Ekonomická fakulta
Faculty of Economics

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých Budějovicích
Ekonomická fakulta
Katedra aplikované matematiky a informatiky

Bakalářská práce

Využití neparametrických regresních metod při modelování ekonomických časových řad

Vypracoval: Pavel Mareš
Vedoucí práce: Mgr. Irena Štrausová, Ph.D.

České Budějovice 2025

JIHOČESKÁ UNIVERZITA V ČESKÝCH BUDĚJOVICÍCH

Ekonomická fakulta

Akademický rok: 2024/2025

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Pavel MAREŠ**
Osobní číslo:	**E22028**
Studijní program:	**B0688A140010 Podniková informatika**
Téma práce:	**Využití neparametrických regresních metod při modelování ekonomických časových řad**
Zadávající katedra:	**Katedra aplikované matematiky a informatiky**

Zásady pro vypracování

Zásady pro vypracování

Bakalářská práce bude obsahovat v teoretické části přehled různých regresních přístupů vhodných pro analýzu ekonomických časových řad (regrese, spliny, neparametrické regresní přístupy jako je např. lokální regrese, případně jiné, zde neuvedené). V praktické části posluchač předvede aplikaci vybraných metod na reálných ekonomických datech. Pro tento účel student využije některý z programovacích jazyků R, nebo Python a provede jejich srovnání z hlediska predikční účinnosti.

Metodický postup:

1) Studium základů použití nástrojů pro analýzu ekonomických časových řad.
2) Studium různých regresních přístupů.
3) K modelování bude využit některý z jazyků R nebo Python
4) Na vybrané problematice ukáže aplikace výše zmíněných metod
5) Srovnání výsledků v ekonomických souvislostech. Provede diskusi a formuluje závěry.

Rozsah pracovní zprávy:	**40 – 50 stran**
Rozsah grafických prací:	**dle potřeby**
Forma zpracování bakalářské práce:	**tištěná**

Seznam doporučené literatury:

FORBELSKÁ, Marie. Stochastické modelování jednorozměrných časových řad. Brno: Masarykova univerzita, 2009.

ARLT, J. a kol. Analýza ekonomických časových řad s příklady. Praha: VŠE, 2002.

ARLT, J. a M. ARLTOVÁ. Ekonomické časové řady. Praha: Professional Publishing, 2009. ISBN 978-80-86929-43-9.

Gujarati, D.N. and Porter, D.C. Basic Econometrics. 5th Edition, McGraw Hill Inc., New York, 2009.

CIPRA, Tomáš. Finanční ekonometrie. 1. vyd. Praha: Ekopress, 2008, 538 s. ISBN 978-80-86929-43-9.
BROCKWELL, P. J. and R. A. DAVIS. Time Series: Theory and Methods. Springer Series in Statistics. New York: Springer, 1991.

KŘIVÝ, I. Analýza časových řad. Ostrava: PřF OU, 2006.

KOZÁK, J., R. HINDLS a J. ARLT. Úvod do analýzy ekonomických časových řad. Praha: VŠE, 1994.

SHUMWAY, R. H. and D. S. STOFFER. Time Series Analysis and Its Applications. New York: Springer, 2000. ISBN 0-387-98950-1.

Hamilton, J.D. Time Series Analysis. Princeton University Press, Princeton, 1994.

MONTGOMERY, Douglas C., Cheryl L. JENNINGS a Murat KULAHCI. Introduction to Time Series Analysis and Forecasting. Hoboken, N.J.: Wiley-Interscience, 2008, xi, 445 s. ISBN 04-716-5397-7.

Vedoucí bakalářské práce:	**Mgr. Irena Štrausová, Ph.D.** Katedra aplikované matematiky a informatiky
Konzultant bakalářské práce:	**doc. Ing. Michael Rost, Ph.D.** Katedra genetiky a biotechnologií
Datum zadání bakalářské práce:	**30. ledna 2025**
Termín odevzdání bakalářské práce:	**10. dubna 2025**

JIHOČESKÁ UNIVERZITA
V ČESKÝCH BUDĚJOVICÍCH
EKONOMICKÁ FAKULTA
Studentská 13
370 05 České Budějovice

doc. RNDr. Zuzana Dvořáková Líšková, Ph.D.
děkanka

Mgr. Irena Štrausová, Ph.D.
vedoucí katedry

V Českých Budějovicích dne 3. února 2025

Prohlášení

Prohlašuji, že jsem autorem této kvalifikační práce a že jsem ji vypracoval pouze s použitím pramenů a literatury uvedených v seznamu použitých zdrojů.

18.4.2025 Pavel Mareš v.r.

Poděkování

Rád bych touto cestou vyjádřil své upřímné poděkování doc. Ing. Michaelu Rostovi, Ph.D. a Mgr. Ireně Štrausové, Ph.D. za jejich neocenitelnou pomoc, odborné vedení a cenné rady během vypracování mé bakalářské práce. Jejich trpělivost, ochota a vstřícnost mi byly velkou oporou a významně přispěly k jejímu úspěšnému dokončení.

Obsah

Seznam symbolů

y_t - vysvětlovaná regresního modelu

$\bar{y}$ - odhadnutá střední hodnota stacionární časové řady y_t

$\widehat{y_t}$ - vypočtená OLS-hodnota (tj. metodou nejmenších čtverců

Y_t - systematická / deterministická složka časové řady

ε_t - reziduální (chybová) složka časové řady

$\hat{\varepsilon}_t$ - odhadnutá reziduální složka

C_t - cyklická složka

S_t - sezónní složka

β_t - vliv změny vysvětlující proměnné

Δ - diferenční operátor

$\mathbf{X}$ - matice nezávislých proměnných

$\boldsymbol{\beta}$ - je vektor parametrů

$\boldsymbol{\varepsilon}$ - je vektor reziduí

NLS - nelineární metoda nejmenších čtverců

OLS - metoda nejmenších čtverců

MSE - střední čtvercová chyba

$\hat{\beta}_t$ - odhadnutá hodnota parametru metodou nejmenších čtverců

$\hat{\beta}_0$ - intercept, odhadnutá konstanta funkce

$f'(x)$ - první derivace funkce v bodě x

$f''(x)$ - druhá derivace funkce v bodě x

a_t, b_t, c_t, d_t - koeficienty splinové funkce

h - šířka pásma

h^* - optimální šířka pásma

$\mu|x_0$ - skutečná očekávaná hodnota (teoretická střední hodnota)

E - operátor střední hodnoty

w_t - váha přiřazená každému pozorování

$K_N(z)$ - Gaussova jádrová funkce

$K_T(z)$ - trikubická jádrová funkce

p - stupeň polynomu

exp - exponenciální funkce reprezentující e^x

1. Úvod

V oblasti ekonomie a finanční analýzy je analýza časových řad klíčová pro pochopení a predikci chování ekonomických veličin v čase. Tato práce se zaměřuje na aplikaci regresních metod při modelování akciových dat, konkrétně na analýzu časových řad akcií společnosti ČEZ. Cílem je využít různé regresní přístupy k predikci budoucího vývoje akciových cen, přičemž se budeme soustředit na neparametrické metody, jako jsou smoothing splines a lokalizované regrese (LOESS a LOWESS).

V teoretické části práce budou představeny různé regresní přístupy vhodné pro analýzu ekonomických časových řad. Nejprve se zaměříme na základy regrese a parametrické metody a následně se budeme věnovat specifickým neparametrickým metodám, jako jsou spliny a lokální regrese. V této části se také seznámíme s teoretickými základy výběru těchto metod pro aplikaci na časové řady ekonomických dat.

Praktická část bude věnována aplikaci vybraných regresních metod na reálná data akcií společnosti ČEZ. Cílem bude provést srovnání těchto metod z hlediska jejich predikční schopnosti na historických datech a analyzovat jejich prediktivní výkonnost v kontextu reálného trhu.

1.1. Cíle práce

Analyzovat vhodnost regresních metod pro modelování časových řad akciových cen, se zaměřením na predikci budoucích hodnot.

Porovnat efektivitu parametrických a neparametrických přístupů, zejména smoothing splines a lokálních regresních metod, z hlediska jejich predikční schopnosti.

Implementovat a otestovat vybrané metody v programovacím jazyce R, analyzovat jejich výkon na historických datech a posoudit jejich využitelnost pro predikci v praxi.

2. Literární přehled

Analýza časových řad nám umožňuje pochopit, jak se různé ukazatele nebo procesy vyvíjejí v čase. Díky ní jsme zároveň schopni modelovat a predikovat různé události skutečného světa. Obecně známým přístupem je například dekompozice časové řady, která umožňuje rozdělit data na trendovou, sezónní a reziduální složku (Montgomery et al., 2008). Dekompozice je sice silný analytický nástroj, ale její lineární povaha a nároky na data ji často omezují.

Dalším přístupem jsou regresní metody, které přinášejí větší variabilitu při modelování vztahů mezi proměnnými. Jednoduchá lineární regrese, založená na minimalizaci čtverců reziduí, je první volbou při analýze trendů (Cipra, 2013). Pokud ale data vykazují složitější vzory chování, pak lineární regrese nestačí a je vhodné zvážit jiné metody, které lépe zachytí nelineární závislosti.

Co se pokročilejších technik týče, spliny představují kompromis mezi hladkostí a přizpůsobením modelu datům. Často využívanými jsou například kubické spliny, které zajišťují spojitost druhých derivací, což vede k plynulým a přirozeně vypadajícím křivkám (Fox, 2015). Vyhlazující spliny jsou ještě o něco lepší, protože zavádějí penalizaci za drsnost, což pomáhá vyhnout se nadměrnému přizpůsobení datům.

Neparametrické metody, jako je LOESS a LOWESS, jsou využívány pro svou schopnost přizpůsobit se různým tvarům dat. Umožňují nám modelovat nelineární vztahy, aniž bychom museli předem definovat konkrétní tvar funkce (Fox, 2015). Tyto metody, se hodí zejména v situacích, kdy data vykazují vysokou volatilitu.

Pro ekonomii jsou neparametrické metody leckdy nenahraditelné. Umožňují analyzovat složité a nepravidelné časové řady, které se v této oblasti často vyskytují. Spojením regresních a neparametrických metod získáváme nástroje, které nám umožňují nejen lépe porozumět datům, ale také spolehlivěji predikovat jejich budoucí vývoj (Horová & Zelinka, 2008).

3. Úvod do analýzy časových řad

Analýza časové řady umožňuje zpracovat posloupnost hodnot určitého ukazatele, závislého na časovém faktoru. Analýza se dělí na dva typy, a to extrapolaci a interpolaci. Interpolace, také nazývaná jako deskriptivní analýza, popisuje historické vzorce v datech a extrapolace neboli prediktivní analýza, umožňuje předpověď budoucích hodnot. Předpokladem je, že každé pozorování y_t je složeno ze složky systematické Y_t a iregulární ε_t (Kozák et al., 1994).

To lze vyjádřit aditivním modelem:

$$y_t = Y_t + \varepsilon_t, \qquad t = 1, \dots, n. \tag{3.1}$$

Za pomoci aditivního modelu jsme schopni provést odhady systematické složky, pomocí níž jsme dále schopni vypočítat odhady iregulární složky:

$$\hat{\varepsilon}_t = y_t - \hat{Y}_t, \qquad t = 1, \dots, n. \tag{3.2}$$

Tato posloupnost je nazývána reziduální složkou časové řady.

Dále je pro úvod do analýzy časových řad vhodné zmínit **kvantitativní metody**, které se zabývají zpracováním historických dat, pomocí statistických a matematických modelů k predikci budoucího trendu. Mezi kvantitativní modely, patří:

- **Regresní modely**, které analyzují vztahy mezi vysvětlovanou a vysvětlující proměnnou.
- **Vyhlazovací modely**, které se zaměřují na eliminaci šumu v datech, jako jsou cyklické a sezónní složky a jsou taktéž užitečná k odhalování trendů.
- **Obecné modely časových řad**, které využívají statistické charakteristiky historických dat k odhadu neznámých parametrů a předpovědi budoucích událostí

(Montgomery et al., 2008).

Také je vhodné zmínit **kvalitativní proměnné**, které se běžně používají v ekonometrických analýzách. Tyto proměnné mají obvykle malý rozsah hodnot a využívají se k zohlednění výjimečných událostí, nebo identifikaci odlehlých pozorování. V regresní analýze jsou označovány jako **dummy proměnné** (Cipra, 2013).

4. Regresní metody

Regresní analýza se zabývá modelováním vztahu mezi závislou proměnnou y a nezávislými proměnnými $x_1, x_2, \dots, x_t$. Cílem je pochopit, jak vstupní proměnné ovlivňují výstup, a použít tento vztah pro predikci. Regresní metody lze rozdělit na lineární a nelineární.

Regrese obvykle neprochází všemi naměřenými body, ale zachycuje jejich celkový trend. Nejběžnějším případem regresní analýzy je lineární regrese, kterou modelujeme pomocí metody nejmenších čtverců (OLS). Tato metoda minimalizuje součet rozdílů čtverců mezi hodnotami v datech a odpovídajícími hodnotami navržené přímky (Cipra, 2013).

„Jak uvádí Cipra (2013, s. 35-37) metoda nejmenších čtverců, hledá odhady parametrů $\boldsymbol{\beta}$ tak, že se vzhledem k těmto parametrům minimalizuje součet čtverců:

$$S = \sum_{t=1}^{n} (y_t - (\beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk}))^2 = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}). \tag{4.1}$$

Optimalizační úloha, kdy minimalizujeme (4.1) přes parametry $\boldsymbol{\beta}$ má řešení:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}. \tag{4.2}$$

Hodnoty $\mathbf{y}$ předpovíme pomocí zkonstruovaného modelu:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \tag{4.3}$$

A hodnoty reziduální složky $\boldsymbol{\varepsilon}$ odhadneme pomocí modelu:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}." \tag{4.4}$$

Dále Cipra (2013) uvádí, že je možné taktéž tyto odhady znázornit geometricky jako zobrazení pozorovaných hodnot $\mathbf{y}$ do lineárního prostoru generovaného sloupci matice $\mathbf{X}$. Projekční matice $\mathbf{H}$ (hat matrix) provádějící zobrazení n-rozměrného vektoru do prostoru generovaného sloupci $\mathbf{X}$ a má tvar:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y} = \mathbf{H}\mathbf{y}, \tag{4.5}$$

Fox (2015) tyto poznatky rozšiřuje ve své kapitole o vážené lineární regresi (WLS), kde se odhadované hodnoty $\hat{\boldsymbol{y}}$ počítají pomocí váhové matice $\mathbf{W}$ a projekční matice $\mathbf{H}_{\mathbf{W}}$:

$$\hat{\mathbf{y}} = \mathbf{H_W y} \qquad \text{kde} \qquad \mathbf{H_W} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}. \tag{4.6}$$

Projekční matice $\mathbf{H_W}$ je analogická tzv. vyhlazovací matici $\mathbf{S}$, používané v neparametrické regresi. Vyhlazovací matice $\mathbf{S}$ transformuje pozorované hodnoty $\mathbf{y}$ na odhadnuté hodnoty $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} = \mathbf{S}\mathbf{y}. \tag{4.7}$$

Jak uvádí Fox (2015),“ vyhlazovací matice $\mathbf{S}$ slouží v neparametrických metodách obdobně jako hat-matrix $\mathbf{H}$ v lineární regresi, přestože se liší svými matematickými vlastnostmi. Matice $\mathbf{H}$ je vždy symetrická a zachovává výsledky při opakované aplikaci ($\mathbf{H}=\mathbf{H}^{\mathrm{T}}$ a $\mathbf{H}=\mathbf{H}\mathbf{H}$), zatímco matice $\mathbf{S}$ tyto vlastnosti obecně nemá, což ji odlišuje.“

Podobně jako v OLS lze v neparametrické regresi definovat ekvivalentní stupně volnosti modelu $\boldsymbol{df_{mod}}$ (degrees of freedom) různými způsoby:

$$\boldsymbol{df_{mod}} = \mathrm{trace}(\mathbf{S}), \qquad \mathrm{trace}(\mathbf{S}\mathbf{S}^{\mathrm{T}}), \qquad \mathrm{trace}(2\mathbf{S} - \mathbf{S}\mathbf{S}^{\mathrm{T}})$$

Tyto definice reflektují vliv vyhlazovací matice na odhadované hodnoty a umožňují analýzu složitosti modelu. Analogicky lze definovat i stupně volnosti chyby pomocí zbytku matice $\mathbf{I} - \mathbf{S}$:

$$\boldsymbol{\varepsilon} = (\mathbf{I} - \mathbf{S})\mathbf{y}. \tag{4.8}$$

kde $\mathbf{I}$ označuje jednotkovou matici. To nám ukazuje, že vyhlazovací matice $\mathbf{S}$ a projekční matice $\mathbf{H}$ sdílejí podobné principy, přičemž každá je přizpůsobena specifickému přístupu – lineární nebo neparametrické regresi (Fox, 2015).

4.1. Lineární regrese

Lineární regresní model můžeme zapsat jako:

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{4.9}$$

kde y_t představuje hodnotu vysvětlované proměnné y v čase t, x_{tk} jsou hodnoty vysvětlujících proměnných, $\beta_0,\ \beta_1, \ldots, \beta_k$ jsou neznámé parametry modelu a ε_t je reziduální složka (Cipra, 2013).

Lineární model lze také zapsat pomocí vektorové notace, což poskytuje přehlednější a efektivnější způsob reprezentace:

$$\mathbf{y} = (1, x_{t1}, x_{t2}, \dots, x_{tk}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{pmatrix} = \boldsymbol{x\beta} + \boldsymbol{\varepsilon}_t \tag{4.10}$$

nebo ve formě maticového zápisu:

$$\boldsymbol{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \tag{4.11}$$

kde:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{pmatrix}, \tag{4.12}$$

X je matice nezávislých proměnných, β je vektor parametrů a ε je vektor reziduí. (Cipra, 2013)

Jsou-li v modelu využity pouze dva parametry, tak specificky mluvíme o modelu regresní přímky zapisované jako:

$$y_t = \alpha + \beta x_t + \varepsilon_t, \qquad t = 1, \dots, n, \tag{4.13}$$

Lineární rovnice pro koeficienty α a β, se nazývají souborem normálních rovnic kde n je počet pozorování:

$$\hat{\alpha} \sum_{t=1}^{n} x_t + \hat{\beta} n = \sum_{t=1}^{n} y_t$$

$$\hat{\alpha} \sum_{t=1}^{n} x_t^2 + \hat{\beta} \sum_{t=1}^{n} x_t = \sum_{t=1}^{n} x_t y_t$$

Řešením tohoto souboru získáme dané koeficienty nejmenších čtverců:

$$\hat{\beta} = \frac{\sum_{t=1}^{n} x_t y_t - n \,.\, \bar{x} \,.\, \bar{y}}{\sum_{t=1}^{n} x_t^2 - n \,.\, \bar{x}^2}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta} \,.\, \bar{x}, \qquad t = 1, \dots, n, \tag{4.14}$$

kde $\bar{y}$ a $\bar{x}$ značí aritmetické průměry (Cipra, 2013).

4.2. Nelineární regrese

Nelineární regrese, také zvaná jako polynomiální regrese, modeluje vztahy nelineárních závislostí mezi proměnnými.

Pro odhad parametrů nelineárního modelu se používá metoda nelineárních nejmenších čtverců (NLS), která minimalizuje součet čtverců reziduí. Parametry $\hat{\beta}_{NLS}$ se odhadují

pomocí minimalizace funkce (4.1). Podmínka pro minimalizaci je získána z první derivace této funkce:

$$\mathbf{G}(\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - f(\mathbf{X}, \boldsymbol{\beta}))=0, \tag{4.15}$$

kde $\mathbf{G}(\boldsymbol{\beta})$ je matice $(n \times k)$ prvních parciálních derivací $\mathbf{f}(\mathbf{X}, \boldsymbol{\beta})$ podle $\boldsymbol{\beta}$ (Cipra, 2013).

Gausův-Newtonův algoritmus je podle Cipry (2013) založen na linearizaci regresní funkce, kdy v příslušném rozvoji kolem počáteční hodnoty parametrů $\hat{\beta}_0$ ponecháme pouze členy řádu jedna

$$f(\mathbf{X}, \boldsymbol{\beta}) \approx f(\mathbf{X}, \hat{\beta}_0) + \mathbf{G}(\hat{\beta}_0)\left(\boldsymbol{\beta} - \hat{\beta}_0\right). \tag{4.16}$$

Dosazením do modelu (4.11) tento model linearizujeme, takže v něm lze provést klasický OLS-odhad

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{G}(\hat{\beta}_0)^{\mathrm{T}}\mathbf{G}(\hat{\beta}_0)\right)^{-1}\mathbf{G}(\hat{\beta}_0)^{\mathrm{T}}(\mathbf{y} - f(\mathbf{X}, \hat{\beta}_0) + \mathbf{G}(\hat{\beta}_0)\hat{\beta}_0) =$$

$$= \hat{\beta}_0 + \left(\mathbf{G}(\hat{\beta}_0)^{\mathrm{T}}\mathbf{G}(\hat{\beta}_0)\right)^{-1}\mathbf{G}(\hat{\beta}_0)^{\mathrm{T}}(\mathbf{y} - f(\mathbf{X}, \hat{\beta}_0)). \tag{4.17}$$

Tím jsme získali první iterační hodnotu, kterou můžeme označit jako $\hat{\beta}_1$. Při pokračování tímto způsobem dál, pak přechod mezi n-tým a $(n+1)$-ním iteračním krokem bude

$$\hat{\beta}_{n+1} = \hat{\beta}_n + \left(\mathbf{G}(\hat{\beta}_n)^{\mathrm{T}}\mathbf{G}(\hat{\beta}_n)\right)^{-1}\mathbf{G}(\hat{\beta}_n)^{\mathrm{T}}(\mathbf{y} - f(\mathbf{X}, \hat{\beta}_n)). \tag{4.18}$$

5. Klasická dekompozice časových řad

Dekompozice časové řady podle Forbelské (2009) „zahrnuje rozklad na deterministickou a náhodnou složku. Klasická dekompozice se opírá o:

- **Aditivní modely**: $y_t = T_t + C_t + S_t + \varepsilon_t, \qquad t = 1, \dots, n.$ (5.1)
- **Multiplikativní modely**: $y_t = T_t \,.\, C_t \,.\, S_t \,.\, \varepsilon_t, \qquad t = 1, \dots, n.$ (5.2)

Jednotlivé složky jsou definovány jako:

- **Trend** T_t: Odraz dlouhodobých změn.
- **Sezónní složka** S_t: Periodické změny, např. vliv ročních období.
- **Cyklická složka** C_t: Periodické změny s periodou delší než jeden rok.
- **Náhodné fluktuace** ε_t: Drobná, nepostižitelné příčiny kolísání.

Proces dekompozice je užitečný pro přípravu dat pro neparametrické regresní metody a umožňuje lepší analýzu trendů a vzorců, což zlepšuje modelování.“ (viz. s. 58 uvedeného textu, kapitola 2)

Model, který kombinuje všechny tyto složky, nazýváme **ryzí aditivní model** (5.1). Tento model předpokládá, že časová řada je sestavena jako součet různých složek, z nichž každá reprezentuje určitý směr pohybu. Předmětem analýzy je tedy rozložení tohoto součtu na jednotlivé komponenty (Kozák et al., 1994).

Pro efektivní analýzu je důležité rozlišovat mezi periodickými a neperiodickými řadami. Periodické řady obsahují cyklické a sezónní složky nebo obě zároveň a vykazují pravidelný vzor chování. Neperiodické řady vykazují náhodné nebo nepravidelné změny a obvykle se vyznačují trendovou složkou, která může být vyjádřena jako monotónně rostoucí nebo klesající funkcí.

Zmíněný trend je klíčový při modelování a predikci časových řad. Trendy nám umožňují rozpoznat dlouhodobé změny v chování sledovaného ukazatele. Základním trendovým modelem je lineární trend:

$$T_t = \beta_0 + \beta_1 t, \qquad t = 1, \dots, n, \tag{5.3}$$

kde β_0 je konstanta (počáteční hodnota) a β_1 je sklon, který udává rychlost změny. Odhady parametrů β_0 a β_1 pro tento model získáme pomocí metody nejmenších čtverců. V případech, kdy se trend nechová lineárně, je třeba použít **nelineární modely**, které lépe

zachycují složitější dynamiku časové řady. Mezi nelineární modely patří například exponenciální růst nebo různé typy polynomů (Cipra, 2013).

Pro naši praktickou část existuje v jazyce R funkce decompose se syntaxí decompose(x, type = c("additive", "multiplicative"), filter = NULL). Funkce nejprve určí trendovou složku pomocí klouzavého průměru, pokud je filtr NULL, použije se symetrické okno se stejnými váhami a odstraní ji z časové řady. Poté se vypočítá sezónní údaj zprůměrováním pro každou časovou jednotku za všechna období. Sezónní údaj je pak vycentrován. Nakonec je chybová složka určena odstraněním trendu a sezónního údaje recyklovaného podle potřeby z původní časové řady.

6. Spliny

Polynomiální regrese může být citlivá na odlehlé hodnoty v datech, což může ovlivnit přesnost modelu. U vyšších polynomiálních stupňů také hrozí, že model bude vykazovat extrémní oscilace nebo nerealistické chování na okrajích dat. Alternativou je použití spline, které rozdělují rozsah na menší intervaly a pro každý interval vytvářejí specifický model. Tím se zlepšuje přesnost modelu a snižuje zkreslení oproti globálním polynomiálním metodám (Fox, 2015).

6.1. Lineární spline

Lineární spline je základní typ spline interpolace, který využívá přímky k propojení sousedních bodů (uzlů) časové řady, což vytváří lomenou čáru, která prochází všemi uzly. Tento přístup poskytuje přesnou interpolaci, ale výsledná křivka může být hrubá a nezachytit plynulost, protože používá lineární funkce. Vyšší stupně spline interpolace, jako kubický spline, zajistí hladší přechody mezi jednotlivými úseky (Mathews & Fink, 1999).

(Mathews & Fink, 1999) definují, že „matematicky lze lineární spline popsat pomocí Lagrangeovy interpolace:

$$S_t(x) = y_t \frac{x - x_{t+1}}{x_t - x_{t+1}} + y_{t+1} \frac{x - x_t}{x_{t+1} - x_t} \qquad \text{pro } x_t \le x \le x_{t+1}, \quad t = 1, \dots, n, \qquad (6.1)$$

Pro znázornění jednotlivých částí lineárního splinu můžeme využít ekvivalentní výraz odvozený z bodového sklonového vzorce pro úsečku:

$$S_t(x) = y_t + d_t(x - x_t), \qquad t = 1, \dots, n, \qquad (6.2)$$

kde $d_t = (y_{t+1} - y_t)/(x_{t+1} - x_t)$. Výsledná funkce lineárního splinu může být formulována následovně:

$$S(x) = \begin{cases} y_0 + d_0(x - x_0) & \text{pro } x \in \langle x_0, x_1 \rangle, \\ y_1 + d_1(x - x_1) & \text{pro } x \in \langle x_1, x_2 \rangle, \\ \vdots & \vdots \\ y_t + d_t(x - x_t) & \text{pro } x \in \langle x_t, x_{t+1} \rangle, \\ \vdots & \vdots \\ y_{n-1} + d_{n-1}(x - x_{n-1}) & \text{pro } x \in \langle x_{n-1}, x_n \rangle, \end{cases} \qquad t = 1, \dots, n. \qquad (6.3)$$

Tvar rovnice (6.3) je vhodnější než tvar rovnice (6.1) pro explicitní výpočet $S(x)$.“

6.2. Kvadratický spline

Kvadratický spline využívá kvadratický polynom na každém subintervalu pro konstrukci interpolovaných funkcí. Nicméně, tato metoda vykazuje několik nevýhod, které ji činí nevhodnou pro analýzu časových řad.

Jedním z problémů je přerušení spojitosti druhé derivace na sudých uzlech, což může mít za následek náhlé změny zakřivení a vést k nežádoucím deformacím grafu. Tato vlastnost není vhodná pro modely, které vyžadují plynulé změny v průběhu časové řady (Mathews & Fink, 1999).

Další nevýhodou je nižší schopnost přizpůsobení kvadratických splinů při zajišťování plynulých přechodů mezi sousedními polynomy. Kvadratické polynomy mají tři volné konstanty, ale pouze dvě podmínky, což není dostatečné pro zajištění spojitosti první derivace, a tento problém může vést k nespolehlivým výsledkům (Burden & Faires, 2010).

Z těchto důvodů nejsou kvadratické spliny vhodnou volbou pro modelování časových řad, kde jsou požadovány hladké a stabilní přechody mezi jednotlivými částmi modelu. Pro zajištění hladkých přechodů mezi sousedními intervaly a spojitosti derivací se místo kvadratických splinů běžně používají kubické spliny, které umožňují spojitost první i druhé derivace.

6.3. Kubický spline

Kubický spline je funkcí S, která interpoluje funkci f na intervalu $\langle a, b\rangle$ pro množinu uzlů $a = x_0 < x_1 < \cdots < x_n = b$. „Kubický interpolační splajn vyhovuje následujícím podmínkám:

a) $S(x)$ je na každém dílčím intervalu $\langle x_t, x_{t+1}\rangle$ polynom třetího stupně, označovaný $S_t(x)$, pro každé $t = 0, 1, \dots, n-1$.
b) $S_t(x_t) = f(x_t)$, pro každé $t = 0, 1, \dots, n-1$.
c) $S_{t+1}(x_{t+1}) = S_t(x_{t+1})$, pro každé $t = 0, 1, \dots, n-2$,
d) $S'_{t+1}(x_{t+1}) = S'_t(x_{t+1})$, pro každé $t = 0, 1, \dots, n-2$.
e) $S''_{t+1}(x_{t+1}) = S''_t(x_{t+1})$, pro každé $t = 0, 1, \dots, n-2$.

“ (Horová & Zelinka, 2008, s. 193)

Dále je požadováno, aby byla splněna jedna z následujících podmínek:

i. $S''(x_0) = S''(x_n) = 0$

ii. $S''(x_0) = f'(x_0)$ a $S'(x_n) = f'(x_n)$

Pokud kubický spline S splňuje první podmínku, nazývá se přirozený spline, v případě druhé podmínky jde o úplný spline.

Pro každý dílčí interval $\langle x_t, x_{t+1} \rangle$ je kubický spline $S(x)$ definován jako polynom třetího stupně:

$$S_t(x) = a_t + b_t(x - x_t) + c_t(x - x_t)^2 + d_t(x - x_t)^3, \quad t = 1, \dots, n-1, \qquad (6.4)$$

kde a_t, b_t, c_t a d_t jsou koeficienty, které se určují pro každý interval tak, aby spline splňoval výše uvedené podmínky spojitosti a derivací (včetně okrajových podmínek). Výsledná funkce kubického splinu by vypadala:

$$S(x) = \begin{cases} a_0 + b_0(x - x_0) + c_0(x - x_0)^2 + d_0(x - x_0)^3 \\ a_1 + b_1(x - x_1) + c_1(x - x_1)^2 + d_1(x - x_1)^3 \\ \vdots \\ a_t + b_t(x - x_t) + c_t(x - x_t)^2 + d_t(x - x_t)^3 \\ \vdots \\ a_{t-1} + b_{t-1}(x - x_{t-1}) + c_{t-1}(x - x_{t-1})^2 + d_{t-1}(x - x_{t-1})^3 \end{cases} \qquad (6.5)$$

kde $a_t = f(x_t)$ a každá funkce $S_t(x)$ je kubický polynom, který interpoluje funkci $f(x)$. (Burden & Faires, 2010)

7. Neparametrické metody

Tyto metody zkoumají podmíněné rozdělení závislé proměnné v závislosti na jedné nebo více vysvětlujících proměnných. Na rozdíl od parametrických metod, které vyžadují specifikaci funkční formy, jako je lineární nebo nelineární model, neparametrické metody nepředpokládají konkrétní tvar tohoto vztahu. Proto neparametrické regrese poskytuje větší schopnost přizpůsobení a je užitečná v případech, kdy není možné přesně určit strukturu vztahu mezi proměnnými (Fox, 2015).

Neparametrická regrese se často používá k analýze vztahu mezi dvěma kvantitativními proměnnými. Tento vztah je zobrazen ve formě bodového grafu, kde se metody neparametrické regrese nazývají „vyhlazovače bodového grafu“ (scatterplot smoothers). Mezi hlavní metody neparametrické regrese patří:

- **Jádrová regrese (Kernel regression):** Tato metoda využívá váhové funkce ke stanovení vztahu mezi proměnnými v okolí každého bodu.
- **Lokálně-polynomická regrese (Local-polynomial regression):** Obecnější přístup než jádrová regrese, která umožňuje použití polynomů vyššího řádu pro lepší přizpůsobení datům.
- **Vyhlazovací spliny (Smoothing splines):** Pokročilejší metoda, která využívá penalizaci složitosti modelu a umožňuje hladké změny ve vztahu mezi proměnnými.

Neparametrickou regresi lze využít jako základ pro rozšířené modely, například neparametrické vícenásobné regrese nebo aditivní modely. Ve vizualizaci dat a interpretaci bodových grafů se často využívá například metoda LOWESS (lokálně-vážená regrese), která umožňuje snadnější pochopení vztahu mezi proměnnými (Fox, 2015).

7.1. Jádrové funkce

Jádrová regrese rozšiřuje metodu lokálního průměrování v neparametrické regresi. Tento přístup se využívá k odhadu funkce $f(x_0)$, tedy závislosti závislé proměnné na konkrétní hodnotě vysvětlující proměnné x_0, a přiděluje větší váhu pozorováním, která jsou této hodnotě blíže.

Principem této metody, jak popisuje Fox (2015) ve svém díle, je přiřazení vah jednotlivým pozorováním pomocí jádrové funkce. Váhy dosahují maxima v bodě x_0 a symetricky klesají směrem od tohoto bodu. To lze vyjádřit vztahem:

$$w_t = K\left(\frac{(x_t - x_0)}{h}\right), \qquad t = 1, \dots, n. \tag{7.1}$$

$K(z)$ značí zmíněnou jádrovou funkci, kde z je normalizovaná vzdálenost mezi x_t a ohniskovou hodnotou x_0. Čím menší je vzdálenost $(x_t - x_0)$, dělená šířkou okna h (bandwith), tím větší váhu w_t pozorování získá. Odhadnutá hodnota v x_0 je poté spočítána jako vážený průměr hodnot y_t

$$\hat{f}(x_0) = \frac{\sum_{t=1}^{n} w_t\, y_t}{\sum_{t=1}^{n} w_t}, \qquad t = 1, \dots, n. \tag{7.2}$$

Fox (2015) ve svém díle také uvádí existenci různých jádrových funkcí, které lze použít pro výpočet vah. Mezi nejčastější patří:

1. **Gaussova jádrová funkce:** U této funkce odpovídá šířka okna $\boldsymbol{h}$ směrodatné odchylce normálního rozdělení, přičemž pozorování vzdálená více než $\boldsymbol{2h}$ mají téměř nulovou váhu.

$$K_N(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \tag{7.3}$$

2. **Trikubická jádrová funkce:** U této funkce je $\boldsymbol{h}$ poloviční šířkou okna, přičemž pozorování mimo toto okno nejsou zohledněna

$$K_T(z) = \begin{cases} (1 - |z|^3)^3 & \text{pro } |z| < 1, \\ 0 & \text{pro } |z| \geq 1. \end{cases} \tag{7.4}$$

3. **Obdélníková jádrová funkce:** Tato funkce přiřazuje stejnou váhu všem pozorováním uvnitř okna šířky $\boldsymbol{h}$ a ignoruje ostatní.

$$K_R(z) = \begin{cases} 1 & \text{pro } |z| < 1, \\ 0 & \text{pro } |z| \geq 1. \end{cases} \tag{7.5}$$

O výběru šířky okna $\boldsymbol{h}$ pak můžeme říct, že je klíčovým parametrem, který určuje hladkost odhadu. Malé hodnoty h zachycují detaily v datech a větší hodnoty zase poskytují

hladší výsledky. Místo pevného $\boldsymbol{h}$ lze použít adaptivní šířku okna, která zahrnuje pevný počet nebo procento dat (span). Pro optimální výsledek je vhodné zvolit šířku okna, která zajistí přiměřenou hladkost a přesnost modelu (Fox, 2015).

7.2. Lokální polynomiální regrese

Lokální regrese je neparametrickou metodou odhadu závislosti mezi proměnnými, která vylepšuje přístupy založené na jádrové funkci a poskytuje větší flexibilitu při modelování nelineárních vztahů. Tato metoda se opírá o využití polynomů pro lokální odhady a použití vážené regrese metodou nejmenších čtverců (Weighted Least Squares, WLS).

WLS rozšiřuje OLS tím, že přiřazuje pozorováním váhy na základě odhadované nebo známé variability chyb. Například model lineární regrese (4.11) má nulovou střední hodnotu chyby ε_t, ale různou rozptylovou složku $\sigma_t^2 = \sigma_\varepsilon^2 / w_t^2$. Váhy w_t jsou inverzní k variaci chyb, a to zajišťuje větší váhu pozorování s nižší variabilitou (Fox, 2015).

„Podle Foxe (2015, s. 304) je pravděpodobnostní funkce modelu je definována jako:

$$L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = \frac{1}{(2\pi)^{n/2}|\mathbf{S}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right), \tag{7.6}$$

kde $\mathbf{S} = \sigma_\varepsilon^2 . \mathrm{diag}(1/w_1^2, \ldots, 1/w_n^2)$. Odhady parametrů maximální věrohodnosti jsou:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}, \qquad \sigma_\varepsilon^2 = \frac{\sum_{t=1}^{n} w_t^2 \varepsilon_t^2}{n}, \tag{7.7}$$

kde $\mathbf{W} = \mathrm{diag}(w_1, w_2, \ldots, w_n)$ je váhová matice. Vážená regrese minimalizuje vážený součet čtverců reziduí:

$$\sum_{t=1}^{n} w_t \varepsilon_t^2 \text{ ."} \tag{7.8}$$

Pro bod x_0 se odhaduje lokální polynomiální model ve tvaru:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1(x_t - x_0) + \hat{\beta}_2(x_t - x_0)^2 + \cdots + \hat{\beta}_p(x_t - x_0)^p + \hat{\varepsilon}_t, \quad t = 1, \ldots, n. \tag{7.9}$$

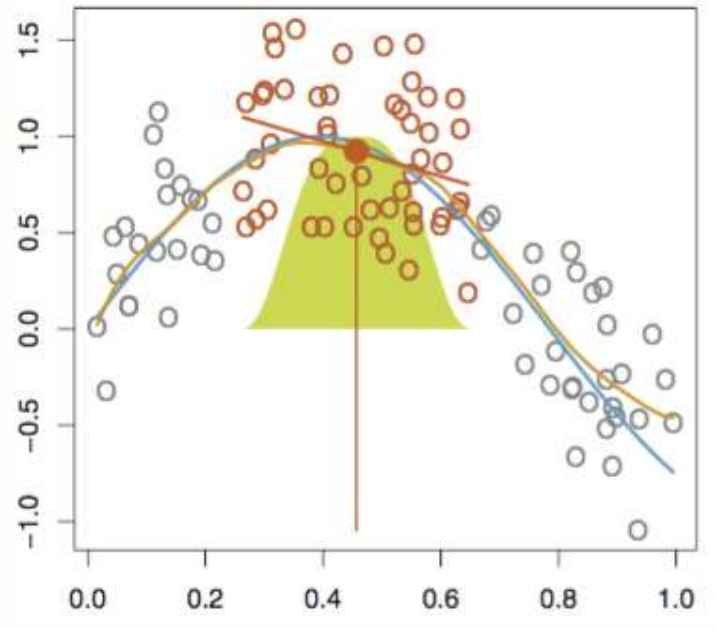
Váhy pro jednotlivá pozorování jsou definovány jádrovou funkcí, jak bylo uvedeno podle vzorce (7.1), přičemž šířka pásma h kontroluje rozsah lokální oblasti a hodnota $\hat{y}_{x_0} = \hat{\beta}_0$ (Fox, 2015).

Podle Foxe (2015) nejčastější volbou stupně polynomu p je lineární aproximace ($p = 1$), která minimalizuje zkreslení a zlepšuje přesnost na okrajích dat. Vyšší stupně polynomů ($p = 2,\ p = 3$) poskytují větší flexibilitu za cenu vyšší variace. Obecně se ale používají polynomy lichých stupňů, pro které existují teoretické výhody. Co se šířky pásma týče tak ta určuje kolik sousedních bodů přispívá k odhadu v x_0. Lze ji zvolit pevnou nebo nastavit pomocí počtu nejbližších sousedů (span).

LOESS (Locally Estimated Scatterplot Smoothing) nebo **LOWESS** (Locally Weighted Scatterplot Smoothing) je metoda lokální polynomiální regrese, která navíc využívá iterativní přístup k vážené regresi. Na rozdíl od běžné lokální regrese nepočítá váhy pouze jednou, ale opakovaně je upravuje na základě okolních dat. Tento postup činí metodu odolnější vůči extrémním hodnotám a umožňuje jí lépe se přizpůsobit nelineárním vzorcům, které by jednorázový polynom nemusel správně zachytit (Xavier Bourret Sicotte, 2018).

Rozvoj těchto metod, je úzce spojen s prací Clevelanda (1979), který přispěl k jejich vylepšení a širokému přijetí v oblasti statistiky. Cleveland ve své studii zdůraznil význam iterativního přizpůsobování vah podle okolních dat, což těmto metodám umožnilo dosáhnout lepších výsledků při analýze dat s komplexními vzorci. Jeho přístup přinesl novou úroveň přesnosti při modelování nelineárních vztahů v datech.

Graf 1: Lokální polynomiální regrese



Zdroj: (Trevor Hastie et. al., 2017)

„Ilustrace lokální regrese na simulovaných datech, kde modrá křivka představuje skutečnou funkci $f(x)$, světle oranžová křivka je odhad lokální regrese. Oranžové body jsou lokalizovány kolem cílového bodu x_0, a váhy (žlutá křivka) klesají s rostoucí vzdáleností od x_0." (Xavier Bourret Sicotte, 2018)

7.2.1. Bližší pohled na šířku pásma lokálně regresního vyhlazovače

Volba šířky pásma h je zásadní pro kvalitu odhadu, protože ovlivňuje rovnováhu mezi zkreslením (bias) a variabilitou (variance). Malá šířka pásma nám umožňuje lepší zachycení detailů, ale vede k vysoké variabilitě odhadu. Velká šířka nám zjednodušuje odhad, ale může způsobit zkreslení.

Střední kvadratická chyba (MSE) odhadu je definována jako:

$$MSE(\hat{y}_0) = \mathrm{E}[(\hat{y}_0 - \mu|x_0)^2] = \mathrm{Variance} + \mathrm{Bias}^2, \qquad (7.10)$$

kde $\hat{y}_0$ je odhadovaná hodnota x_0, $\mu|x_0$ je teoretická střední hodnota a E představuje operátor střední hodnoty. Bias a variance lokálně-lineárního odhadu v bodě x_0 lze modelovat:

$$\mathrm{Bias}(\hat{y}|x_0) \approx \frac{h^2}{2} s_K^2 f''(x_0), \qquad (7.11)$$

$$\text{Variance}(\hat{y}|x_0) \approx \frac{\sigma_\varepsilon^2 a_K^2}{nhp_x(x_0)}. \tag{7.12}$$

Konstanty s_K^2 a a_K^2 jsou závislé na jádrové funkci, $f''(x_0)$ je zakřivení regresní funkce a $p_x(x_0)$ nám značí hustotu dat v bodě x_0. Pokud je v daném bodě k dispozici málo dat, pak dostáváme malou hustotu $p_x(x_0)$ a to nám zvyšuje varianci (Fox, 2015).

„Optimální šířka pásma h^* minimalizuje MSE a zajišťuje vyváženost mezi zkreslením a variancí. Pro lokálně-lineární odhad je optimální šířka pásma:

$$h^*(x_0) = \left[\frac{\sigma_\varepsilon^2 a_K^2 s_K^4}{nhp_x(x_0)[f''(x_0)]^2}\right]^{\frac{1}{5}}. \tag{7.13}$$

Pokud je $f''(x_0) = 0$, optimální šířka h^* je nekonečná, což odpovídá globálnímu lineárnímu odhadu“ (Fox, 2015, s. 537).

Ve srovnání s metodou nejbližších sousedů, která zohledňuje počet dat ($nhp_x(x_0)$), ale nezohledňuje lokální zakřivení $f''(x_0)$ docházíme k závěru, že pevná šířka pásma může vést k nepřesnostem v oblastech s velkými změnami regresní funkce (Fox, 2015).

7.2.2. Výběr rozsahu křížovou validací

Jednou z možností odhadu šířky pásma je formální odhad její optimální hodnoty h^*. Tento odhad lze provádět buď pro každou hodnotu x_0 proměnné x_t, kde se hodnotí $\hat{y}|x$, nebo lze určit optimální průměrnou hodnotu, která se použije pro pevnou šířku pásma. Podobný postup lze využít i u metody LOWESS. Jednodušším způsobem odhadu optimální šířky pásma nebo rozsahu s je pomocí křížové validace, kterou lze použít u obou zmíněných typů odhadů (Fox, 2015).

Křížová validace hodnotí regresní funkci na základě pozorování x_t. Základním principem je vynechání t-ého pozorování z lokální regrese v ohniskové hodnotě x_t a tím dosahujeme nezávislého odhadu na skutečné hodnotě y_t. Odhad očekávané hodnoty $\text{E}(y|x_t)$, získaný po vynechání tohoto pozorování, označujeme jako $\hat{y}_{-t}|x_t$. Tvar křížové funkce je následující:

$$CV(s) = \frac{1}{n}\sum_{t=1}^{n}(\hat{y}_{-t}(s) - y_t)^2, \tag{7.14}$$

kde $\hat{y}_{-t}(s)$ je odhad $\hat{y}_{-t}|x_t$ pro daný rozsah s. Naším cílem je nalézt hodnotu s, která minimalizuje tuto funkci (Fox, 2015).

Výpočet CV může být výpočetně náročný, protože je nutné model n-krát opakovaně odhadnout pro každou hodnotu s. Jako řešení tohoto problému, vznikly přibližné metody, jako je generalizovaná křížová validace (GCV), jak popisuje Wahba (1985). GCV je definováno jako:

$$GCV(s) = \frac{n\,.SS(s)}{[df_{res}(s)]^2}, \tag{7.15}$$

kde $SS(s)$ je reziduální součet čtverců a $df_{res}(s)$ jsou „ekvivalentní" reziduální stupně volnosti pro model lokální regrese s rozsahem s. GCV poskytuje velmi přesnou aproximaci k původní funkci CV (Fox, 2015).

7.2.3. Stupně volnosti v lokální regresi

Fox (2015) se zmiňuje, že v lokálně-polynomiální regresi jsou odhadnuté hodnoty $\hat{y}_t$ vážené součty pozorovaných hodnot y. To lze vyjádřit vzorcem:

$$\hat{y}_t = \sum_{k=1}^{n} s_{tk} y_k, \tag{7.16}$$

jehož váhy s_{tk} lze seskupit do vyhlazovací matice $\mathbf{S}$, která je matice o rozměrech $(n \times n)$ a obsahuje váhové koeficienty. Jak jsme již zmínili v kapitole 4. matice je analogická k projekční matici $\mathbf{H}$ (4.5) v lineární regresi, která transformuje pozorované hodnoty na jejich odhady (4.7).

Stupně volnosti v regresním modelu chápeme jako jejich míru složitosti, tedy počet parametrů, které byly použity k přizpůsobení dat. V lokální regresi nemáme explicitní parametry jako v lineární regresi, ale je možné použít následující definice:

1. $\boldsymbol{df_{mod}} = \mathrm{trace}(\mathbf{S})$: Je nejjednodušší na výpočet. Sleduje diagonální prvky matice $\mathbf{S}$, které určují, kolik informací z původních dat bylo použito k přizpůsobení modelu. V lineární regresi by se jednalo o počet parametrů.
2. $\mathrm{trace}(\mathbf{S}\mathbf{S}^{\mathrm{T}})$: Tato varianta sleduje vztah mezi $\mathbf{S}$ a její transponovanou maticí $\mathbf{S}^{\mathrm{T}}$. Je užitečná v případech, kdy je potřeba analyzovat přesnější vliv jednotlivých pozorování na model.

3. $\text{trace}(\mathbf{2S} - \mathbf{SS}^{\mathrm{T}})$: Tato definice vyjadřuje komplexnější vztahy mezi přizpůsobenými hodnotami a reziduální chybou. Je málo využívaná, kvůli její výpočetní náročnosti.

Nejčastěji se používá první definice, která udává, že stupně volnosti modelu jsou dány součtem diagonálních prvků matice $\mathbf{S}$. To odpovídá množství informací, které model využívá a vyjadřuje vztah mezi $\widehat{\boldsymbol{y}}$ a váhami použitými při jejich výpočtu (Fox, 2015).

Stupně volnosti jsou zásadní pro odhad rozptylu chyby modelu σ_ε^2:

$$\mathbf{S}_{\boldsymbol{\varepsilon}}^2 = \frac{\sum_{t=1}^{n} \varepsilon_t^2}{df_\epsilon} = \frac{\sum_{t=1}^{n} \varepsilon_t^2}{n - df_{mod}}. \tag{7.17}$$

df_ϵ a $n - df_{mod}$ je počet reziduálních stupňů volnosti. Stupně volnosti nám zároveň pomáhají pochopit, kolik z variability dat je model schopen vysvětlit (Fox, 2015).

7.3. Vyhlazující spliny

Na rozdíl od regresních splinů, které se řadí mezi parametrické regresní modely, vyhlazující spliny vznikly jako řešení problému nalezení funkce $\hat{f}(x)$ s dvěma spojitými derivacemi, které minimalizují penalizovaný součet čtverců reziduí v neparametrické regresi.

V této metodě h představuje vyhlazovací parametr, který lze přirovnat šířce pásma, který jsme blíže popsali v předešlé kapitole.

$$SS^*(\lambda) = \sum_{t=1}^{n} [(y_t - f(x_t)]^2 + \lambda \int [f''(x)]^2 \, dx. \tag{7.18}$$

První člen v rovnici představuje součet čtverců reziduí. Druhý člen pak slouží k penalizaci za drsnost funkce. Ta nabývá vysokých hodnot, pokud integrovaná druhá derivace $f(x)$ vykazuje výrazné změny, a to nám naznačuje vysokou nepravidelnost funkce. Koncové hodnoty integrálu pokrývají rozsah dostupných hodnot: $x_{min} < x_{(1)}$ a $x_{max} > x_{(n)}$ (Fox, 2015).

Fox (2015, s. 549) se zmiňuje, že „v extrému, pokud je konstanta vyhlazení nastavena na $\lambda = 0$ (a pokud jsou všechny hodnoty x_t odlišné), pak $\hat{f}(x)$ jednoduše interpoluje data.

Na opačném extrému, pokud je λ velmi velké, pak bude $\hat{f}(x)$ vybrána tak, že $\hat{f}''(x)$ bude všude nulové, což znamená globální lineárnímu přizpůsobení pomocí metody nejmenších čtverců.“

Funkce $\hat{f}(x)$, která minimalizuje rovnici (7.18), je přirozený kubický spline v jedinečných pozorovaných hodnotách x. To by mohlo naznačovat, že je zapotřebí n parametrů, ale penalizace za drsnost zavádí dodatečná omezení, která snižují ekvivalentní počet parametrů pro vyhlazovací spline a zároveň brání funkci $\hat{f}(x)$ v interpolaci dat (Fox, 2015).

Při výběru vyhlazovací konstanty λ se uplatňují podobné přístupy jako při volbě šířky pásma nebo rozpětí u lokálně-polynomických vyhlazovačů. Těmito přístupy jsou například křížová validace nebo zobecněná křížová validace. V praxi je však běžné nastavit hodnotu vyhlazovací konstanty λ nepřímo, například nastavením ekvivalentního počtu parametrů pro zvolený vyhlazovač (Fox, 2015).

Vyhlazovací spliny mají drobné výhody oproti lokálně-polynomiálním vyhlazovačům. Oba vyhlazovače jsou lineární, takže je lze vyjádřit ve formě $\hat{y} = \mathbf{S}y$ pro vhodně definovanou matici vyhlazovače $\mathbf{S}$, ta má ovšem o něco lepší vlastnosti pro vyhlazovací spliny. Výhodnou vlastností vyhlazovacích splinů, která pro lokálně-polynomiální vyhlazovače neplatí je že, pokud se použijí jako stavební bloky aditivního regresního modelu, algoritmus zpětného přizpůsobení (backfitting), je zaručeně konvergentní (Fox, 2015).

8. Metriky predikční účinnosti

K hodnocení kvality predikční účinnosti modelů slouží různé statistické metriky. Ty jsou založeny na analýze odchylek mezi skutečnými hodnotami a hodnotami odhadovanými model, přičemž cílem je minimalizovat chyby a optimalizovat přesnost modelu.

Neboli jak uvádí Arlt a další (2002, s. 27): „Po odhadu parametrů trendového modelu časové řady y_t (např. metodou nejmenších čtverců) analyzujeme rozdíly mezi skutečnými hodnotami y_t a odhadnutými hodnotami trendu $\hat{T}_t$, tzv. rezidua ($\widehat{\varepsilon_t} = y_t - \hat{y}_t = y_t - \hat{T}_t$), která představují odhad nesystematické složky."

Cipra (2013) jako základní metriky chyb předpovědi uvádí:

1. **Součet čtvercových chyb (SSE – Sum of Squared Errors):**

$$SSE = \sum_{t=n+1}^{n+h} (y_t - \hat{y}_t)^2, \tag{8.1}$$

kde h je počet předpovězených hodnot. SSE se často objevuje v regresní analýze, zdůrazňuje větší chyby, které mají větší vliv na výsledek.

2. **Střední čtvercová chyba (MSE – Mean Squared Error):**

$$MSE = \frac{1}{h} \sum_{t=n+1}^{n+h} (y_t - \hat{y}_t)^2, \tag{8.2}$$

umožňuje srovnání různých modelů na základě průměrné chyby. Je citlivá na extrémní hodnoty a často slouží jako standardní ukazatel přesnosti.

3. **Akaikeho informační kritérium (AIC):**

$$AIC = 2k - \ln(L), \tag{8.3}$$

kde k je počet parametrů modelu a L je likelihood (pravděpodobnost) modelu, což je míra toho, jak dobře model vysvětluje data. AIC je jednou z nejběžněji používaných metrik pro modelování časových řad. Tento ukazatel penalizuje složitější modely, čímž pomáhá vybrat ten nejvhodnější model s ohledem na jeho přesnost. Nižší hodnota AIC značí lepší model, který poskytuje kompromis mezi přesností a složitostí.

9. Metodika

Cílem praktické části je porovnat metody zmíněné v teoretické části z hlediska jejich predikční účinnosti. Porovnání je založeno na metrikách získaných z analýzy. Výpočty, vizualizace a metriky jsou generovány pomocí skriptů v R, které jsou dostupné v příloze práce.

Metodický postup:

Vstupními daty je denní vývoj cen akcií společnosti ČEZ, a.s., získaný z oficiálních stránek ve formátu CSV. Dataset obsahuje dva sloupce: „Datum“ a „Cena“. Před aplikací metod byla provedena kontrola kvality dat, odstranění případných NA hodnot, úprava formátování a doplnění chybějících hodnot způsobených víkendovým uzavřením burzy. Následně byla data rozdělena na trénovací a testovací sadu, přičemž všechny metody pracovaly se stejným trénovacím a testovacím oknem.

Na trénovacích datech byly aplikovány vybrané regresní metody, které vytvořily modely sloužící k predikci hodnot v testovací sadě. Postup analýzy skriptů, popsaný v kapitole 9.3, je pro všechny metody stejný. Uchován byl model s nejlepšími hodnotami metrik predikční účinnosti.

U metod jako LOESS bylo možné upravit šířku okna, zatímco u splinů se nastavoval počet uzlů, aby se předešlo overfittingu či underfittingu a zajistilo se správné zobecnění dat. U smoothing splines byly explicitně nastaveny parametry, jako je penalizační parametr a stupně volnosti. Pokud nebylo možné vhodné parametry určit přímo, byla k jejich optimalizaci použita křížová validace.

Na základě vytvořených modelů byly generovány predikce pro testovací sadu a porovnány se skutečnými hodnotami. Tyto rozdíly sloužily k výpočtu metrik predikční účinnosti, jako jsou MSE, SSE a AIC. Skripty zároveň generovaly grafy znázorňující přizpůsobení modelů časové řadě a porovnání predikovaných a skutečných hodnot.

Výsledky analýzy umožnily porovnat jednotlivé metody z hlediska jejich predikční účinnosti a vhodnosti pro modelování dané ekonomické časové řady. Interpretace těchto výsledků je uvedena v následujících částech práce.

9.1. Použitá data

V praktické části této bakalářské práce jsou analyzována data o cenách akcií společnosti ČEZ, a. s., získaná přímo z oficiálního internetového portálu společnosti ČEZ. Dataset obsahuje historické hodnoty uzavíracích cen akcií a byl vybrán s ohledem na jeho vhodnost pro analýzu ekonomických časových řad, zejména vzhledem k jeho nahodilosti a dostatečnému množství dat, které umožňuje aplikaci neparametrických regresních metod.

Data jsou exportována ve formátu CSV (Comma-Separated Values), který je široce využíván ve statistických softwarových nástrojích. Tento formát umožňuje snadný import a následné zpracování velkých datasetů, přičemž zachovává jednoduchost a přístupnost.

Historicky jsou data o akciích společnosti ČEZ dostupná od 17. 3. 2003 až do současnosti, přičemž záznamy jsou pořizovány na denní bázi s výjimkou víkendů, kdy je burza uzavřena. Pro tuto analýzu byl zvolen dataset obsahující údaje za rok 2024, což odpovídá nejaktuálnějším dostupným datům a umožňuje relevantní zhodnocení vývoje cen akcií v nedávném období.

Dataset obsahuje dvě proměnné:

- **Datum**: Určuje konkrétní den, ke kterému se váže hodnota ceny akcie. Formát této proměnné je standardní (DD-MM-YYYY).
- **Cena**: Reprezentuje uzavírací cenu akcií v daný den, vyjádřenou v českých korunách za jednu akcii.

Takto jednoduše strukturovaný soubor nám umožňuje přehlednou manipulaci a zpracování dat v prostředí R, který si představíme v následující kapitole.

Načtení dat proběhne použitím funkce implementované v R, read.csv().

9.2. Programovací jazyk R

Pro tuto bakalářskou práci je vytvořena sada skriptů, která je schopna aplikovat regresní přístupy nebo neparametrické metody na importovaná data ve formátu CSV. Aplikace je naprogramována ve vývojovém prostředí R-4.4.2 pro Windows. Programovací jazyk je možné stáhnout na jejich oficiálních stránkách https://cran.r-project.org/.

„R je systém pro statistické výpočty, analýzu dat a grafiku. Skládá se z jazyka a běhového prostředí s grafikou, debuggerem, přístupem k určitým funkcím systému a schopností spouštět programy uložené v souborech skriptů.“ (R Core Team, 2024)

Co vyzdvihuje R nad ostatními programovacími jazyky je jeho široká podpora statistických metod, bohatá knihovna balíčků a nástroje pro vizualizaci dat. R je robustním nástrojem pro modelování, predikci a vizualizaci analýzy časových řad (Crawley, 2013).

Pro práci s časovými řadami v R existuje řada specializovaných knihoven, mezi něž patří:

- **forecast** – metody a nástroje pro zobrazení a analýzu jednorozměrných předpovědí včetně exponenciálního vyhlazování pomocí modelů stavového prostoru a automatického modelování ARIMA (R Core Team, 2024).
- **lubridate** – poskytuje jednoduché nástroje pro práci s daty a časy, včetně manipulace s datovými formáty, časovými zónami a časovými intervaly (R Core Team, 2024).
- **zoo** a **xts** – umožňují efektivní práci s indexovanými daty a zaměřují se zejména na nepravidelné časové řady číselných vektorů, matic a faktorů (R Core Team, 2024).
- **stats** – zpřístupňuje základní statistické funkce a modely, včetně regresních analýz a analýz trendů (R Core Team, 2024).

V R jsou časové řady často reprezentovány jako objekty typu **ts** (time series) nebo jako data frame (**data.frame**). Importování dat z CSV souborů se provádí pomocí funkce read.csv():

```
data <- read.csv("export.csv")
```

Pro manipulaci s daty jsou velmi užitečné knihovny **dplyr** a **tidyr**, které umožňují efektivní čištění, transformaci a agregaci dat (Crawley, 2013).

R podporuje různé regresní techniky, včetně lineární a nelineární regrese. Například pro lineární regresi lze použít funkci **lm()**:

```
model <- lm(y ~ x, data = imported_data)
summary( model)
```

Pro aplikaci splinů nebo nelineárních regresí existuje knihovna **mgcv**, která umožňuje použití Generalized Additive Models (**GAM**), které využívají spojení lineárních a nelineárních funkcí, kde nelineární části modelu jsou reprezentovány pomocí splinů. Tento přístup je užitečný pro analýzu dat, kde vztahy mezi proměnnými nejsou jednoduše lineární, což umožňuje přizpůsobit model specifickým charakteristikám dat (Crawley, 2013).

Příklad použití splinů v mgcv:

```
library(mgcv)
model <- gam(y ~ s(x), data = imported_data)
summary(model)
```

Funkce $s()$ v tomto modelu označuje splinovou funkci, která umožňuje nelineární přizpůsobení k datům.

Pro práci s polynomiálními spliny lze také použít balíček splines, který poskytuje funkce pro tvorbu a aplikaci B-splinů nebo natural splinů. Funkce bs() v balíčku splines umožňuje tvorbu B-splinů, což jsou typy splinů, které jsou užitečné pro hladké křivky a mohou být použity jako regresní modely (Crawley, 2013).

Příklad použití B-splinu v regresním modelu:

```
library(splines)
model_spline <- lm(y ~ bs(x, degree = 3), data = imported_data)
summary(model_spline)
```

Jak již bylo zmíněno R poskytuje i nástroje pro vizualizaci, zejména použitím knihovny **ggplot2**, která umožňuje efektivně vizualizovat časové řady a interpretaci výsledků:

```
library(ggplot2)
ggplot(data, aes(x=Date, y=Price))+
  geom_line()+
  labs(title="Cenový vývoj", x = "Datum", y = "cena")
```

Pro diagnostiku modelů lze použít funkci plot() k zobrazení reziduí a kontrolu jejich náhodnosti (Crawley, 2013).

9.3. Vysvětlení metodiky

Pro analýzu dat bylo vytvořeno pět samostatných skriptů, z nichž každý plní specifický úkol v procesu modelování a predikce časových řad. Následující sekce popisují jednotlivé kroky metodiky včetně zdůvodnění použitých parametrů.

Skript 1 - Iterativní hledání optimálního modelu pro predikci

První skript slouží k nalezení nejvhodnějšího predikčního modelu pomocí iterativního snižování velikosti tréninkového datasetu. Tento přístup umožňuje analyzovat, jak model reaguje na různé rozsahy tréninkových dat, a posoudit jeho schopnost generalizace. Iterace také pomáhá identifikovat, zda vybrané parametry vedou ke stabilním výsledkům napříč různými vzorky dat.

Skript 2 - Optimalizace parametrů modelu

Druhý skript provádí optimalizaci parametrů modelu, přičemž opět využívá iterativní snižování velikosti tréninkového datasetu. Cílem je identifikovat takové hodnoty parametrů, které nejlépe vystihují strukturu časové řady a minimalizují riziko přefitování nebo podfitování. Hlavní optimalizované parametry zahrnují počet uzlů u spline metod a hodnoty span/spar u metod LOESS/LOWESS a smoothing splines. Uložený model je ten, který dosahuje nejnižší chyby fitu na tréninkových datech. Správnost nastavení parametrů je ověřována prostřednictvím vizualizací a predikčních metrik.

Skript 3 - Vyvážení mezi tréninkovou a testovací chybou

Třetí skript, podobně jako předchozí dva, využívá iterativní snižování velikosti tréninkového datasetu, aby našel rovnováhu mezi chybou MSE na tréninkových datech, chybou MSE na testovacích datech a hodnotou AIC. Tento přístup zajišťuje, že model nebude ani přefitovaný (overfitting), ani příliš obecný (underfitting). Optimalizace těchto metrik vede leckdy k robustnějšímu modelu s lepší predikční schopností na testovacích datech.

Skript 4 - Aplikace na kompletní dataset

Čtvrtý skript aplikuje optimalizovaný model na celý dataset, čímž ověřuje jeho výkon na všech dostupných datech. Tento krok umožňuje nejen zhodnotit stabilitu modelu, ale také analyzovat průměrnou chybu predikce při různých nastaveních parametrů.

Skript 5 - Sliding window pro vyhodnocení průběžné chyby

Pátý skript využívá metodu sliding window k vyhodnocení vývoje chyby modelu v čase. Trénovací okno je nastaveno na 3 měsíce, po nichž model predikuje vývoj ceny na následujících 7 dní. Po každé iteraci se okno posune o 7 dní, což simuluje reálné podmínky predikce a umožňuje posoudit, jak se model přizpůsobuje změnám v časové řadě. Tento přístup poskytuje detailní přehled o stabilitě modelu a jeho predikční přesnosti v průběhu celého roku.

9.3.1. Parametry modelů a zdůvodnění jejich volby

Rozdělení dat, predikční horizont, minimální počet dat pro tvorbu modelu

Data byla rozdělena tak, že poslední měsíc byl skryt a z těchto dat se vybral počet odpovídající stanovenému predikčnímu horizontu (7 dní). Tento horizont byl zvolen, protože delší predikční období může vést k extrémním hodnotám predikcí, zejména u denních dat z burzy, kde jsou na začátku a konci týdne časté fluktuace. Skrytím celého měsíce poskytujeme prostor pro testování predikcí na delších horizontech.

Pro trénování modelů bylo stanoveno minimální množství dat na 90 hodnot. Tento limit zajišťuje, že model dokáže zachytit klíčové trendy a výkyvy, zatímco příliš krátká data by mohla omezit jeho schopnost generalizace.

Počet uzlů u spline metod

Pro lineární spliny byl zvolen rozsah 4–6 uzlů, pro kubické spliny 2–6 uzlů. Příliš malý počet uzlů může ignorovat strukturu dat, zatímco příliš velký může vést k přefitování. Vyšší spodní hranice pro lineární spliny byla zvolena, protože nižší počet uzlů ignoroval trendy v některých úsecích dat. U kubických splinů byla naopak nižší hranice stanovena na 2 uzly, protože při více než dvou uzlech model často vykazoval chybné predikce. Tento rozsah byl výsledkem experimentální analýzy a představuje vhodný kompromis.

Parametry LOESS/LOWESS a smoothing splines

Pro metody LOESS/LOWESS a smoothing splines byl zvolen rozsah parametru span/spar 0.3–0.7, což zajistilo dostatečnou schopnost přizpůsobení modelů bez přefitování. Při nastavení parametru span na 0.2 se model LOESS příliš přizpůsobil tréninkovým datům,

což vedlo k nižší chybě MSE na tréninkové sadě, ale zároveň k vyšší chybě MSE na testovacích datech, což naznačovalo přefitování. Zvolený interval tedy představuje kompromis mezi přesností modelu a jeho schopností zobecnění.

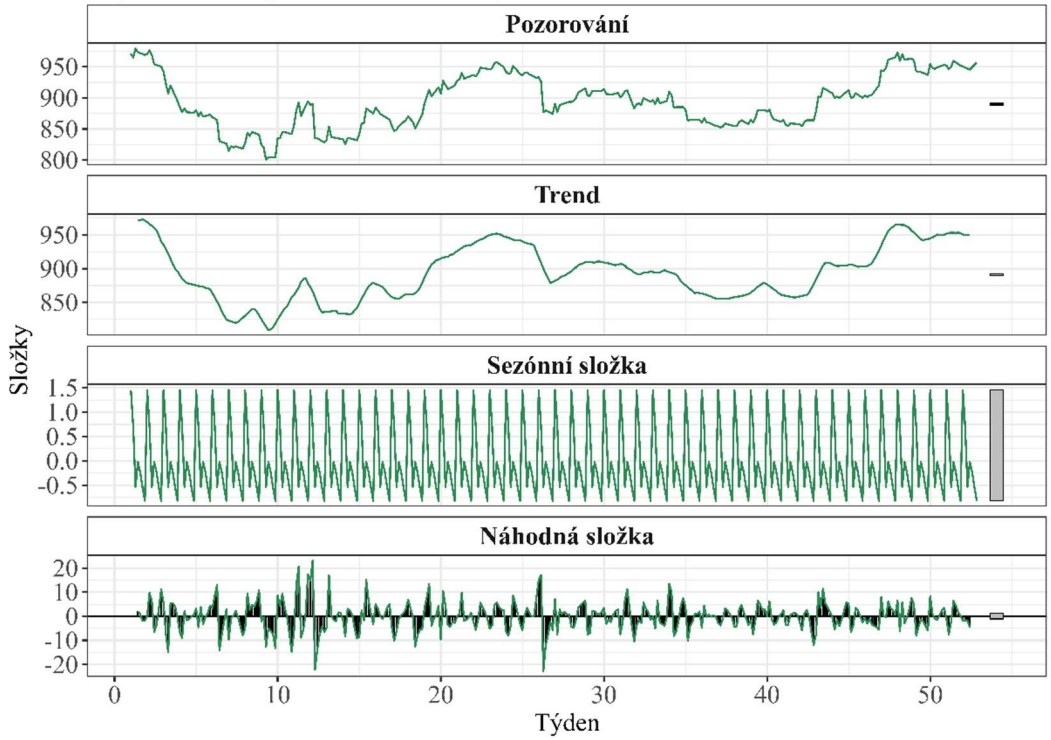
Predikce pomocí lokálních regresních metod

Metody LOESS a LOWESS slouží k vyrovnávání dat, ale neobsahují parametr pro extrapolaci. Při použití pro predikci je model nejprve natrénován na historických tréninkových datech, a poté predikuje hodnoty pro testovací data na základě podobného chování. Místo lineární extrapolace model „prodlouží“ historická data pomocí lokálního přizpůsobení, kdy vychází z informací o chování okolních bodů. Predikce je tak založena na identifikaci podobných vzorců, nikoliv na aplikaci univerzálního trendu. Predikce je tedy omezená na krátký horizont sedmi dní, protože dlouhodobá extrapolace může být nepřesná bez globálního trendu.

10. Testování – Denní ceny akcií ČEZ, a. s. za rok 2024, H = 7

Abychom lépe porozuměli struktuře časové řady denních cen akcií ČEZ v roce 2024, provedli jsme její dekompozici na tři základní složky: trendovou, sezónní a náhodnou.

Graf 2: Dekompozice časové řady denních cen akcií ČEZ v roce 2024



Zdroj: Data z portálu ČEZ, zpracování autor

Trendová složka

Trendová složka ukazuje dlouhodobý vývoj cen akcií. V průběhu roku 2024 bylo možné pozorovat pokles ceny, následovaný stabilizací a růstem (viz graf 2). Tento trend může být ovlivněn makroekonomickými faktory, jako jsou změny v energetickém sektoru, regulace nebo výsledky ČEZ.

Sezónní složka

Sezónní složka vykazuje pravidelné cyklické vzory během roku (viz graf 2), pravděpodobně ovlivněné obchodními cykly, jako jsou týdenní fluktuace nebo reakce na pravidelné ekonomické zprávy.

Náhodná složka

Náhodná složka zachycuje krátkodobé fluktuace (viz graf 2), které nelze vysvětlit trendem nebo sezónností. Zvýšená volatilita naznačuje vliv externích šoků, jako jsou významné události na trhu nebo změny v sentimentu investorů.

10.1. Vyhodnocení analýzy lineární regrese

Při experimentální analýze jsme snížili minimální počet dat pro tvorbu lineární regrese na 60 záznamů. Pokud bychom na tuto metodu kladli stejné nároky jako na ostatní (tedy minimální počet 90 záznamů), model by se přizpůsobil celému datasetu a výsledky by byly výrazně horší.

Tabulka 1: Výsledky analýzy lineární regrese

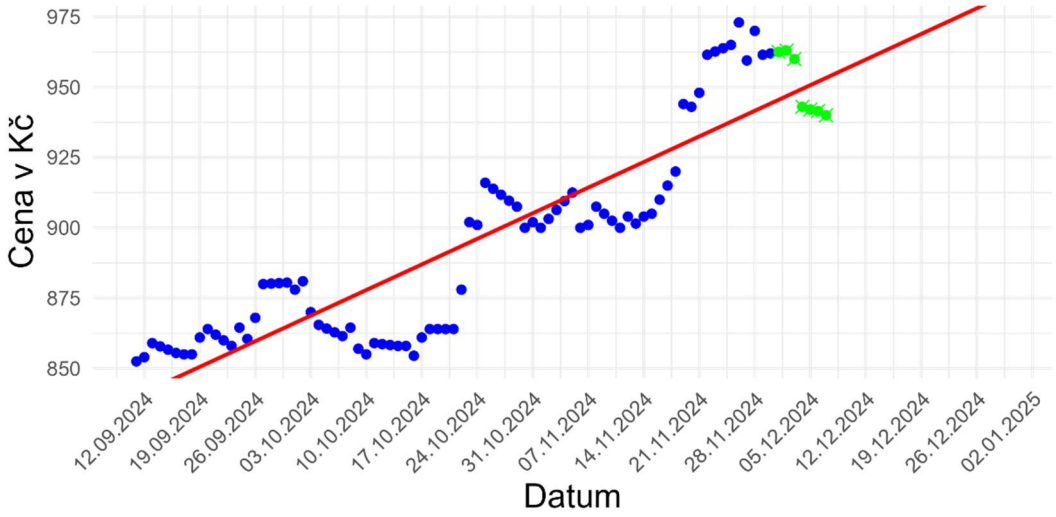
	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat využitý k tvorbě modelu
Skript 1	156.7283	1097.098	278.3326	22544.94	700.1484	81
Skript 2	331.359	2319.513	221.3721	13282.32	511.6389	60
Skript 3	156.7283	1097.098	278.3326	22544.94	700.1484	81
Skript 4	2507.7	17553.9	1539.919	514332.9	3405.239	334
Skript 5	969.7819	6788.473	427.0967	38438.71	791.5665	334

Zdroj: Autor

V rámci analýzy lineární regrese jsme vyhodnotili různé přístupy k výběru tréninkových dat a predikční přesnosti modelu. Výsledky ukazují, že nejlepší predikční výkon dosahuje model při použití posledních 81 hodnot tréninkové sady, kdy data vykazují rostoucí trend (viz graf 3). V tomto případě je MSE predikce 156.7283, MSE tréninkové chyby 278.3326 a AIC 700.1484 (viz tabulka 1). Tato nastavení odpovídají skriptům 1 a 3, přičemž stejné výsledky potvrzují, že skript 1 představuje nejlepší model s optimálním poměrem mezi chybou fitu a predikcemi.

Skript 2, vykazuje MSE predikce 331.359 a MSE tréninkové chyby 221.3721 při použití 60 záznamů. Tyto hodnoty jsou nízké ve srovnání s ostatními metodami, ale je třeba si uvědomit, že byly dosaženy díky velmi krátkému období, které vykazuje monotónní trend, který je nutný pro kvalitní výsledky pomocí lineární regrese.

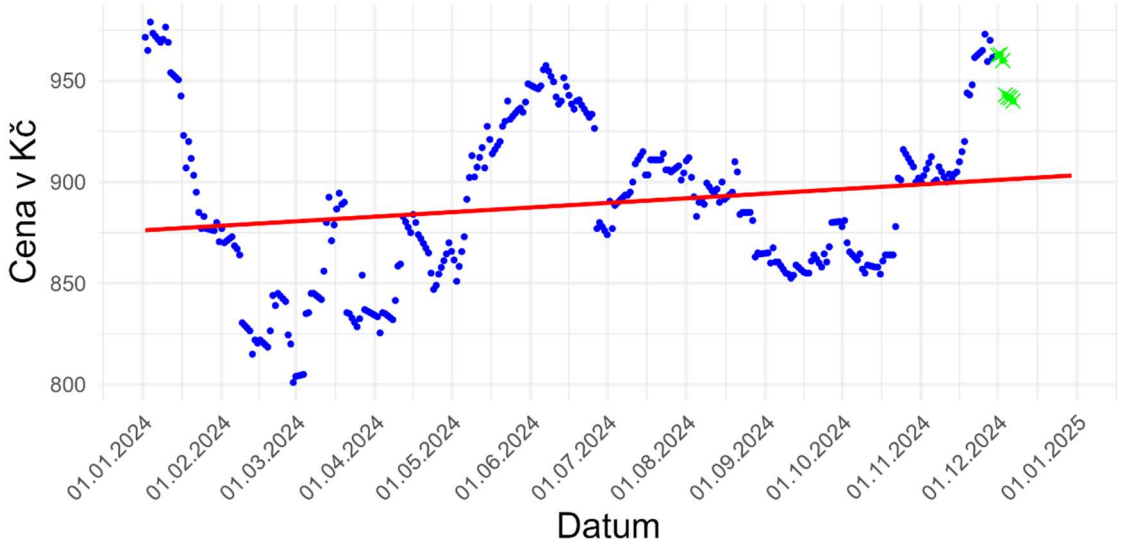
Graf 3: Nejlepší predikce lineární regrese



Zdroj: Data z portálu ČEZ, zpracování autor

Pokud aplikujeme lineární regresi na celou časovou řadu bez přizpůsobení, výsledky se výrazně zhorší. Skript 4, který modeluje celou řadu bez ohledu na změny trendu, vykazuje MSE predikce 2507.7, MSE tréninkové chyby 1539.919 a AIC 3405.239. Tyto vysoké chyby dokazují, že model není schopen adekvátně zachytit změny ve struktuře dat (viz graf 4).

Graf 4: Lineární regrese aplikovaná na celých datech

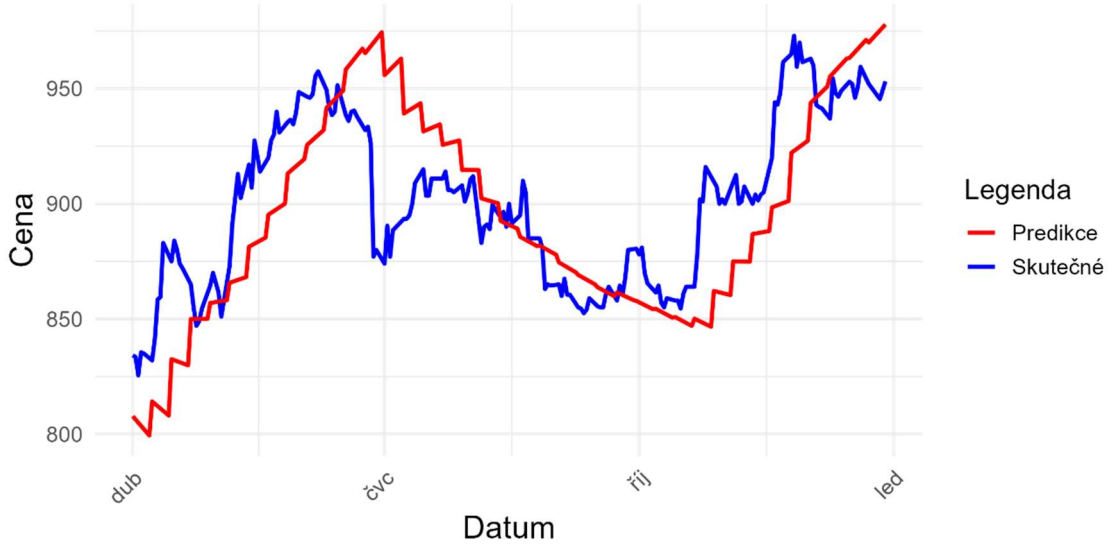


Zdroj: Data z portálu ČEZ, zpracování autor

Pro reálnější simulaci predikce je použit pátý skript s metodou sliding window, kde model trénujeme na tříměsíčních intervalech a testujeme na následujících sedmi dnech. Tento

přístup umožňuje získat průměrnou chybu modelu za celý rok (viz graf 5). Výsledky ukazují, že MSE predikce je 969.7819, MSE tréninkové chyby 427.0967 a AIC 791.5665. Tyto výsledky jsou horší než u pokročilejších metod, protože lineární regrese není schopná zachytit nelineární vzory a rychlé změny trendu, které jsou běžné u ekonomických časových řad.

Graf 5: Sliding window - Skutečné hodnoty a hodnoty predikce lineární regrese



Zdroj: Data z portálu ČEZ, zpracování autor

Zároveň zmíněné výsledky ukazují na omezenou predikční schopnost lineární regrese, která nebere v úvahu sezónní a cyklické vlivy. Pokročilejší metody, jako jsou spliny nebo nelineární regrese, by mohly lépe modelovat složité struktury těchto dat.

10.2. Vyhodnocení analýzy nelineární regrese

Tabulka 2: Výsledky analýzy nelineární regrese

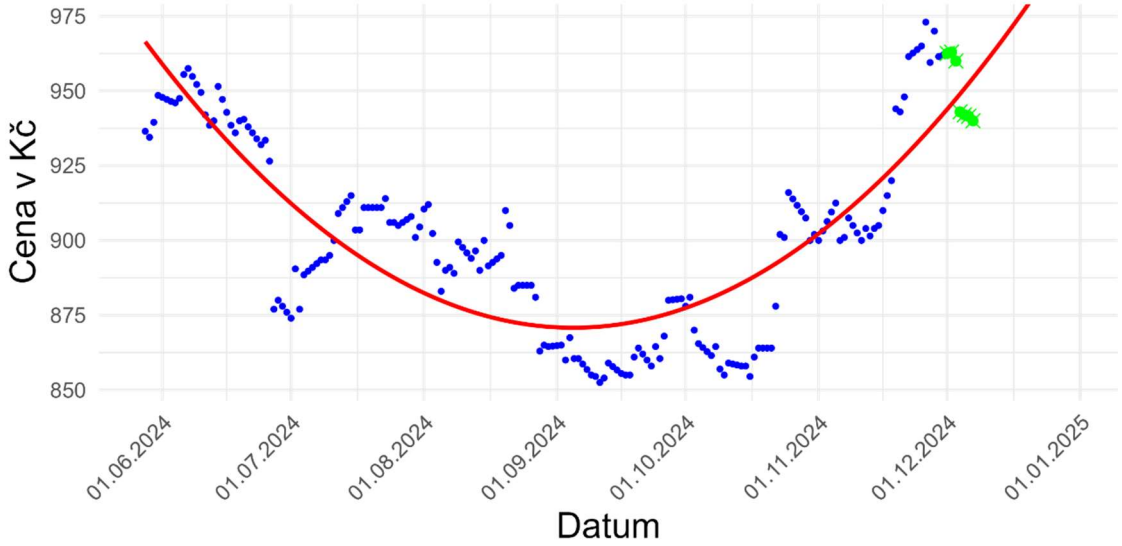
	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat využitý k tvorbě modelu
Skript 1	178.5188	1249.632	315.4144	58982.49	1625.663	187
Skript 2	1295.075	9065.522	134.9155	15245.45	889.8774	113
Skript 3	178.5188	1249.632	315.4144	58982.49	1625.663	187
Skript 4	1584.225	11089.58	1519.886	507642.1	3402.865	334
Skript 5	838.8574	5872.002	277.1809	24946.28	764.3864	334

Zdroj: Autor

Pro analýzu nelineární regrese jsme použili kvadratický model, který zachycuje křivkové trendy v datech, což je užitečné pro predikci s nelineárními vzory v tréninkových datech. Výsledky (viz tabulka 2) ukazují, že nelineární regrese poskytuje výrazně lepší predikci než lineární regrese při aplikaci na větší rozsah dat.

Při použití tréninkových dat z posledních 187 dní viz. graf 6 dosahuje kvadratická regrese MSE predikce 178.5188 a MSE tréninkové chyby 315.4144. Model lépe zachycuje nelineární vzory v datech, ale vykazuje vyšší hodnoty AIC (1625.663) než lineární regrese. To je důsledek použití širšího rozsahu dat a přidání kvadratického členu do modelu.

Graf 6: Nejlepší predikce pomocí nelineární regrese



Zdroj: Data z portálu ČEZ, zpracování autor

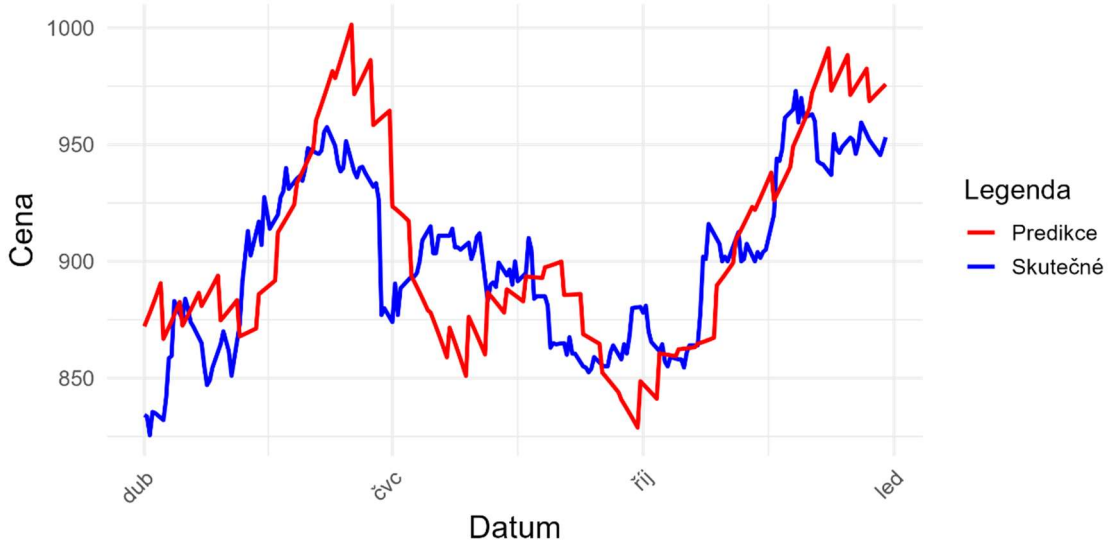
Skript 2, který se soustředí na model přizpůsobující se co nejlépe datům, trénuje na menším počtu záznamů (113), což vede k nižší MSE tréninkové chyby (134.9155). Vysoké hodnoty MSE predikce (1295.075) však naznačují, že model se přizpůsobuje konkrétnímu tréninkovému období a nezvládá při tom poskytovat efektivní hodnoty predikce.

Skript 4 modeluje celou časovou řadu bez ohledu na změny trendu, což vede k velmi vysokým hodnotám MSE predikce (1584.225) a AIC (3405.239). To ukazuje na neschopnost modelu správně zachytit změny trendu.

Při použití skriptu 5, jsou výsledky nelineární regrese lepší než u lineární regrese viz. graf 7. MSE predikce se zlepšilo na 838.8574, MSE tréninkové chyby na 277.1809, a AIC kleslo na 764.3864. Tyto výsledné metriky značí, že tento model lépe popisuje změny v trendech a je schopen se lépe přizpůsobit dynamickým změnám v časové řadě než lineární regrese.

Nicméně pokud budeme uvažovat tvar zvolené kvadratické funkce a strukturu dat, jsou tyto lepší výsledky prakticky bezvýznamné, protože jediný moment, kdy metoda s touto funkcí predikuje hodnoty blízké skutečnosti je, pokud tréninková data připomínají konvexní či konkávní tvar.

Graf 7: Sliding window - Skutečné hodnoty a hodnoty predikce nelineární regrese



Zdroj: Data z portálu ČEZ, zpracování autor

Porovnání lineární a nelineární regrese

Lineární regrese je efektivní metodou pro predikci budoucích hodnot v případě, kdy data vykazují monotónní nárůst, či klesání hodnot ve stejné velikosti. Nelineární regrese je zase použitelnou metodou jedině v případě, pokud známe strukturu dat a jsme schopni na základě této informace, aplikovat funkci, která je schopná se struktuře dat přizpůsobit.

10.3. Vyhodnocení analýzy lineární spline

Při analýze lineárního splinu je kladen důraz na počet uzlů, který ovlivňuje přizpůsobení spline křivky. Výsledky ukazují vliv počtu uzlů a velikosti tréninkových dat na predikční a tréninkovou chybu.

Tabulka 3: Výsledky analýzy lineárního splinu

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Počet knotů
Skript 1	193.8167	1356.717	478.694	159405.1	3022.657	334	4
Skript 2	3067.039	21469.27	55.65284	5231.367	666.8031	94	5
Skript 3	193.8167	1356.717	478.694	159405.1	3022.657	334	4
Skript 4a	398.7082	2790.957	375.0901	125280.1	2936.673	334	4 - 6
Skript 4b	193.8167	1356.717	478.2241	159726.8	3022.657	334	4
Skript 5	709.4892	4966.424	100.7363	9066.267	673.472	334	4 - 6 nejlepší 6

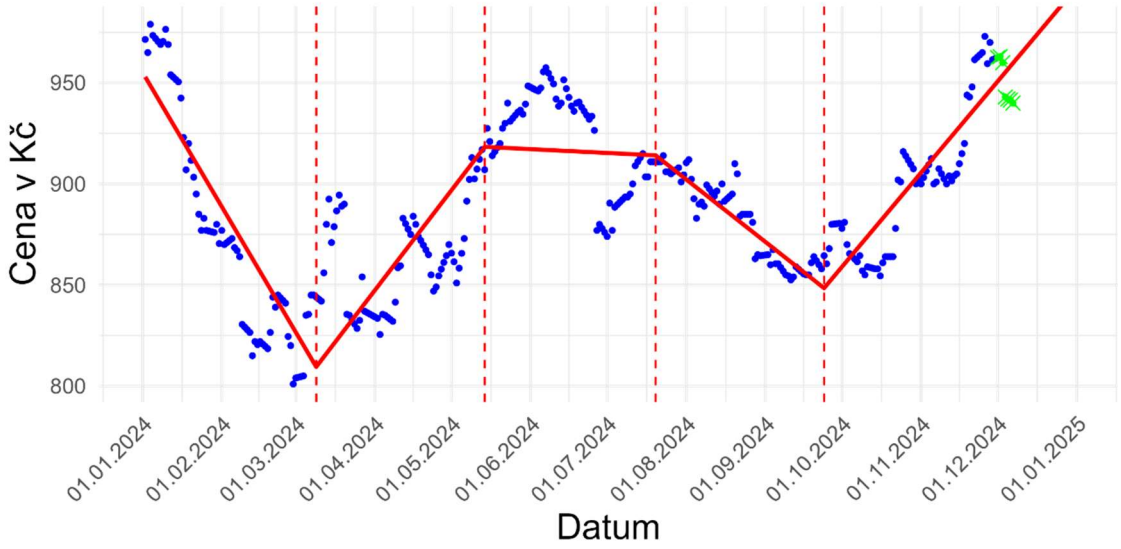
Zdroj: Autor

Skript 4a poskytuje průměrné výsledky pro počet uzlů v rozmezí 4 až 6, zatímco Skript 4b se zaměřuje na optimální nastavení, které bylo určeno v předchozím skriptem 4a.

Skripty 1, 3 a 4b vedly ke stejným výsledkům, což naznačuje, že model se specificky lépe přizpůsobuje při větším rozsahu tréninkových dat. MSE predikce dosahuje 193.8167 a MSE tréninkové chyby 478.694, přičemž experimentální analýza ukázala, že s tréninkovým obdobím delším než šest měsíců byla metoda výrazně efektivnější.

Další testy ukázaly, že pokud se model pokusíme donutit k využití kratšího období blížícího se polovině dostupných dat, predikční výsledky se výrazně zhoršují. Ačkoliv jsme v analýze nepracovali s méně než čtyřmi uzly, při dvou uzlech by model dosáhl ještě lepšího MSE = 156.5176, avšak za cenu ignorování struktury a trendu dat.

Graf 8: Nejlepší predikce lineárního splinu

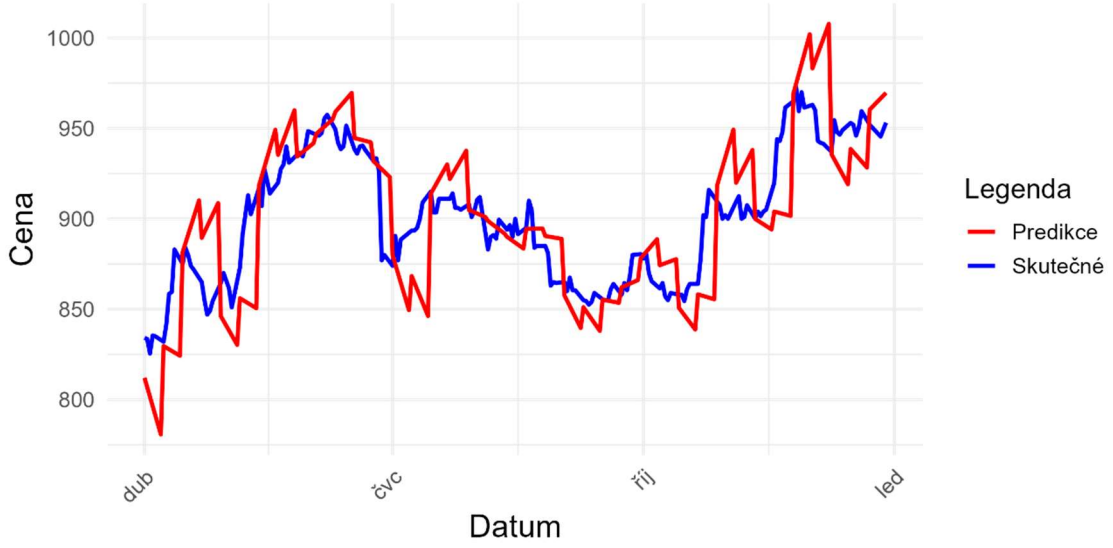


Zdroj: Data z portálu ČEZ, zpracování autor

Skript 4a dosahuje průměrné MSE predikce 398.7082, což ukazuje, že jakékoliv nastavení uzlů v rozsahu 4 až 6 vede k lepším výsledkům než lineární a nelineární metody aplikované na celý dataset. Skript 4b, který pracuje s optimálním počtem uzlů (viz graf 8) určeným na základě analýzy ve Skriptu 4a, vykazuje ještě lepší výsledky s nižšími hodnotami MSE predikce i tréninkové chyby.

Skript 5 prochází stejným rozsahem uzlů jako Skript 4a a následně počítá metriky pro sliding window s optimálním nastavením počtu uzlů. Výsledky viz. graf 9 ukazují nižší hodnotu MSE predikce (709.4892) než u předchozích metod. Optimální nastavení počtu uzlů na 6 a zvětšení tréninkového okna na 120 údajů vede k průměrnému MSE 560.3133, což je srovnatelné s výsledky predikcí neparametrických metod. Tento model dosahuje také nejnižšího průměrného AIC (673.472) při analýze celého roku, což dokazuje jeho efektivitu i přes zvýšenou komplexnost.

Graf 9: Sliding window - Skutečné hodnoty a hodnoty predikce lineárního splinu

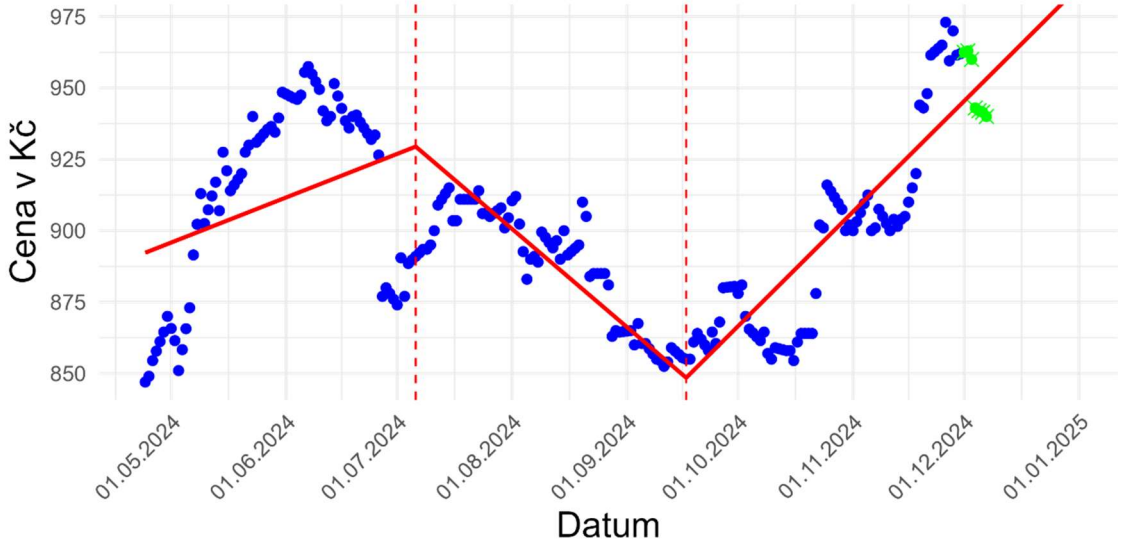


Zdroj: Data z portálu ČEZ, zpracování autor

V porovnání s lineární regresí, která vykazuje vyšší MSE predikce (969.7819) a AIC (791.5665), je model lineárního splinu efektivnější v přizpůsobivosti a zachycení trendu časové řady. Stejně tak nelineární regrese dosáhla lepších výsledků než lineární regrese, ale byla méně přesná než lineární spline. Výkon lineárního splinu však závisí na posledním uzlovém bodě, což znamená, že metody lineární regrese a lineárního splinu mohou vést k shodným výsledkům, pokud je pro lineární aproximaci použito stejné množství dat od posledního uzlového bodu.

Poznámka: Během analýzy bylo zjištěno, že pokusy o rozšíření nebo snížení počtu uzlů mimo rozsah 4 až 6 z většiny vedly k nespolehlivým výsledkům. To bylo způsobeno výběrem menšího tréninkového rozsahu, což vedlo k neadekvátnímu přizpůsobení modelu a zhoršení predikčních schopností. Optimální rozsah uzlů pro vyvážený výkon modelu je tedy 4 až 6.

Graf 10: Lineární regrese hledající nejlepší predikci při nastavení rozsahu knotů 2:8



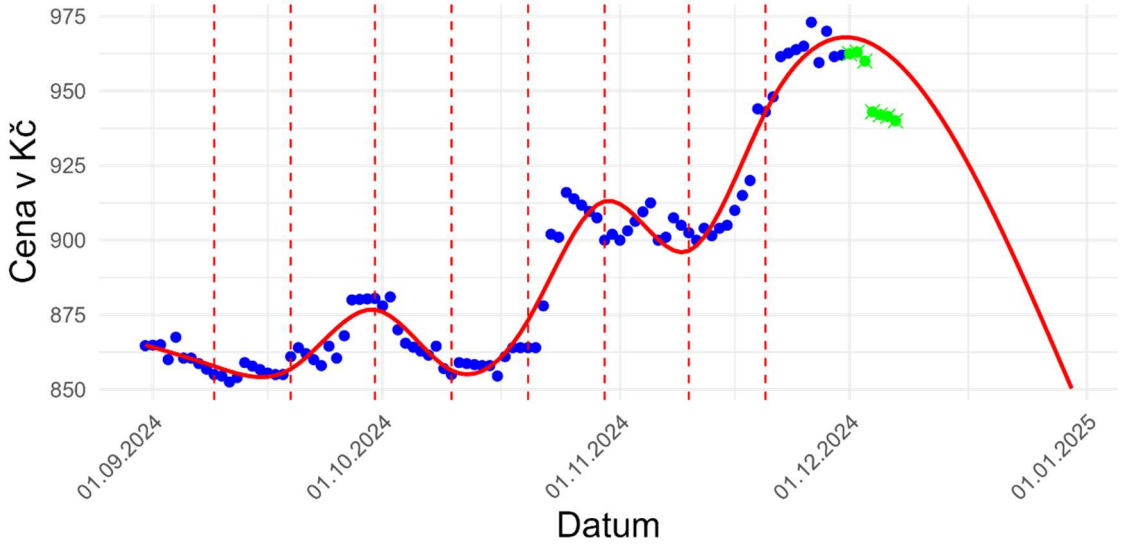
Zdroj: Data z portálu ČEZ, zpracování autor

Pokud by bylo lineárnímu splinu umožněno nastavení uzlů v rozmezí 2 až 8, došlo by k buď příliš hrubému zobecnění, které by ignorovalo podstatné trendy, nebo k nadměrnému přizpůsobení modelu, který by reagoval na náhodné fluktuace v datech. Při 2 uzlových bodech, viz. graf 10, model zachycuje pouze základní trend a ignoruje výrazné změny ve struktuře dat.

10.4. Vyhodnocení analýzy kubický spline

Při analýze kubického splinu byl zohledněn vliv počtu uzlů na predikční a tréninkovou chybu. Rozsah uzlů byl omezen na 2 až 6, protože při vyšší spodní hranici nebyl model schopen dosáhnout nižší predikční chyby než MSE = 600. Zvýšení horní hranice na osm uzlů také nevedlo k zlepšení, protože model při iterativním snižování dat ve skriptech 1 až 3 vždy volil nastavení osmi uzlů a nejmenší možné období pro tréninková data, což vedlo k přílišnému přizpůsobení tréninkovým hodnotám, jak můžeme vidět na grafu 11.

Graf 11: Ukázka kubického splinu při nastavení osmi uzlových bodů



Zdroj: Data z portálu ČEZ, zpracování autor

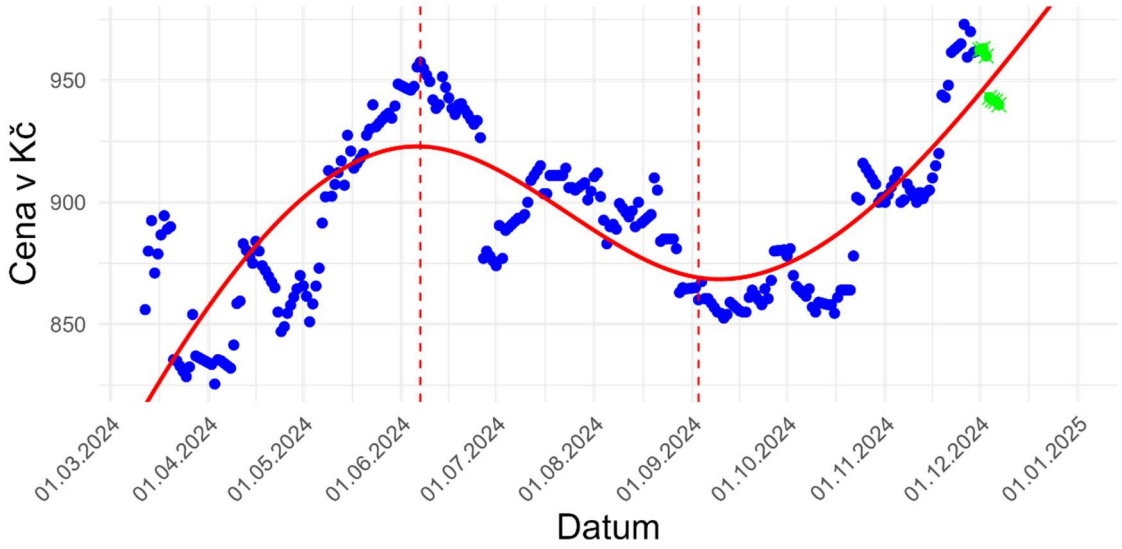
Tabulka 4: Výsledky analýzy kubického splinu

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Počet knotů
Skript 1	169.6475	1187.532	485.9871	128300.6	2401.928	264	2
Skript 2	4985.96	34901.72	73.05218	8254.896	829.8846	113	6
Skript 3	169.6475	1187.532	485.9871	128300.6	2401.928	264	2
Skript 4a	1220.47	8543.289	592.3088	197831.1	3023.342	334	2 - 6
Skript 4b	602.9629	4220.741	356.2373	118983.3	2924.3	334	4
Skript 5	715.9007	5011.305	186.0998	16748.98	725.3117	334	2 - 6 Výběr: 3

Zdroj: Autor

Nejlepších predikčních výsledků bylo dosaženo při použití dvou uzlů (Skript 1), kde MSE predikce dosáhlo hodnoty 169.6475 a MSE tréninkové chyby 485.9871. Vyšší hodnota AIC (2401.928) potvrzuje, že kubický spline je složitější metodou než lineární spline, což je očekávané vzhledem k vyššímu počtu parametrů. Skript 3 potvrzuje efektivitu tohoto nastavení pro predikci a přizpůsobení se tréninkovým datům (viz Graf 12).

Graf 12: Nejlepší predikce kubického splinu

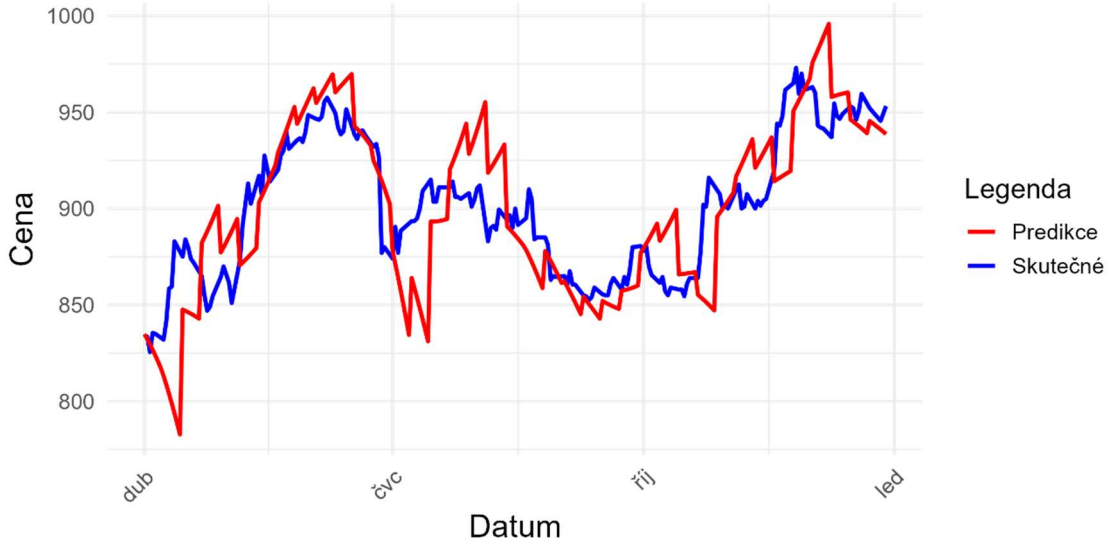


Zdroj: Data z portálu ČEZ, zpracování autor

Skript 4a vykázal vyšší predikční chybu (MSE = 1220.47) než lineární spline. Změnění rozsahu uzlů na 4 až 6, jako u lineárního splinu, nepřineslo zlepšení (MSE = 901.7334) a při nastavení 4–8 uzlů se MSE ještě mírně zvýšilo (MSE = 999.5695). Skript 4b, který používá optimální nastavení uzlů, dosáhl hodnoty MSE predikce 602.9629, což je přibližně trojnásobná chyba lineárního splinu při této metodice a hodnoty AIC 2924.3. Tento model se ukazuje jako relativně složitý a nedosahuje přesnosti potřebné pro modelování na velkém datovém souboru.

Skript 5, testující model s optimálním nastavením tří uzlů, dosáhl MSE predikce 715.9007, což je téměř shodné s hodnotou lineárního splinu ve Skriptu 5 (709.4892). Tento výsledek je lepší než u lineární a nelineární regrese, ale horší než u lineárního splinu. Kubický spline vykázal horší výsledky ale i v metrikách, jako je MSE tréninkových dat a AIC, přičemž hlavním faktorem byla vysoká volatilita v březnu, dubnu a červenci. V těchto měsících byly hodnoty MSE tréninkového data kubického splinu výrazně vyšší (např. 287.55, 259.31, 375.02 a 351.69) než u lineárního splinu (194.14, 184.94, 215.80 a 182.56) viz. Graf 13. Výsledky kubického splinu se nezlepšily ani při experimentování s rozsahem tréninkového okna.

Graf 13: Sliding window - Skutečné hodnoty a hodnoty predikce kubického splinu



Zdroj: Data z portálu ČEZ, zpracování autor

10.5. Vyhodnocení analýzy LOESS

Tabulka 5: Výsledky analýzy LOESS

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Span	Jádrová funkce
Skript 1	138.8619	972.0331	302.2691	70428.7	2024.556	233	0.3	Trikubické
Skript 2	405.7432	2840.202	29.02716	2670.498	471.1923	92	0.3	Obdélníkové
Skript 3	151.5921	1061.145	128.6171	18649.47	1063.211	189	0.5	Trikubické
Skript 4a	713.6064	4995.245	391.9951	130926.4	2960.705	334	0.3 – 0.7	Gaussovo
Skript 4b	193.275	1352.925	449.0068	149968.3	3065.626	334	0.6	Gaussovo
Skript 5	921.4778	6369.861	55.2243	4970.187	530.1324	334	0.3 – 0.7 Výběr: 0.3	Exponenciální
Skript 5	829.1722	4587.699	165.794	14921.46	675.9581	334	0.3 – 0.7 Výběr: 0.7	Obdélníkové

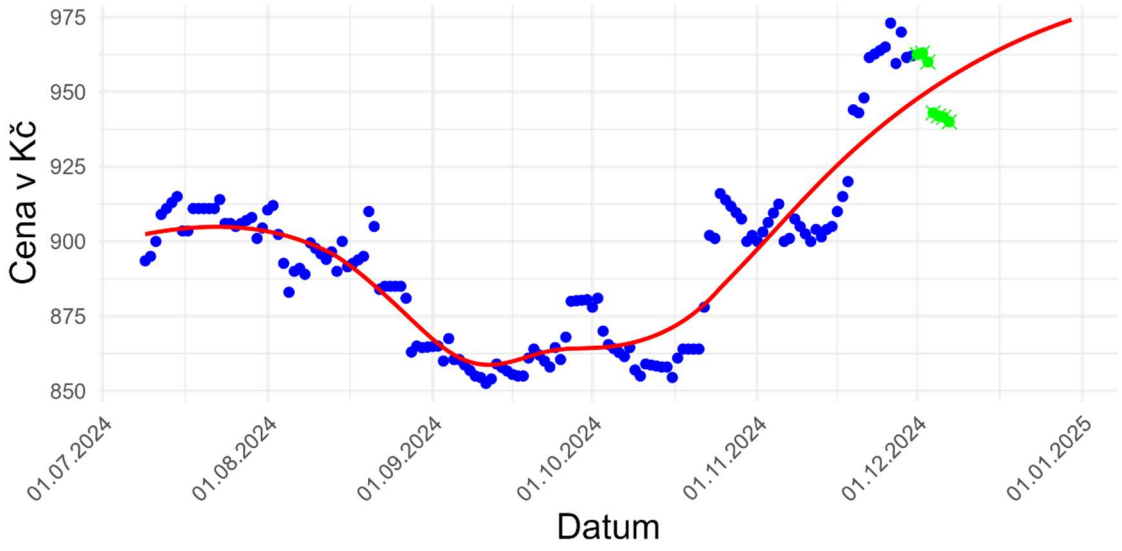
Zdroj: Autor

V rámci analýzy LOESS jsme testovali vliv parametrů na predikční a tréninkovou chybu. Zaměřili jsme se na velikost lokálního okna (span) v rozsahu 0.3 až 0.7 a efektivitu různých jádrových funkcí, které určují váhy bodů podle jejich vzdálenosti od bodu predikce. Zkoumali jsme trikubickou jádrovou funkci, exponenciální jádrovou funkci, obdélníko-

vou jádrovou funkci, Gaussovu jádrovou funkci a lineární jádrovou funkci. V rámci LOESS jsme se soustředili na kvadratickou aproximaci (degree = 2), protože metoda LOWESS využívá pouze lineární aproximaci (degree = 1). Pokud bychom v LOESS použili také lineární aproximaci, výsledky by byly v některých případech shodné s těmi, které poskytuje LOWESS.

Tabulka 5 ukazuje, že nejlepších predikčních výsledků (MSE = 138.8619) jsme dosáhli při span = 0.3 a použití trikubické jádrové funkce pro váhování v rámci lokální regrese. Tato jádrová funkce se osvědčil i ve skriptu 3, který optimalizoval vyvážení mezi predikční přesností (MSE = 151.5921) a kvalitou fitu (MSE = 128.6171), což je skutečně lepší hodnota než u nastavení skriptu 1 (MSE = 302.2691). Skript 3 pracoval s menším datovým rozsahem (189 záznamů) a použil span = 0.5, což vedlo k nejlepším výsledkům skriptu 3 ze všech metod (viz Graf 14). LOESS se jeví jako efektivní metoda pro přizpůsobení modelu různým datovým rozsahům při zachování vysoké predikční účinnosti.

Graf 14: Nejlepší predikce a fit metody LOESS



Zdroj: Data z portálu ČEZ, zpracování autor

Schopnost metody přizpůsobit se datům, byla otestována skriptem 2, jehož metodika vedla k výsledkům chybovosti fitu (MSE = 29.02716) a predikční chyby (MSE = 405.7432), při použití obdélníkového jádrové funkce a rozsahu tréninkových dat 92 záznamů. Tyto výsledky mohou naznačovat přefitování modelu, ale i při použití celého rozsahu dat a

různých hodnotách spanu, dosahuje LOESS průměrného MSE fitu 391.9951 (viz Tabulka 5, Skript 4a). Obě hodnoty MSE fitu jsou nejnižší v porovnání s výsledky ostatních metod.

V analýze vlivu parametru span ve skriptu 4a jsme zjistili, že se s rostoucím spanem zvyšuje chyba fitu, zatímco predikční chyba klesá, až do hodnoty span = 0.7, kde se tento trend začíná narušovat. Například při span = 0.3 byla MSE predikce 1206.358, zatímco při span = 0.6 klesla na 193.275 (viz Tabulka 6). Tyto výsledky ukazují, že vyšší hodnoty spanu pomáhají modelu lépe se přizpůsobit trendům v datech.

Tabulka 6: Skript 4a pro LOESS - MSE výsledky s různým nastavením spanu

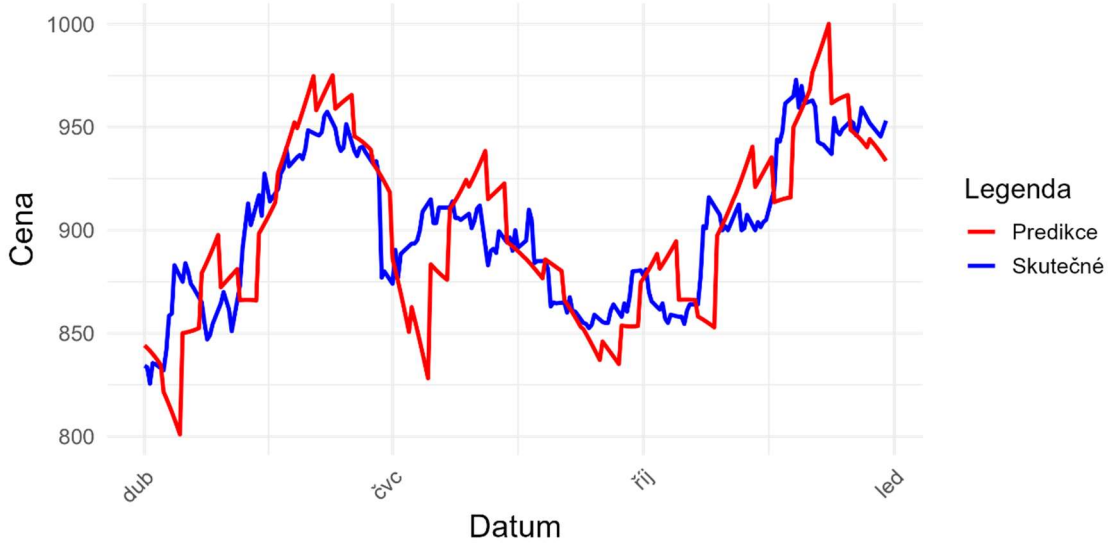
MSE predikce	MSE predikce	Span
1206.358	228.9019	0.3
1171.633	267.644	0.4
566.6114	350.9151	0.5
193.275	449.0068	0.6
430.1551	663.5077	0.7

Zdroj: Autor

Skript 4b využívá optimální hodnotu parametru span = 0.6, kterou určil skript 4a, a jako nejvhodnější váhovou funkci vybral Gaussovu jádrovou funkci. I za těchto podmínek metoda LOESS překonává parametrické metody. Skript 1, který iterativně přizpůsoboval rozsah dat, dosáhl podobných výsledků jako skript 4a.

Nejlepší nastavení parametrů LOESS v rámci skriptu 5 poskytuje průměrnou predikční chybu (MSE = 829.1722) a průměrnou chybu fitu (MSE = 165.794) za celý rok (viz. graf 15). Ačkoli tyto hodnoty nejsou nejlepší ve srovnání s výsledky jiných metod, zvýšená chybovost je způsobena příliš malým trénovacím oknem. Při rozšíření tréninkového okna o pouhých 30 záznamů (na 120 záznamů) dokáže LOESS dosáhnout nejnižších metrik chybovosti, nebo-li nejlepších výsledků, v porovnání s parametrickými metodami (viz Tabulka 7).

Graf 15: Sliding window - Skutečné hodnoty a hodnoty predikce metody LOESS



Zdroj: Data z portálu ČEZ, zpracování autor

Tabulka 7: LOESS - Výsledky skriptu 5 při zvětšení tréninkového okna

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Span	Jádrová funkce
Skript 5	697.6083	4366.872	203.4525	24414.3	947.476	334	0.7	Obdélníkové

Zdroj: Autor

Pokud se podíváme na hodnoty AIC, zjistíme jejich nárůst u všech skriptů LOESS v porovnání s jinými metodami. Tento nárůst potvrzuje vyšší komplexnost neparametrických metod, která vede k větší penalizaci za složitost modelu. I přesto, že metoda preferuje širší rozsah dat, její predikční účinnost zůstává lepší než u parametrických metod, a to i při menším rozsahu dat, jak ukázaly výsledky skriptu 2.

Zajímavé je sledovat, jak nastavení parametrů metody LOESS ovlivní výsledky. To můžeme pomocí tabulky 8, kde jsou zapsány druhé nejlepší výsledky dosažené s jiným nastavením parametrů. V porovnání s předchozími výsledky (viz Tabulka 5) vykazuje skript 4a vyšší průměrnou chybu predikce na celých datech při různých nastaveních. Avšak při optimálním nastavení spanu na 0.7 a použití obdélníkové jádrová funkce dosahuje skript 4b ještě lepších výsledků. Je tedy nutné brát v úvahu, že volba váhové funkce, může vést ke zlepšení či zhoršení ve dvou různých rovinách, a to, pokud zkoumáme průměrnou chybovost metody nebo pokud se zaměřujeme pouze na optimální nastavení.

Tabulka 8: Druhé nejlepší výsledky analýzy LOESS

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Span	Jádrová funkce
Skript 1	185.9741	1301.818	481.6184	149783.3	2892.65	311	0.7	Exponenciální
Skript 2	615.3748	4307.624	30.59188	2814.453	478.4858	92	0.3	Exponenciální
Skript 3	306.1104	2142.773	152.5738	14036.79	701.6133	92	0.3	Lineární
Skript 4a	781.0292	5467.204	349.8577	116852.5	2917.186	334	0.3 – 0.7	Obdélníkové
Skript 4b	183.3432	1283.402	529.7017	176920.4	3148.429	334	0.7	Obdélníkové
Skript 5	921.4778	6369.861	55.2243	4970.187	530.1324	334	0.3 – 0.7 Výběr: 0.3	Exponenciální

Zdroj: Autor

Zajímavý je i výběr jádrových funkcí. Nejlepší výsledky obvykle poskytuje exponenciální a obdélníkové jádro. To není překvapující, protože obdélníková funkce přiřazuje stejnou váhu všem bodům v rámci určitého intervalu a nulovou váhu bodům mimo tento interval, čímž eliminuje nadměrné přizpůsobení šumu. Exponenciální funkce je doporučována pro modelování finančních dat, protože přikládá větší váhu bodům blíže k bodu predikce, zatímco vzdálenější body mají menší vliv.

10.6. Vyhodnocení analýzy LOWESS

LOWESS je variantou LOESS s lineární aproximací (degree = 1). Použité jádrové funkce a nastavení parametrů zůstávají shodné s těmi u LOESS.

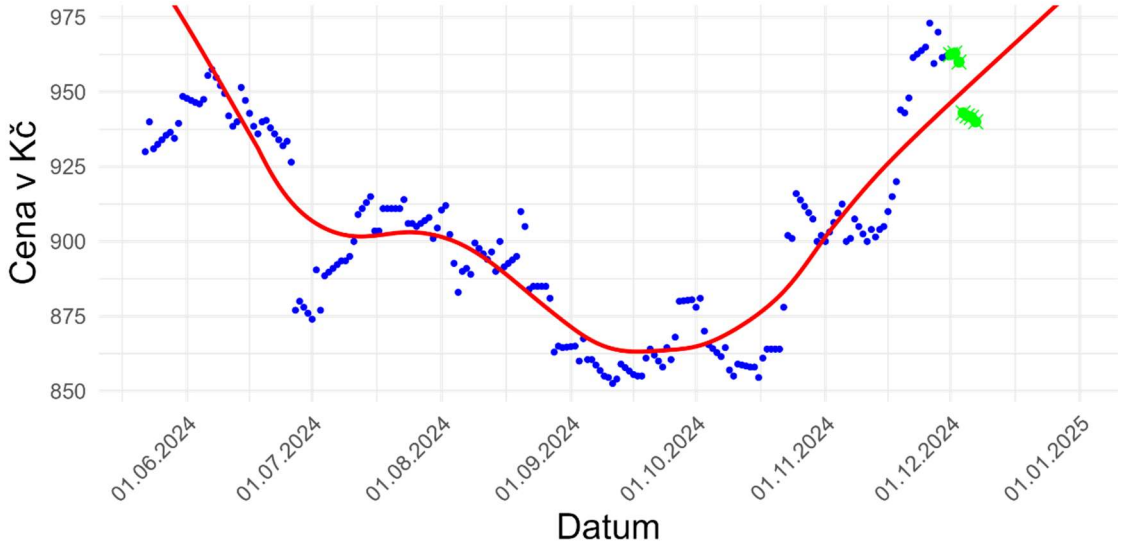
Tabulka 9: Výsledky analýzy LOWESS

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Span	Jádrová funkce
Skript 1	155.7083	1089.958	296.0635	57140.26	1676.568	193	0.3	Trikubické
Skript 2	2113.651	14795.56	63.81185	5870.69	579.5196	92	0.3	Obdélníkové
Skript 3	170.7882	1195.517	174.5764	22869.51	1020.523	131	0.6	Lineární
Skript 4a	1193.407	8353.848	851.4622	284388.4	3340.484	334	0.3 – 0.7	Obdélníkové
Skript 4b	203.3245	1423.271	386.9947	129256.2	2991.164	334	0.3	Obdélníkové
Skript 5	510.7004	3550.356	240.5339	21648.05	730.6819	334	0.3 – 0.7 Výběr: 0.7	Obdélníkové

Zdroj: Autor

Při porovnání výsledků skriptu 1 s LOESS pozorujeme mírné zhoršení predikční chyby (MSE = 155.7083, o 16.8467 vyšší než u LOESS), ale zároveň mírné zlepšení chybovosti fitu (MSE = 296.0635, o 6.2056 nižší než u LOESS). Tyty rozdíly jsou zanedbatelné. Obě metody preferovaly trikubickou jádrovou funkci a obdobný rozsah dat (193 pro LOWESS a 233 pro LOESS), přičemž ignorovaly volatilní období první poloviny roku 2024 a soustředily se na stabilnější druhou polovinu (viz Grafy 14 a 16).

Graf 16: Nejlepší predikce metody LOWESS



Zdroj: Data z portálu ČEZ, zpracování autor

Ve skriptu 2 metoda LOWESS vykazuje pětinásobně vyšší predikční chybu (MSE = 2113.651) a dvojnásobnou chybu fitu (MSE = 63.81185) oproti LOESS, přičemž obě metody používají stejné jádro (obdélníkovou funkci) a identické nastavení parametru span. LOWESS i přes to, poskytuje lepší výsledky než parametrické metody.

V metodice skriptu 3, LOWESS opět vykazuje zanedbatelnou vyšší chybu predikce o 19.1961 než u LOESS. Použitá jádrová funkce byla lineární, s nastavením spanu na 0.6. Jedinou výraznou změnou je, že LOWESS použila menší rozsah dat pro trénink, konkrétně o 58 hodnot, což naznačuje, že dokáže generovat dobré výsledky predikce i při menším tréninkovém souboru. Naopak LOESS se lépe přizpůsobuje finančním datům díky nelineárnímu modelování.

Tabulka 10: Druhé nejlepší výsledky analýzy LOWESS

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Span	Jádrová funkce
Skript 1	162.1218	1134.852	369.1855	105587	2542.079	286	0.4	Obdélníkové
Skript 2	2243.432	15704.02	63.36691	5829.756	578.557	92	0.3	Exponenciální
Skript 3	155.7083	1089.958	296.0635	57140.26	1676.568	193	0.3	Trikubické
Skript 4a	1800.717	12605.02	902.4374	301414.1	3364.849	334	0.3 – 0.7	Exponenciální
Skript 4b	178.4839	1249.388	391.2446	130675.7	2996.636	334	0.3	Exponenciální
Skript 5	550.4603	3808.152	100.2547	9022.927	609.3241	334	0.3 – 0.7 Výběr: 0.3	Exponenciální

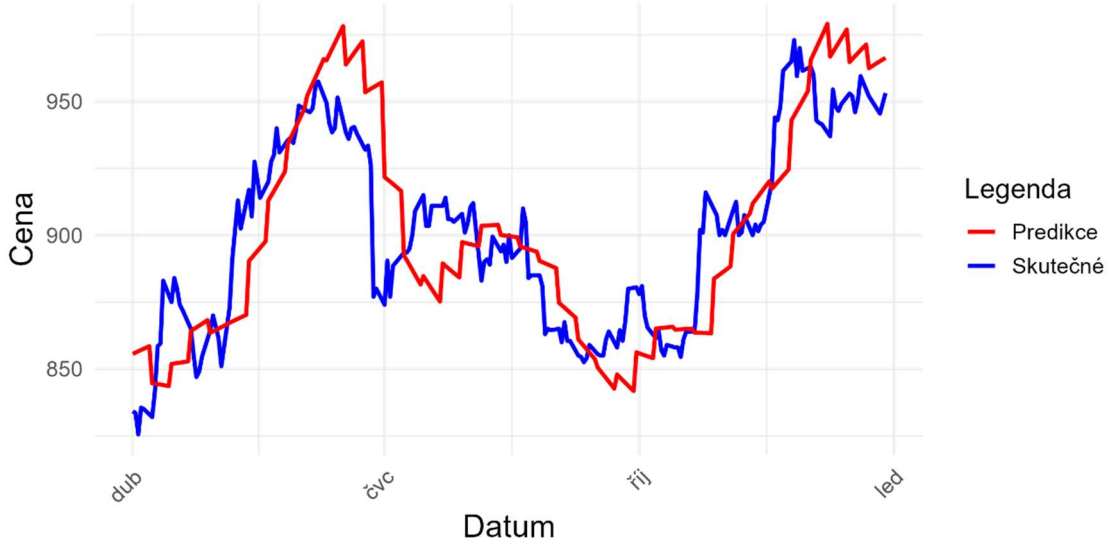
Zdroj: Autor

Tabulka 10 ukazuje, že druhé nejlepší výsledky pro LOWESS jsou téměř srovnatelné s nejlepšími výsledky, na rozdíl od LOESS, kde změny parametrů modelu vedli k výrazně větším změnám výsledků. Experimentální analýza ukazuje, že LOWESS si zachovává spolehlivost i při různých nastaveních parametrů.

U výsledků skriptu 4 je patrné, že LOWESS predikuje hůře než její protějšek s lokální kvadratickou aproximací o hodnotu MSE = 479.8006. Pokud však snížíme rozsah tréninkových dat na polovinu (167 záznamů), LOWESS vykazuje průměrnou chybovost MSE = 350.2511 což je trojnásobně nižší hodnota než u LOESS. Naopak u LOESS by chybovost při vynechání malého počtu počátečních záznamů vzrostla na MSE = 1019.515.

Při analýze metodikou skriptu 5, poskytovala LOWESS nejlepší výsledky při nastavení obdélníkové jádrové funkce s hodnotou span 0.7, stejně jako LOESS. Při tréninkovém okně 90 záznamů a průměrování predikcí za 39 týdnů dosáhla LOWESS průměrné chybovosti MSE = 510.7004 (viz graf 17). Kromě výsledků smoothing splines se jedná o výsledky, které se nejvíce blíží realitě z dosud popsaných metod. Nejbližší výsledky reálnému světu z parametrického přístupu poskytl lineární spline, s MSE = 709.4892 a MSE fitu 100.7363.

Graf 17: Sliding window - Skutečné hodnoty a hodnoty predikce metody LOESS



Zdroj: Data z portálu ČEZ, zpracování autor

Experimentování s rozsahem tréninkového okna ukazuje, že výsledky predikce LOWESS se zlepšují při zvětšování trénovací sady. Při 150 tréninkových datech dosahuje průměrné chybovosti MSE = 104.4917 (viz tabulka 11), což s jistotou naznačuje, že model poskytuje výsledky, které jsou v použitelné pro predikování v praxi, pokud jsou zvoleny optimální období a nastavení parametrů. S těmito výsledky můžeme říct, že LOWESS poskytuje dobré predikce jak při malých, tak velkých rozsazích dat.

Tabulka 11: LOWESS - Výsledky skriptu 5 při zvětšení tréninkového okna

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Span	Jádrová funkce
Skript 5	104.4917	731.4417	188.0035	28200.53	1172.245	334	0.3	Obdélníkové

Zdroj: Autor

Metoda také vykazuje stabilní výsledky při různých nastaveních parametrů. Exponenciální funkce a hodnoty span 0.3 vedly k druhým nejlepším výsledkům u obou metod, přičemž třetí nejlepší výsledky byly získány při použití Gaussovi funkce. Průměrná chybovost predikce pro LOWESS nepřesáhla MSE = 614.5649, zatímco u LOESS dosahovaly druhé nejlepší výsledky predikce hodnoty MSE = 921.4778.

10.7. Vyhodnocení analýzy smoothing spline

Poslední analyzovanou metodou jsou smoothing splines, což je neparametrická metoda, která se liší od lineárních a kubických splinů. Místo pevně definovaných uzlů využívá penalizaci složitosti křivky a automaticky vybírá optimální hladkou křivku, která nejlépe odpovídá datům. Parametr spar (normalizovaná penalizace) je nastavován v rozsahu 0.3 až 0.7, zatímco stupně volnosti a penalizační parametr λ jsou určeny metodou REML na základě maximální věrohodnosti. Výsledky predikcí pomocí smoothing splines vedou k závěru, že tato metoda je nejlepší z hlediska efektivnosti predikce budoucích hodnot, a že ruční nastavování parametrů není efektivní.

Tabulka 12: Nejlepší výsledky analýzy smoothing spline

	MSE predikce	SSE predikce	MSE tréninku	SSE tréninku	AIC	Počet dat	Spar	Df
Skript 1	168.7023	1180.916	262.6561	52793.88	1709.616	201	0.7	3.02
Skript 2	578.1921	4047.345	124.3662	17535.64	1098.12	141	0.3	3.417
Skript 3	168.7023	1180.916	262.6561	52793.88	1709.616	201	0.7	3.02
Skript 4a	530.5731	3714.012	689.0615	230146.5	3139.197	334	0.3 – 0.7	3.75
Skript 4b	284.0763	1988.534	553.474	184860.3	3069.824	334	0.3	4.178
Skript 5	439.1778	3074.244	181.8096	16362.86	722.9297	334	0.3 – 0.7 Výběr: 0.7	3.38

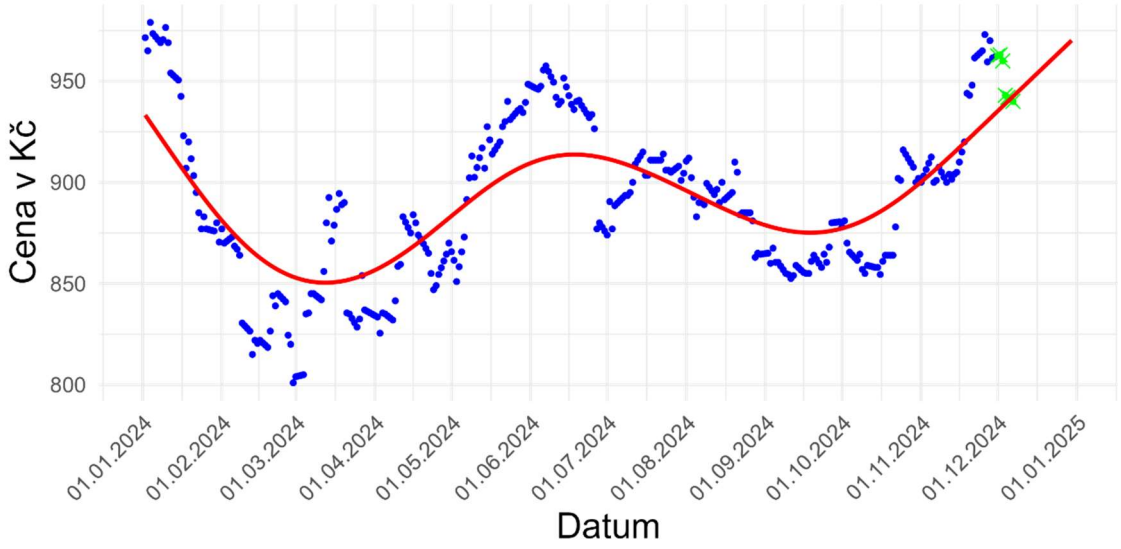
Zdroj: Autor

Porovnáme-li výsledky skriptu 1 s ostatními metodami, smoothing splines vykazují mírně horší predikční výsledky (MSE = 168.7023) než některé neparametrické metody, ale stále se jedná o nižší chybovost než výsledek jakékoliv parametrické metody. Tento výsledek byl dosažen při nízké chybovosti fitu (MSE = 262.6561) a nastavení spar na 0.7 a stupni volnosti 3.02. Stejně jako u metod LOESS a LOWESS výsledky smoothing splines jsou výrazně lepší při zvolení rozsahu tréninkových dat od poloviny dubna, kdy volatilita začíná klesat. Výsledky skriptu 1 odpovídají výsledkům skriptu 3.

Výsledky skriptu 2, byly srovnatelné s LOESS. Ačkoli skript 2 kladl důraz na minimalizaci chyby fitu (MSE = 124.3662), smoothing splines si udržely relativně nízkou predikční chybu (MSE = 578.1921), a to i při širším rozsahu tréninkových dat než jiné metody, které použily menší tréninková okna. To naznačuje, že smoothing splines dokážou dobře přizpůsobit model datům a zároveň udržet schopnost predikce budoucích hodnot.

Skript 4a testoval různé hodnoty spar (0.3 až 0.7) při modelování celého roku. Průměrné predikční chyby MSE = 530.5731 a průměrné chyba fitu MSE = 689.0615, bylo dosaženo při nastavení stupňů volnosti na 3.75. Jedná se o nejlepší odhady budoucích hodnot, při naivní aplikaci na celý rozsah dat, kromě odhadů lineárního splinu, který měl ještě nižší MSE o 131.8649. U této metody byla ojediněle průměrná chyba fitu vyšší než predikční chyba.

Graf 18: Smoothing spline aplikovaný na celých datech

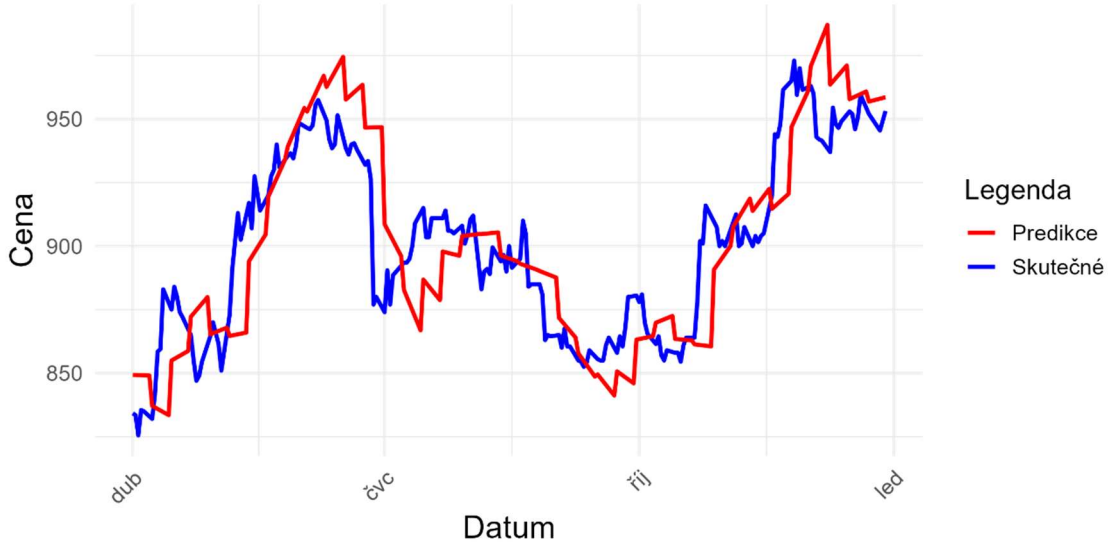


Zdroj: Data z portálu ČEZ, zpracování autor

Po nalezení optimální hodnoty spar (0.3) pro celý dataset skriptem 4a byly získány výsledky skriptu 4b: predikční chyba MSE = 284.0763 a chyba fitu MSE = 553.474 při stupních volnosti 4.178. Nicméně i přesto, že smoothing splines vykazují při optimálním nastavení parametrů vyšší chybovost než jiné metody, poskytují stabilnější výsledky při různých hodnotách spar, jak jsme mohli vidět u skriptu 4a.

Jak ukazuje graf 19, výsledky smoothing splines vykazují nejnižší průměrnou predikční chybu (MSE = 439.1778) a chybu fitu (MSE = 181.8096) ze všech analyzovaných metod při aplikaci metodiky sliding window na celý rok.

Graf 19: Sliding window - Skutečné hodnoty a hodnoty predikce smoothing spline



Zdroj: Data z portálu ČEZ, zpracování autor

Optimální nastavení parametrů modelu, pro dosažení těchto výsledů, je hodnota spar 0.7 a stupně volnosti nastavené na hodnotu 3.38. Jedná se o nejnižší chybovost při predikci 39 separátních týdnů ze všech analyzovaných metod. Jedinou metodou, která tyto výsledky předčila, byla metoda LOWESS při přenastavení tréninkového okna na 120 záznamů.

Dále experimentování s parametry ukázalo, že smoothing splines si udržují nízkou chybu predikce a dobré přizpůsobení datům i při různých hodnotách spar a velikosti tréninkového okna. Tyto výsledky ukazují, že smoothing splines jsou robustní metodou, schopnou efektivně modelovat a predikovat hodnoty i při různém nastavení parametrů.

11. Závěr

Regresní přístupy mohou být užitečné při predikci budoucího vývoje ekonomických časových řad, konkrétně cen akcií společnosti ČEZ. Různé metody však vyžadují odlišné nastavení vstupních parametrů i volbu rozsahu tréninkových dat. Proto je nezbytné upravit parametry modelu na základě experimentální analýzy testovacích dat, aby bylo dosaženo optimální rovnováhy mezi schopností modelu zobecňovat a jeho predikční efektivitou. Zároveň je důležité zohlednit charakter jednotlivých metod a podle toho stanovit realistický predikční horizont, ve kterém budou výsledky stále smysluplné.

Parametrické regresní přístupy obecně dosahovaly horších výsledků než neparametrické metody. Ačkoliv lineární regrese vykazovala relativně solidní predikční výkonnost, je třeba si uvědomit, že se jedná o velmi jednoduchý přístup, který pouze extrapoluje trend v datech, a proto je vhodný pouze pro krátkodobé predikce. Naproti tomu nelineární regrese se ukázala jako nevhodná pro predikci budoucích hodnot. Mezi parametrickými metodami se k výsledkům LOESS, jako nejméně efektivní neparametrické metodě, nejvíce přiblížily splinové metody, jelikož rozdělují časovou řadu na menší segmenty a v rámci nich provádějí lokální přizpůsobení pomocí polynomů. Ze spline přístupů dosáhl nejlepších výsledků lineární spline, který při optimálním nastavení uzlů a volbě tréninkového okna dosáhl třetí nejnižší chybovosti podle metodiky skriptu 5.

Z neparametrických metod se nejlépe osvědčila metoda smoothing splines, která dosáhla vysoké přesnosti díky optimalizaci hladkosti křivky prostřednictvím penalizačního parametru. Ten efektivně vyvažuje přesnost fitu a schopnost generalizace na nové hodnoty. Tato metoda dokázala zachytit hlavní trend časové řady, aniž by se nadměrně přizpůsobovala krátkodobým fluktuacím, což vedlo k nižší predikční chybě oproti ostatním neparametrickým přístupům. Srovnatelně dobré výsledky vykazovala i metoda LOWESS, především díky schopnosti přidělovat datům váhy prostřednictvím různých jádrových funkcí, přičemž nejvhodnější se ukázaly být obdélníková a exponenciální jádrová funkce. Zajímavé bylo také porovnání metod LOESS a LOWESS. V praxi se LOWESS ukázala jako robustnější jak při malém, tak při velkém rozsahu dat. Zároveň byla schopna poskytovat výrazně efektivnější predikce při optimálním nastavení tréninkového okna podle metodiky skriptu 5, čímž předčila všechny ostatní metody (viz tabulka 11).

12. Summary and keywords

This thesis is devoted to the application of non-parametric regression methods in modeling of economic time series. The stock prices of the Czech energy company ČEZ are analyzed using non-parametric methods, such as local regression, classical regression, and splines, to predict their values based on past observations over time. These methods are processed using the R programming language in a created software application, that applies these methods to real economic data. Emphasis is placed on comparing the prediction accuracy of individual methods and their suitability for different types of time series. The thesis aims to evaluate the predictive effectiveness of these methods in forecasting economic indicators and to provide recommendations for the most appropriate approaches in the analysis of economic time series.

Keywords: Non-parametric methods, time series, stock prices, prediction accuracy, prediction effectivenes

13. Seznam použitých zdrojů

Kozák, J., Hindls, R., & Artl, J. (1994). *Úvod do analýzy ekonomických časových řad* [Učební texty]. Vysoká škola ekonomická v Praze.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2008). Introduction to Time Series Analysis and Forecasting: WILEY SERIES IN PROBABILITY AND STATISTICS. John Wiley & Sons. Inc. Hoboken. New Jersey.

Křivý, I. (2012). ANALÝZA ČASOVÝCH ŘAD: Inovace výuky informatických předmětů ve studijních programech Ostravské univerzity [Skripta, Ostravská univerzita]. https://web.osu.cz/~Bujok/files/ancas.pdf

Cipra, T. (2013). FINANČNÍ EKONOMETRIE (druhé, upravené vydání). Ekopress.

Forbelská, M. (2009). STOCHASTICKÉ MODELOVÁNÍ JEDNOROZMĚRNÝCH ČASOVÝCH ŘAD. Masarykova univerzita.

Fox, J. (2015). *APPLIED REGRESSION ANALYSIS and GENERALIZED LINEAR MODELS* (third edition). SAGE Publications.

Mathews, J. H., & Fink, K. D. (1999). *Numerical Methods: Using MATLAB* (3.rd ed.). Prentice Hall.

Burden, R. L., & Faires, J. D. (2010). *Numerical Analysis* (9th). Cengage Learning.

Horová, I., & Zelinka, J. (2008). *Numerické metody* (2. vydání). Masarykova univerzita.

Crawley, M. J. (2013). *The R Book*. John Wiley & Sons.

R Core Team. (2024). R: A Language and Environment for Statistical Computing. Retrieved November 8, 2024, from https://www.R-project.org/

Xavier Bourret Sicotte. (2018). *Locally Weighted Linear Regression (Loess)*. Retrieved November 11, 2024, from https://xavierbourretsicotte.github.io/loess.html#Deriving-the-vectorized-implementation

Arlt, J., Arltová, M., & Rublíková, E. (2004). *Analýza ekonomických časových řad s příklady*. Oeconomica.

ČEZ a.s. (c 2025). *Vývoj cen akcií*. Retrieved February 6, 2025, from https://www.cez.cz/cs/pro-investory/akcie/vyvoj-cen-akcii

14. Seznam grafů

15. Seznam tabulek

16. Seznam příloh

Příloha 1: CD-R obsahující skripty pro analýzu v programovacím jazyce R 4.4.2