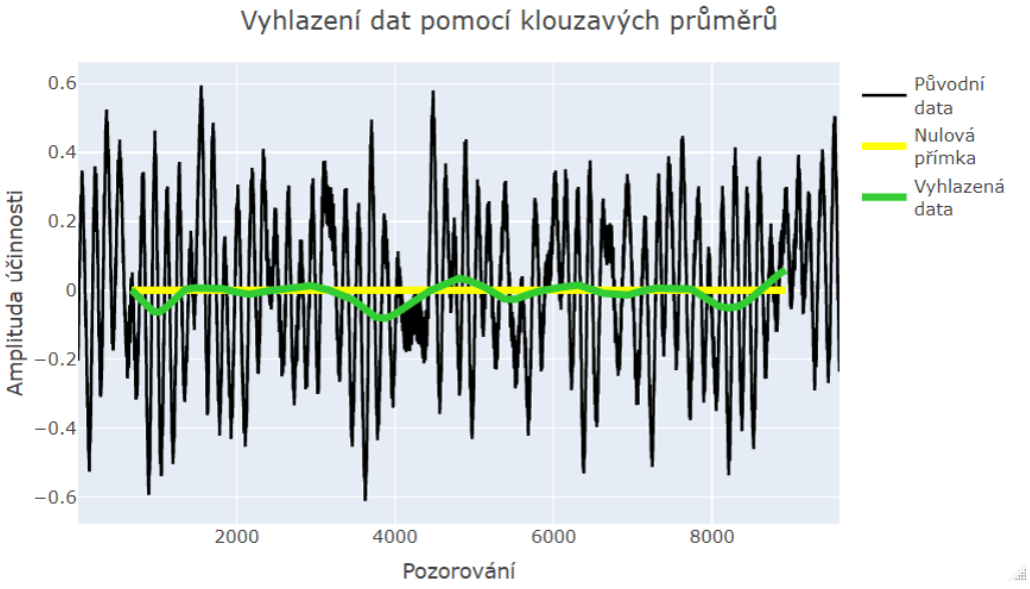
Obrázek 13: Vyhlazení dat4b pomocí klouzavých průměrů

2.3.2. Ustálené řady po vynechání prvních 2000 pozorování

Bohužel ne o všech řadách ale můžeme prohlásit, že jsou ustálené s konstantní střední hodnotou. U řad 1b, 2b a 2a jsme toho však schopni dosáhnout, pokud vynecháme část pozorování ze začátku řady. Konkrétně pro řady 1b a 2b stačí vynechat 2000 pozorování a řady budou ustálené. V případě dat2a 2000 pozorování nestačí a musíme jich vynechat 4000. Poté už je ovšem ustálenost zřejmá.



Obrázek 14: Vyhlazení dat1b pomocí klouzavých průměrů

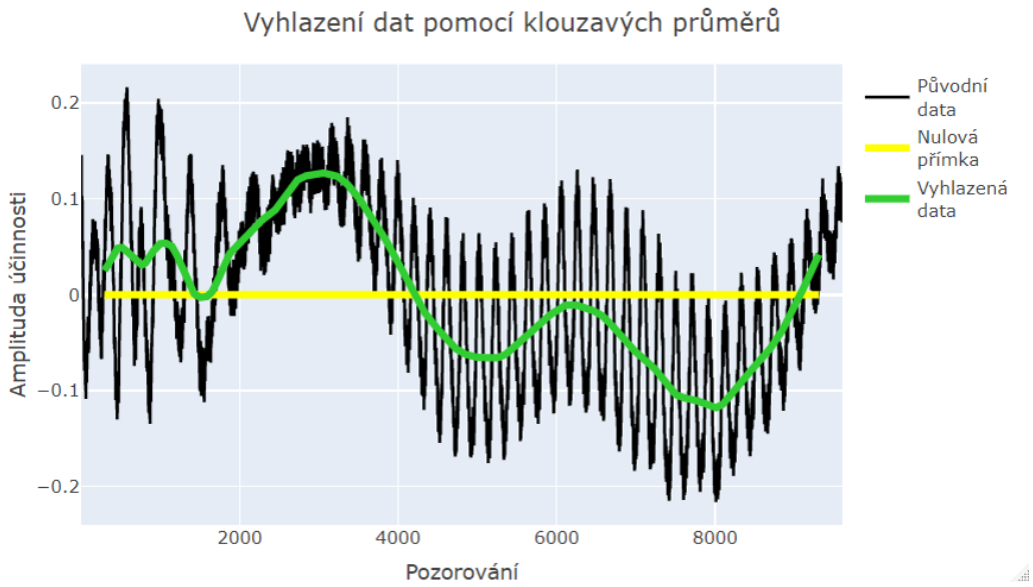


Obrázek 15: Vyhlazení dat1b bez prvních 2000 pozorování pomocí klouzavých průměrů

Obrázek 14 zobrazuje vyhlazení celé časové řady. Obrázek 15 pak ukazuje vyhlazení zkrácené řady. Vyhlazená zkrácená časová řada 1b má rozpětí dat pouhých 0.14, což činí jen 11.62 % z celkového rozpětí 1.20. Průměr hodnot vyhlazené řady, a tím i asymptotická střední hodnota, činí 86.13 %. Případ pro řady 2a a 2b bude shrnut v tabulce na konci kapitoly.

2.3.3. Neustálené časové řady

Některé časové řady jsou ale neustálené a nepomůže nám ani zkracování řady. V takovém případě je určování asymptotické střední hodnoty velmi obtížné až nemožné. Typickým příkladem neustálenosti je řada 1c. Vyhlazení této řady po vynechání prvních 2000 pozorování je následující:



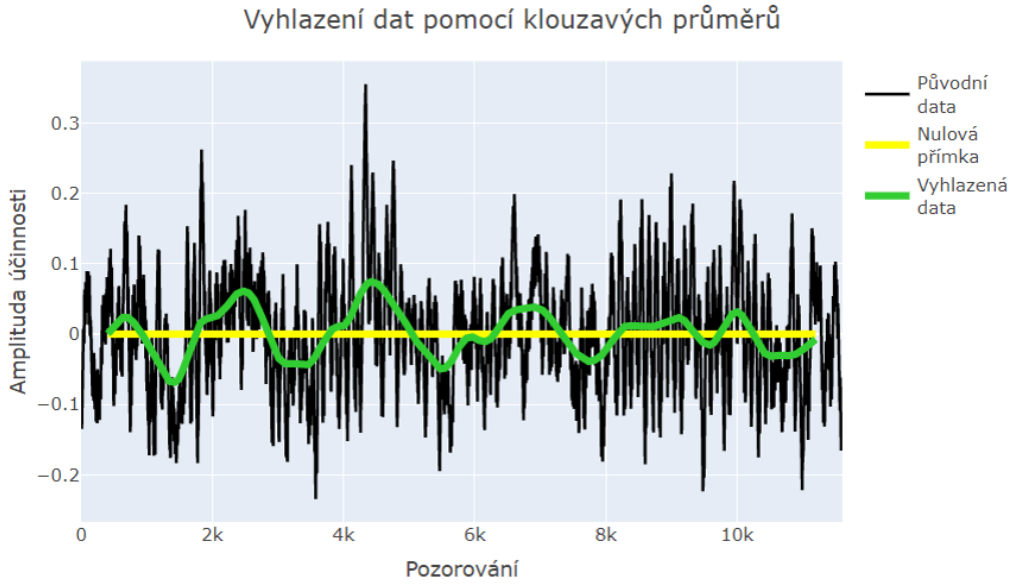
Obrázek 16: Vyhlazení dat1c bez prvních 2000 pozorování pomocí klouzavých průměrů

Z obrázku 16 je patrné, že v datech je dominantní frekvence 3. Navíc z tabulky 2 víme, že vůbec nejdůležitější je zde frekvence 1. To naznačuje neustálenost procesu a vyhlazení dat to jen potvrdilo. Byť je důležité pozna-

menat, že se pohybujeme v rámci malé amplitudy, tak i přesto musíme řadu považovat za neustálenou a nemůžeme její asymptotickou střední hodnotu určovat na základě průměru dat. Pro úplnost doplňme, že původní řada má rozpětí dat 0.43 a vyhlazená 0.25, což činí 56.71 % původního rozpětí. Tento fakt jen potvrzuje neustálenost řady. Mezi neustálené řady můžeme rovněž zahrnout řadu 3b.

2.3.4. Řady vyžadující individuální posouzení

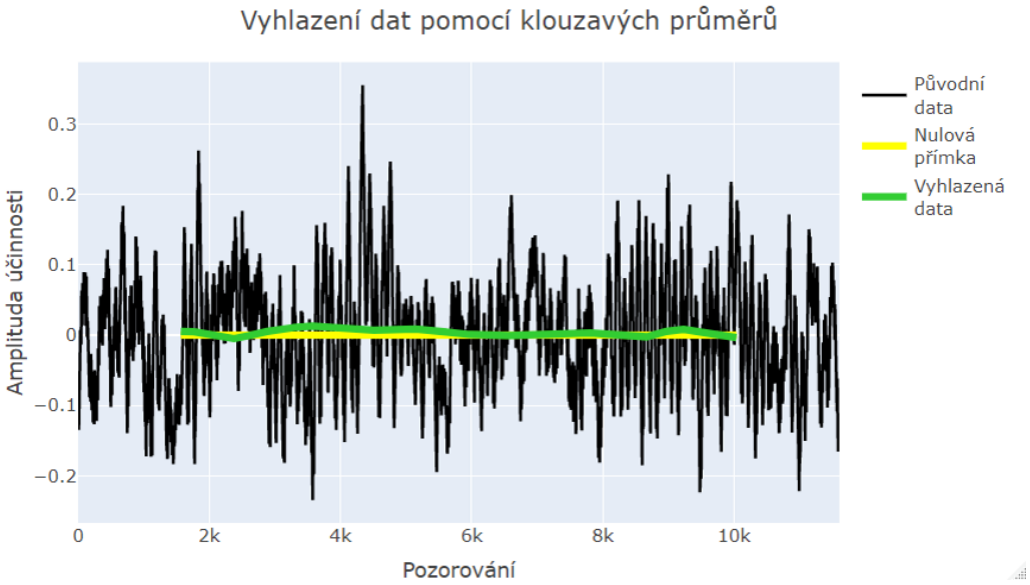
Dosud jsme nezmínili časové řady 3c a 2c. Posouzení ustálenosti je u nich komplikovanější a možná také subjektivnější než tomu je u předchozích řad. Proto je u nich vhodná expertní intervence a individuální posouzení. Podíváme-li se na vyhlazení dat3c, dostaneme následující obrázek:



Obrázek 17: Vyhlazení dat3c pomocí klouzavých průměrů

Amplituda se zde sice snížila, ale ne tak výrazně jako v kapitolách 2.3.1 a 2.3.2. Konkrétně se rozpětí dat snížilo z 0.59 na 0.14, což činí 24.33 % původního rozpětí. V datech je stále zřejmá periodicita, konkrétně frekvence

5 nebo 6. Pro vyhlazení dat na obrázku 17 byla použita klouzavá okna určená z frekvencí 107, 30, 645 a 41. Standardně nechceme z důvodu přílišného zkrácení délky vyhlazené řady používat frekvence ≤ 10, ale pokud bychom zde výjimečně povolili k vyhlazení také frekvenci 6, získali bychom krásně vyhlazenou řadu takřka až na konstantní přímku. Tento stav ilustruje obrázek 18.

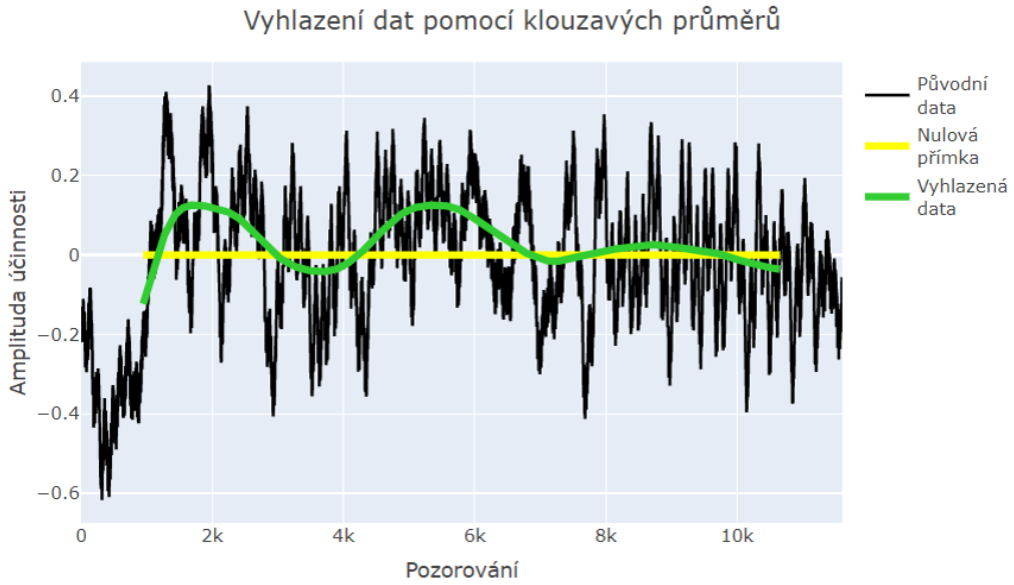


Obrázek 18: Vyhlazení dat3c pomocí klouzavých průměrů včetně využité frekvence 6

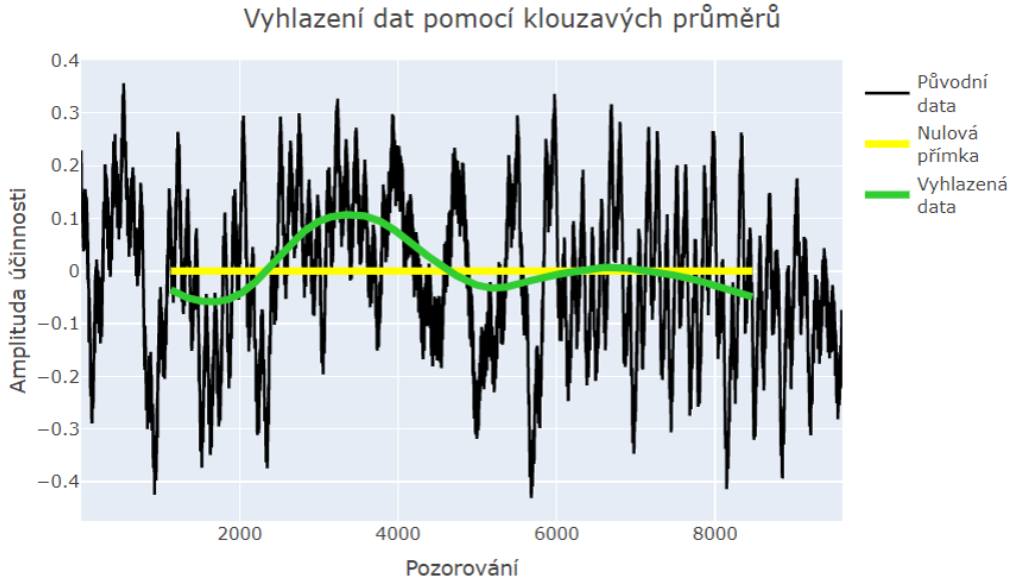
Vyhlazení je zde opravdu takřka dokonalé, neboť z rozpětí původní řady o hodnotě 0.59 jsme získali na vyhlazené řadě naprosto minimální rozpětí 0.02, což je jen 3.11 % z původního rozpětí. Přihlédneme-li k tomu, že viditelná periodicita na obrázku 17 je poměrně pravidelná, můžeme si dovolit o této řadě prohlásit, že je ustálená a její asymptotickou střední hodnotu můžeme odhadovat pomocí průměru vyhlazené řady, což činí 88.32 % účinnosti.

Druhou problematickou časovou řadu, tedy dat2c, můžeme vyhladit a vykreslit pomocí obrázku 19. Z něj je patrná významnost frekvence 3, což

ostatně plyne i z tabulky 3.



Obrázek 19: Vyhlazení dat2c pomocí klouzavých průměrů



Obrázek 20: Vyhlazení dat2c bez prvních 2000 pozorování pomocí klouzavých průměrů

Zkrácení řady o prvních 2000 pozorování nám pomůže jen částečně. Z obrázku 20 můžeme vyčíst, že v signálu stále hraje roli frekvence 3. Víme již, že nízké frekvence signalizují neustálenost řady. Ovšem z vyhlazení se zdá, že se řada postupně ustaluje, neboť perioda přibližně mezi pozorováními 7000 a 10000 v obrázku 19, respektive mezi 5000 a 8000 v obrázku 20, má jasně nejmenší amplitudu.

Graficky bychom tedy mohli expertně usoudit, že se řada ustaluje. Podíváme-li se na rozpětí (zaměříme se na zkrácenou řadu) původních dat, to činí 0.79. Pro vyhlazená data je to 0.17, tedy 21.07 %. Což je o něco více, než v případě řad, o kterých jsme prohlásili, že jsou jednoznačně ustálené, ale výrazně méně než u neustálených. Přihlédneme-li také k tomu, že se v rámci této řady pohybujeme celkově na malé amplitudě, můžeme i o této řadě prohlásit, že je ustálená. Predikce asymptotické střední hodnoty by i v tomto případě mohla být poměrně přesná pomocí aritmetického průměru vyhlazené řady, což činí 83.81 %.

2.3.5. Shrnutí ustálenosti řad

V předchozích částech kapitoly jsme ukázali vyhlazení pro vybrané datové soubory. Vyhlazení těch zbylých je možné vyzkoušet si v příloze pojmenované *spousteni_funkci.R*. O všech řadách, včetně těch jejichž vyhlazení zde nebylo vykresleno, jsme prohlásili, zda je považujeme za ustálené či nikoliv. Konkrétní výsledky vyhlazování zachycuje tabulka 6.

V tabulce jsou zachyceny jednotlivé datové soubory, frekvence, které byly využity pro výpočet délky klouzavého okna, rozpětí původního datového souboru, rozpětí vyhlazeného datového souboru po aplikaci klouzavých průměrů se zvolenými frekvencemi, kolik procent činí rozpětí vyhlazených dat z celkového rozpětí původních dat, dále hodnocení ustálenosti a průměrné hodnoty

vypočítané z vyhlazených dat. Zkratky pro použitá data vycházejí z názvů datových souborů. Zdvojením písmene je označeno zkrácení datového souboru o prvních 2000 pozorování, ztrojení pak označuje zkrácení dat o 4000 pozorování. Označení 2a tedy znamená druhý datový soubor při nejslabším průtoku a použití všech 11605 pozorování, 2aa značí ten stejný datový soubor jen zkrácený o 2000 pozorování zepředu, 2aaa pak označuje zkrácení téhož souboru o 4000 prvních pozorování.

Použitá data	Použité frekvence	Původní rozpětí	Vyhlazené rozpětí	Procento rozpětí	Hodnocení ustálenosti	Průměr vyhlazení
1a	412, 57, 12	6.79	0.94	13.91	ustálená	77.58
1b	76, 16, 13, 63	1.52	0.31	20.15	nejisté	86.11
1bb	63, 56, 13, 43	1.20	0.14	11.62	ustálená	86.13
1c	57, 11	0.75	0.24	31.64	neustálená	87.25
1cc	47, 26, 1601	0.43	0.25	56.71	neustálená	87.23
2a	20, 17, 289	2.05	0.67	32.95	neustálená	85.75
2aa	16, 14, 18, 235	2.05	0.60	29.17	neustálená	85.84
2aaa	13, 11, 186	1.85	0.31	16.93	ustálená	85.88
2b	93, 24	1.45	0.81	55.92	neustálená	87.38
2bb	77, 20, 18, 15	1.09	0.15	13.57	ustálená	87.43
2c	17, 100, 19, 29	1.04	0.25	23.88	nejisté	83.82
2cc	14, 81, 24, 16, 28	0.79	0.17	21.07	nejisté	83.81
3a	69, 413, 82, 599	7.76	1.10	14.23	ustálená	79.44
3b	61, 44, 72, 49	3.79	1.24	32.66	neustálená	86.85
3bb	36, 48, 41, 57	3.79	1.22	32.25	neustálená	86.83
3c	107, 30, 645, 41	0.59	0.14	24.33	nejisté	88.32
3c	107, 30, 645, 41, 6	0.59	0.02	3.12	ustálená	88.32
4a	308, 53, 381, 931, 1238, 76	3.00	0.33	11.15	ustálená	81.29
4b	13,11,20,25,16,18	1.64	0.10	6.37	ustálená	85.93
4c	13, 16, 54, 108	2.38	0.34	14.28	ustálená	83.48

Tabulka 6: Vyhlazení dat, jeho kvalita a hodnocení ustálenosti procesů

Máme-li shrnout ustálenost na základě tabulky 6, můžeme prohlásit, že řady 1a, 3a, 4a, 4b a 4c jsou ustálené i při použití všech datových bodů. Ve všech těchto případech shodně dosahuje vyhlazení takové kvality, že rozpětí vyhlazených dat činí méně než 15 % rozpětí původní datové sady. V případě řad 1b a 2b nám k ustálenosti pomůže vynechání prvních 2000 pozorování, pro řadu 2a je k ustálenosti nutno vynechat 4000 pozorování. Řady 1c a 3b jsou neustálené i při zkracování řady, procento rozpětí vyhlazených dat z rozpětí původních dat zde činí vždy více než 30 %. Posouzení ustálenosti řad 2c a 3c vyžaduje speciální péči a ideálně expertní posudek. Procento rozpětí v obou případech činí něco přes 20 %. Po důkladnější analýze můžeme prohlásit, že řady 2c i 3c jsou ustálené.

Pokud bychom chtěli stanovit nějaké objektivní kritérium, na základě kterého bychom mohli co nejvíce automatizovat hodnocení ustálenosti, nabízí se kritérium procenta rozpětí, tedy kolik procent z původního rozpětí dat činí rozpětí vyhlazené časové řady. Na základě dat, která máme k dispozici, by takové kritérium mohlo být například takové, že pokud procento rozpětí bude nižší než 15 %, prohlásíme řadu automaticky za ustálenou. Pokud by toto procento bylo větší než 30 %, označíme řadu za neustálenou. Bude-li se procento pohybovat v intervalu od 15 % do 30 %, můžeme řadu předat expertovi k důkladnější analýze a posouzení. Důležité je testovat ustálenost i na zepředu zkrácených časových řadách, neboť první pozorování v některých případech mohou ustálenost ovlivnit.

Kritéria nastavená procenty v předchozím odstavci vycházejí z dat, která jsme měli pro analýzu k dispozici. Správnost tohoto přístupu a nastavení by bylo třeba otestovat na více datových souborech. Samozřejmě to není jediný možný přístup jak hodnotit ustálenost řady. Můžeme se také rozhodovat čistě na základě nejdůležitějších frekvencí získaných pomocí FFT či například

dle hodnot amplitudy původních dat. Přístup popsaný výše však z těchto možností nabízí asi nejobjektivnější posouzení. Vždy však můžeme expertně přihlédnout i k jiným okolnostem.

V případě, kdy označíme řadu za ustálenou, můžeme její asymptotickou střední hodnotu odhadovat pomocí průměru vyhlazené časové řady. Tento průměr dává velmi podobné (nikoliv stejné) výsledky jako průměr původní nevyhlazené časové řady. Vždy se liší maximálně o setiny procenta účinnosti. Odhad asymptotické střední hodnoty bychom pochopitelně mohli dělat i průměrem celého datového souboru, ovšem vyhlazená data odfiltrovala šum a jejich průměr by tak mohl dávat o něco přesnější výsledky.

2.4. Využití modifikovaného exponenciálního trendu

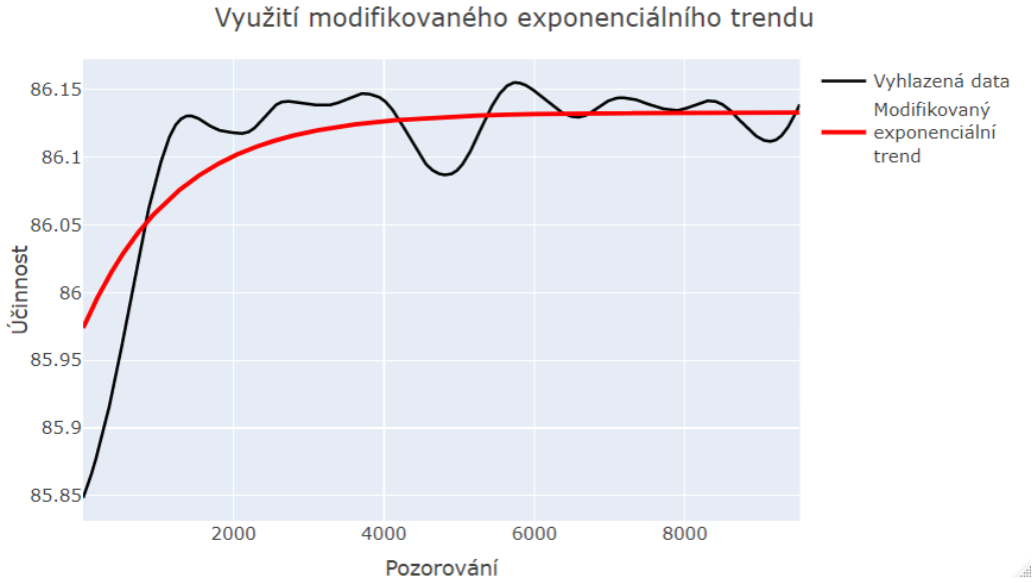
Jednou z dalších metod nabízejících se k hodnocení ustálenosti procesu a odhadu asymptotické střední hodnoty je využití modifikovaného exponenciálního trendu, který byl teoreticky popsán v kapitole 1.2.1. Pro jednotlivá data budeme využívat k odhadu parametrů metodu částečných součtů.

Je důležité říci, že modifikovaný exponenciální trend rozhodně nebude vhodný pro všechny časové řady, které máme k dispozici. Abychom mohli metodu použít, nesmí v řadě platit ani jedna ze situací $S_1 > S_2 < S_3$, nebo $S_1 < S_2 > S_3$. Tedy součet S_2 nesmí být ani největší, ani nejmenší ze součtů S_1, S_2, S_3, jinak bychom nemohli provést odhad $\hat{\beta}$ dle vzorce (1.16).

Za účelem co nejjednoduššího použití modifikovaného exponenciálního trendu vznikla pro účely diplomové práce funkce `mod_exp_trend()`. Jejím úkolem je odhadnout parametry modifikovaného exponenciálního trendu pomocí metody částečných součtů, vykreslit křivku trendu s těmito parametry a samotné parametry vypsat. V případě, kdy by součet S_2 byl největší, nebo nejmenší z částečných součtů, funkce sice provede veškeré úkony, ale vypíše

hlášku varující o tomto faktu, a že v tomto případě použití modifikovaného exponenciálního trendu zřejmě není vhodné. Odhad $\hat{\beta}$ dle vzorce (1.16) by byl proveden vzorcem v absolutní hodnotě. Funkce je součástí přiloženého kódu *funkce.R*.

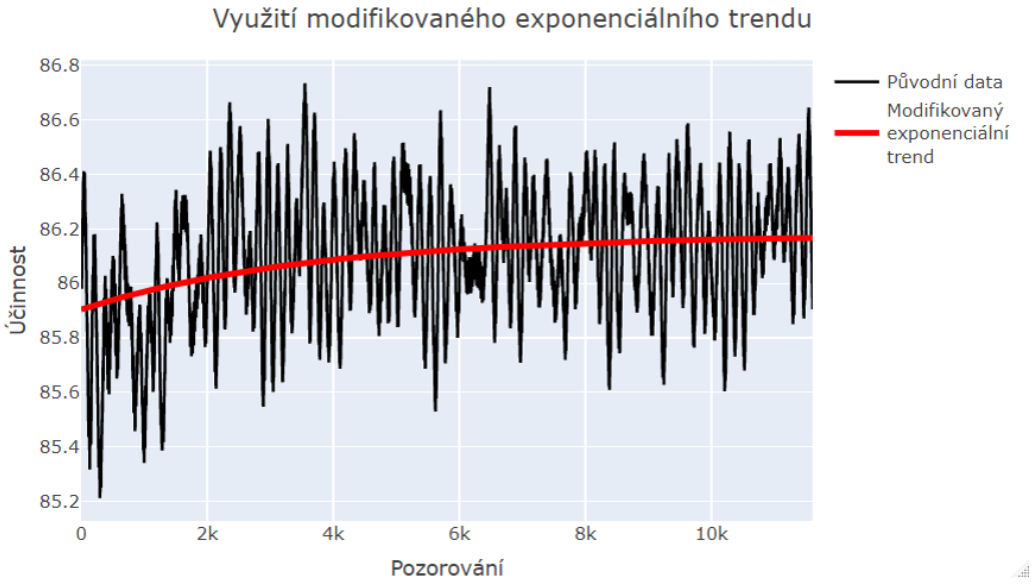
Pro velkou část námi analyzovaných časových řad bohužel není využití modifikovaného exponenciálního trendu vhodným nástrojem z výše zmíněného důvodu extrémnosti S_2 v rámci částečných součtů. Takovými časovými řadami jsou řady 1a, 2a, 2c, 3a, 3b, 3c a 4c a nebudeme se jimi tedy zabývat. V ostatních časových řadách má smysl podívat se na výsledky modifikovaného exponenciálního trendu. Začněme řadou 1b, která je dobrým příkladem využití modifikovaného exponenciálního trendu. Vyzkoušet aplikaci tohoto trendu můžeme postupně na vyhlazená i na nevyhlazená původní data. Nejprve se podívejme na případ pro vyhlazená data:



Obrázek 21: Modifikovaný exponenciální trend pro vyhlazenou řadu 1b

Z obrázku 21 je patrných několik věcí. Modifikovaný exponenciální trend správně vyhodnotil, že konec řady je ustálenější než její začátek a zvolil tvar

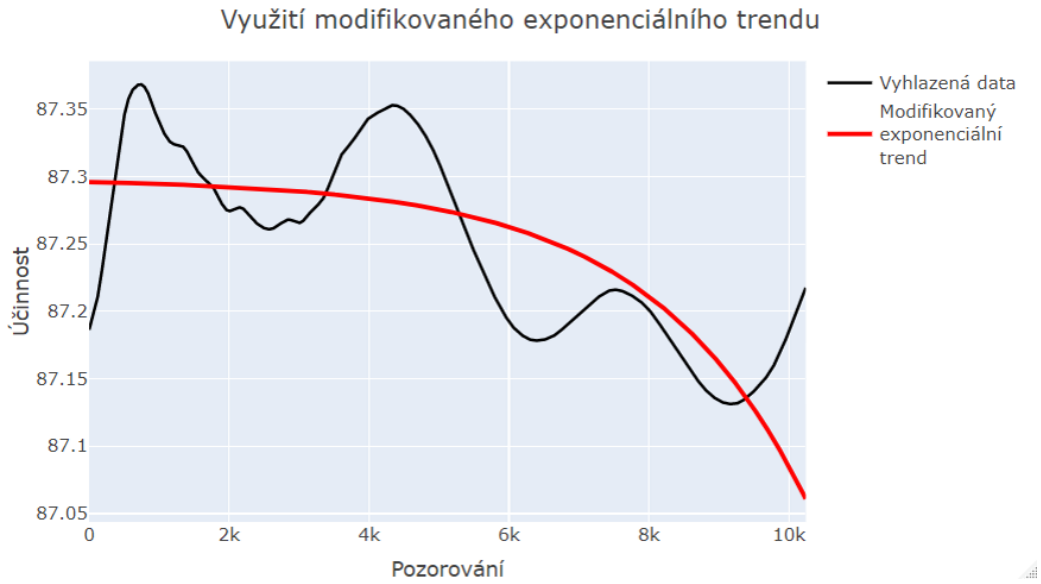
křivky, který poměrně dobře kopíruje data. Červená trendová křivka se zde ustaluje, a to konkrétně na hodnotě 86.13 %. Tuto hodnotu vyčteme z parametrů modifikovaného exponenciálního trendu odhadnutých pro tato data pomocí metody částečných součtů. Konkrétně nás zde zajímá primárně parametr γ, jehož interpretací je asymptota. Jeho hodnota je v tomto případě, jak již bylo zmíněno, 86.13. Pomocí této hodnoty můžeme odhadovat asymptotickou střední hodnotu. Nahlédneme-li do tabulky 6, zjistíme, že se tato hodnota plně shoduje s průměrem vyhlazení této řady po zkrácení o prvních 2000 pozorování. Tento odhad asymptotické střední hodnoty by tedy mohl být velmi přesný. Můžeme se podívat také na použití modifikovaného exponenciálního trendu přímo na původní nevyhlazená data:



Obrázek 22: Modifikovaný exponenciální trend pro původní řadu 1b

I zde se zdá, že trendová křivka dobře zachytila chování časové řady. Odhad parametru γ pro tuto křivku je $\hat{\gamma} = 86.18$. Opět se tedy pohybujeme ve velmi podobných hodnotách a jen jsme se ujistili, že řada je opravdu ustálená a odhad asymptotické střední hodnoty byl zřejmě velmi přesný.

Podíváme-li se na opačný extrémní případ, a to na řadu 1c, o které jsme prohlásili, že je neustálená, zjistíme, že nám modifikovaný exponenciální trend potvrdí naši domněnku o neustálenosti. Obrázek 23 zachycuje aplikaci tohoto trendu na vyhlazenou řadu 1c. Tvar trendové křivky jasně naznačuje neustálenost řady. Pro nevyhlazenou řadu zde nelze doporučit tento typ trendu použít, neboť S_2 je největší z částečných součtů.

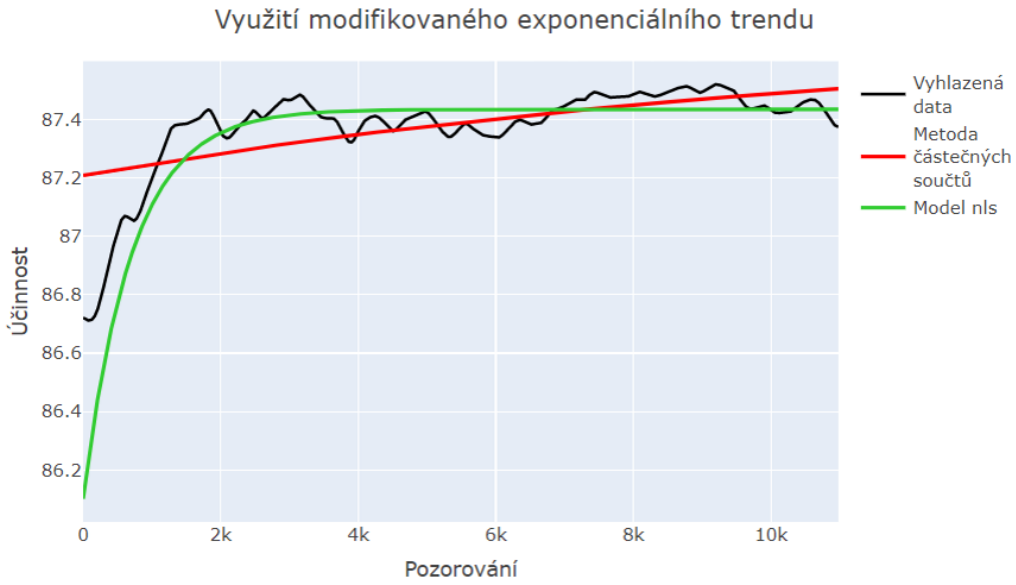


Obrázek 23: Modifikovaný exponenciální trend pro vyhlazenou řadu 1c

Některé časové řady jsou k proložení dat modifikovaným exponenciálním trendem vhodné, ovšem občas může být pro metodu částečných součtů složité odhadnout parametry přesně, obzvláště pak při malé amplitudě dat. Alternativně lze pro odhad parametrů a následné vykreslení trendu využít nelineární regresní model, respektive nelineární metodu nejmenších čtverců. V programu *R* je tato metoda prováděna v rámci funkce `nls()` (nonlinear least squares). Vstupem pro tuto funkci jsou data, pro která chceme parametry odhadovat, předpis trendu (v našem případě tedy předpis (1.10)) a vhodné počáteční odhady parametrů. Jako počáteční odhady využijeme

odhady získané metodou částečných součtů a pomocí funkce `nls()` tyto odhady zkusíme upřesnit. Pro snadnější použití metody vznikla pro účely této diplomové práce funkce `mod_exp_trend2()`, kterou čtenář může najít v příloze *funkce.R*.

Dobrou ukázkou toho, kdy nám nelineární regrese pomůže odhadnout parametry lépe, je řada 2b. Metoda částečných součtů v těchto datech sice dokázala identifikovat správně, že se řada ustaluje, ovšem srovnáme-li v obrázku 24 zelenou křivku vypočtenou pomocí nelineární regrese a červenou křivku vypočtenou metodou částečných součtů, musíme prohlásit, že nelineární regrese dává podle všeho přesnější výsledky.

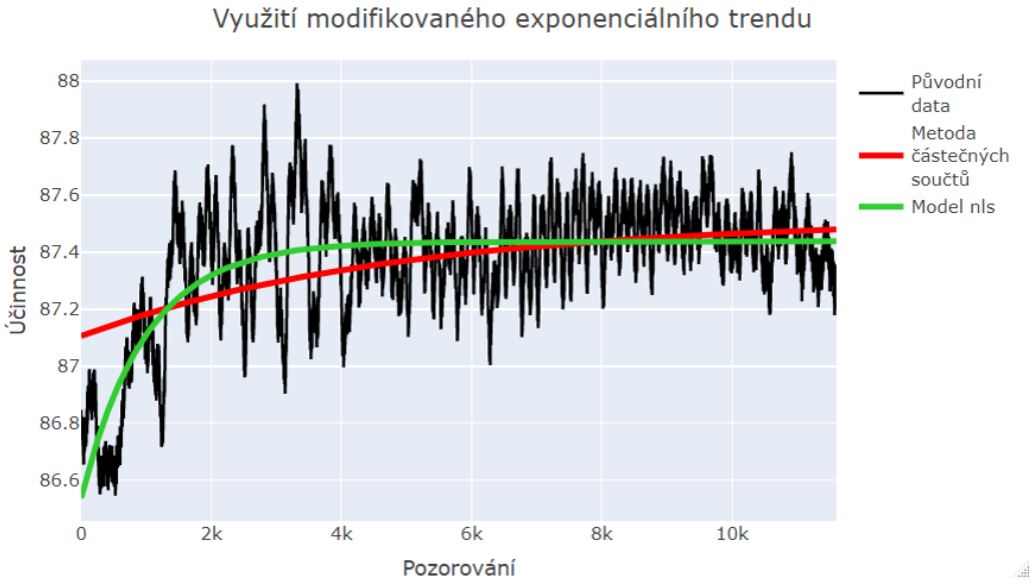


Obrázek 24: Modifikovaný exponenciální trend pro vyhlazenou řadu 2b

Odhad asymptoty pomocí metody částečných součtů zde je $\hat{\gamma} = 87.71$. Pomocí funkce `nls()` pak odhad činí $\hat{\gamma} = 87.43$. Z tabulky 6 vidíme, že se tento výsledek plně shoduje s odhadem učiněným na základě aritmetického průměru.

Pro úplnost se můžeme podívat i na případ s původními daty, který je ilu-

strován obrázkem 25. I zde funguje metoda nelineárních nejmenších čtverců lépe a přesněji kopíruje data. Odhady zde činí $\hat{\gamma} = 87.52$ pro částečné součty a $\hat{\gamma} = 87.44$ pro nelineární regresi. Potvrzuje se nám tedy, že odhad asymptotické střední hodnoty pro tato data bude velmi přesný.



Obrázek 25: Modifikovaný exponenciální trend pro původní řadu 2b

Modifikovaný exponenciální trend můžeme aplikovat také na řady 4a a 4b (byť u 4b pouze na původní řadu, na vyhlazenou nikoliv). Vykreslovat je zde již nebudeme, ovšem čtenář má možnost si vykreslení sám zkusit v příloze *spousteni_funkci.R*. Asymptoty ustálených řad, pro které má využití modifikovaného exponenciálního trendu opodstatnění, zahrneme alespoň do tabulky 7. V tabulce jsou uvedena použitá data a ke každým z nich pak odhad parametru γ metodou částečných součtů i pomocí funkce `nls()`. V posledním sloupci jsou pak uvedeny průměry příslušných řad, kterými jsme v minulé kapitole odhadovali asymptotickou střední hodnotu. V případě řad 1b a 2b jsou to průměry bez prvních 2000 pozorování, pro řady 4a a 4b pak průměry celých řad.

Použitá data	Odhad γ částečnými součty	Odhad γ pomocí nls	Odhad asymptoty průměrem
Vyhlazené 1b	86.13	86.13	86.13
Původní 1b	86.18	86.15	86.14
Vyhlazené 2b	87.71	87.43	87.43
Původní 2b	87.52	87.44	87.42
Vyhlazené 4a	81.25	81.24	81.29
Původní 4a	81.26	81.25	81.30
Původní 4b	85.90	85.92	85.94

Tabulka 7: Asymptoty vybraných časových řad s využitím modifikovaného exponenciálního trendu

Z tabulky je zřejmé, že ve chvíli, kdy se dá rozumně využít modifikovaný exponenciální trend, tak výsledná asymptota, určená na základě metody částečných součtů i pomocí funkce `nls()`, nabývá velmi podobných hodnot jako prostý aritmetický průměr ustálených řad. Funkce `nls()` pak dává většinou o něco přesnější výsledky.

U časových řad 1b, 2b, 4a a 4b se nám tedy s využitím modifikovaného exponenciálního trendu podařilo potvrdit výsledky analýzy z kapitoly 2.3.5. Zmíněné časové řady jsou opravdu ustálené a jejich asymptotickou střední hodnotu můžeme odhadovat hodnotami z tabulky 7.

Závěr

V diplomové práci jsme se zaměřili na analýzu časových řad, ve kterých hraje významnou roli periodická složka. Po objasnění teoretického pozadí problematiky jsme se pustili do stěžejní části práce, kterou byla praktická analýza simulovaných dat ze společnosti Sigma. Analyzovali jsme všech dvanáct časových řad, zachycujících účinnosti čerpadel, které jsme měli k dispozici. K analýze byl využit statistický software *R*.

Prvním krokem, který bylo třeba učinit, byla vhodná vizualizace dat. K tomu jsme využili v softwaru *R* knihovnu *plotly*, která umožňuje vykreslení interaktivních grafů. Díky tomu jsme mohli data libovolně přibližovat a zachytit vizuálně periodicitu, která by při jiném vykreslení nebyla zřejmá. Poté, co jsme získali základní představu o datech, jsme se pustili do detekce nejvýznamnějších frekvencí a period pomocí metody rychlé Fourierovy transformace. Zjistili jsme, že některé časové řady mají významné dlouhé periody, což naznačuje neustálenost procesu.

Délky nejvýznamnějších period jsme následně využili jako délky klouzavých oken k vyhlazení dat pomocí metody klouzavých průměrů. Za účelem co nejjednoduššího a co nejvíce automatizovaného provedení vyhlazení vzniklo pro účely diplomové práce několik užitečných funkcí, které jsou součástí přílohy nazvané *funkce.R* a jsou jedním z přínosů diplomové práce.

K rozhodnutí o ustálenosti řad jsme vedle vizuálního posouzení vyhlazení využili zejména objektivní kritérium kvality vyhlazení, a to snížení rozpětí dat ve vyhlazené řadě vzhledem k rozpětí dat původní řady. Na základě kvality vyhlazení dat jsme mohli pozorovat různé typy výsledků. Pět časových řad jsme označili za jednoznačně ustálené, neboť kvalita jejich vyhlazení byla vysoká a rozpětí vyhlazených dat dosahovalo jen nízkého procenta původního rozpětí dat. Tři řady jevily známky ustalování, ovšem ke kompletnímu pro-

hlášení řady za ustálenou bylo nutné odstranit část dat ze začátku řady, která vlivem simulací ustálená nebyla. Další dvě řady jsme označili za neustálené, protože vyhlazením se nepovedlo dostatečně snížit rozpětí dat a navíc v nich hrály významnou roli dlouhé periody, které signalizují neustálenost. Vyhlazení zbývajících dvou řad dávalo nejednoznačné výsledky a za ustálené jsme je prohlásili až po zvážení konkrétního bližšího kontextu.

U ustálených řad jsme mohli odhadovat asymptotickou střední hodnotu. Odhad jsme provedli prostým aritmetickým průměrem vyhlazených dat. Na některé řady jsme poté aplikovali modifikovaný exponenciální trend a z odhadu parametrů jsme vyčetli asymptotu. Ta se velmi podobala odhadům zjištěným pomocí aritmetického průměru a potvrdili jsme tedy další metodou, že takový odhad asymptotické střední hodnoty by mohl být velmi přesný. Zhodnocením ustálenosti časových řad a odhadem asymptotických středních hodnot jsme tedy splnili stanovené cíle diplomové práce. Data jsou vhodná pro další analýzu a mohlo by být například užitečné vyzkoušet pomocí různých metod predikce střední hodnoty na postupně zkracovaných datech.

Celková analýza byla složitější, než jsem si před začátkem práce dovedl představit, nicméně byla značně obohacující. Přinesla mi, krom kýženého výsledku v podobě úspěšně provedené analýzy dat, také řadu cenných zkušeností z praktické práce s daty a programování v softwaru *R*. Jsem vděčný, že jsem si datovou analýzu tohoto typu mohl vyzkoušet a věřím, že nabyté zkušenosti využiji také v profesním životě.

Literatura

[1] HINDLS, Richard; HRONOVÁ, Stanislava; SEGER, Jan a FISCHER, Jakub. *Statistika pro ekonomy.* 7. vyd. Praha: Professional Publishing, 2006. ISBN 80-86946-16-9.

[2] HYNDMAN, Rob J. a ATHANASOPOULOS, George. *Forecasting: principles and practice.* Online. 3rd ed. Melbourne, Australia.: OTexts, 2021. Dostupné z: https://OTexts.com/fpp3. [cit. 2024-03-25].

[3] CIPRA, Tomáš. *Analýza časových řad s aplikacemi v ekonomii.* Praha: Státní nakladatelství technické literatury, 1986.

[4] SEGER, Jan a HINDLS, Richard. *Statistické metody v tržním hospodářství.* Praha: Victoria Publishing, 1995. ISBN 80-7187-058-7.

[5] CRYER, Jonathan D. a CHAN, Kung-Sik. *Time series analysis: with applications in R.* 2nd ed. Springer texts in statistics. New York, N.Y.: Springer, 2008. ISBN 978-0-387-75958-6.

[6] BRUNTON, Steven L. a KUTZ, Jose Nathan. *Data-driven science and engineering: machine learning, dynamical systems, and control.* Cambridge: Cambridge University Press, 2019. ISBN 978-1-108-42209-3.

[7] SHUMWAY, Robert H. a STOFFER, David S. *Time series analysis and its applications: with R examples.* 4th ed. Springer texts in statistics. Cham: Springer, 2017. ISBN 978-3-319-52451-1.

[8] CHATFIELD, Chris a XING, Haipeng. *The analysis of time series: An introduction with R.* 7th ed. Chapman and Hall/CRC, 2019. ISBN 978-1-138-06613-7.

[9] BRACEWELL, Ronald N. *The Fourier transform and its applications.* 3rd ed. Boston: McGraw-Hill, 2000. ISBN 978-0073039381.

[10] PRESS William H., TEUKOLSKY Saul A., VETTERLING William T. a FLANNERY Brian P. *Numerical Recipes in C: The Art of Scientific Computing.* 2nd ed. Cambridge: Cambridge University Press, 1992. ISBN: 978-0521431088.

[11] R Core Team. *R: A Language and Environment for Statistical Computing.* Online. Vienna, Austria: R Foundation for Statistical Computing, 2022 Dostupné z: https://www.R-project.org/. [cit. 2024-03-26].

[12] HYNDMAN, Rob J. a KHANDAKAR, Yeasmin. *Automatic Time Series Forecasting: The forecast Package for R.* Journal of Statistical Software. Online. Vol. 27, iss. 3, July 2008. Dostupné z: 10.18637/jss.v027.i03. [cit. 2024-03-26].

[13] RITZ, Christian a STREIBIG, Jens Carl. *Nonlinear Regression with R.* Springer New York, NY, 2008. ISBN 978-0-387-09615-5.

[14] ZEILEIS, Achim a GROTHENDIECK, Gabor. *zoo: S3 Infrastructure for Regular and Irregular Time Series.* Journal of Statistical Software. Online. Vol.14, iss.6, May 2005. Dostupné z: 10.18637/jss.v014.i06. [cit. 2024-04-03].

Seznam kódů v příloze

- *nacteni.R* - načtení a úprava datových sad
- *vizualizace.R* - vykreslení všech datových sad
- *funkce.R* - všechny naprogramované funkce využité při analýze
- *spousteni_funkci.R* - ukázky spouštění všech funkcí z přílohy *funkce.R* s různými vstupními parametry