UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Modelování a predikce poptávky v supply chain managementu

Katedra matematické analýzy a aplikací matematiky
Vedoucí bakalářské práce: **Mgr. Ondřej Vencálek, Ph.D.**
Vypracovala: **Bc. Anna Bartolotti**
Studijní program: N1103 Aplikovaná matematika
Studijní obor: Aplikace matematiky v ekonomii
Forma studia: prezenční
Rok odevzdání: 2022

BIBLIOGRAFICKÉ IDENTIFIKACE

Autor: Bc. Anna Bartolotti

Název práce: Modelování a predikce poptávky v supply chain managementu

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

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

Rok obhajoby práce: 2023

Abstrakt: Diplomová práce se zabývá predikcí denní poptávky v maloobchodním prodeji. Cílem je vytvořit model pro poptávku a vypočítat predikce denních prodejů. Výsledky jsou porovnány se současným přístupem nástroje Veritico, který vychází z měsíčních predikcí. V první části práce je představena obecná problematika predikování poptávky v dodavatelských řetězcích. Následně jsou popsány matematické metody potřebné k sestavení modelu poptávky. Nakonec jsou tyto modely ověřeny na reálných datech a výsledky denních predikcí jsou porovnány s výsledky Veritica.

Klíčová slova: poptávka, forecast, predikce, časové řady, regresní analýza, ARIMA modely

Počet stran: 66

Počet příloh: 0

Jazyk: Český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Anna Bartolotti

Title: Demand modeling and forecasting in supply chain management

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: Mgr. Ondřej Vencálek, Ph.D.

The year of presentation: 2023

Abstract: This master's thesis deals with the prediction of daily demand in retail sales. The goal is to create a demand model and calculate daily sales prediction. The results are compared with the current approach of the Veritico tool which is based on monthly predictions. The general issue of prediction in supply chain management is presented in the first part of the thesis. The mathematical methods which are needed to create the demand model are described in next section. Finally, these models are validated on real data and the daily prediction results are compared with Veritico results.

Key words: demand, forecast, time series, regression analysis, ARIMA models

Number of pages: 66

Number of appendices: 0

Language: Czech

Prohlášení

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

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Seznam tabulek

Poděkování

Ráda bych poděkovala vedoucímu práce Mgr. Ondřeji Vencálkovi, Ph.D. za odborné vedení, ochotu a trpělivost zodpovědět mé dotazy a za veškerý jeho čas, který mi věnoval během vypracovávání mé diplomové práce. Dále bych chtěla poděkovat společnosti Logio s.r.o. za poskytnutí zajímavého tématu a potřebných dat k vypracování této práce.

Úvod

Mohutný rozvoj digitalizace v současné době umožňuje sběr velkého množství dat a následně jejich využití v oblastech rozhodování a plánování.

V řízení dodavatelských řetězců to není jiné a právě data se dají využít ke zpřesnění a zrychlení všech oblastí tohoto procesu. Tato diplomová práce se zabývá oblastí forecastingu, tedy předpovědí prodejů, která je pro účinný a plynulý průběh celého dodavatelského řetězce nezbytná.

Cílem práce je vytvoření modelu poptávky a vypočítat denní predikce prodejů produktů nabízených v maloobchodu s potravinami a základní drogérií. Výsledky těchto predikcí budou porovnány s predikcemi vypočítanými nástrojem Veritico od společnosti Logio, jehož jednou z hlavních funkcí je navrhnout správné výše objednávek. Předpověď poptávky je jedním ze stěžejních bodů takového výpočtu a zpřesnění predikcí by pomohlo ke zlepšení celého výpočetního procesu Veritica.

Kapitola 1

Teoretické pozadí problematiky

Vzhledem k současné globalizaci a celkové rychlosti dnešní doby si firmy velmi dobře uvědomují, že už nestačí provozovat a řídit podnikání jen na jednom místě. S přesunem a rozmístěním jednotlivých segmentů výroby do různých částí světa se dá ušetřit značné množství nákladů, ale o to větší důraz je třeba vkládat do efektivního plánování.

1.1. Dodavatelský řetězec

Anglický výraz *supply chain* můžeme podle [2] přeložit například jako *dodavatelský řetězec*, *dodavatelsko–odběratelský řetězec* nebo *plně integrovaný logistický řetězec*. V této práci budeme používat doslovný překlad *dodavatelský řetězec*, který patří k nejčastěji používanému výrazu v odborných publikacích.

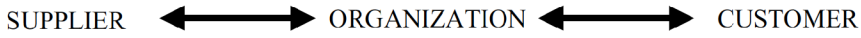
Hugos, 2018 [3] zmiňuje, že pro dodavatelský řetězec existuje mnoho definic, ale obecně bychom ho mohli podle [1] popsat jako soubor tří a více organizací, mezi kterými proudí materiální, informační a finanční toky. Jde o dynamické propojení všech subjektů, které přímo či nepřímo ovlivňují splnění požadavků koncového zákazníka. Zahrnuje veškeré procesy a aktivity podílející se na uvedení produktu nebo služeb na trh.

Je důležité, že tyto materiální, informační a finanční toky proudí mezi jednotlivými články řetězce oboustranně. Zahrnují tak například i vrácení zboží, reklamaci nebo servis. Jeden z hlavních principů dodavatelského řetězce je nabývání

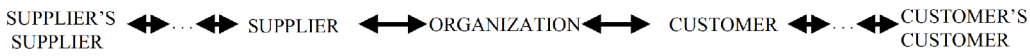
hodnoty produktu s každým dalším článkem řetězce, kterým prostupuje. [4]

Mezi klasickým logistickým konceptem a dodavatelským řetězcem je zásadní rozdíl. Logistický koncept se zaměřuje pouze na činnosti v rámci jedné organizace, zatímco dodavatelský řetězec pracuje s celou sítí subjektů, které mají za cíl dostat výsledný produkt ke koncovému zákazníkovi. [3]

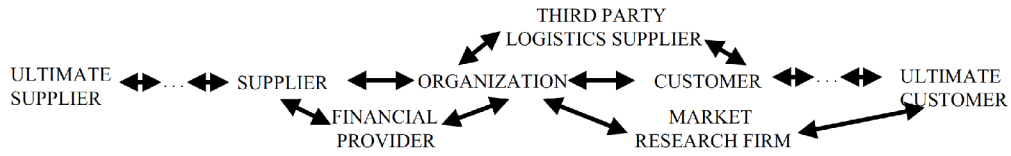
1.1.1. Obecné schéma dodavatelského řetězce



(a) Jednoduchý dodavatelský řetězec



(b) Rozšířený dodavatelský řetězec



(c) Kompletní dodavatelský řetězec

Obrázek 1.1: Schéma dodavatelského řetězce, zdroj: [1]

Dodavatelský řetězec se může skládat z různého množství článků, které do něj vstupují. Může se jednat o přímé činitele nebo o třetí strany, které na výsledný produkt působí nepřímo. [3]

Nejjednodušší schéma, které můžeme názorně vidět na obrázku 1.1a, je tvořeno ze tří členů – dodavatele, prodejce a zákazníka. [1] Může jít například o maloobchod, který odebírá zeleninu od přímého dodavatele a prodává ji konečnému spotřebiteli.

Rozšířený dodavatelský řetězec, na obrázku 1.1b, bude širší o další subdodavatele nebo zákazníky zákazníků. Jako příklad můžeme uvést velkoobchod, který své zboží nakupuje od dodavatelů, kteří mají také své dodavatele, a dál je prodává do maloobchodů, kde je prodají konečnému zákazníkovi.

Na posledním obrázku 1.1c můžeme vidět, jak moc komplexní může dodavatelský řetězec být. V takovém případě už do systému vstupují i třetí strany, což může být například finanční sponzor nebo marketingová agentura.

1.2. Řízení dodavatelských řetězců

Řízení dodavatelských řetězců je stále častěji využívanou moderní strategií řízení obchodu. Pro anglický výraz *supply chain management*, neboli řízení dodavatelských řetězců, existuje mnoho definic, které se drobně liší v jednotlivých odborných publikacích.

Obecně jde o proces řízení, jehož cílem je zajistit plynulý tok materiálu, dat a financí mezi dodavatelem a koncovým zákazníkem. Jednotlivé subjekty vstupující do řetězce jsou navzájem průběžně koordinovány a díky tomu je tak zajišťen fungující systém, ve kterém spolu články spolupracují.

V klasickém pojetí logistiky se každý z těchto částí řetězce soustředil na své cíle a nesdílel informace mimo svou organizaci, ale se zrychlováním globálního trhu a se zvyšující se konkurencí je pro všechny výhodnější začít podnikání propojovat. [4]

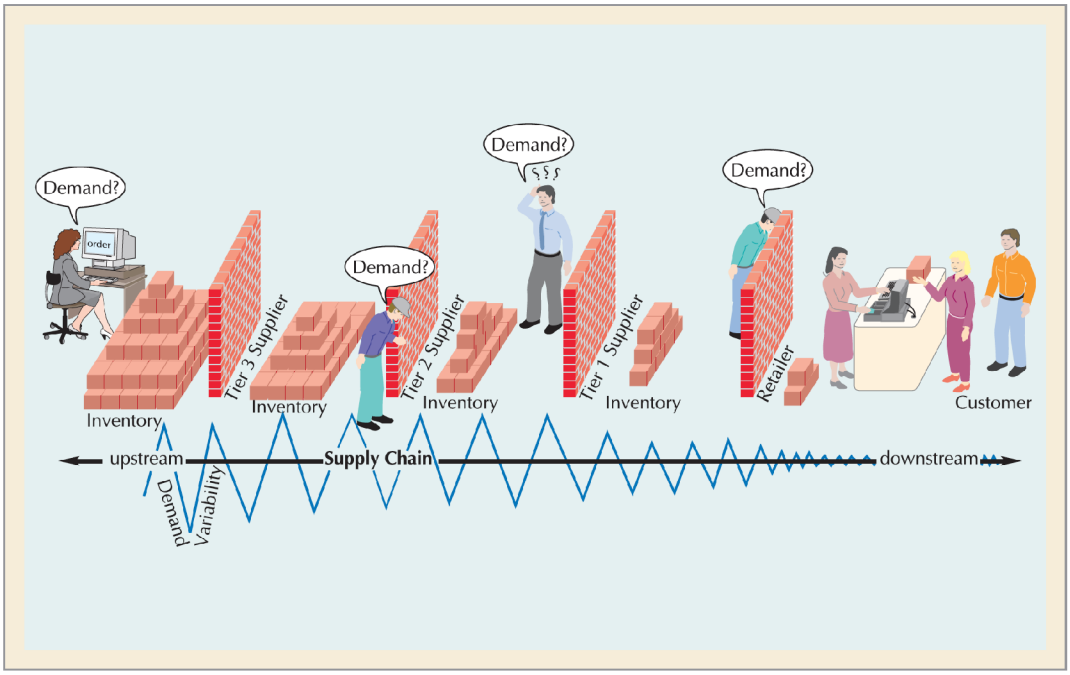
Podle [4] je cesta k efektivnímu řízení dodavatelských řetězců daná čtyřmi faktory: sdílení informací, komunikace, kooperace a důvěra mezi jednotlivými subjekty. Čím více aktuálních informací se mezi sebou v rámci řetězce sdílí, tím rychleji mohou jednotlivci reagovat na případné změny.

Pomocí efektivního využití zdrojů a vytvoření synergie mezi jednotlivými články dodavatelského řetězce tak lze maximalizovat hodnotu produktu současně s minimalizací nákladů.[2]

1.3. Predikce poptávky v řízení dodavatelských řetězců

Aby byl tok všech částí dodavatelského řetězce plynulý, musí na sebe jednotlivé proudy navazovat jak časově, tak hmotně. Plánování je tak nedílnou součástí procesu a základem pro většinu plánování v řízení dodavatelských řetězců je informace o budoucí poptávce. Teprve na základě těchto odhadů se tvoří plány na rozpočty, produkci, pracovní kapacity nebo transport zboží. [5]

V případě chybějícího odhadu budoucí poptávky může nastat tzv. *efekt biče*. Efekt biče popisuje situaci, kdy nepřesné informace o poptávce koncového zákazníka mohou způsobit mnohonásobně větší nejistotu v poptávce směrem zpět po dodavatelském řetězci. V důsledku toho to pro dodavatele znamená vytváření větších zásob, špatné plánování a zbytečné výdaje.



Obrázek 1.2: Efekt biče, zdroj: [4]

Na obrázku 1.2 je znázorněna situace, kdy poptávka koncového spotřebitele je poměrně stabilní a skladové zásoby tak mohou být malé. Při jakékoliv změně, o které ostatní subjekty řetězce nebudou dostatečně informováni předem, mohou

dodavatelé inklinovat ke zvyšování nebo naopak snižování zásob, aby tuto změnu zvládli vyvážit. Další články řetězce budou mít o to víc přehnanou reakci na změnu poptávky, čím se nacházejí blíž začátku dodavatelského řetězce. [4]

Chceme-li se efektu biče vyvarovat, je důležité, aby mezi sebou jednotlivé články dodavatelského řetězce sdílely informace, především ty o předpokládané poptávce koncového zákazníka. Díky tomu mohou eliminovat nejistotu a zbytečné výdaje.

Metody predikce poptávky

V oblasti řízení dodavatelských řetězců se používají různé metody pro predikci poptávky. Můžeme je rozdělit na kvalitativní a kvantitativní metody.

Kvalitativní metody jsou založené na úsudku experta, který zná daný trh a zákazníka a dokáže na základě zkušeností odhadnout hodnotu budoucí poptávky. Další možností je průzkum trhu a zákazníků pomocí dotazníků. Tyto metody jsou často využívány například v případě predikcí poptávky u nového produktu, o kterém nejsou dostatečně obsáhlá historická data.

Kvantitativní metody využívají k predikci dostupná data, která se týkají například historických prodejů, vlastností produktu nebo externích faktorů, které poptávku mohou ovlivňovat.

V rámci této diplomové práce se budeme zabývat pouze kvantitativními metodami, z nichž si vybrané detailněji představíme v kapitole 2 a následně je aplikujeme v praxi v kapitole 3.

1.4. Maloobchod a retail

Maloobchod je podnik, který nakupuje zboží od dodavatele, např. velkoobchodu nebo od výrobce, a dále ho prodává koncovému zákazníkovi – spotřebiteli. Je tak specifickou částí dodavatelského řetězce, jelikož tvoří poslední člen před koncovým zákazníkem.

Maloobchod musí pečlivě vyvažovat své zásoby, aby uspokojil potřeby svých zákazníků a zároveň u toho eliminoval nadbytečné náklady na skladování. Zna-

mená to dostatečně rozmanitý sortiment, který pokryje široký výběr druhů zboží napříč cenovými kategoriemi, a zároveň dostatečně a efektivně naplněné sklady, které v těchto typech obchodů nebývají příliš velké. [5]

Maloobchodní prodej můžeme dělit na:

- **Potravinářský maloobchod**

  Maloobchod, který nabízí převážně potraviny, označujeme jako potravinářský, ačkoliv sortiment často zahrnuje i nepotravinářské zboží. Jde často o takové produkty, které jsou nakupovány na denní bázi, a proto je tento typ maloobchodu velmi rozšířený.

- **Nepotravinářský maloobchod**

  Do nepotravinářského maloobchodu spadá velké množství typů prodeje, například prodej elektroniky, oblečení nebo automobilů.

Maloobchodní společnosti, které mají zajištěný komplexní dodavatelský řetězec, se nazývají *retail*. Mnoho takových společností například vlastní svůj velkoobchod, dopravu a další části dodavatelského řetězce, které jim usnadní celý proces. [6]

Ačkoliv je poptávka v maloobchodě nestabilní a není možné ji předpovědět se stoprocentní jistotou, její predikce je nejdůležitější informací pro všechy ostatní články dodavatelského řetězce, protože poptávka koncového zákazníka ovlivňuje veškeré plánování dalších subjektů.

V případě, že by predikce poptávky byla velmi nepřesná, pro maloobchod může nastat jedna z následujících situací:

- **Pesimistický/podhodnocený odhad**

  Odhad, který poptávku spíše podhodnocuje a vrací menší hodnotu poptávky, než jaká ve skutečnosti nastane, způsobí vyprodání zásob. Díky tomu

by společnost přišla o ušlé zisky, případně o zákazníky, kteří by se rozhodli obrátit se na konkurenci s dostatečně naskladněným zbožím.

- **Optimistický/nadhodnocený odhad**

  V případě optimistického neboli nadhodnoceného odhadu může nastat situace, kdy výše poptávky bude ve skutečnosti menší, než jaká byla predikována, a na skladě zůstane značné množství neprodaných kusů. Takový případ způsobí například zbytečné výdaje za skladování nadměrných zásob a v případě krátké trvanlivosti zboží také výdaje ve formě odpisů.

Poptávka je ovliňována velkým množstvím externích faktorů a se zvyšujícím se množstvím značek a druhů zboží má zákazník daleko větší výběr mezi více produkty stejného typu. Navzdory dnešním pokročilým výpočetním technikám a matematickým metodám tak predikování poptávky nebylo nikdy těžší.

Kapitola 2

Matematické metody

V této kapitole si představíme metody a pojmy potřebné pro analýzu dat a modelování poptávky, které využijeme v kapitole 3. Informace k této kapitole jsme čerpali především ze zdrojů [15], [14], [13], [11], [12], [17].

2.1. Základní pojmy

Standardizace

Standardizace je jednou z mnoha možností jak transformovat data. Každou transformaci je důležité provádět s rozmyslem, jelikož tím z dat odebereme informaci o jednotce a převedeme tak různé proměnné na stejné měřítko.

Nechť $\overline{x}$ je průměrná hodnota proměnné a s_x je její směrodatná odchylka. Standardizaci na nulový průměr a jednotkovou směrodatnou odchylku pak provedeme následující operací:

$$x_{std} = \frac{x - \overline{x}}{s_x}$$

Akaikeho informační kritérium

Informační kritéria slouží pro porovnání různých modelů, takže samy o sobě nám nedávají žádnou informaci a musíme je vždy srovnávat s hodnotou informačního kritéria konkurenčního modelu. Informační kritéria se snaží vybalancovat vychýlení odhadů a rozptyl parametrů. Jsou vypočítávána z maxima věrohodnostní funkce, které se používá k nalezení odhadu parametru θ. Cílem je

přitom maximalizovat pravděpodobnost (reps. věrohodnost), že hodnoty pocházejí z předpokládaného rozdělení.

Akaikeho informační kritérium je definováno následovně:

$$AIC = -2lnL(\widehat{\theta}) + 2q$$

kde q je počet odhadovaných parametrů v modelu, $\widehat{\theta}$ je maximálně věrohodný odhad parametru θ a $L(\widehat{\theta})$ je hodnota maxima věrohodnostní funkce.

Časová řada

Časová řada $y_1, ..., y_t$ je soubor pozorování chronologicky uspořádaných v čase. Pozorování y_t jsou realizace posloupnosti náhodných veličin Y_t v čase t, které svými vlastnostmi klasifikují výslednou časovou řadu na intervalovou nebo okamžikovou. $\{y_t, t \in T\}$.

2.1.1. Klouzavé průměry

Metoda klouzavých průměrů vychází z předpokladu, že každá „rozumná“ funkce může být aproximována polynomem. Vyrovnáním časové řady klouzavým průměrem potlačíme krátkodobá kolísání a jsme tak schopni vidět dlouhodobý vývoj.

Jednoduché klouzavé průměry aproximují časovou řadu lokálním lineárním trendem. Pro zvolenou délku časového okna $2m+1$ proložíme data lineární funkcí

$$y_{t+\tau} = \beta_0(t) + \beta_1(t)\tau, \qquad \tau = -m, \ldots, m$$

Z formulace jednoduchého klouzavého průměru je zřejmé, že v tomto případě časovou řadu vyhlazujeme pomocí prostého aritmetického průměru hodnot ze zvoleného časového okna.

Parametry β_0 a β_1 odhadujeme pomocí metody nejmenších čtverců

$$\min \sum_{\tau=-m}^{m} (y_{t+\tau} - \beta_0 - \beta_1\tau)^2$$

Podrobný postup odhadu parametrů najde čtenář v příslušné literatuře [13].

Vyhlazená hodnota přes toto okno se pak vypočítá pomocí:

$$\widehat{T_t} = \frac{1}{2m+1} \sum_{\tau=-m}^{m} y_{t+\tau}.$$

2.2. Regresní analýza

Regresní analýza slouží k nalezení a kvantifikování vztahu mezi proměnnými. Díky ní můžeme vysvětlit hodnoty závisle proměnné pomocí hodnot vysvětlujících proměnných. V této kapitole si popíšeme, jak vytvořit lineární regresní model.

Zavedeme si značení:

Y_i	i-té pozorování závisle proměnné
x_{ik}	k-tá vysvětlující proměnná příslušná i-tému pozorování (nezávisle proměnná)
$\beta_0, \beta_1, \ldots, \beta_k$	neznámé regresní parametry
ε_i	i-tá náhodná odchylka
p	počet neznámých regresních parametrů

2.2.1. Klasický model vícenásobné lineární regrese

Vícenásobnou lineární regresi využijeme v případě, kdy chceme vysvětlovanou proměnnou popsat více vysvětlujícími proměnnými.

Formálně můžeme lineární regresní model s k vysvětlujícími proměnnými zapsat takto:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n$$

Předpokládáme, že pro vektor náhodných odchylek platí

- $E(\varepsilon_i) = 0, \quad \forall i$
- $var(\varepsilon_i) = \sigma^2, \quad \forall i$
- $cov(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j$

Dále předpokládáme, že

- $n > p$ ... (počet pozorování je větší než počet neznámých regresních parametrů),
- $h(\mathbf{X}) = p$ ... (sloupce matice $\mathbf{X}$ jsou lineárně nezávislé).

Nechť $\mathbf{Y}$ je $(n \times 1)$ náhodný vektor vysvětlované proměnné o n pozorování

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix},$$

$\mathbf{X}$ je $(n \times p)$ matice vysvětlujících proměnných, kterou nazýváme matice plánu

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix},$$

$\boldsymbol{\beta}$ je $(p \times 1)$ vektor neznámých regresních parametrů, kde $p = k + 1$ v případě modelu s absolutním členem a $p = k$ v případě modelu bez absolutního členu

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

a $\boldsymbol{\varepsilon}$ je $(n \times 1)$ vektor náhodných chyb

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Maticový zápis lineárního regresního modelu pak bude mít tvar:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Předpokládáme, že pro vektor náhodných odchylek platí:

- $E(\boldsymbol{\varepsilon}) = \mathbf{0}_n$
- $var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

2.2.2. Odhad regresních parametrů

K odhadu regresních parametrů $\boldsymbol{\beta}$ se nejčastěji používá metoda nejmenších čtverců. Ta hledá odhady parametrů pomocí minimalizace součtu druhých mocnin odchylek od skutečné hodnoty.

Označme si $\widehat{Y}_i$ jako odhad i-té hodnoty vysvětlované proměnné Y:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_k x_{ik}.$$

Rozdíl odhadu a skutečné hodnoty nazveme rezidui, které vypočítáme jako

$$e_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_k x_{ik}).$$

Úloha minimalizace součtu čtvercové chyby má pak tvar

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} e_i^2 = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})\right)^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

a řešení splňuje soustavu normálních rovnic

$$(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}.$$

Odhad $\widehat{\boldsymbol{\beta}}$ parametrů $\boldsymbol{\beta}$ určíme jako

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'\mathbf{Y}$$

Vlastnosti odhadů

V případě, že model splňuje předpoklady zmíněné v kapitole 2.2.1, odhad $\widehat{\boldsymbol{\beta}}$ parametrů $\boldsymbol{\beta}$ je

1. nestranným odhadem:

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

2. nejlepším nestranným lineárním odhadem:

$$var(\widetilde{\boldsymbol{\beta}}) - var(\widehat{\boldsymbol{\beta}}) \text{ je pozitivně semidefinitní matice pro každý jiný nestranný lineární odhad } \widetilde{\boldsymbol{\beta}}$$

2.2.3. Umělé proměnné

Umělé proměnné v regresní analýze mohou zastupovat kvalitativní proměnné nebo kvantitativní proměnné agregované do skupin. Příkladem kvalitativní proměnné je například pohlaví nebo dosažené vzdělání. Agregované kvantitativní proměnné mohou popisovat například věkové skupiny rozdělené po 10 letech, tj. věk 0 až 9, 10 až 19, 20 až 29, atd. Díky této transformaci můžeme kategorické proměnné využít pro tvorbu regresního modelu a odhadnout parametry pomocí metody nejmenších čtverců.

Uvažujme umělou proměnnou p, která označuje, zda pozorování spadá do dané kategorie, nebo ne. Jako názorný příklad můžeme uvažovat kategorii žena, kdy umělá proměnná bude nabývat hodnoty 1 v případě, že se jedná o ženu, a naopak hodnoty 0, pokud se jedná o muže.

$$p_i = \begin{cases} 1 & \text{i-té pozorování je žena} \\ 0 & \text{i-té pozorování je muž (referenční kategorie)} \end{cases}$$

Model lineární regrese, kde vysvětlovaná proměnná y závisí na vysvětlující proměnné x a umělé proměnné p, by pak měl podobu:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 p_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

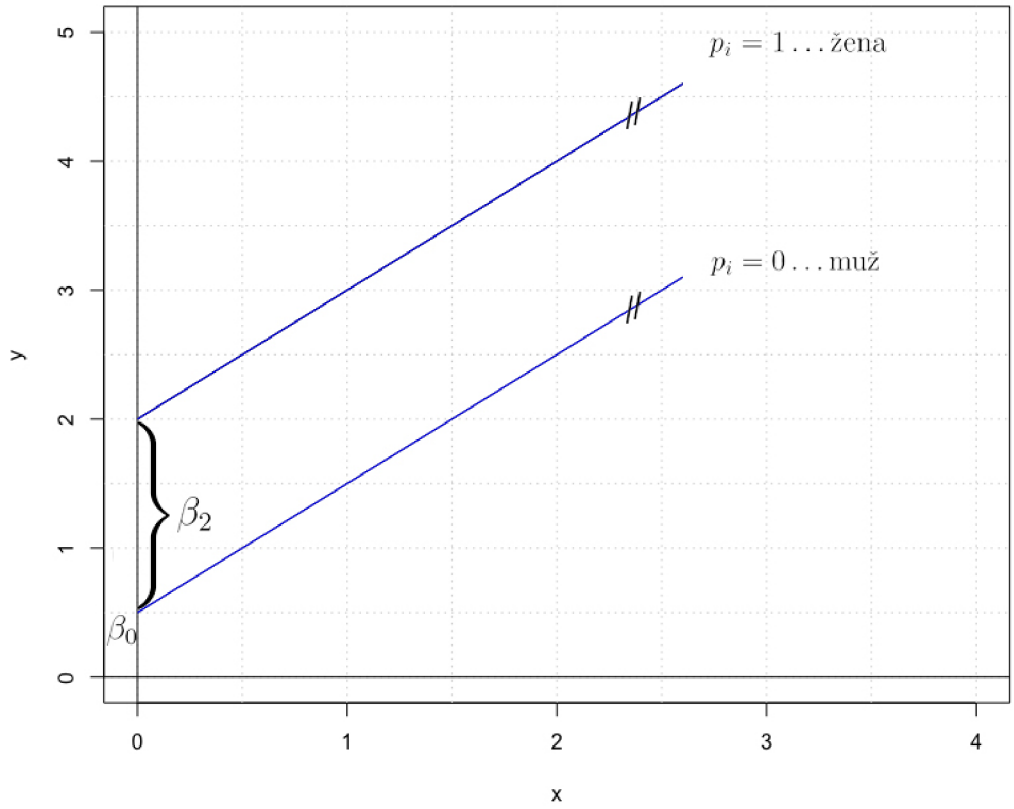
Pro jednotlivé kategorie nám po dosazení za hodnotu p_i vzniknou různé modely:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{pro muže}$$

$$Y_i = (\beta_0 + \beta_2) + \beta_1 x_i + \varepsilon_i \qquad \text{pro ženu}$$

Očekává se, že regresní přímka pro obě kategorie bude mít stejný sklon, který je daný parametrem β_1. V případě jednotkové změny proměnné x se střední

hodnota vysvětlované proměnné změní právě o velikost parametru β_1. Při dané hodnotě x je rozdíl mezi očekávanou hodnotou vysvětlované proměnné pro muže a ženy daný parametrem β_2. Vizuálně si interpretaci parametrů naznačíme na obrázku 2.1.



Obrázek 2.1: Interpretace umělých proměnných

V případě kategoriální proměnné s více než dvěma kategoriemi postupujeme obdobně. Obecně platí, že v případě m kategorií tvoříme $(m-1)$ umělých proměnných pro model s absolutním členem, protože v případě vytvoření m umělých proměnných by nám vznikly lineráně závislé sloupce v matici $\mathbf{X}$. Jedna kategorie je tak referenční a všechny umělé proměnné pro tuto kategorii nabývají hodnoty 0.

Uvažujme kategoriální proměnnou vzdělání, která má tři kategorie: ZŠ, SŠ a VŠ. Vytvoříme si 2 umělé proměnné p_1 a p_2, které budou nabývat hodnot:

Vzdělání	p_1	p_2
ZŠ	1	0
SŠ	0	1
VŠ	0	0

Tabulka 2.1: Umělé proměnné

Z tabulky vidíme, že kategorie VŠ je pro náš regresní model referenční. Model pak bude mít tvar

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 p_{i1} + \beta_3 p_{i2} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

a pro jednotlivé kategorie můžeme zapsat:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 p_{i1} + \varepsilon_i \qquad \text{pro ZŠ}$$

$$Y_i = \beta_0 + \beta_1 x_i + \beta_3 p_{i2} + \varepsilon_i \qquad \text{pro SŠ}$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{pro VŠ}$$

2.3. Box-Jenkinsova metodologie

Jeden z oblíbených přístupů k analýze časových řad popularizovali Box a Jenkins a dnes ho tak známe jako Box-Jenkinsova metodologie. Jeho základem je modelování náhodné složky jako systému korelovaných náhodných veličin v čase. Pro tento přístup je typické, že umí rychle reagovat na změny v průběhu časové řady, a je proto výhodné ho použít v případě, že se časová řada velmi těžko modeluje pomocí klasických dekompozičních přístupů.

Pro predikci časových řad se kombinuje autoregresní proces (AR) a proces klouzavých součtů (MA). Tento přístup je možné použít pouze na stacionární časové řady.

Stacionarita

Stacionární časová řada se vyznačuje svým ustáleným pravděpodobnostním chováním v čase. Uvažujme náhodný proces $\{Y_t, t \in T\}$, $t \ldots$ čas, pro který platí:

- $E(Y_t) = \mu$ $\quad \forall t = \ldots, -1, 0, 1, \ldots$
- $var(Y_t) = \sigma^2$ $\quad \forall t = \ldots, -1, 0, 1, \ldots$
- $cor(Y_t, Y_{t+h}) = cor(Y_s, Y_{s+h})$ pro libovolný čas t, s a časovou vzdálenost h.

Náhodný proces s takovými vlastnostmi nazveme slabě stacionární časovou řadou.

Bílý šum

Bílým šumem nazveme posloupnost nezávislých stejně rozdělených náhodných veličin, které mají nulovou střední hodnotu a konstantní rozptyl. Značíme

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Autokorelační funkce (ACF)

Autokorelační funkce je funkce, vyjadřující korelaci dvou náhodných veličin v rámci jednoho procesu Y, které jsou od sebe vzdálené o časový úsek k, v závislosti právě na vzdálenosti k.

Uvažujeme slabě stacionární časovou řadu $\{Y_t, t \in T\}$, pak autokorelační funkci lze vyjádřit jako

$$\begin{aligned} \rho : \mathbf{Z} &\to \langle -1, 1 \rangle \\ k &\to cor(Y_t, Y_{t+k}), \quad k \in \mathbf{Z}. \end{aligned}$$

Díky předpokladu stacionarity závisí velikost korelace pouze na časové vzdálenosti dvou veličin k a nikoliv na čase t.

Autokovariační funkci slabě stacionární časové řady lze vyjádřit jako

$$\gamma_k = cov(Y_t, Y_{t+k}) = E(Y_t - \mu)(Y_{t+k} - \mu), \quad k \in \mathbf{Z},$$

kde μ označuje střední hodnotu procesu. Autokorelační funkci zapíšeme jako

$$\rho_k = \frac{cov(Y_t, Y_{t+k})}{\sqrt{varY_t \cdot varY_{t+k}}} = \frac{\gamma_k}{\gamma_0}, \qquad k \in \mathbf{Z},$$

a označujeme ji ACF.

Vzhledem ke stacionaritě procesu platí, že

1. $\gamma_k = \gamma_{-k}$ a $\rho_k = \rho_{-k}$, tj. stačí zkoumat $k \geq 0$
2. $\rho_0 = 1$
3. $|\rho_k| \leq 1$

Jelikož jsou obecně parametry μ, γ_0 a ρ_k neznámé, autokorelační funkci musíme odhadnout z dostupných dat. Za dodržení předpokladu stacionarity můžeme odhadnout parametr μ pomocí výběrového průměru. Předpokládejme, že máme n hodnot časové řady, výběrový průměr získáme jako

$$\overline{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t.$$

Odhad autokovariační funkce c_k je dán

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} (Y_t - \overline{Y})(Y_{t+k} - \overline{Y})$$

a z toho získáme odhad autokorelační funkce jako

$$r_k = \frac{c_k}{c_0} = \frac{\sum_{t=1}^{n-k} (Y_t - \overline{Y})(Y_{t+k} - \overline{Y})}{\sum_{t=1}^{n} (Y_t - \overline{Y})^2}.$$

Hodnoty odhadnuté autokorelační funkce vykresujeme do tzv. korelogramu. Tyto hodnoty využíváme pro identifikaci modelu, kdy hledáme identifikační bod k_0, což je nejmenší k takové, pro které platí $\rho_k \neq 0$ a zároveň pro všechny hodnoty $k > k_0$ platí $\rho_k = 0$.

Parciální autokorelační funkce (PACF)

Parciální autokorelační funkce nám popisuje korelaci mezi dvěma veličinami Y_t a Y_{t+k}, která je očištěná od vlivu veličin $Y_{t+1}, \ldots, Y_{t+k-1}$.

Parciální korelační funkci označíme:

$$\begin{aligned} \rho : Z &\to \langle -1, 1 \rangle \\ k &\to pcor(Y_t, Y_{t+k}), \quad k \in Z, \end{aligned}$$

kde $pcor(Y_t, Y_{t+k})$ je parciální korelační koeficient veličin Y_t a Y_{t+k} při pevně daných $Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}$.

2.3.1. Proces klouzavých součtů (MA)

Proces klouzavých součtů (MA) modeluje současnou hodnotu veličiny pomocí současné a minulé hodnoty náhodné chyby, která má vlastnosti bílého šumu.

MA(q) proces klouzavých součtů řádu q můžeme zapsat předpisem

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q},$$

kde

t ..... $t \in \mathbf{Z}$ je čas

Y_t ..... pozorovaná hodnota v čase t

ε_t ..... $\varepsilon_t \sim WN(0, \sigma^2)$ náhodná chyba v čase t.

MA procesy mají vlastnosti, které definují stacionaritu. Pro proces MA(q) platí:

- $E(Y_t) = 0$
- $var(Y_t) = \sigma^2(1 + \theta_1^2 + \ldots + \theta_q^2)$
- $\rho_k = \begin{cases} \frac{\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \ldots + \theta_q^2} & \text{pro } k \leq q \\ 0 & \text{pro } k > q \end{cases}$

2.3.2. Autoregresní proces (AR)

Autoregresní proces (AR) k modelování současné hodnoty veličiny využívá lineární kombinaci jejích minulých hodnot a současnou hodnotu náhodné chyby. AR(p) značí autoregresní proces řádu p, kde p popisuje z kolikátého pozorování zpět si model bere informaci pro odhad současné hodnoty.

Autoregresní proces řádu p AR(p) můžeme zapsat ve formě

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t,$$

kde

t ..... $t \in \mathbf{Z}$ je čas

Y_t ..... je pozorovaná hodnota v čase t

ε_t ..... $\varepsilon_t \sim WN(0, \sigma^2)$ je náhodná chyba v čase t.

2.3.3. Smíšené procesy

Smíšený proces ARMA(p,q)

Kombinací procesů AR(p) a MA(q) vzniká autoregresní proces klouzavých součtů ARMA(p,q)

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t.$$

Smíšený integrovaný proces ARIMA(p,d,q)

V případě, že pracujeme s procesem, který není stacionární, lze jej pomocí diferencí stacionarizovat. Proces $\{Y_t, t \in T\}$, jehož d-tá diference je procesem ARMA(p,q) se označuje jako integrovaný smíšený proces ARIMA(p,d,q).

2.3.4. Identifikace modelu

Při hledání optimálního modelu využíváme hodnoty odhadnuté autokorelační funkce a parciální korelační funkce. Ty si vykreslíme do korelogramu a hledáme

identifikační bod k_0, což je nejmenší k takové, pro které platí $\rho_k \neq 0$ a zároveň $\rho_k = 0$ pro všechny hodnoty $k > k_0$.

V případě, že ACF má identifikační bod, jedná se o proces MA(q), kde $q = k_0$. Pokud ACF nemá identifikační bod, hledáme ho pro PACF. Časová řada se řídí procesem AR(p), jestli pro PACF existuje identifikační bod a platí, že $p = k_0$.

Jestliže neexistuje k_0 pro ACF ani PACF, jedná se o proces ARMA(p,q) nebo je nutné řadu diferencovat.

Obecně pro procesy AR(p), MA(q) a ARMA(p,q) platí:

	ACF	**PACF**
AR(p)	k_0 neexistuje	$k_0 = p$
MA(q)	$k_0 = q$	k_0 neexistuje
ARMA(p,q)	k_0 neexistuje	k_0 neexistuje

Tabulka 2.2: Existence identifikačního bodu

2.4. Regresní model s ARIMA chybami

Výše představenou Box-Jenkinsovu metodologii lze využít v lineární regresi k modelování chyb modelu. Předpokládejme vícenásobný lineární regresní model

$$Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \eta_t,$$

kde chyba η_i může být modelována pomocí smíšeného procesu ARMA(p,q)

$$\eta_t = \phi_1 \eta_{t-1} + \phi_2 \eta_{t-2} + \ldots + \phi_p \eta_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t.$$

2.5. Míry přesnosti odhadu

Každá předpověď musí být zvalidována, abychom dokázali posoudit, zde je dostatečně přesná či nikoliv. Je důležité zmínit, že hodnocení predikce validuje pouze statistickou přesnost a různé metriky mohou dávat různé odpovědi na otázku, jaký model je lepší.

V této kapitole uvedeme nejčastěji používané míry pro validaci přesnosti predikce a vybereme míru, která je pro náš případ nejlépe použitelná. Označme si y_i jako skutečnou hodnotu i-tého pozorování a $\widehat{y}_i$ jako predikovanou hodnotu.

Accuracy

Ve společnosti Logio s.r.o. se pro hodnocení předpovědí nejčastěji používá metrika Accuracy. Tato míra je obecně daná předpisem:

$$Accuracy = 1 - \frac{\sum_{i=1}^{n} |y_i - \widehat{y}_i|}{\sum_{i=1}^{n} y_i}.$$

Statisticky je tato míra poměrně problematicky interpretovatelná, jelikož může nabývat hodnot od minus nekonečna do jedné. Záporná přesnost tak matematicky nedává žádný význam. Společnost Logio si tuto míru upravila pro své vlastní účely a její formulace je následující:

$$Accuracy = \begin{cases} 1 - \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} y_i} & \text{pro } \sum_{i=1}^{n} y_i \geq \sum_{i=1}^{n} \hat{y}_i \\ 1 - \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} \hat{y}_i} & \text{pro } \sum_{i=1}^{n} y_i < \sum_{i=1}^{n} \hat{y}_i \end{cases}$$

Čím je hodnota Accuracy bližší jedné, tím je predikce vnímána jako přesnější a lepší.

MAPE - Střední absolutní procentuální odchylka

Jednou z nejpoužívanějších metrik je MAPE – Mean Absolute Percentage Error neboli Střední absolutní procentuální odchylka. Tato míra může nabývat hodnot od nuly do nekonečna a udává průměrnou procentuální absolutní odchylku predikce od skutečnosti.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \widehat{y}_i|}{y_i}$$

Čím je hodnota MAPE menší, tím je predikce lepší. Ačkoliv je tato míra poměrně snadno interpretovatelná, pro naše účely ji nemůžeme použít, jelikož se v datech

vyskytují i případy, kdy je velikost prodejů v testovací sadě rovna nule a v takových případech nelze míra MAPE definovat.

MAE - Střední absolutní odchylka

V případě nulových skutečných hodnot je možné využít míry, ve kterých odchylky nejsou škálované skutečnými hodnotami. Tyto míry jsou tím pádem ale závislé na jednotce proměnné. Míra MAE – Mean Absolut Error je neškálovaná metrika a není tak možné ji použít k porovnání různých proměnných mezi sebou. V našem případě predikování prodejů různých produktů tedy hodnotu MAE nemůžeme srovnávat mezi jednotlivými produkty. Tato metrika nabývá nezáporných hodnot a můžeme ji interpretovat jako průměrnou absolutní odchylku odhadu od skutečnosti.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i|$$

Jelikož je cílem této práce porovnat zde vytvořené výsledky s výsledky Veritica, budeme mezi sebou vždy porovnávat pouze odhady stejných produktů. V našem případě je tedy bezpečné tuto metriku použít.

Relativní míry přesnosti odhadu

Další možností pro ohodnocení přesnosti odhadu podle [15] je využít relativní míru. Označme si MAE_b jako míru vypočítanou ze základního modelu, se kterou budeme výsledky ostatních metod porovnávat. Relativní MAE pak můžeme zapsat jako

$$RelMAE = \frac{MAE}{MAE_b}.$$

V případě, že je hodnota $RelMAE < 1$, naše zkoumaná metoda má menší MAE než základní metoda a můžeme tím pádem říct, že má lepší přesnost. Naopak pokud je $RelMAE > 1$, odhad zkoumanou metodou je horší.

2.6. Fourierova transformace

Fourierova transformace je matematická metoda, která pomocí tzv. fourierových řad dokáže aproximovat funkci. Myšlenka této transformace vychází ze skutečnosti, že každá periodická funkce může být vyjádřena pomocí součtu trigonometrických funkcí sin a cos.

Nechť máme libovolnou L periodickou funkci f. Tuto funkci můžeme vyjádřit pomocí:

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

kde koeficienty A_k a B_k lze vypočítat jako normovaný skalární součin funkce f a příslušné trigonometrické funkce:

$$A_k = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{2\pi kx}{L}\right)\mathrm{d}x = \frac{1}{\| \cos\left(\frac{2\pi kx}{L}\right) \|^2}\langle f(x), \cos\left(\frac{2\pi kx}{L}\right)\rangle,$$

$$B_k = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{2\pi kx}{L}\right)\mathrm{d}x = \frac{1}{\| \sin\left(\frac{2\pi kx}{L}\right) \|^2}\langle f(x), \sin\left(\frac{2\pi kx}{L}\right)\rangle.$$

Skrz limitní proces $L \to \infty$ lze použít metodu i v případě neperiodických funkcí.

Předpis funkce $f(x)$ lze formulovat pro obor komplexních čísel. Nechť $f(x) \in \mathbf{C}$ je libovolná 2π periodická funkce. Jelikož platí

$$e^{ikx} = \cos(kx) + i\sin(kx),$$

můžeme funkci $f(x)$ zapsat jako

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

2.6.1. Diskrétní Fourierova transformace (DFT)

V případě diskrétních dat využijeme k aproximaci diskrétní fourierovu transformaci. Předpokládejme, že máme n napozorovaných hodnot f_j, $j = 0, \ldots, n-1$.

Pomocí přímé DFT můžeme tyto hodnoty převést na Fourierovy koeficienty $\hat{f}_K$ podle vztahu:

$$\hat{f}_K = \sum_{j=0}^{n-1} f_j e^{i2\pi j \frac{K}{n}}, \quad K = 0, \ldots, n-1.$$

Hodnoty $\hat{f}_K$ jsou Fourierovy koeficienty příslušící daným frekvencím. Říkají nám, jak velké frekvence musíme zahrnout do výpočtu, abychom co nejlépe aproximovali hodnoty f_j. Tyto koeficienty využijeme při transformaci našich napozorovaných dat. Transformaci nazýváme jako zpětná DFT a je dána vztahem:

$$f_K = \frac{1}{n} \sum_{K=0}^{n-1} \hat{f}_K e^{i2\pi j \frac{K}{n}}, \quad K = 0, \ldots, n-1.$$

Označme

$$\omega_n = e^{\frac{-2\pi i}{n}},$$

pak maticový zápis přímé DFT má tvar:

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

Kapitola 3

Modelování časových řad a predikce poptávky

Cílem této dipomové práce je vypočítat predikce denních prodejů porovnat je s předpovědí, kterou počítá nástroj Veritico. Nástroj Veritico bude popsán v kapitole 3.1.1 včetně nastínění jeho postupu výpočtu predikcí.

Modely poptávky a výpočty predikcí v této práci byly vytvořeny ve vývojovém prostředí R Studio, které využívá programovací jazyk R. [16]

3.1. Společnost Logio s.r.o.

Obrázek 3.1: Logo společnosti Logio s.r.o.

Logio s.r.o. je technologicko-poradenská společnost, která se na českém trhu vyskytuje od roku 2004. Zabývá se převážně oblastí dodavatelských řetězců a nabízí svým zákazníkům komplexní řešení této problematiky. Společnost vznikla

spojením dvou konzultačních firem, jedna se specializovala na logistické technologie a druhá vlastnila software pro plánování a řízení dodavatelských řetězců. V současné době je tak Logio schopné poskytovat poradenské služby i softwarové řešení napříč celou oblastí řízení dodavatelských řetězců. [7]

Nabízené služby Logia se dají rozdělit do dvou hlavních směrů. Konzultantská část se zaměřuje na poradenství v oblasti logistiky a business inteligence. Mají na starost projekty například v rámci automatizace, optimalizace skladů a výroby nebo reportingu.

Druhá část je sofwarová a má na starost implementaci a vývoj nástrojů, které pomohou zákazníkům efektivně řídit a plánovat dodavatelské řetězce. Patří mezi ně také Veritico, které si blíže představíme v podkapitole 3.1.1.

Společnost tvoří přes 150 zaměstnanců s hlavním sídlem v Praze a s pobočkami v Brně a Hradci Králové. Mezi největší zákazníky společnosti patří například Ahold Česká republika, Dr. Max, Škoda Auto a.s. nebo Plzeňský prazdroj a.s.

3.1.1. Veritico

Obrázek 3.2: Logo nástroje Veritico

Nástroj Veritico byl vyvinut s cílem zajistit organizacím efektivně řízený dodavatelský řetězec. Pomáhá společnostem správně řídit zásoby a zajistit účinné plánování. [8]

Veritico se skládá ze tří součástí a každá má na starost jinou oblast řízení dodavatelských řetězců.

- **Veritico stock:**

  Tento produkt má na starost co nejlépe odhadnout budoucí poptávku pomocí dostupných historických dat. Na základě těchto výpočtů pak určí op-

timální skladové zásoby a optimální výše objednávek na několik období dopředu. Jsou tak zajištěny dostatečně naplněné, ale zároveň nepřehlcené sklady.

- **Veritico price:**

  Veritico price se zabývá cenotvorbou a zajišťuje optimální výši cen a slev. Navrhuje správné výše promoakcí a vybírá produkty, na kterých jsou promoakce efektivní. Dále umí navrhovat optimální cenu pro udržení správných marží nebo pro vyprodání aktuálně nesezónního sortimentu.

- **Veritico plan:**

  Poslední část Veritica se stará o správné a efektivní obchodně-provozní plánování. Může se jednat o dlouhodobé (3 - 12 měsíců) nebo krátkodobé (0 - 3 měsíce) plány, které slouží ke strategickému řízení bysnysu. Díky Veriticu jsou takové plány datově podložené a aktualizované s každou změnou, která nastane.

Největší přínos má pro firmu samozřejmě kompletně řízený systém pomocí všech částí Veritica, ale mohou se implementovat i jednotlivé části samostatně.

Předpověď prodejů ve Veritico

Stejně jako je informace o budoucí poptávce základem pro veškeré plánování organizace, predikce poptávky je základní informací pro všechny další moduly Veritica. Je proto důležité mít k dispozici dostatečně přesné odhady.

V současné době je modul pro odhad budoucích prodejů navrhnutý na měsíční výpočty. Poptávka se vypočítá na nadcházející měsíc a následně se pomocí koeficientů rozpočítá na jednotlivé dny. Pro nejbližší nadcházející dny má Veritico také modul fast-adapt, který pomocí regresní analýzy dokáže v případě výrazné změny trendu tuto odchylku detekovat a upravit tak predikované hodnoty pomocí lineární regrese.

Jedním z cílů této práce je porovnat přístupy měsíčního a denního výpočtu. Pro společnosti prodávající nepotravinové zboží nemusí být nutně kritické znát

odhad poptávky detailně na dny, jelikož zboží nemá trvanlivost a nemusí se objednávat tak často. V potravinářském maloobchodu se ale často objednává v rozmezí týdnů, často i dnů a každá přesnější informace je tak velmi cenná.

3.2. Data

K dispozici máme data denních prodejů 489 produktů z maloobchodního prodeje potravin a základní drogerie v období od 1.7. 2019 do 30.10. 2022. Celkem jde o 540 834 záznamů. Jelikož jsou data citlivým údajem společnosti, není možné je přiložit k práci.

Maloobchod nabízí sice přes 30 000 produktů, ale z důvodu omezené výpočetní síly nebylo možné vytvořit predikce pro všechny položky. Vybrali jsme si proto vzorek 489 produktů, které měly v období od 1. 7. 2021 do 30. 10. 2022 nejvyšší prodané množství.

Je nutné poznamenat, že každý produkt má jiné měrné jednotky, ačkoliv se kusy i hmotnost zaznamenávají jednotně jako proměnná *amount*. Naším cílem je odhadnout budoucí hodnotu právě této proměnné s přesností na dny.

Produkty, o kterých máme informace, jsou rozdělené do 9 kategorií. V tabulce 3.1 vidíme počty produktů v jednotlivých kategoriích.

Dále máme ke každému produktu informaci, zda byl v promoakci, případně jak vysoká byla sleva. Proměnná *promo* nabývá hodnoty od 0 do 1, kdy 0 znamená, že produkt nebyl v promo akci a hodnota 1 by znamenala, že na něj byla poskytnuta 100% sleva.

Jelikož náš datový soubor zahrnuje období, kdy se svět potýkal s nákazou COVID-19 a mnoho věcí nemělo standardní průběh, rozhodli jsme se tuto informaci zahrnout do modelu predikce.

Z externích zdrojů [9] jsme získali data o vývoji Strigency Indexu, který byl navržen projektem Oxford COVID-19 Government Response Tracker. Tento index, který byl publikován v březnu 2021 [10], popisuje přísnost vládních protiepidemických opatření a nabývá hodnot od 0 do 100. Je vypočítaný na základě devíti ukazatelů, které popisují uzavření škol, uzavření pracovišť, zrušení veřejných akcí,

Kategorie	Počet produktů
Drogerie a kosmetika	14
Maso a ryby	10
Mléčné a chlazené	140
Mražené	9
Nápoje	73
Ovoce a zelenina	100
Pekárna a cukrárna	45
Trvanlivé	67
Uzeniny a lahůdky	31

Tabulka 3.1: Počty produktů v kategoriích

omezení veřejných shromáždění, omezení hromadné dopravy, požadavky na nevycházení z domova, veřejné informační kampaně, omezení vnitrostátního pohybu a přísnost mezinárodních cestovních kontrol. Z dostupných dat jsme si vybrali ta, která popisovala Stringency Index v České republice.

Dále jsme si na základě data prodeje vytvořili umělé proměnné pro dny v týdnu (ut = úterý, st = středa, ct = čtvrtek, pa = pátek, so = sobota a ne = neděle), které nabývají hodnoty 1, pokud záznam odpovídá danému dnu v týdnu, nebo 0, pokud to byl jiný den. Pondělí jsme nechali jako referenční kategorii. Příklad vytvořených umělých proměnných pro prvních 10 dní z datového souboru vidíme v tabulce 3.2.

Druhý datový soubor, který máme k dispozici, obsahuje předpovězené hodnoty prodejů pomocí nástroje Veritico. Tato data máme z období od 14. 7. 2022 do 30. 9. 2022 a naše výsledky budeme porovnávat právě s hodnotami z tohoto datasetu.

Datum	*ut*	*st*	*ct*	*pa*	*so*	*ne*
2019-07-01	0	0	0	0	0	0
2019-07-02	1	0	0	0	0	0
2019-07-03	0	1	0	0	0	0
2019-07-04	0	0	1	0	0	0
2019-07-05	0	0	0	1	0	0
2019-07-06	0	0	0	0	1	0
2019-07-07	0	0	0	0	0	1
2019-07-08	0	0	0	0	0	0
2019-07-09	1	0	0	0	0	0
2019-07-10	0	1	0	0	0	0

Tabulka 3.2: Vytvořené umělé proměnné

Dostupná data si rozdělíme na dvě sady. První sada – trénovací – bude obsahovat období od 1. 7. 2019 do 13. 7. 2022 a budeme na ní trénovat vytvořené modely. Druhá sada, kterou nazveme kontrolní, s daty z období od 14. 7. 2022 do 30. 9. 2022 nám bude sloužit pouze ke změření kvality a přesnosti predikce.

3.3. Vizualizace dat

Jak jsme již zmiňovali v podkapitole 3.2, proměnná *amount* nemá jednotné měrné jednotky a prodané množství je tak v různých měřítkách. Abychom mohli vývoj prodaného množství jednotlivých produktů porovnávat mezi sebou, je potřeba z dat odstranit informaci o jednotce, respektive rozptylu.

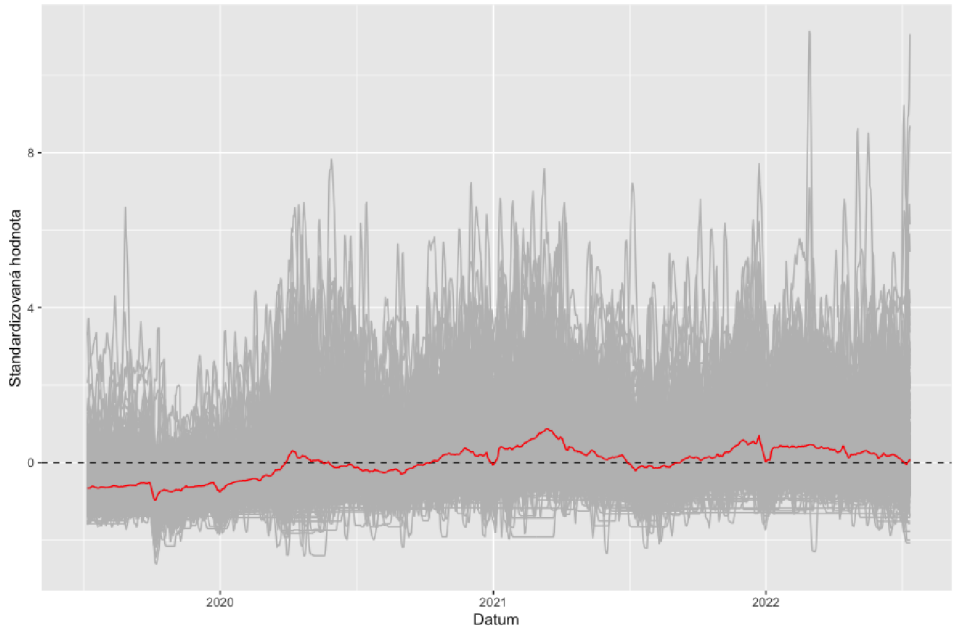
Data jsme standardizovali pomocí postupu, zmíněného v kapitole 2.1. Od každého denního prodaného množství produktu odečteme jeho střední hodnotu a

vydělíme směrodatnou odchylkou, které jsme vypočítali z trénovacího datasetu. Tak získáme hodnoty, které mezi sebou napříč produkty můžeme porovnávat.

Obecný vývoj prodejů

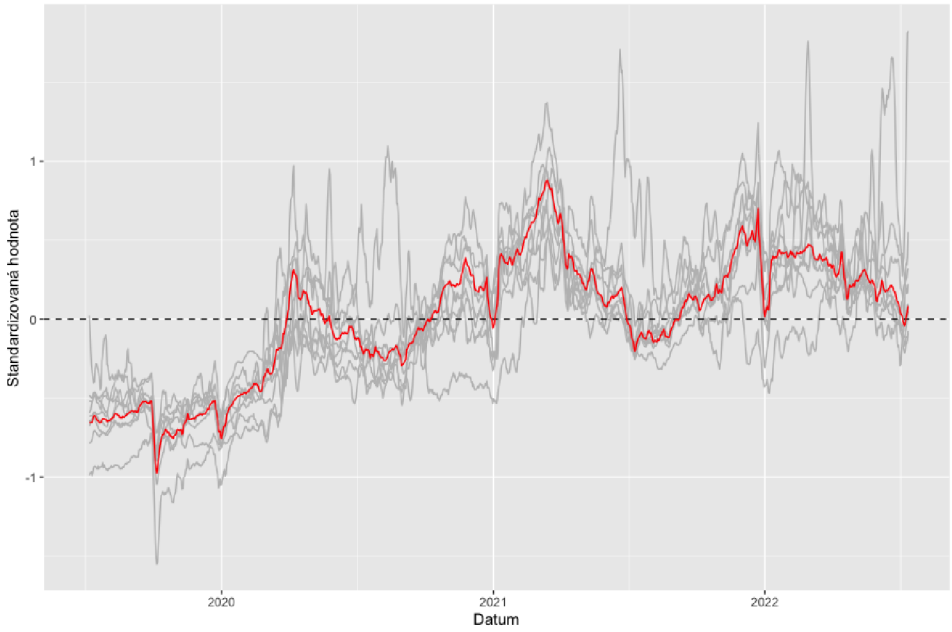
Standardizované hodnoty zprůměrujeme pro každý den a získáme tak časovou řadu zobrazující obecný vývoj prodejů daného maloobchodu. Abychom lépe viděli směr, kterým vývoj směřuje, data vyhladíme pomocí klouzavých průměrů přes 7 dní.

Na obrázku 3.3 vidíme vývoj standardizovaných prodejů všech produktů za trénovací období, kde červená křivka značí průměr těchto hodnot vyhlazených přes 7 dní klouzavým průměrem.



Obrázek 3.3: Vývoj standardizovaných prodejů produktů

Produkty máme rozdělené do devíti kategorií, v rámci nichž si můžeme vykreslit celkový vývoj prodejů, jako jsme to udělali u všech produktů zároveň. Na obrázku 3.4 vidíme šedé křivky pro každou z kategorií. Červená křivka je, stejně jako na předchozím grafu, křivka pro celkový průměr standardizovaných prodejů vyhlazená přes 7 dní.



Obrázek 3.4: Vývoj standardizovaných prodejů kategorií

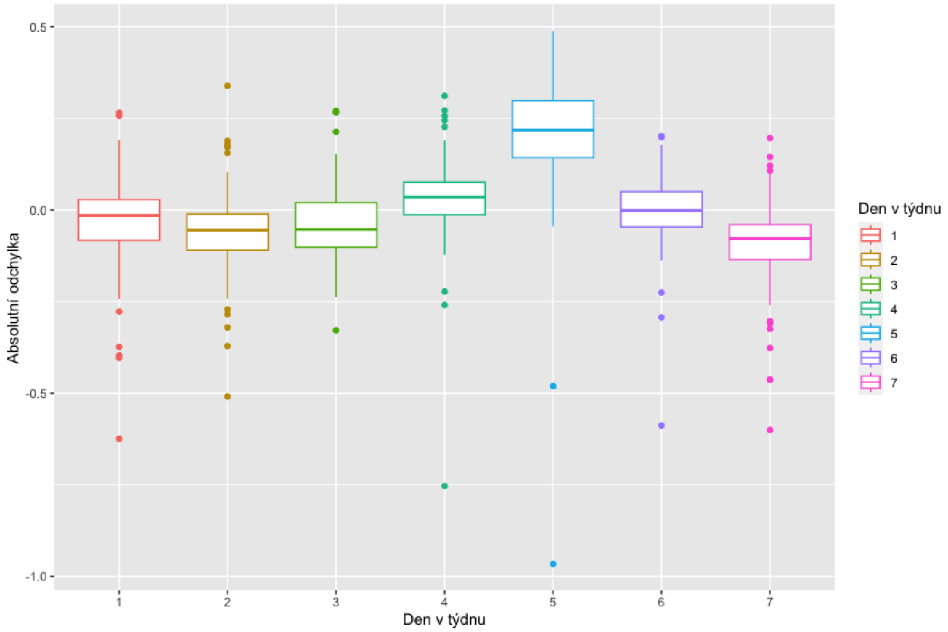
Efekt dne v týdnu

V denních datech se naskytuje možnost hledat týdenní sezónnost. Abychom se podívali, zda existuje vliv dne v týdnu na velikost prodeje, vypočítali jsme si odchylku standardizované hodnoty od hodnoty vypočítané klouzavým průměrem přes 7 dní. Získáme tak informaci, zda pro daný den v týdnu byla hodnota prodeje výrazně odlišná od týdenního průměru.

Hodnoty těchto odchylek jsme zobrazili pomocí boxplotů v grafu 3.5. Můžeme vidět, že obecně jsou prodeje v pátek vyšší než v ostatních dnech v týdnu a v neděli naopak lehce pod průměrem.

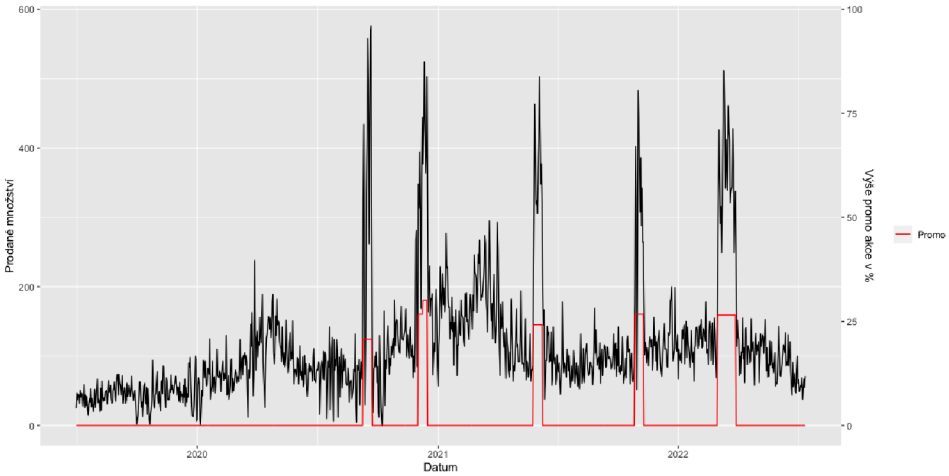
Promo akce

V datech máme dostupnou také informaci o tom, kdy byl produkt v akci a jaká byla poskytnuta sleva. Je zřejmé, že tato proměnná bude mít na výši prodejů vliv. V datovém souboru máme 103 produktů, u kterých se v trénovací sadě ani jednou nevyskytla promo akce. Většina zkoumaných produktů tak byla alespoň jednou v akci. Jako ilustrační příklad, na kterém si vliv promo akce na



Obrázek 3.5: Efekt dne v týdnu

výši prodejů ukážeme, vezmeme produkt *Mléko Selské 1l*. Průběh prodejů tohoto produktu zároveň s velikostí promoakcí vidíme na obrázku 3.6.



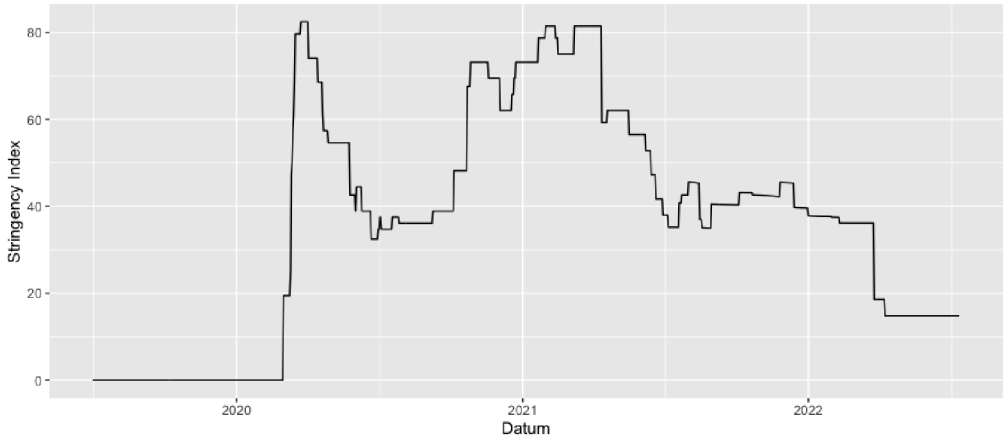
Obrázek 3.6: Efekt promo akce na produktu *Mléko Selské 1l*

Levá osa y značí výše prodejů a pravá osa y byla seškálována, aby odpovídala výši promo akce v procentech. Hodnoty proměnné *promo* byly pro účely této vizualizace vynásobeny 100, aby odpovídaly procentům. Je zřejmé, že proměnná

promo má na prodeje tohoto konkrétního produktu vliv.

Stringency Index

Stringency Index popisuje přísnost vládních protiepidemických opatření. Do modelu ho zahrnujeme z důvodu možného nestandardního chování nákupní síly během covidového období. Vývoj hodnoty Stringency Indexu pro Českou republiku v průběhu let zobrazuje graf 3.7.



Obrázek 3.7: Stringency Index

3.4. Tvorba modelů

Pro predikci budoucího prodaného množství jsme se rozhodli využít modely vícenásobné lineární regrese s ARIMA chybami a klasické modely vícenásobné lineární regrese.

Naším cílem je vytvořit predikce pro 489 produktů, u kterých je zřejmé, že se nechovají stejně. Optimální model budeme vybírat pro každý produkt zvlášť na základě hodnoty Akaikeho informačního kritéria. Pro názornou ukázku si zde postup ilustrujeme na jednom konkrétním produktu.

3.4.1. Model vícenásobné lineární regrese s ARIMA chybami

Model vícenásobné lineární regrese s chybami modelovanými procesem ARIMA (dále budeme značit lm_arima) má na vstupu vysvětlující proměnné a časovou řadu predikované proměnné. V R Studiu je možné tvořit modely pomocí funkce `arima()` nebo `auto.arima()` s vysvětlovanými proměnnými přidanými přes argument funkce xreg.

Uvažujeme vysvětlující proměnné z tabulky 3.3, které mohou být použity v modelu.

Proměnná	**Popis**
SI	hodnota Stringency Indexu
$promo$	výše promoakce
$categ$	odhadnutá standardizovaná hodnota průměrného prodaného množství kategorie produktu (více si popíšeme v další podkapitole)
dny	matice umělých proměnných pro dny v týdnu (proměnné ut, st, ct, pa, so, ne)
$fourier$	matice hodnot získaných Fourierovou transformací pro aproximaci sezónnosti časové řady

Tabulka 3.3: Vysvětlující proměnné

Velikost matice proměnné *fourier* závisí na parametru K, který označuje, pomocí kolika frekvencí je časová řada aproximována. V R Studiu existuje pro Fourierovu tranformaci funkce `fourier()` z balíčku {forecast}. Argument funkce K udává, kolik členů sinus a cosinus má funkce vrátit. Pro časovou řadu o délce n je tak vytvořena matice $(n \times 2K)$ transformovaných hodnot. Pro použití Fourierovy transformace pomocí funkce `fourier()` je nutné, aby byla časová řad definovaná jako sezónní se zvolenou periodou.

V rámci hledání nejlepšího modelu jsme Fourierovu transformaci provedli pro každou kombinaci vysvětlujících proměnných 5krát. Hledali jsme $K \in \{1, 2, 3, 4, 5\}$ takové, které odpovídalo modelu s nejmenší hodnotou AIC. Parametr K jsme omezili shora číslem 5, z důvodu omezené výpočetní síly.

Model pro kategorie

Vycházíme z myšlenky, že se produkty v rámci kategorií chovají podobně a že předpověď vývoje celé kategorie by mohla přispět k predikci jednotlivých produktů.

Pro každou kategorii jsme predikovali průměrné standardizované hodnoty prodeje. V případě dat agregovaných do kategorií jsme mohli uvažovat pouze vysvětlující proměnné uvedené v tabulce 3.4.

Proměnná	**Popis**
SI	hodnota Stringency Indexu
dny	matice umělých proměnných pro dny v týdnu (proměnné ut, st, ct, pa, so, ne)
$fourier$	matice hodnot získaných Fourierovou transformací pro aproximaci sezónnosti časové řady

Tabulka 3.4: Možné vysvětlující proměnné pro modelování proměnné *categ*

Celkem jsme pro každou kategorii vybírali z 23 možných kombinací regresorů:

- 3 kombinace získáváme z možných kombinací proměnných SI a dny
- 5 možností fourierových regresorů

Uvažujeme-li, že do modelu může být zahrnut i pouze jeden regresor, dostáváme dohromady $3 \cdot 5 + 3 + 5 = 23$ kombinací regresorů. Výsledný model byl vybrán na základě nejmenší hodnoty AIC. Modely chyb ARIMA(p,d,q) byly vybrány automaticky pomocí funkce `auto.arima()`.

Finální výběr regresorů a procesů pro tvorbu modelů časových řad kategorií můžeme vidět v tabulce 3.5.

Kategorie	**Regresory**	**Model chyb**
Ovoce a zelenina	dny, fourier(K=1)	ARIMA(5,1,5)
Nápoje	SI, dny	ARIMA(1,1,1)
Drogerie a kosmetika	SI, dny, fourier(K=2)	ARIMA(1,1,2)
Mléčné a chlazené	dny	ARIMA(5,1,4)
Uzeniny a lahůdky	dny	ARIMA(5,1,3)
Trvanlivé	SI, dny, fourier(K=5)	ARIMA(1,1,2)
Pekárna a cukrárna	SI, dny, fourier(K=1)	ARIMA(5,1,4)
Maso a ryby	dny, fourier(K=1)	ARIMA(2,1,2)
Mražené	dny, fourier(K=1)	ARIMA(1,1,1)

Tabulka 3.5: Modely pro predikci kategorií

Pomocí modelu jsme predikovali budoucí průměrné standardizované hodnoty prodejů kategorie na příslušné období dopředu a uložili jsme si je jako proměnnou *categ*.

Pro názornou ukázku použijeme například kategorii *Nápoje*. Vysvětlujícími proměnnými byly proměnné SI a umělé proměnné označující dny v týdnu. Pro modelování chyby byl vybraný model ARIMA(1,1,1). Hodnotu *categ* v čase t jsme modelovali pomocí:

$$categ_t = \beta_1 SI_t + \beta_2 ut_t + \beta_3 st_t + \beta_4 ct_t + \beta_5 pa_t + \beta_6 so_t + \beta_7 ne_t + \eta_t,$$

kde chyba η_t je modelovaná pomocí procesu ARIMA(1,1,1):

$$\Delta\eta_t = \phi_1 \Delta\eta_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t,$$

kde

$$\Delta\eta_{t-1} = \eta_{t-1} - \eta_{t-2}.$$

Model s odhadnutými hodnotami regresních parametrů má pak podobu:

$$categ_t = 0.0026SI_t + 0.0345ut_t + 0.0722st_t + 0.1619ct_t + 0.3180pa_t + 0.0851so_t + 0.0585ne_t + \eta_t$$

a

$$\Delta\eta_t = 0.6261\Delta\eta_{t-1} - 0.9556\varepsilon_{t-1} + \varepsilon_t.$$

Model pro produkty

U tvorby modelů pro časové řady jednotlivých produktů jsme postupovali obdobně. Pro každý produkt jsme vytvořili modely pomocí všech možných kombinací regresorů.

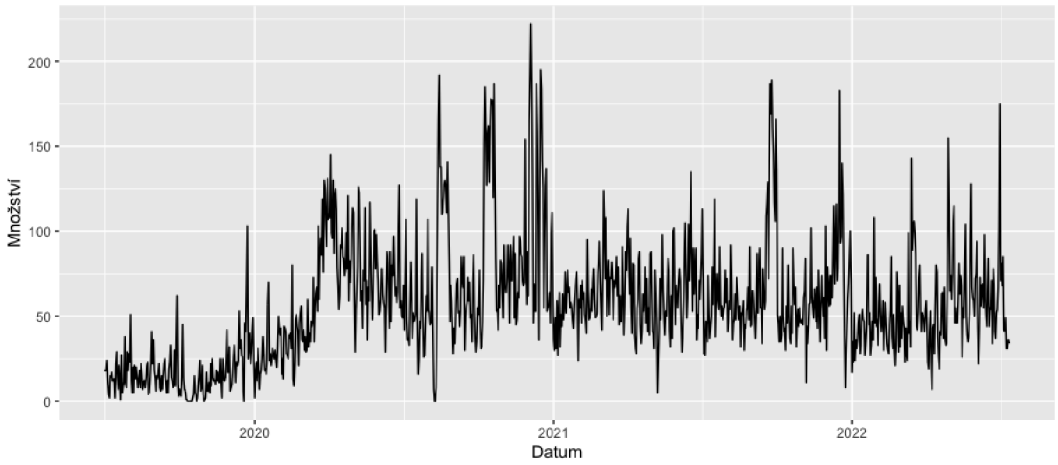
- z regresorů *SI*, *promo*, *categ*, *dny* získáme 15 různých kombinací ($2^4-1 = 15$)
- proměnná *fourier* může mít 5 různých velikostí

Uvažujeme, že model může být tvořen jen jedním regresorem a získáváme tak celkem $15 \cdot 5 + 15 + 5 = 95$ možných modelů, které mezi sebou porovnáme pomocí hodnot AIC.

K nalezení optimálního modelu chyb jsme využili funkci `auto.arima()`, která hledá nejlepší model ARIMA(p,d,q) podle hodnoty AIC modelů.

Jelikož je naším úkolem přepovědět hodnoty prodejů pro 489 produktů, postup budeme ilustrovat pouze na jednom konkrétním produktu. Vybereme si například produkt *Pivo ležák, plech 0,5l*, který spadá do kategorie *Nápoje*.

Na obrázku 3.8 vidíme časovou řadu prodeje ilustračního produktu – plechovkového piva.



Obrázek 3.8: Vývoj prodeje produktu *Pivo ležák, plech 0,5l*

Nejlepší model pro prodané množství tohoto produktu v čase t má tvar:

$$amount_t = \beta_1 SI_t + \beta_2 ut_t + \beta_3 st_t + \beta_4 ct_t + \beta_5 pa_t + \beta_6 so_t + \beta_7 ne_t + \beta_8 promo_t + \beta_9 categ_t + \eta_t,$$

kde η_t je modelovaný pomocí procesu ARIMA(1,1,1):

$$\Delta\eta_t = \phi_1 \Delta\eta_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t,$$

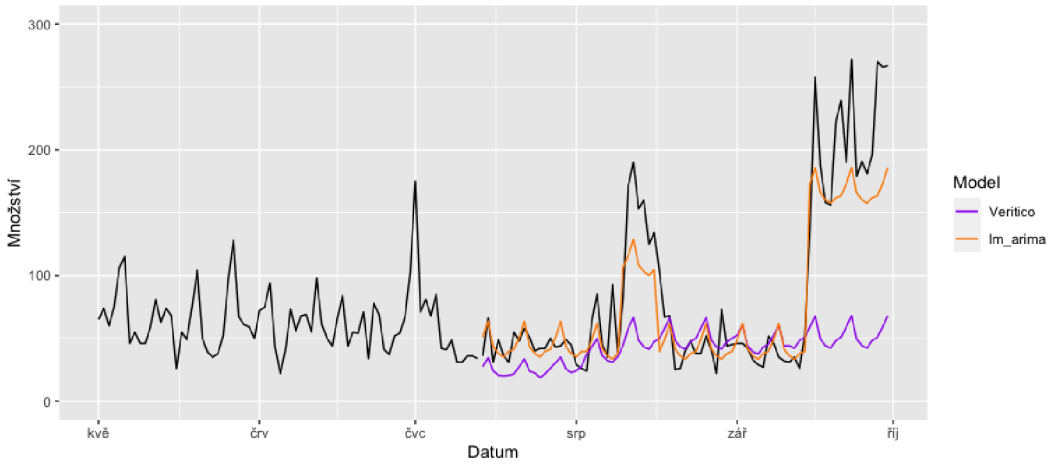
Model s odhadnutými regresními parametry je pak ve tvaru:

$$amount_t = 0.3891 SI_t + 3.3046 ut_t + 3.6032 st_t + 10.1210 ct_t + 17.6243 pa_t + 5.3007 so_t + 4.7761 ne_t + 479.9804 promo_t + 34.0337 categ_t + \eta_t,$$

kde

$$\Delta\eta_t = 0.3636 \Delta\eta_{t-1} - 0.9554 \varepsilon_{t-1} + \varepsilon_t.$$

Na obrázku 3.9 máme detailnější pohled na prodeje od začátku května 2022 až do konce září 2022. Do grafu jsme vykreslily oranžovou křivku odhadnuté hodnoty pomocí výše zmíněného modelu a fialovou křivku odhadnutých hodnot nástrojem Veritico.



Obrázek 3.9: Predikce prodeje produktu *Pivo ležák, plech 0,5l* – Veritico a model lm_arima

3.4.2. Model vícenásobné lineární regrese

Predikce jsme vytvořili i pomocí modelů klasické lineární regrese (dále budeme značit jako lm). Model jsme v tomto případě hledali pomocí fce `step()`, která krokovou metodou vybírá optimální model. Znamená to tedy, že ze všech možných kombinací regresorů vybere tu kombinaci, která odpovídá modelu s nejmenší hodnotou AIC.

Pro výši prodejů v čase t našeho ilustračního produktu *Pivo ležák, plech 0,5l* má výsledný model tvar:

$$amount_t = \beta_0 + \beta_1 SI_t + \beta_2 ut_t + \beta_3 st_t + \beta_4 ct_t + \beta_5 pa_t + \beta_6 so_t + \beta_7 ne_t + \beta_8 promo_t + \beta_9 categ_t + \beta_{10} fourierC(1)_t + \beta_{11} fourierS(2)_t + \beta_{12} fourierC(2)_t + \beta_{13} fourierS(3)_t + \beta_{14} fourierC(3)_t + \beta_{15} fourierS(4)_t + \beta_{16} fourierC(4)_t,$$

kde proměnné $fourierS(j)$, resp. $fourierC(j)$ pro $j = \{1, \ldots, 5\}$ označují vybrané členy z Fourierovy transformace.

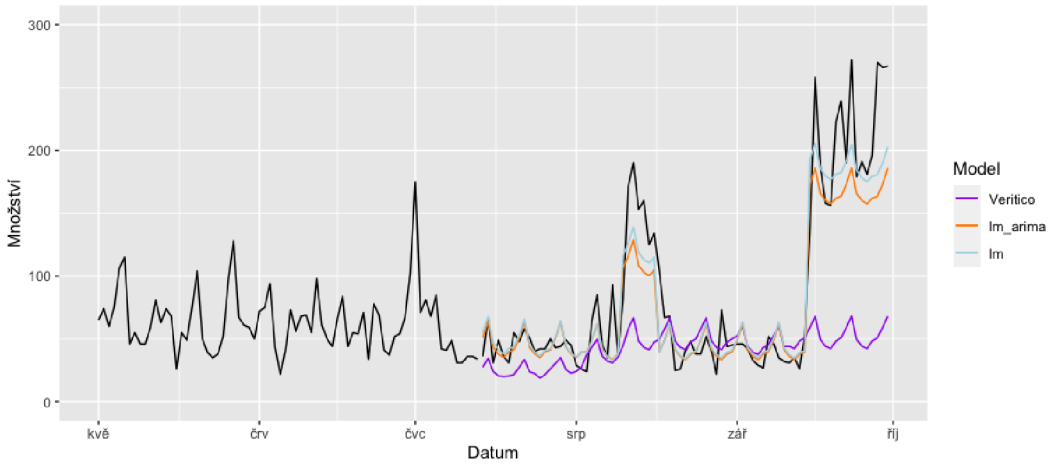
$$fourierS(j) = \sin \tfrac{2\pi ji}{365} \quad \text{a} \quad fourierC(j) = \cos \tfrac{2\pi ji}{365}$$

kde $(i) \in \{1, \ldots, 365\}$ je index dne v roce.

Po dosazení odhadnutých regresních parametrů získáváme:

$$amount_t = 23.495 + 0.568SI_t + 3.994ut_t + 4.849st_t + 13.072ct_t + 23.660pa_t + 7.077so_t + 3.670ne_t + 553.449promo_t + 15.708categ_t + 3.244fourierC(1)_t + -3.816fourierS(2)_t + 3.1452fourierC(2)_t + 1.946fourierS(3)_t - 1.674fourierC(3)_t - 3.612fourierS(4)_t + 1.702fourierC(4)_t$$

Hodnoty predikce si vykreslíme i s předchozími výsledky do jednoho grafu a můžeme je porovnat s předchozím modelem. Na obrázku 3.10 jsou fialovou křivkou znázorněny predikce Veritica, oranžovou křivkou predikce pomocí modelu lm_arima a světle modrou křivkou odhadnuté hodnoty získané modelem lm.



Obrázek 3.10: Predikce prodeje produktu *Pivo ležák, plech 0,5l* – Veritico, model lm_arima a lm

3.5. Výsledky

Přesnost predikce budeme hodnotit v rámci čtyř různých časových období pomocí metrik zmíněných v kapitole 2.5. Pro každý produkt jsme si tak vypočítali míry pro model lm_arima i pro model lm. Označíme si $RelMAE$:

$$RelMAE_{lm\_a} = \frac{MAE_{lm\_a}}{MAE_V}, \quad \text{resp.} \quad RelMAE_{lm} = \frac{MAE_{lm}}{MAE_V}$$

jako poměr metrik, kde MAE_{lm_a} je střední absolutní odchylka získaná z predikcí modelu lm_arima, MAE_{lm} je získaná z predikcí modelu lm a MAE_V je

získaná z predikcí Veritica. Jako lépe předpovězené produkty, vnímáme takové, pro které je

$$RelMAE_{lm\_a} < 1, \text{ reps. } RelMAE_{lm} < 1.$$

Dále jsme si vypočítali metriku accuracy podle upravené formulace Logia, kterou jsme uváděli v kapitole 2.5. Označíme si ACC_V jako hodnotu accuracy pro predikce vypočítané Veriticem, ACC_{lm_a} jako hodnotu accuracy pro model lm_arima a ACC_{lm} pro model lm.

Při použití této metriky považujeme za úspěšné ty případy, kdy platí

$$ACC_V < ACC_{lm\_a}, \text{ resp. } ACC_V < ACC_{lm}.$$

3.5.1. Porovnání výsledků přes 18 dní

Jelikož jsme si datový soubor rozdělili na trénovací a testovací v půlce července, budeme zkoumat i přesnost predikce na období menší než měsíc. Nástroj Veritico pracuje s měsíčními odhady, které se rozpadávají přes koeficienty do dnů, takže predikce na měsíc červenec musela být vypočítána v červnu. Naše modely by tak v tomto případě mohly využít skutečnost, že mají k dispozici více informací, než měl při predikci nástroj Veritico.

Budeme tvořit predikce na nadcházejících 18 dní, tedy na období od 14. 7. 2022 do 31. 7. 2022. V datovém souboru se vyskytuje 13 produktů, u nichž jsme napočítali hodnotu $MAE_V = 0$. Jedná se o produkty, které v kontrolních 18 dnech neměly žádné prodeje a predikce Veritica byly správně rovny 0. Veličina $RelMAE_{lm\_a}$, reps. $RelMAE_{lm}$ by ale v takovém případě nebyla definována, takže tyto produky vyřadíme z celkového hodnocení.

Z vyřazených produktů jich 12 pocházelo z kategorie *Ovoce a zelenina* a 1 produkt z kategorie *Nápoje*. Každý model lm_arima i lm zvládl v jednom z případů správně předpovědět tyto nulové prodeje a celkem u dvou produktů je tak hodnota MAE_{lm_a} nebo MAE_{lm} rovna nule.

Po vyřazení výše zmíněných produktů máme k dispozici pro hodnocení finální počet 476 produktů. V tabulce 3.6 vidíme souhrnné počty a procenta označující

pro kolik produktů jsme vytvořili lepší predikce než Veritico. V řádcích máme hodnoty podle modelů lm_arima a lm a ve sloupci je hodnocení podle metriky MAE a Accuracy.

	dle MAE		dle Accuracy	
Model	**počet**	**%**	**počet**	**%**
lm_arima	260	54.62 %	246	51.7 %
lm	200	42.02 %	206	43.3 %

Tabulka 3.6: Počty lepších predikcí přes 18 dní

V tabulce 3.7 vidíme top 5 nejčastěji zahrnutých kombinací regresorů, které byly použity v modelech lm_arima a lm pro produkty, které měly lepší predikci než Veritico. Regresory jsou zde vypsané bez informace o členech Fourierovy transformace. Sloupec „počet“ označuje počet produktů, které měly příslušné kombinace regresorů zahrnuté ve svém modelu.

Regresory	**počet**
dny, categ, promo	38
dny, categ, promo, fourier	31
categ, promo	28
dny, categ	26
categ, promo, fourier	21

(a) model lm_arima

Regresory	**počet**
SI, dny, categ, promo, fourier	120
SI, dny, categ, fourier	37
dny, categ, promo, fourier	20
SI, dny, promo, fourier	11
dny, categ, fourier	6

(b) model lm

Tabulka 3.7: Kombinace regresorů v modelech lépe predikovaných produktů

Je zajímavé, že ve výsledných modelech lm byla proměnná *fourier* zahrnuta vždy a v modelech lm_arima se vyskytovala pouze u 43 % produktů. Proměnná

categ se v modelech lm_arima objevila u 97 % produktů a v modelech lm u 94 %. Je z toho zřejmé, že se produkty v rámci kategorií opravdu chovají podobně.

Vypočítané predikce jsme zhodnotili i přes kategorie. Procentuální zastoupení produktů s lepší predikcí najdeme v tabulce 3.8.

	dle MAE		dle Accuracy		
	model		model		Produktů
Kategorie	lm_arima	lm	lm_arima	lm	v kategorii
Drogerie a kosmetika	42.9 %	50 %	64.3 %	50 %	14
Maso a ryby	**90 %**	0 %	40 %	10 %	10
Mléčné a chlazené	45.7 %	37.9 %	38.6 %	37.1 %	140
Mražené	66.7 %	**77.8 %**	100 %	100 %	9
Nápoje	61.1 %	**58.3 %**	68.1 %	66.7 %	72
Ovoce a zelenina	60.2 %	42 %	47.7 %	37.5 %	88
Pekárna a cukrárna	**68.9 %**	40 %	66.7 %	35.6 %	45
Trvanlivé	46.3 %	38.8 %	45.2 %	44.8 %	67
Uzeniny a lahůdky	51.6 %	32.3 %	51.6 %	32.3 %	31

Tabulka 3.8: Kategorie – procento lepších predikcí přes 18 dní

Nejlepších výsledků podle metriky MAE jsme dosáhli pomocí modelů lm_arima v kategorii *Maso a ryby*, kde jsme predikovali 90 % produktů lépe. Modely lm měly největší úspěch v kategorii *Mražené*, kde se podařilo predikovat lépe 77.78 % produktů. Je nutné ale zdůraznit, že kategorie *Maso a ryby* obsahuje pouze 10 produktů a kategorie *Mražené* pouze 9 produktů. Vzorek tak nemusí být dostatečně obsáhlý, abychom mohli vyslovit závěry pro celou kategorii.

Když se podíváme na kategorie s více než 20 produkty, modelům lm_arima se dařilo v kategorii *Pekárna a cukrárna*, kde je 68.89 % lepších predikcí a modely

lm měly 58.3 % produktů s lepší predikcí v kategorii *Nápoje*.

Obecně vidíme, že kromě kategorií *Drogerie a kosmetika* a *Mražené*, měly modely lm_arima celkově větší podíl kvalitnějších predikcí než modely lm.

Podle metriky Accuracy jsou například v kategorii *Mražené* všechny predikce lepší než ty vypočítané Veriticem, ale naopak kategorie *Maso a ryby* má u modelů lm_arima menší procentuální úspěšnost než podle MAE.

3.5.2. Porovnání výsledků přes 30 dní

Další období, přes které budeme hodnotit kvalitu predikce je 30 dní, protože odpovídá průměrné délce jednoho měsíce, což je doba, pro kterou Veritico vytváří predikce. Jelikož Veritico vytváří predikce v měsíčních hodnotách, tj. vždy pro konkrétní měsíc a ne 30 dní, může být v jeho a v denním přístupu poměrně významný rozdíl.

Predikci budeme počítat pro dny od 14. 7. 2022 do 12. 8. 2022. Pro toto období se nulová hodnota MAE_V vyskytla u 9 produktů a jedná se o stejné produkty, které měly nulovou hodnotu MAE_V i pro období 18 dní. Celkově budeme za toto období hodnotit predikce 480 produktů.

Souhrnné výsledky za 30 dní vidíme v tabulce 3.9, kde jsou počty produktů, pro které jsme vytvořili lepší predikce a jejich procentuální zastoupení.

	dle MAE		dle Accuracy	
Model	**počet**	**%**	**počet**	**%**
lm_arima	295	61.5 %	280	58.3 %
lm	232	48.3 %	246	51.2 %

Tabulka 3.9: Počty lepších predikcí přes 30 dní

Celkově je vidět, že se s prodlouženým časovým obdobím podařilo modelům lm_arima i lm vypočítat lepší predikce ve větším procentu případů než u 18denního období. Detail výsledků podle kategorií zobrazje tabulka 3.10.

	dle MAE		dle Accuracy		
	model		model		Produktů
Kategorie	lm_arima	lm	lm_arima	lm	v kategorii
Drogerie a kosmetika	50 %	57.1 %	64.3 %	71.4 %	14
Maso a ryby	**90 %**	10 %	40 %	10 %	10
Mléčné a chlazené	55 %	43.6 %	52.1 %	42.1 %	140
Mražené	77.8 %	**77.8 %**	100 %	100 %	9
Nápoje	**80.6 %**	**70.8 %**	83.3 %	77.8 %	72
Ovoce a zelenina	55.4 %	43.5 %	50 %	46.7 %	92
Pekárna a cukrárna	77.8 %	42.2 %	73.3 %	42.2 %	45
Trvanlivé	49.3 %	52.2 %	46.3 %	58.2 %	67
Uzeniny a lahůdky	58.1 %	32.3 %	48.4 %	32.3 %	31

Tabulka 3.10: Kategorie – procento lepších predikcí přes 30 dní

S delším časovým obdobím získáváme z modelů lm_arima oproti Veriticu lepší predikce především u produktů z kategorie *Nápoje*. Stejně jako v předchozím období mají modely lm větší procento kvalitnějších předpovědí u kategorie *Drogerie a kosmetika* a tentokrát také u kategorie *Trvanlivé*. V kategorii *Mražené* mají tyto dva modely stejný poměr lepších predikcí.

3.5.3. Porovnání výsledků přes 49 dní

Období 49 dní odpovídá datům 14. 7. 2022 až 30. 8. 2022. V tomto časovém úseku už máme celý měsíc a zajímá nás, zda v tom případě bude mít Veritico výhodu.

Z hodnocení vyřadíme 8 produktů, které mají opět nulovou hodnotu MAE_V, takže tentokrát hodnotíme predikce 481 produktů.

Celkové počty a procenta lepších predikcí najdeme v tabulce 3.11.

	dle MAE		dle Accuracy	
Model	**počet**	**%**	**počet**	**%**
lm_arima	308	64 %	293	60.9 %
lm	250	52 %	257	53.4 %

Tabulka 3.11: Počty lepších predikcí přes 49 dní

Oproti předchozímu období se počty lepších predikcí sice zvedly, ale není to tak výrazné zlepšení jako jsme viděli mezi 18denním a 30denním obdobím. Detailní výsledky po kategoriích najdeme v tabulce 3.12.

	dle MAE		dle Accuracy		
	model		model		Produktů
Kategorie	lm_arima	lm	lm_arima	lm	v kategorii
Drogerie a kosmetika	50 %	50 %	57.1 %	78.6 %	14
Maso a ryby	**90 %**	10 %	40 %	20 %	10
Mléčné a chlazené	58.6 %	44.3 %	54.3 %	40.7 %	140
Mražené	66.7 %	**66.7 %**	88.9 %	100 %	9
Nápoje	**81.9 %**	**73.6 %**	86.1 %	80.6 %	72
Ovoce a zelenina	61.3 %	52.7 %	55.9 %	51.6 %	93
Pekárna a cukrárna	68.9 %	40 %	68.9 %	33.3 %	45
Trvanlivé	56.7 %	59.7 %	53.7 %	62.7 %	67
Uzeniny a lahůdky	61.3 %	45.2 %	51.6 %	48.4 %	31

Tabulka 3.12: Kategorie – procento lepších predikcí přes 49 dní

V tomto období mají modely lm větší procento lepších predikcí oproti modelům lm_arima pouze v kategorii *Trvanlivé*.

3.5.4. Porovnání výsledků přes 79 dní

Poslední období je od 14. 7. 2022 do 30. 9. 2022, tj. je dlouhé 79 dní. V tomto období jsme z hodnocení vyřadili jen 5 produktů z důvodu nulové hodnoty MAE_V. Do finálního hodnocení tedy zahrneme 484 produktů.

	dle MAE		dle Accuracy	
Model	**počet**	**%**	**počet**	**%**
lm_arima	314	64.9 %	288	59.5 %
lm	274	56.6 %	273	56.4 %

Tabulka 3.13: Počty lepších predikcí přes 79 dní

Je vidět, že procenta lepších predikcí se oproti předchozímu období v případě modelů lm_arima už moc nezvedají. Pro modely lm je ale nárůst procent úspěšnějších predikcí stejný jako mezi předchozími obdobími. Detaily po kategoriích jsou zobrazeny v tabulce 3.14.

Modely lm_arima mají v tomto období větší procento lepších predikcí skoro ve všech kategoriích. Pouze v kategorii *Trvanlivé* mají, stejně jako v předchozích dvou obdobích, modely lm větší poměr lepších predikcí.

	dle MAE		dle Accuracy		
	model		model		Produktů
Kategorie	lm_arima	lm	lm_arima	lm	v kategorii
Drogerie a kosmetika	**78.6 %**	50 %	64.3 %	71.4 %	14
Maso a ryby	**90 %**	30 %	50 %	30 %	10
Mléčné a chlazené	62.9 %	51.4 %	57.9 %	47.9 %	140
Mražené	77.8 %	66.7 %	88.9 %	100 %	9
Nápoje	76.4 %	**73.6 %**	84.7 %	84.7 %	72
Ovoce a zelenina	58.3 %	56.2 %	51.0 %	51.0 %	96
Pekárna a cukrárna	62.2 %	40 %	46.7 %	28.9 %	45
Trvanlivé	53.7 %	59.7 %	49.3 %	61.2 %	67
Uzeniny a lahůdky	77.4 %	**67.7 %**	67.7 %	64.5 %	31

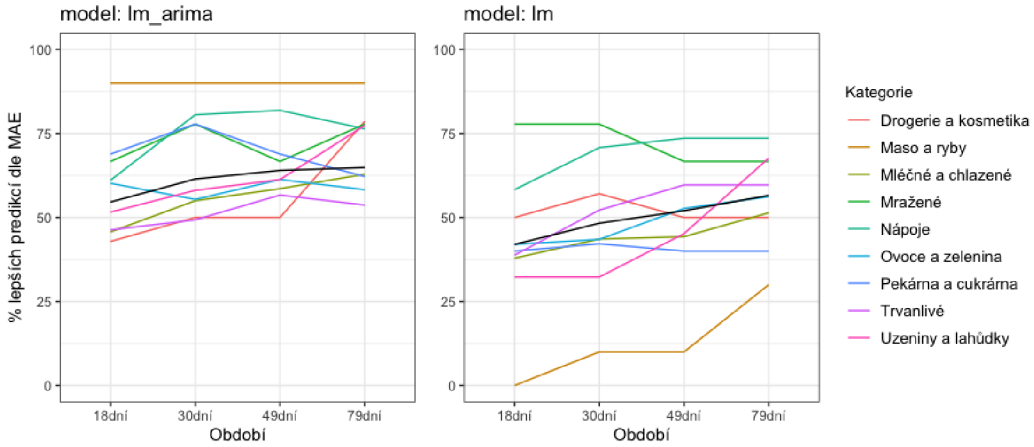
Tabulka 3.14: Kategorie – procento lepších predikcí přes 79 dní

3.6. Celkové hodnocení

V předchozích kapitolách jsme hodnotili kvality predikce z modelů lm_arima a lm a porovnávali jsme je s kvalitou predikce nástroje Veritica. Porovnávání úspěšnosti bylo provedeno pomocí metrik MAE a Accuracy.

Hodnocení dle MAE

Na následujících grafech 3.11 můžeme vidět, jak se poměr lepších předpovědí podle metriky MAE měnil napříč různě dlouhými obdobími. Detailně jsou procenta vykreslena barevně pro každou kategorii a černá křivka zobrazuje celkovou procentuální úspěšnost.



Obrázek 3.11: Procenta predikcí, která byla podle metriky MAE lepší než predikce Veritica v závislosti na délce predikovaného období

Procento kvalitnějších predikcí se s prodlužujícím časovým obdobím většinou spíše zvyšuje nebo zůstává stejné. Obecně by se dalo říci, že modely lm_arima dávaly více lepších predikcí než modely lm a byly úspěšnější ve většině kategorií.

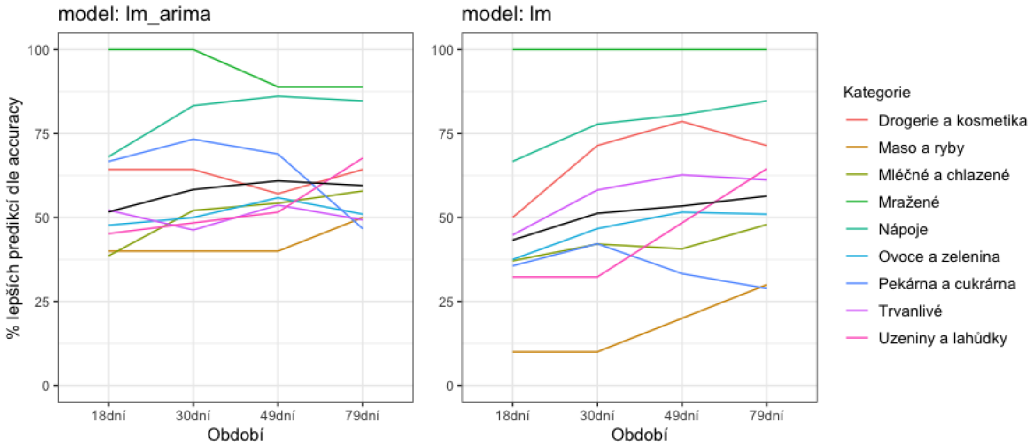
U modelu lm_arima můžeme říci, že v rámci časových úseků delších než 30 dní byl poměr úspěšnějších predikcí v každé kategorii větší než 50 %. Prodeje produktů v kategorii *Maso a ryby* jsou stabilně lépe předpovězeny pomocí modelu lm_arima v 90 % případech. U kategorie *Pekárna a cukrárna* nastal zlom ve 30denním období a s delším obdobím se začal poměr lepších predikcí snižovat.

Model lm má nejhorší poměr kvalitnějších předpovědí u kategorie *Maso a ryby*, kterou jsme pomocí modelu lm_arima naopak předpověděli lépe než Veritico ve většině případů, avšak s prodlužujícím se časovým obdobím už predikce z modelu lm dosahovaly úspěchu alespoň kolem 25 %. U kategorie *Uzeniny a lahůdky* nastal zlom u 30denního období a poměr úspěšnosti předpovědí se začal s delším obdobím prudce zvyšovat.

Ačkoliv jsme očekávali, že denní přístup k predikcím bude mít úspěšnost hlavně v kratších časových obdobích, opak je pravdou a i 79 dní dopředu zvládneme pomocí toho přístupu předpovědět ve většině případů lépe.

Hodnocení dle Accuracy

Na obrázku 3.12 máme vykreslené procentuální úspěšnosti modelů v kategoriích v průběhu období, tentokrát podle metriky Accuracy. Barevné značení kategorií je stejné jako na předchozím grafu, černá křivka opět značí celková procenta úspěšnosti.



Obrázek 3.12: Procenta predikcí, která byla podle metriky Accuracy lepší než predikce Veritica v závislosti na délce predikovaného období

Podle této metriky dává více lepších predikcí stále model lm_arima oproti modelu lm, ačkoliv se procentuální úspěšnost, oproti té dle hodnot MAE, lehce zmenšila. Je zajímavé, že kategorie *Maso a ryby*, která byla dle metriky MAE

konstantně úspěšnější v 90 %, má dle Accuracy poměrně nízké procento lepších predikcí. Je to pravděpodobně způsobeno tím, že výpočet Accuracy podle upravené formulace Logia ořezává extrémní hodnoty. V případě, že odchylky predikcí od skutečných hodnot budou hodně velké, míra Accuracy už nezaznamená, jak velké byly, a její hodnota bude rovna 0 (v případě záporné Accuracy se jako hodnota bere také 0). Pokud tedy Veritico i modely lm_arima předpověděly hodnoty, které se výrazně lišily od skutečné hodnoty, Accuracy bude pro všechny modely nulová. Žádná předpověď tak nebude vyhodnocena jako lepší, respektive všechny budou vnímány jako stejně špatné.

Závěr

Cílem této diplomové práce bylo vytvořit model poptávky a následná predikce denních prodejů maloobchodu s potravinami a základní drogerií. Dále bylo cílem porovnat výsledné hodnoty predikcí s aktuálním přístupem nástroje Veritico. Ten počítá předpovědi v měsíční agregaci, a pak je pomocí koeficientů rozpočítá na denní hodnoty.

V první kapitole jsme si vysvětlili, co si představit pod pojmem řízení dodavatelských řetězců a jak je pro tuto oblast řízení důležité vytvářet předpovědi o budoucí poptávce. Ve druhé kapitole jsme si představili matematické metody, které byly využity při tvorbě modelů poptávky a pomocí kterých jsme vypočítávali predikce. Popsali jsme si regresní analýzu, Box-Jenkinsovu metodologii a představili jsme čtenáři, jak lze využít tyto dva přístupy v jednom modelu pomocí lineární regrese s chybami modelovanými ARIMA procesy. Nakonec jsme čtenáři představili Fourierovu transformaci, kterou jsme využili pro aproximaci sezónnosti časové řady.

V poslední kapitole jsme ukázali, jak byly pro každý produkt vytvořeny dva modely poptávky. Pomocí těch jsme následně předpověděli budoucí hodnoty prodejů 489 produktů na čtyři různě dlouhá časová období. Výsledky predikcí jsme porovnali s predikcemi nástroje Veritico a podle metrik MAE a Accuracy jsme vyhodnotili, v kolika procentech byly predikce námi vytvořenými modely přesnější.

Ve výsledku se nám povedlo pomocí modelu lm_arima předpovědět budoucí prodeje u více než 50 % produktů lépe, než zvládlo Veritico, a to jak v případě validace pomocí míry MAE, tak i Accuracy. V případě různých predikčních období jsme s prodlužujícím se časovým úsekem většinou zaznamenali větší procento

lepších predikcí. Výjimku tvoří kategorie *Pekárna a cukrárna*, u které začalo procento úspěšnosti modelu lm_arima od 30denního období dál výrazně klesat.

Modely představené v této práci by mohly být obohaceny o další proměnné, které z důvodu omezené výpočetní síly nebo nedostupnosti dat nebylo možné do modelů zahrnout. Velmi významnou proměnnou by mohla být například informace o vyprodání produktu nebo stavu skladů. Pomocí ní bychom mohli zjistit skutečnou výši poptávky navzdory tomu, že byly produkty z nějakého důvodu nedostupné.

Vytvořené modely byly validovány pouze na vzorku 489 produktů, které měly v posledním roce největší prodané množství. Nelze tak vyvodit závěry o tom, jak by oproti Veriticu obstály predikce pro produkty, které nemívají vysoké prodeje.

Literatura

[1] Menzer, John T.: *Supply Chain Management.* Thousand Oaks, CA: SAGE, 2001.

[2] Pernica, P.: *Logistika pro 21. století: (Supply chain management).* Praha: Radix, 2005.

[3] Hugos, Michael H.: *Essentials of supply chain management.* Hoboken, NJ: John Wiley & Sons, Inc., 2018.

[4] Russell, Roberta S. and Taylor, Bernard W.: *Operations management: Creating value along the supply chain, 7th edition.* John Wiley & Sons, 2010.

[5] Waller, D. L.: *Operations management: A supply chain approach, 2nd edition.* London: Thomson Learning, 2003.

[6] Pražská, L., Jindra J.: *Obchodní podnikání.* Praha: Management Press, 1997.

[7] Logio [online]. [cit. 2022-10-30]. Dostupné z: https://logio.cz/

[8] Veritico [online]. [cit. 2022-10-30]. Dostupné z: https://veritico.cz/

[9] COVID-19: Stringency Index. Our World in Data [online]. [cit. 2022-11-30]. Dostupné z:https://ourworldindata.org/covid-stringency-index

[10] Hale, T., Angrist, N., Goldszmidt, R. et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nat Hum Behav 5, 529–538 (2021). https://doi.org/10.1038/s41562-021-01079-8

[11] Standardizace dat. Matematická biologie [online]. Institut biostatistiky a analýzy Lékařské fakulty Masarykovy univerzity [cit. 2022-11-28]. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat–vicerozmerne-metody-pro-analyzu-dat–vicerozmerna-rozdeleni-pravdepodobnosti–transformace-dat–standardizace-dat

[12] Hron, K., Kunderová P.: *Základy počtu pravděpodobnosti a metod matematické statistiky (2. dopl. vydání).* Univerzita Palackého v Olomouci, Olomouc, 2015

[13] Cipra, T.: *Finanční ekonometrie. 2. upr. vyd.* Ekopress, Praha, 2013

[14] The ARIMAX model muddle. Rob J Hyndman [online]. [cit. 2022-12-05]. Dostupné z: https://robjhyndman.com/hyndsight/arimax/

[15] Hyndman, Rob J., Koehler, Anne B. Koehler. Another look at measures of forecast accuracy. International Journal of Forecasting [online]. 2006, 22(4), 679-688 [cit. 2022-12-02]. ISSN 0169-2070. Dostupné z:https://www.sciencedirect.com/science/article/pii/S0169207006000239

[16] R: A language and environment for statistical computing: R Foundation for Statistical Computing [online]. Vienna, Austria: R Core Team, 2021 [cit. 2022-12-06]. Dostupné z: https://www.R-project.org/

[17] Brunton, Steven L., Kutz, Jose N. *Data-driven science and engineering: machine learning, dynamical systems, and control.* [online] Cambridge University Press, Cambridge, 2019. Dostupné z: http://databookuw.com/databook.pdf