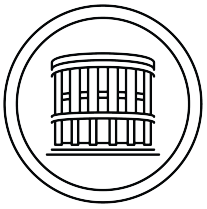
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



NEGAUSSOVSKÉ ČASOVÉ RADY

DIPLOMOVÁ PRÁCA

2026 Bc. Natália ZIFČÁKOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

NEGAUSSOVSKÉ ČASOVÉ RADY

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	Matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Jozef Kováč, PhD.

Bratislava 2026 **Bc. Natália ZIFČÁKOVÁ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

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

Názov: Negaussovské časové rady
Non-Gaussian Time Series

Anotácia: Pri klasických ARMA modeloch pre časové rady sa bežne uvažuje nejaká forma normality (napríklad pri odhade parametrov metódou maximálnej vierohodnosti, predikčných intervaloch a podobne). V niektorých typoch časových radov ale predpoklad normality nie je úplne na mieste - napríklad ak máme časový rad počtov (nezáporné celé čísla) alebo dokonca binárnych hodnôt. V takýchto prípadoch môže byť priamočiare použitie ARMA modelov nevhodné. Cieľom práce je spraviť prehľad modelov časových radov pre vybrané špecifické typy dát, ilustrovať ich na reálnych dátach a prípadne porovnať porovnateľné modely.

Vedúci: Mgr. Jozef Kováč, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 09.01.2025

Dátum schválenia: 15.01.2025

prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Čestne vyhlasujem, že celú diplomovú prácu na tému „Negaussovské časové rady“, vrátane všetkých jej príloh a obrázkov, som vypracovala samostatne, a to s použitím literatúry uvedenej v priloženom zozname a nástrojov umelej inteligencie. Vyhlasujem, že nástroje umelej inteligencie som použila v súlade s príslušnými právnymi predpismi, akademickými právami a slobodami, etickými a morálnymi zásadami za súčasného dodržania akademickej integrity a že ich použitie je v práci vhodným spôsobom označené.

.....

Bc. Natália Zifčáková

Poďakovanie Touto cestou by som sa chcela poďakovať vedúcemu diplomovej práce Mgr. Jozefovi Kováčovi, PhD. za ochotu, odborné rady a usmernenie pri písaní tejto práce.

Abstrakt

Bc. ZIFČÁKOVÁ, Natália : Negaussovské časové rady [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; vedúci práce: Mgr. Jozef Kováč, PhD, Bratislava, 2026, 62 s.

Cieľom tejto diplomovej práce bolo spraviť prehľad modelov časových radov pre vybrané špecifické typy dát, ilustrovať ich na reálnych dátach a prípadne porovnať porovnateľné modely. Práca je rozdelená do štyroch kapitol. Prvá kapitola je venovaná krátkemu zhrnutiu základných pojmov časových radov. Druhá kapitola sa venuje časovým radom počtov. V tejto kapitole definujeme tri modely – INAR, ACP a PAR model a uvedieme ich vlastnosti. Venujeme sa distribúcii odhadnutých parametrov, tvorbe predikcií a následnej diagnostike modelov, na základe ktorej je vybraný najvhodnejší model pre konkrétne dáta. Tretia kapitola sa zameriava na časové rady proporcií, kde definujeme βARMA model. Štvrtá kapitola je venovaná binárnym časovým radom, kde analyzujeme BAR model. V oboch kapitolách sa zameriavame na distribúciu odhadnutých parametrov a tvorbe predikcií.

Kľúčové slová: časové rady počtov, časové rady proporcií, binárne časové rady, predikcie, diagnostika modelu.

Abstract

Bc. ZIFČÁKOVÁ, Natália: Non-Gaussian Time Series [Master's Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Jozef Kováč, PhD., Bratislava, 2026, 62 p.

The aim of this master's thesis was to make an overview of time series models for selected specific data types, to illustrate their application using real-world data, and possibly to compare comparable models. The thesis is divided into four chapters. The first chapter is devoted to a brief summary of the fundamental concepts of time series analysis. The second chapter is devoted to count time series. In this chapter, we define three models, the INAR, ACP and PAR models and their properties. We focus on the distribution of estimated parameters, forecasting, and model diagnostics, based on which the most appropriate model is selected for the given data. The third chapter focuses on the time series of proportions, where we define the βARMA model. The fourth chapter is devoted to binary time series, where we analyse the BAR model. In both chapters, we focus on the distribution of estimated parameters and forecasting.

Keywords: count time series, proportion time series, binary time series, forecast, model diagnostic.

Obsah

Úvod **10**

1 Základné pojmy časových radov **12**

2 Časové rady počtov **16**

2.1 Integer-Valued Autoregressive Model - INAR model 16

2.1.1 Binomický thinning operátor 16

2.1.2 INAR(1) 19

2.1.3 INAR(p) 22

2.2 Autoregressive Conditional Poisson Model - ACP model 23

2.2.1 ACP(p, q) model 23

2.3 Odhad parametrov 25

2.3.1 Model ACP 26

2.3.2 Model PAR 27

2.3.3 Model INAR 28

2.4 Predikcie 29

2.4.1 Model ACP 30

2.4.2 Model PAR 32

2.4.3 Model INAR 33

2.5 Pearsonove rezíduá 36

2.6 Predikcie pre sezónne dáta 39

2.7 Predikcie pomocou ARMA modelu 42

3 Časové rady proporcií **44**

3.1 Beta autoregressive moving average model - βARMA model 44

3.2 Odhad parametrov 46

3.3 Predikcie 47

3.4 Predikcie pomocou ARMA modelu 50

4 Binárne časové rady **52**

4.1 Binárny autoregresný model - BAR model 52

4.2 Odhad parametrov 53

4.3 Predikcie 55

Záver **58**

Zoznam použitej literatúry **60**

Príloha **62**

Úvod

Časové rady patria medzi základné nástroje štatistického modelovania, ktoré nám umožňujú opísať a predpovedať vývoj javu v čase na základe jeho minulých hodnôt. Tradičné prístupy, ako ARMA a ARIMA modely, sú postavené na predpoklade normálneho rozdelenia. V mnohých reálnych aplikáciách však tento predpoklad nie je splnený, keďže analyzované dáta vykazujú diskrétnosť, ohraničenie alebo nadobúdajú len binárne hodnoty. Tieto obmedzenia motivujú prechod k negaussovským časovým radom, ktoré umožňujú flexibilnejšie a realistickejšie modelovanie skutočných dát.

Negaussovské časové rady nachádzajú široké uplatnenie v rôznych oblastiach. V epidemiológií sa využívajú pri sledovaní počtu nových prípadov ochorenia. V poisťovníctve slúžia na monitorovanie počtu nových škodových udalostí. Ďalším príkladom sú napríklad volebné preferencie alebo oblačnosť v meteorológii, ktoré často bývajú vyjadrené v percentách. Zaujímavou aplikáciou je aj kontrola kvality výroby, kde 1 označuje chybný výrobok a 0 označuje bezchybný výrobok.

Negaussovským časovým radom je venovaných veľa kníh a vedeckých článkov. Táto oblasť sa aj v súčasnosti aktívne rozvíja a neustále vznikajú nové prístupy a metódy, ktoré rozširujú jej teoretické aj aplikačné možnosti. My sa v tejto práci zameriame na Integer-Valued Autoregressive Model, skrátene INAR model, Autoregressive Conditional Poisson Model, skrátene ACP model a Poisson Autoregressive model, skátene PAR model, ktorý je špeciálnym prípadom ACP modelu. Tieto modely sú kategorizované ako časové rady počtov. Časové rady proporcií budeme modelovať pomocou Beta Autoregressive Moving Average modelu, ktorého skratka je βARMA model. V prípade binárnych časových radov sa budeme zaoberať binárnym autoregresným modelom, skrátene BAR model.

V tejto práci budeme vychádzať z [1], [2], [12], [17], [18] a [20], ktoré sa venujú teórii INAR modelu. Naopak, zdroje [4], [9] a [19] sa venujú teoretickým poznatkom ACP a PAR modelu. V prípade časových radov proporcií budeme vychádzať najmä z [15]. V prípade binárnych časových radov budeme čerpať poznatky z [10] a [11].

Hlavným cieľom tejto diplomovej práce je spraviť prehľad modelov časových radov pre vybrané špecifické typy dát, ilustrovať ich na reálnych dátach a prípadne porovnať porovnateľné modely.

Práca je rozdelená do štyroch hlavných kapitol. Prvá kapitola je venovaná základným

pojmom časových radov. V druhej kapitole sa venujeme modelom časových radov počtov. Definujeme v nej jednotlivé modely spolu s ich vlastnosťami. V jednotlivých podkapitolách sa venujeme odhadu parametrov modelov, kde simulujeme každý model s nami zadanými parametrami a riešime predikcie z teoretického aj praktického hľadiska. Na záver druhej kapitoly sa venujeme diagnostike modelov, kde uvádzame Pearsonove rezíduá. Tretia kapitola je venovaná časovým radom proporcií, teda časovým radom vyjadreným v percentách. Taktiež sa v tejto kapitole venujeme simulácii modelu a následnému odhadu parametrov. Na reálnych dátach rátame predikcie a taktiež uvádzame aj ich teoretické odvodenie. V prípade nezáporných časových radov a časových radov proporcií porovnávame úspešnosť daných modelov s ARMA modelom, pričom je použitá vhodná transormácia dát, aby sa daný typ dát dal aplikovať na ARMA model. V poslednej štvrtej kapitole sa venujeme časovým radom núl a jednotiek. Aj v tomto prípade analyzujeme odhad parametrov nami simulovaného modelu. Zameriavame sa na analýzu predikcií z teoretického aj praktického hľadiska s ich aplikáciou na reálne dáta.

1 Základné pojmy časových radov

V tejto kapitole uvedieme len základné pojmy časových radov, na ktorých budeme stavať v ďalších kapitolách tejto práce. Budeme vychádzať z kníh [3] a [13].

Časový rad je chronologicky usporiadaná postupnosť hodnôt určitého náhodného javu, ktorú možno matematicky vyjadriť ako $\{X_t\}_{t\in\mathbb{Z}}$. Pod pojmom časový rad sa často myslia aj priamo číselné dáta x_t, $t = 1, ..., T$, ktoré predstavujú realizácie náhodných premenných $X_1, ..., X_t$.

Pod pojmom **stredná hodnota** označujeme funkciu μ_t, ktorá je definovaná ako $\mu_t = = E(X_t)$. **Autokovariančná funkcia**, definovaná ako $\gamma(t, t+k) = Cov(X_t, X_{t+k})$, meria kovarianciu medzi hodnotou časového radu v čase t a hodnotou v čase $t + k$. Jednou z kľúčových vlastností, ktorú chceme dosiahnuť v časových radoch, je **stacionarita**. Tá je dôležitá napríklad pri tvorbe predikcií, pretože stabilný charakter časového radu umožňuje jednoduchšie predikovanie budúcich hodnôt procesu. Proces sa považuje za **stacionárny**, ak je jeho stredná hodnota konštantná, disperzia je konečná a autokovariančná funkcia $Cov(X_t, X_k) = \gamma(t, k)$ závisí len od časového oneskorenia $|t - k|$ pre ľubovoľné $t, k \in \mathbb{Z}$. V opačnom prípade sa proces považuje za **nestacionárny**.

Ďalšími významnými pojmami sú **autokorelačná a parciálna autokorelačná funkcia**, skrátene ACF a PACF funkcie. Funkcia ACF je definovaná ako

$$\rho(\tau) = Corr(X_t, X_{t+\tau}) = \frac{Cov(X_t, X_{t+\tau})}{\sqrt{D(X_t)(DX_{t+\tau})}}, \tag{1.1}$$

pričom platí, že $\rho(0) = 1$ a $\rho(-\tau) = \rho(\tau)$. ACF slúži na opis časových závislostí medzi prvkami časového radu. Parciálna korelácia medzi X_t a X_{t-k} je korelácia, ktorá pretrváva, ak sa eliminujú všetky ostatné vplyvy $X_{t-1}, X_{t-2}, ..., X_{t-k+1}$.

Náhodný proces u_t, nazývaný ako **biely šum**, je charakterizovaný nasledujúcimi vlastnosťami

$$\begin{aligned} E(u_t) &= 0 \quad \forall t, \\ D(u_t) &= \sigma^2 \quad \forall t, \\ Cov(u_t, u_s) &= 0 \quad \forall t \neq s. \end{aligned} \tag{1.2}$$

V prípade bieleho šumu ani ACF, ani PACF nevykazujú žiadnu významnú koreláciu pri žiadnom oneskorení okrem nulového.

Pred definovaním jednotlivých modelov predstavíme **operátor posunu**. Nech x_t je hodnota časového radu v čase t. Keď aplikujeme operátor posunu na túto hodnotu, táto operácia nám zabezpečí posun o jednu periódu dozadu, teda platí

$$Lx_t = x_{t-1}. \tag{1.3}$$

Všeobecne platí

$$L^k x_t = x_{t-k}, \tag{1.4}$$

pre $k = \ldots, -1, 0, 1, \ldots$

Jedny z najznámejších modelov časových radov sú autoregresný model, skrátene AR model a model kĺzavého priemeru, ktorého skratka je MA model. Ako už názov napovedá, v AR modeli hodnota procesu v čase t závisí od minulých hodnôt tohto procesu. Biely šum pridáva tomuto procesu náhodnú variabilitu. Model **AR(p)** je definovaný predpisom

$$x_t = \delta + \alpha_1 x_{t-1} + \ldots \alpha_p x_{t-p} + u_t, \tag{1.5}$$

kde u_t je biely šum, členy x_{t-i}, $i = 1, ..., p$ sa nazývajú autoregresné členy a δ, α_i, $i = 1, ..., p$ sú parametre. Využitím operátora posunu (1.3) vieme model AR(p) prepísať do tvaru

$$x_t = \delta + \alpha_1 L x_t + \ldots \alpha_p L^p x_t + u_t, \tag{1.6}$$

a následnou úpravou výrazu dostávame

$$(\alpha_1 L + \ldots \alpha_p L^p) x_t = \delta + u_t. \tag{1.7}$$

Operátor posunu nám umožňuje určiť podmienku stacionarity pre AR(p) model. Tento model je stacionárny, ak korene polynómu $\alpha_1 z + \ldots \alpha_p z^p$ ležia mimo jednotkového kruhu, t.j. ak pre nejaké z^* platí $\alpha_1 z^* + \ldots \alpha_p z^{*p} = 0$, potom $|z^*| > 1$.

Na rozdiel od autoregresného modelu, model kĺzavého priemeru pri modelovaní hodnoty procesu v čase t berie do úvahy hodnoty bielych šumov z minulosti. Model **MA(q)** je definovaný predpisom

$$x_t = \delta + u_t - \beta_1 u_{t-1} - \ldots - \beta_q u_{t-q}, \tag{1.8}$$

kde u_{t-i}, $i = 1, ..., q$ sú biele šumy a δ, β sú parametre. Ak tento model opäť prepíšeme pomocou operátora posunu (1.3) a členy preusporiadame, model prejde do tvaru

$$x_t - \delta = (1 - \beta_1 L - \ldots - \beta_q L^q) u_t. \tag{1.9}$$

Model MA(q) je vždy stacionárny. Pri MA modeloch je dôležitou podmienkou invertovateľnosť. Tá zabezpečuje, že MA model má jednoznačnú parametrizáciu, keďže dve rôzne hodnoty parametrov môžu viesť k rovnakej ACF. Model MA je invertovateľný, ak korene polynómu $1 - \beta_1 z - \ldots - \beta_q z^q$ ležia mimo jednotkového kruhu. Ak je model MA(q) invertovateľný, dá sa zapísať ako AR(∞).

Zovšeobecnením AR a MA modelov je ARMA model, ktorý modeluje hodnotu procesu v čase t pomocou minulých hodnôt tohto procesu a bielych šumov. Model **ARMA(p, q)** je definovaný nasledovne

$$x_t = \delta + \alpha_1 x_{t-1} + \cdots + \alpha_p x_{t-p} + u_t - \beta_1 u_{t-1} - \ldots - \beta_q u_{t-q}. \tag{1.10}$$

Tento proces môžeme opäť prepísať pomocou operátora posunu (1.3) a dostávame

$$(\alpha_1 L + \ldots \alpha_p L^p) x_t = \delta + (1 - \beta_1 L - \ldots - \beta_q L^q) u_t. \tag{1.11}$$

Označme zátvorky na oboch stranách ako

$$\alpha(L) = \alpha_1 L + \ldots \alpha_p L^p, \tag{1.12}$$

$$\beta(L) = 1 - \beta_1 L - \ldots - \beta_q L^q. \tag{1.13}$$

Podmienkou je, že $\alpha(L)$ a $\beta(L)$ nemôžu mať rovnaké korene. Model ARMA(p, q) je stacionárny, ak korene polynómu autoregresnej časti $\alpha(L)$ ležia mimo jednotkového kruhu, podobne ako v prípade AR(p) modelu.

Pri modeloch typu ARMA je najskôr potrebné určiť ich rád, čo sa zvyčajne robí pomocou ACF a PACF funkcií alebo na základe informačných kritérií, ako sú AIC alebo BIC.

V praxi sa zvykne overovať nekorelovanosť rezíduí modelu pomocou **Ljung-Boxovho testu**, ktorý testuje nasledovnú hypotézu

$$H_0 : \rho(1) = \rho(2) = ... = \rho(\tau) = 0,$$

kde τ je počet lagov a $\tau < T$. Testová štatistika je definovaná nasledovne

$$Q = T(T+2) \sum_{i=1}^{\tau} \frac{\hat{\rho}(i)^2}{T-i}. \tag{1.14}$$

Ak riešime rezíduá modelu, táto testová štatistika sa riadi Chí-kvadrát rozdelením s $\tau - m$ stupňami voľnosti, kde m je počet parametrov modelu. Hodnoty $\hat{\rho}(i)$ predstavujú odhady autokorelačnej funkcie získané z dát.

ARMA modely sú založené na predpoklade normálneho rozdelenia. Podmienka normality v dátach z reálneho sveta však nemusí byť vždy zaručená. Napríklad pri nezáporných celočíselných časových radoch by biele šumy a parametre museli byť tiež nezáporné celé čísla, preto modelovanie týchto časových radov nie je vhodné pomocou ARMA modelov. Práve kvôli tomuto obmedzeniu normality boli vytvorené špeciálne modely pre jednotlivé typy dát.

2 Časové rady počtov

Táto kapitola bude venovaná časovým radom počtov. Sú to časové rady, ktorých hodnoty sú kladné celé čísla. Zameriame sa na dva hlavné modely, a to Integer-Valued Autoregressive Model, skrátene INAR model a Autoregressive Conditional Poisson model, ktorého skratka je ACP model. V prípade ACP modelu sa uvádza aj Poisson Autoregressive model, skrátene PAR model, ktorý je špeciálnym prípadom ACP modelu.

V tejto kapitole postupne definujeme jednotlivé modely, uvedieme ich vlastnosti a pozrieme sa na odhad parametrov a na predikcie z teoretickej aj praktickej stránky. Taktiež ukážeme aplikáciu sezónnych dát, kde využijeme kovariáty a na záver kapitoly porovnáme výkonnosť ACP a ARMA modelov na simulovaných dátach.

2.1 Integer-Valued Autoregressive Model - INAR model

V tejto časti budeme čerpať najmä z [1], [2], [3], [12], [18] a [20]. Prvé dva články sú od autorov Alzaid a Al-Osh, ktorí v roku 1987 publikovali článok, v ktorom predstavili INAR(1) model a o tri roky neskôr spolu publikovali článok, ktorý sa zaoberal jeho rozšírenou verziou, modelom INAR(p). V článku [18] sú bližšie popísané vlastnosti INAR(1) modelu, kniha [3] a články [12] a [20] sa zaoberajú tzv. binomickým thinning operátorom (z angl. binomial thinning operator) a teóriou INAR(1) modelu.

2.1.1 Binomický thinning operátor

Pred definovaním samotného INAR modelu vysvetlíme operáciu $\alpha \circ X$. Výraz $\circ$ sa nazýva tzv. binomický thinning operátor.

Definícia 2.1.
Nech X je nezáporná celočíselná náhodná premenná definovaná na $\{0, 1, \dots\}$, potom pre každé $\alpha \in [0, 1]$ je operátor $\circ$ definovaný ako

$$\alpha \circ X = \sum_{i=1}^{X} Y_i, \tag{2.15}$$

kde Y_i je i.i.d postupnosť Bernoulliho náhodných premenných, ktoré sú nezávislé od X, pričom platí

$$\mathrm{P}(Y_i = 1) = 1 - \mathrm{P}(Y_i = 0) = \alpha. \tag{2.16}$$

Teda Y_i nadobúda hodnotu 1 s pravdepodobnosťou α a hodnotu 0 s pravdepodobnosťou $1-\alpha$. Binomický thinning operátor $\alpha \circ X$ môže vygenerovať 0, ak $X = 0$.

Z definície operátora $\circ$ je zrejmé, že:

$$0 \circ X = 0, \tag{2.17}$$

$$1 \circ X = X, \tag{2.18}$$

$$\beta \circ (\alpha \circ X) = (\beta\alpha) \circ X, \quad \text{pre ľubovoľné } \beta \in [0,1]. \tag{2.19}$$

Tvrdenie 2.1.
Stredná hodnota, variancia a kovariancia operátora $\alpha \circ X$ sú dané nasledovnými vzťahmi

$$\mathrm{E}(\alpha \circ X) = \alpha\,\mathrm{E}(X), \tag{2.20}$$

$$\mathrm{Var}(\alpha \circ X) = \alpha^2 \mathrm{Var}(X) + \alpha(1-\alpha)\mathrm{E}(X), \tag{2.21}$$

$$\mathrm{Cov}(\alpha \circ X, X) = \alpha \mathrm{Var}(X). \tag{2.22}$$

V nasledujúcej časti dokážeme vzťahy (2.20), (2.21) a (2.22). Využijeme v nich základné vlastnosti binomického rozdelenia, a to $E(X) = np$ a $D(X) = np(1-p)$.

Kedže v definícii (2.15) vystupuje X ako počet, $\alpha \circ X$ vytvoríme tak, že spravíme $n = X$ nezávislých Bernoulliho pokusov, každý s pravdepodobnosťou úspechu $p = \alpha$, teda z toho vyplýva

$$\alpha \circ X \mid X \sim Bin(X, \alpha). \tag{2.23}$$

Dôkaz (2.20):
V nasledujúcej časti dokážeme platnosť vzťahu pre strednú hodnotu. V dôkaze využijeme vzorec pre podmienenú strednú hodnotu

$$E(X) = E[E(X|Y)]. \tag{2.24}$$

V prvom kroku aplikujeme tento vzorec pre $\alpha \circ X$ a dostávame

$$E(\alpha \circ X) = E[E(\alpha \circ X \mid X)].$$

Z (2.23) vyplýva, že $E(\alpha \circ X \mid X) = \alpha X$ a následným využitím linearity strednej hodnoty dostávame výsledok

$$E(\alpha \circ X) = E(E(\alpha \circ X \mid X)) = E(\alpha X) = \alpha E(X). \quad \square$$

Dôkaz (2.21):

V tejto časti dokážeme platnosť vzťahu pre varianciu. V tomto prípade využijeme vzorec pre podmienenú varianciu v tvare

$$Var(X) = E[Var(X \mid Y)] + Var[E(X \mid Y)]. \tag{2.25}$$

Tento vozorec upravíme pre $\alpha \circ X$ a dostávame

$$Var(\alpha \circ X) = E[Var(\alpha \circ X \mid X)] + Var[E(\alpha \circ X \mid X)].$$

Využitím vzťahu (2.23) vieme vyjadriť prvý člen ako $E[Var(\alpha \circ X \mid X)] = (1-\alpha)\alpha X$. V druhom člene využijeme vzorec $E(\alpha \circ X \mid X) = \alpha X$ už odvodený z dôkazu (2.20) a následným využitím linearity strednej hodnoty a násobením konštantou vo variancii dostávame

$$\begin{aligned} \mathrm{Var}(\alpha \circ X) &= E\Big[\mathrm{Var}(\alpha \circ X \mid X)\Big] + \mathrm{Var}\Big[E(\alpha \circ X \mid X)\Big] \\ &= E\Big[\alpha(1-\alpha)X\Big] + \mathrm{Var}(\alpha X) \\ &= \alpha(1-\alpha)E(X) + \alpha^2 \,\mathrm{Var}(X). \quad \square \end{aligned}$$

Dôkaz (2.22):

Nakoniec dokážeme vzťah pre kovarianciu. Využijeme známy vzorec pre kovarianciu v tvare $Cov(X,Y) = E(XY) - E(X)E(Y)$. Opäť tento vzorec aplikujeme pre $\alpha \circ X$ a X a dostávame

$$Cov(\alpha \circ X, X) = E\Big[(\alpha \circ X)X\Big] - E(\alpha \circ X)E(X).$$

Rozoberieme oba výrazy postupne. Vo výraze $E[(\alpha \circ X)X]$ opäť aplikujeme vzorec pre podmienenú strednú hodnotu (2.24), čím dostávame $E[(\alpha \circ X)X] = E[E((\alpha \circ X)X \mid X)]$. Kedže X je v tomto prípade fixné, môžeme ho vyňať a využiť z dôkazu (2.20) už odvodený výraz $E(\alpha \circ X \mid X) = \alpha X$, teda pre prvý výraz dostávame

$$\begin{aligned} E\Big[(\alpha \circ X)X\Big] &= E\Big[E\Big((\alpha \circ X)X \mid X\Big)\Big] \\ &= E\Big[X\, E\Big(\alpha \circ X \mid X\Big)\Big] \\ &= E(\alpha X^2) \\ &= \alpha E(X^2). \end{aligned}$$

V druhom výraze $E(\alpha \circ X)E(X)$ využijeme opäť vlastnosť (2.20), teda platí

$$E(\alpha \circ X)E(X) = \alpha E(X)E(X) = \alpha E(X)^2.$$

Na záver už len využijeme vzorec pre varianciu v tvare $Var(X) = E(X^2) - E(X)^2$ a dostávame

$$\begin{aligned} \mathrm{Cov}(\alpha \circ X, X) &= E\Big[(\alpha \circ X)X\Big] - E(\alpha \circ X)\, E(X) \\ &= \alpha E(X^2) - \alpha E(X)^2 \\ &= \alpha\, \mathrm{Var}(X). \quad \square \end{aligned}$$

2.1.2 INAR(1)

Tento model je špeciálnym typom autoregresného modelu určeného pre časové rady s celočíselnými nezápornými hodnotami.

Definícia 2.2.
Model INAR(1) je definovaný ako

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t, \tag{2.26}$$

kde $t = 0, \pm 1, \pm 2, \ldots$, $\alpha \in [0, 1]$, X_t je hodnota celočíselného časového radu v čase t a ε_t je hodnota nezávislého a identicky rozdeleného (i.i.d.) náhodného procesu s celočíselnými nezápornými hodnotami, pričom X_{t-1} a ε_t sú navzájom nezávislé v každom čase. Najčastejšie sa predpokladá, že ε_t má Poissonovo rozdelenie alebo iné diskrétne rozdelenie ako napr. negatívne binomické alebo Bernoulliho.

Ak vzorec (2.26) vyjadríme rekurzívne, dostávame

$$X_t = \alpha^k \circ X_{t-k} + \sum_{j=0}^{k-1} \alpha^j \circ \varepsilon_{t-j}. \tag{2.27}$$

Model INAR(1), definovaný rovnicou (2.26), vyjadruje, že zložky procesu v čase t, sú

(i) preživšie prvky procesu z času $t-1$, pričom každý prežíva s pravdepodobnosťou α,

(ii) prvky, ktoré vstúpili do systému v intervale $(t-1, t]$ ako inovačný člen ε_t.

Vlastnosti INAR(1)

V tejto časti odvodíme vlastnosti INAR(1) modelu všeobecne bez nejakého predpokladu na rozdelenie ε_t. Predpokladajme, že ε_t má strednú hodnotu μ a disperziu σ^2, pričom X_{t-1} a ε_t sú nezávislé pre všetky t.

Stredná hodnota

Pre nepodmienenú strednú hodnotu platí

$$\begin{aligned} \mathrm{E}(X_t) &= \mathrm{E}(\alpha \circ X_{t-1}) + \mathrm{E}(\varepsilon_t) \\ &= \alpha\,\mathrm{E}(X_{t-1}) + \mu \\ &= \alpha(\alpha\,\mathrm{E}(X_{t-2}) + \mu) + \mu \\ &= \alpha^2\mathrm{E}(X_{t-2}) + \alpha\mu + \mu \\ &\;\;\vdots \\ &= \alpha^t\mathrm{E}(X_0) + \mu\sum_{j=0}^{t-1}\alpha^j. \end{aligned} \tag{2.28}$$

Pre podmienenú strednú hodnotu platí

$$\begin{aligned} \mathrm{E}(X_t \mid X_{t-1}) &= \mathrm{E}(\alpha \circ X_{t-1} + \varepsilon_t \mid X_{t-1}) \\ &= \mathrm{E}(\alpha \circ X_{t-1} \mid X_{t-1}) + \mathrm{E}(\varepsilon_t \mid X_{t-1}) \\ &= \alpha X_{t-1} + \mu. \end{aligned} \tag{2.29}$$

Disperzia

Pre nepodmienenú disperziu platí

$$\begin{aligned} \mathrm{Var}(X_t) &= \alpha^2\mathrm{Var}(X_{t-1}) + \alpha(1-\alpha)\mathrm{E}(X_{t-1}) + \sigma^2 \\ &= \alpha^{2t}\mathrm{Var}(X_0) + (1-\alpha)\sum_{j=1}^{t}\alpha^{2j-1}\mathrm{E}(X_{t-j}) + \sigma^2\sum_{j=1}^{t}\alpha^{2(j-1)}. \end{aligned} \tag{2.30}$$

V prvom kroku odvodenia sme využili vlastnosť variancie (2.21) a následne sme sa rekurzívne dopracovali k vyjadreniu variancie.

Pre podmienenú disperziu platí

$$\begin{aligned} \mathrm{Var}(X_t \mid X_{t-1}) &= \mathrm{Var}(\alpha \circ X_{t-1} + \varepsilon_t \mid X_{t-1}) \\ &= \mathrm{Var}(\alpha \circ X_{t-1} \mid X_{t-1}) + \mathrm{Var}(\varepsilon_t \mid X_{t-1}) \\ &= \alpha(1-\alpha)X_{t-1} + \sigma^2. \end{aligned} \tag{2.31}$$

Pri odvodení vzorca pre podmienenú disperziu sme v predposlednom kroku využili vlastnosť (2.23), teda platí $\mathrm{Var}(\alpha \circ X_{t-1} \mid X_{t-1}) = \alpha(1-\alpha)X_{t-1}$ a z definovania všeobecných vlastností ε_t vyplýva, že $\mathrm{Var}(\varepsilon_t) = \sigma^2$, keďže ε_t a X_{t-1} sú nezávislé.

Stacionarita

Model INAR(1) je stacionárny, ak $\alpha < 1$. Potom pre jeho strednú hodnotu a varianciu platia nasledovné vzťahy

$$\mathrm{E}(X_t) = \frac{\mu}{1-\alpha}, \tag{2.32}$$

$$\mathrm{Var}(X_t) = \frac{\alpha\mu + \sigma^2}{1-\alpha^2}. \tag{2.33}$$

Tieto vzťahy dostaneme úpravou (2.28) a (2.30), pričom využijeme vzorce pre súčet geometrických radov.

V prípade, že ε_t má Poissonovo rozdelenie s parametrom λ, pre Poissonovo rozdelenie platí, že $\mu = \mathrm{E}(\varepsilon_t) = \lambda$ a $\sigma^2 = \mathrm{Var}(\varepsilon_t) = \lambda$. Potom výrazy (2.32) a (2.33) prejdú do nasledovného tvaru

$$\mathrm{E}(X_t) = \frac{\lambda}{1-\alpha}, \tag{2.34}$$

$$\mathrm{Var}(X_t) = \frac{\lambda(1+\alpha)}{1-\alpha^2}. \tag{2.35}$$

Ďalej budeme predpokladať stacionaritu procesu.

Proces INAR(1) vykazuje podobné distribučné vlastnosti ako autoregresný AR(1) model. Ak $X_t \overset{d}{=} X_{t-1}$ a X_{t-1} je nezávislé od ε_t, z platnosti (2.26) pre všetky $\alpha \in (0,1)$ vyplýva, že rozdelnie X_t je jednoznačne určené rozdelením ε_t. Ak má ε_t Poissonovo rozdelenie, aj X_t sa riadi Poissonovým rozdelením.

Kovariancia

Pre kovarianciu tohto modelu platí

$$\begin{aligned}\gamma(k) &= \mathrm{Cov}(X_{t-k}, X_t) \\ &= \mathrm{Cov}(X_{t-k}, \alpha^k \circ X_{t-k}) + \mathrm{Cov}\left(X_{t-k}, \sum_{j=0}^{k-1} \alpha^j \circ \varepsilon_{t-j}\right). \end{aligned} \tag{2.36}$$

$$= \alpha^k \,\mathrm{Cov}(X_{t-k}, X_{t-k}) + \sum_{j=0}^{k-1} \alpha^j \,\mathrm{Cov}(X_{t-k}, \varepsilon_{t-j}) \tag{2.37}$$

$$= \alpha^k \operatorname{Var}(X_{t-k}) \quad = \alpha^k \gamma(0) \tag{2.38}$$

V (2.36) sme využili linearitu kovariancie a rekurzívny vzťah (2.27). V (2.37) sme využili poznatok, že X_{t-j} a ε_{t-j} sú nezávislé pre $j < k$, teda $Cov(X_{t-j}, \varepsilon_{t-j}) = 0$.

Autokorelácie

Pre autokorelácie INAR(1) modelu platí

$$\rho(k) = \alpha^k. \tag{2.39}$$

2.1.3 INAR(p)

V tejto podkapitole priblížime všeobecný model INAR(p). Budeme vychádzať z [1].

Definícia 2.3.

Všeobecný model INAR(p) je definovaný nasledovne

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t, \tag{2.40}$$

kde $t = 0, \pm 1, \pm 2, \ldots$, $\{\varepsilon_t\}_{t \in \mathbb{Z}}$ je postupnosť nezáporných celočíselných i.i.d. náhodných premenných, pričom X_{t-1} a ε_t sú navzájom nezávislé v každom čase.

Z definície binomického thinning operátora (2.15) vyplýva podmienka na parametre α_i, a to $\alpha_i \in [0, 1]$ pre všetky $i = 1, ...p$. Ak platí podmienka $\sum_{i=1}^{p} \alpha_i < 1$, tak model INAR($p$) je stacionárny.

Vlastnosti INAR(p)

Predpokladajme, že ε_t má strednú hodnotu μ a disperziu σ^2, pričom X_{t-1} a ε_t sú nezávislé pre všetky t. Ak je model INAR(p) stacionárny, tak pre jeho strednú hodnotu platí

$$\mathrm{E}(X_t) = \frac{\mu}{1 - \sum_{i=1}^{p} \alpha_i}. \tag{2.41}$$

Autokovariancia INAR(p) modelu je daná vzťahom

$$\gamma(k) = Cov(X_{t-k}, X_t) = \begin{cases} \sum_{i=1}^{p} Cov(X_{t-k}, \alpha_i \circ X_{t-i}) + \sigma^2 & \text{, ak} \quad k = 0 \\ \sum_{i=1}^{p} Cov(X_{t-k}, \alpha_i \circ X_{t-i}) & \text{, inak.} \end{cases}$$

Podrobnejšie vysvetlíme odovodenie vzťahu pre autokovarianciu INAR(p) modelu. Dosadením predpisu modelu (2.40) do vzorca pre autokovarianciu dostávame

$$Cov(X_{t-k}, X_t) = Cov(X_{t-k}, \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t). \tag{2.42}$$

Využitím linearity autokovariancie môžeme rozložiť tento výraz na jednotlivé členy a dostávame

$$Cov(X_{t-k}, X_t) = \sum_{i=1}^{p} Cov(X_{t-k}, \alpha_i \circ X_{t-i}) + Cov(X_{t-k}, \varepsilon_t). \tag{2.43}$$

Ak $k \neq 0$, tak v druhom člene výrazu (2.43) dostávame nulu, pretože vychádzame z predpokladu, že X_{t-k} a ε_t sú navzájom nezávislé pre všetky t. Ak $k = 0$, druhý člen opäť vieme rozpísať pomocou predpisu (2.40) a dostávame

$$Cov(X_t, \varepsilon_t) = Cov(\alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t, \varepsilon_t). \tag{2.44}$$

Využitím linearity autokovariancie opäť môžeme rozložiť tento výraz a dostávame

$$Cov(X_t, \varepsilon_t) = \sum_{i=1}^{p} Cov(\alpha_i \circ X_{t-i}, \varepsilon_t) + Cov(\varepsilon_t, \varepsilon_t). \tag{2.45}$$

Prvý člen (2.45) je nulový, pretože X_{t-i} a ε_t pre $i = 1, \ldots, p$ sú navzájom nezávislé pre všetky t. Nakoniec z vlastností kovariancie v rámci druhého členu (2.45) dostávame

$$Cov(\varepsilon_t, \varepsilon_t) = D(\varepsilon_t) = \sigma^2. \tag{2.46}$$

2.2 Autoregressive Conditional Poisson Model - ACP model

V tejto časti budeme čerpať najmä z článkov [4] a [9] a knihy [19], ktoré sa zaoberajú teóriou ACP modelu.

2.2.1 ACP(p, q) model

Nech X_t je proces nezáporných celých čísel. Nech $\mathcal{F}_t$ označuje informácie dostupné až do času t vrátane. V najjednoduchšom modeli sú počty generované Poissonovým rozdelením s parametrom λ_t, t.j.

$$X_t \mid \mathcal{F}_{t-1} \sim \text{Poiss}(\lambda_t).$$

Definícia 2.4.

Model ACP(p, q) pre $p \geq 1$ a $q \geq 0$ modelujeme nasledovným predpisom

$$\lambda_t = \omega + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}, \tag{2.47}$$

pričom $\omega > 0$, $\alpha_1, ...\alpha_p$, $\beta_1, ...\beta_q \geq 0$.

Ak $q = 0$, model (2.47) sa nazýva Poisson Autoregressive model, skrátene PAR model.

V literatúre sa stretneme s nejednoznačnosťou značenia pre ACP model, pričom niekedy sa nazýva aj Integer-valued Generalized AutoRegressive Conditional Heteroskedasticity, skrátene INGARCH model.

Vlastnosti ACP modelu

V tejto časti odvodíme vlastnosti platné pre ACP(p,q) model.

Stredná hodnota

Za predpokladu, že súčet autoregresných koeficientov je menší ako 1, t.j.

$$\sum_{j=1}^{max(p,q)} \alpha_j + \beta_j < 1 \tag{2.48}$$

platí, že proces ACP(p, q) je stacionárny. V prípade, ak je niektorý z parametrov menší ako $max(p, q)$, chýbajúce členy sa považujú za nulové. Pre strednú hodnotu ACP(p, q) procesu v tomto prípade platí

$$\mathrm{E}(X_t) = \lambda = \frac{\omega}{1 - \left(\sum_{j=1}^{p} \alpha_j + \sum_{j=1}^{q} \beta_j \right)} = \frac{\omega}{1 - \sum_{j=1}^{max(p,q)} (\alpha_j + \beta_j)}. \tag{2.49}$$

V prípade odvodenia tohto vzorca sme využili stacionaritu modelu a $E(X_t) = \mu$.

Disperzia

Pre disperziu ACP(1,1) platí

$$\mathrm{Var}(X_t) = \sigma^2 = \frac{\mu(1 - (\alpha_1 + \beta_1)^2 + \alpha_1^2)}{1 - (\alpha_1 + \beta_1)^2} \geq \mu. \tag{2.50}$$

Model ACP vykazuje nadmernú disperziu aj keď využíva podmienené Poissonovo rozdelenie s rovnakou strednou hodnotou a varianciou. Nadmerná disperzia vzniká v dôsledku

časovej závislosti medzi pozorovaniami a jej miera rastie so zvyšujúcou sa hodnotou parametra α_1 a v menšej miere aj s β_1.

Autokorelácie

Pre autokorelácie ACP(1,1) platí

$$\rho(s) = (\alpha_1 + \beta_1)^{s-1} \frac{\alpha_1(1 - \beta_1(\alpha_1 + \beta_1))}{1 - (\alpha_1 + \beta_1)^2 + \alpha_1^2}. \tag{2.51}$$

Vzorce pre disperziu a autokorelácie sú dokázané v [9].

2.3 Odhad parametrov

V tejto podkapitole sa zameriame na odhady parametrov jednotlivých modelov. V prvom rade definujeme vierohodnostné funkcie modelov a z praktického hľadiska sa v softvéri R pozrieme na odhady parametrov a ich distribúcie. Budeme vychádzať z [3] a [5].

Odhady parametrov sa počítajú pomocou metódy maximálnej vierohodnosti (MLE). Vierohodnostná funkcia pre ACP, resp. pre PAR model je definovaná nasledovne

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{T} P(X_t \mid \mathcal{F}_{t-1}) = \prod_{t=1}^{T} \frac{\lambda_t^{x_t} e^{-\lambda_t}}{x_t!}, \tag{2.52}$$

kde $\boldsymbol{\theta}$ je vektor odhadovaných parametrov a λ_t je definovaná podľa daného modelu pomocou predpisu (2.47). V prípade ACP modelu odhadujeme parametre ω, λ_i, $i = 1, ...p$ a β_j, $j = 1, ...q$. V prípade PAR modelu odhadujeme parametre ω a α_i, $i = 1, ...p$. Vychádzajúc z [5], vierohodnostná funkcia INAR(1) modelu je definovaná ako

$$L(\boldsymbol{\theta}) = \prod_{t=2}^{T} P(X_t \mid X_{t-1}), \tag{2.53}$$

kde podmienená pravdepodobnosť $P(X_t \mid X_{t-1})$ je definovaná nasledovne

$$P(X_t \mid X_{t-1}) = \sum_{i=0}^{min(x_{t-1}, x_t)} \binom{x_{t-1}}{i} \alpha_1^i (1 - \alpha_1)^{x_{t-1} - i} \frac{e^{-\lambda} \lambda^{x_t - i}}{(x_t - i)!}. \tag{2.54}$$

V softvéri R sa na odhadovanie parametrov pre INAR a ACP, resp. PAR modely dajú použiť balíčky `spINAR` a `tscount`. V balíčku `spINAR` existuje funkcia `spinar_est_param`,

ktorá odhaduje parametre INAR(p) modelu, ale je obmedzená len do rádu $p = 2$. V prípade ACP a PAR modelov existuje na odhad parametrov funckia `tsglm`. Na ilustrovanie fungovania týchto funkcií sme ku každému modelu s fixnými hodnotami parametrov vygenerovali 2000 časových radov s dĺžkou $n = 200$, následne sme odhadli parametre a pozreli sme sa na ich distribúciu.

2.3.1 Model ACP

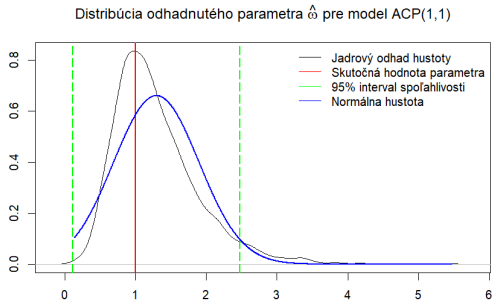
V rámci modelu ACP(1,1) sme simulovali model podľa predpisu (2.47) s parametrami $\omega = 1, \alpha = 0.4$ a $\beta = 0.5$. V prvom kroku simulácie ACP(1,1) modelu sme vypočítali počiatočné hodnoty λ_1 a X_1. Pre výpočet λ_1 sme využili vzorec (2.49) v tvare $\lambda_1 = \frac{\omega}{1-\alpha-\beta}$. Následne sme pomocou funkcie `rpois` s parametrami 1 a λ_1 vygenerovali jednu náhodnú realizáciu X_1, ktorá sa riadi Poissonovým rozdelením s parametrom λ_1. Keď sme už mali vypočítané inicializačné hodnoty v prvej iterácii, pokračovali sme v simulovaní hodnôt λ_t a X_t pre $t = 2, ..., n$. Hodnoty λ_t sme generovali podľa predpisu ACP modelu (2.47) a hodnoty X_t sme opäť v každej iterácii generovali z Poissonovho rozdelenia s parametrom λ_t. Na odhad koeficientov sme využili funkciu `tsglm` z balíka `tscount`, kde sme parameter `model` nastavili na `past_obs` =1 a `past_mean` =1, parameter `link` sme nastavili na *identity*, čím modelujeme proces bez akejkoľvek transformácie. K nastaveniu parametra `link` si bližšie povieme v podkapitole 2.6. Parameter `distr` sme nastavili na *poisson*.

Priemerné hodnoty a disperzie odhadnutých parametrov sú zhrnuté v tabuľke nižšie.

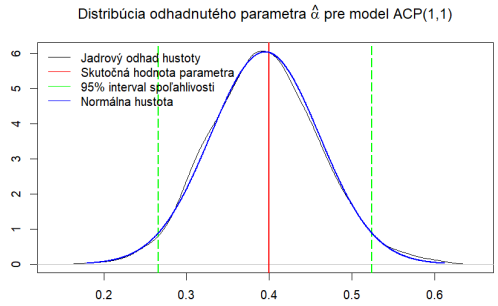
Tabuľka 1: Stredná hodnota a disperzia parametrov pre ACP(1,1) model

Parameter	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$
pôvodná hodnota	1	0.4	0.5
priemer	1.29	0.39	0.47
disperzia	0.36	0.004	0.009

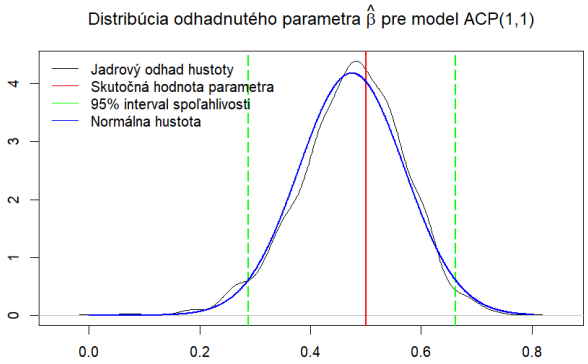
Následne sme vykreslili distribúcie jednotlivých odhadnutých parametrov, pričom sme do grafu vykreslili aj skutočné hodnoty parametrov, s ktorými sme generovali daný proces, 95%-ný interval spoľahlivosti a na porovnanie aj hustotu normálneho rozdelenia s parametrami odhadnutými z dát.



Obr. 1: Distribúcia parametra $\hat{\omega}$ v modeli ACP



Obr. 2: Distribúcia parametra $\hat{\alpha}$ v modeli ACP



Obr. 3: Distribúcia parametra $\hat{\beta}$ v modeli ACP

Z tabuľky 1 a obrázka 1 je vidieť, že priemer parametra $\hat{\omega}$ je síce mierne odlišný od skutočnej hodnoty. V prípade parametrov α a β sú výsledky v súlade s predpokladmi a žiadne extrémne odchýlky sa nevyskytujú, čo je vidieť aj z obrázkov 2 a 3.

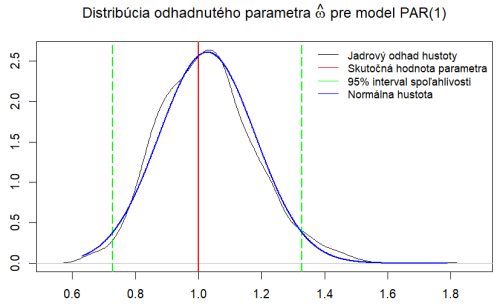
2.3.2 Model PAR

V rámci modelu PAR(1) sme nasimulovali model podľa predpisu (2.47) pre $q = 0$ s parametrami $\omega = 1$ a $\alpha = 0.6$. Postup bol veľmi podobný postupu ako v prípade ACP modelu, len λ_1 sme vypočítali pomocou vzorca $\lambda_1 = \frac{\omega}{1-\alpha}$ a z funkcie `tsglm` sme z parametra `model` odstránili `past_mean` $= 1$. Priemerné hodnoty a disperzie odhadnutých parametrov uvádzame v nasledujúcej tabuľke.

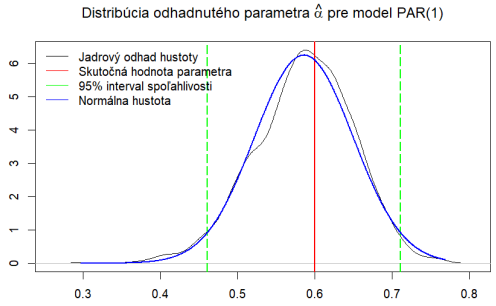
Tabuľka 2: Stredná hodnota a disperzia parametrov pre PAR(1) model

Parameter	$\hat{\omega}$	$\hat{\alpha}$
pôvodná hodnota	1	0.6
priemer	1.03	0.59
disperzia	0.023	0.004

Následne sme vykreslili distribúcie jednotlivých odhadnutých parametrov, pričom sme do grafu vykreslili aj skutočné hodnoty parametrov, s ktorými sme generovali daný proces, 95%-ný interval spoľahlivosti a na porovnanie aj hustotu normálneho rozdelenia s parametrami odhadnutými z dát.



Obr. 4: Distribúcia parametra $\hat{\omega}$ v modeli PAR



Obr. 5: Distribúcia parametra $\hat{\alpha}$ v modeli PAR

V prípade modelu PAR je distribúcia parametrov centrovaná okolo nami zadaných hodnôt s nízkou disperziou, čo je zrejmé aj z tabuľky 2 a obrázkov 4 a 5.

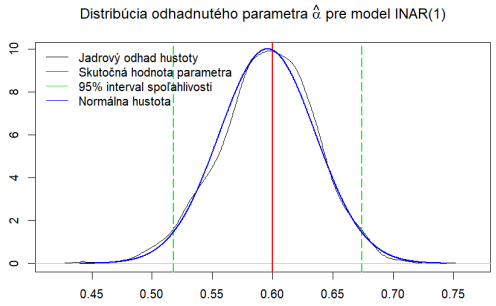
2.3.3 Model INAR

V rámci modelu INAR(1,1) sme simulovali model podľa predpisu (2.26) s parametrami $\lambda = 5$ a $\alpha = 0.6$. V prvom kroku sme vygenerovali počiatočnú hodnotu X_1 pomocou funkcie `rpois` s parametrami 1 a λ. V ďalších iteráciách sme generovali hodnoty X_t pre $t = 2, ..., 200$ pomocou predpisu (2.26). Thinning operátor $\alpha \circ X$ sme generovali pomocou funkcie `rbinom` s parametrami 1, X_{t-1} a α. Hodnoty ε_t sme generovali opäť pomocou funkcie `rpois` s parametrami 1 a λ. Priemerné hodnoty a disperzie odhadnutých parametrov uvádzame v tabuľke nižšie.

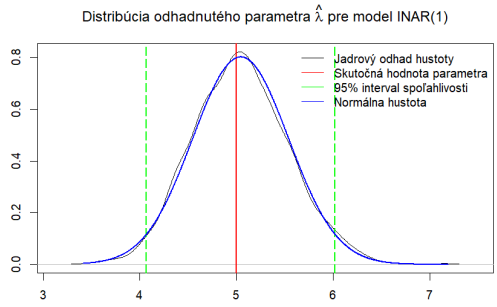
Tabuľka 3: Stredná hodnota a disperzia parametrov pre INAR(1) model

Parameter	$\hat{\alpha}$	$\hat{\lambda}$
pôvodná hodnota	0.6	5
priemer	0.596	5.04
disperzia	0.002	0.25

Opäť sme vykreslili distribúcie jednotlivých odhadnutých parametrov, pričom sme do grafu vykreslili aj skutočné hodnoty parametrov, s ktorými sme generovali daný proces, 95%-ný interval spoľahlivosti a na porovnanie aj hustotu normálneho rozdelenia s parametrami odhadnutými z dát.



Obr. 6: Distribúcia parametra $\hat{\alpha}$ v modeli INAR



Obr. 7: Distribúcia parametra $\hat{\lambda}$ v modeli INAR

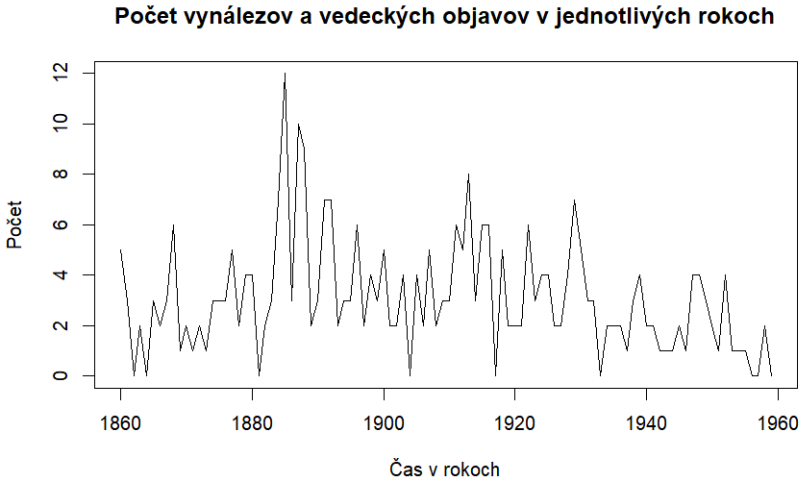
Z obrázkov 6 a 7 vidíme, že distribúcie odhadnutých parametrov $\hat{\alpha}$ a $\hat{\lambda}$ sú centrované okolo skutočných hodnôt parametrov. Z tabuľky 3 je vidieť, že v prípade parametra $\hat{\lambda}$ je disperzia distribúcie vyššia.

2.4 Predikcie

Táto podkapitola bude venovaná tvorbe predikcií z teoretického aj praktického hľadiska. Čerpali sme najmä z [6] a [14], ktoré sa zaoberajú dokumentáciou balíkov `spINAR` a `tscount` v softvéri R.

Dataset, na ktorom sme riešili predikcie, sa nazýva `discoveries`. Tieto dáta reprezentujú počet „veľkých“ vynálezov a vedeckých objavov v každom roku v období od 1860

až po rok 1959 vrátane, priebeh je vidieť na obrázku 8. Priemer týchto dát je 3.1. Tento dataset je možné nájsť v balíku `tscount`. Zdrojom pôvodných dát je kniha *The World Almanac and Book of Facts, 1975 Edition*. Dataset pozostáva zo 100 záznamov, ktoré sme rozdelili do dvoch častí. Na prvých 90 záznamoch tvoríme predikcie a zvyšných desať záznamov využívame na porovnanie s vypočítaným 95%-ným predikčným intervalom.



Obr. 8: Graf významných objavov v každom roku v období od 1860 po rok 1959

2.4.1 Model ACP

Výpočet predikcií ukážeme pre ACP(1,1) model. Pri odvodení predikcií ACP(1,1) modelu vychádzame z predpokladu $X_t \mid \mathcal{F}_{t-1} \sim \text{Poiss}(\lambda_t)$, teda platí $\text{E}(X_t \mid \mathcal{F}_{t-1}) = \lambda_t$.

Jednokrokové predikcie odvodíme iteratívne pre rastúcu veľkosť kroku h. Počítajme:

$$\begin{aligned}
h = 1: \quad \lambda_{t+1} &= \hat{\omega} + \hat{\alpha}_1 x_t + \hat{\beta}_1 \lambda_t \\
h = 2: \quad \lambda_{t+2} &= \hat{\omega} + \hat{\alpha}_1 \hat{x}_{t+1} + \hat{\beta}_1 \hat{\lambda}_{t+1} \\
&= \hat{\omega} + \hat{\alpha}_1 \hat{\lambda}_{t+1} + \hat{\beta}_1 \hat{\lambda}_{t+1} \\
&= \hat{\omega} + (\hat{\alpha}_1 + \hat{\beta}_1) \hat{\lambda}_{t+1} \\
&\vdots \\
h = h: \quad \lambda_{t+h} &= \hat{\omega} + \hat{\alpha}_1 \hat{x}_{t+h-1} + \hat{\beta}_1 \hat{\lambda}_{t+h-1} \\
&= \hat{\omega} + \hat{\alpha}_1 \hat{\lambda}_{t+h-1} + \hat{\beta}_1 \hat{\lambda}_{t+h-1} \\
&= \hat{\omega} + \hat{\alpha}_1 \hat{\lambda}_{t+h-1} + \hat{\beta}_1 \hat{\lambda}_{t+h-1} \\
&= \hat{\omega} + (\hat{\alpha}_1 + \hat{\beta}_1) \hat{\lambda}_{t+h-1}.
\end{aligned}$$

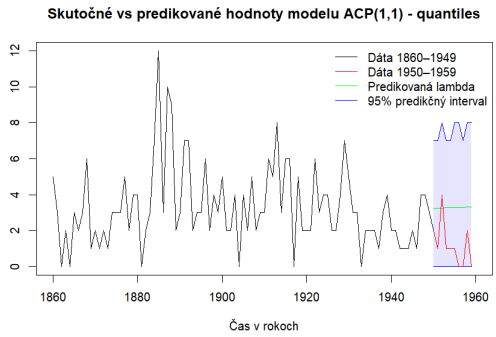
V kroku $h = 1$ máme k dispozícii hodnoty x_t a λ_t, ktoré dosadíme do predpisu a ostatné parametre odhadneme. V ďalších krokoch už nepoznáme žiadne hodnoty, preto ich nahradíme ich odhadmi. Hodnotu x_{t+h-1} odhadneme pomocou $\hat{\lambda}_{t+h-1}$.

Výpočet v softvéri R: Na výpočet 95%-ného predikčného intervalu pre ACP a PAR model sme využili funkciu `predict`. Táto funkcia obsahuje viac parametrov, pričom uvádzame len vybrané z nich spolu s ich predvolenými nastaveniami:

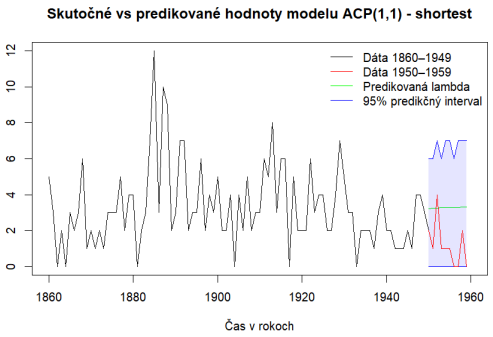
- `level = 0.95`
- `method = "bootstrap"`
- `B = "1000"`.

Funkcia vypočíta 95%-ný predikčný interval pomocou metódy bootstrap, kde počet opakovaných bootstrapových výberov je nastavený na 1000. Taktiež pomocou tejto funkcie získame 10 budúcich hodnôt $\hat{\lambda}_t$. Funkcia má aj parameter `type`, ktorý sme v prvom prípade nastavili na *quantiles* a v druhom prípade na *shortest*.

Varianta *quantiles* určuje intervaly ako dolný a horný kvantil rozdelenia, napr. 2.5% a 97.5% pre 95%-ný interval. Každý bod intervalu je vypočítaný nezávisle od ostatných možných hodnôt. Výsledný interval môže byť širší. Varianta *shortest* generuje predikčný interval s najmenšou možnou šírkou pri danej úrovni spoľahlivosti.



Obr. 9: Predikcie s predikčným intervalom - funkcia `predict`, typ *quantiles*



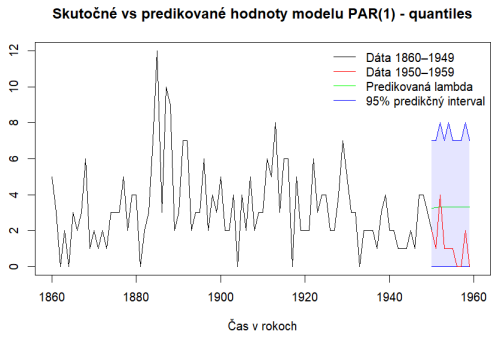
Obr. 10: Predikcie s predikčným intervalom - funkcia `predict`, typ *shortest*

Pri oboch nastaveniach parametra `level` = 95% predikčné intervaly z obrázkov 9 a 10 pokrývajú skutočné pozorované hodnoty. V rámci predikčných intervalov sme porovnali aj celkové priemerné šírky pri odlišných nastaveniach parametra `type`. Výpočty boli založené na simulácii 1000 realizácií hornej a dolnej hranice predikčného intervalu z funkcie `predict`. V každej iterácií sme vypočítali šírku predikčného intervalu ako rozdiel medzi hornou a dolnou hranicou. Vznikla nám matica rozmeru 1000x10. Napokon sme vypočítali celkovú priemernú šírku ako $\frac{1}{1000.10}\sum_{i=1}^{1000}\sum_{j=1}^{10}\check{s}\acute{i}rka_{i,j}$. V prípade varianty *quantiles* vyšla celková priemerná šírka predikčného intervalu približne 7.4, v prípade varianty *shortest* bola celková priemerná šírka menšia, a to približne 6.8.

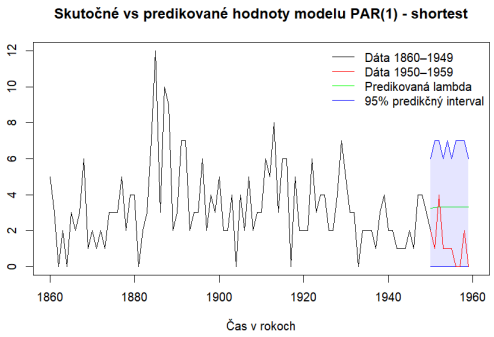
2.4.2 Model PAR

V prípade odvodenia predikcií pre PAR(1) model je postup analogický ako pri modeli ACP(1,1), s tým rozdielom, že do úvahy neberieme parameter λ_{t+h-1}.

Výpočet v softvéri R: V prípade modelu PAR sme postupovali presne ako v prípade modelu ACP(1,1) modelu. Parametre modelu sme odhadovali pre PAR(1) model.



Obr. 11: Predikcie s predikčným intervalom - funkcia `predict`, typ *quantiles*



Obr. 12: Predikcie s predikčným intervalom - funkcia `predict`, typ *shortest*

Z obrázkov 11 a 12 vidíme, že v prípade PAR(1) modelu sa skutočné dáta nachádzajú vo vnútri 95%-ného predikčného intervalu. Aj v prípade PAR modelu sme porovnali celkové priemerné šírky intervalov. Postupovali sme presne ako v prípade ACP modelu. Predikčný interval pre variantu *quantiles* vyšla približne 7.3, pri variante *shortest* vyšla celková priemerná šírka približne 6.74.

2.4.3 Model INAR

Bodová predikcia je odhad podmienenej strednej hodnoty procesu na základe informácií, ktoré máme k dispozícii. Vychádzajúc z [17], uvádzame nasledovné tvrdenie.

Tvrdenie 2.1

Za predpokladu $\mathrm{E}[\varepsilon_t] = \lambda$, pre INAR(1) predikciu na h krokov dopredu platí

$$\hat{X}_{t+h} \mid X_t = \mathrm{E}\left[X_{t+h} \mid X_t\right] = \alpha^h \left[X_t - \frac{\lambda}{1-\alpha}\right] + \frac{\lambda}{1-\alpha} \quad h = 1, 2, 3 \ldots \tag{2.55}$$

Odvodíme toto tvrdenie. Využitím vzorca pre podmienenú strednú hodnotu (2.29)

môžeme postupne rátať:

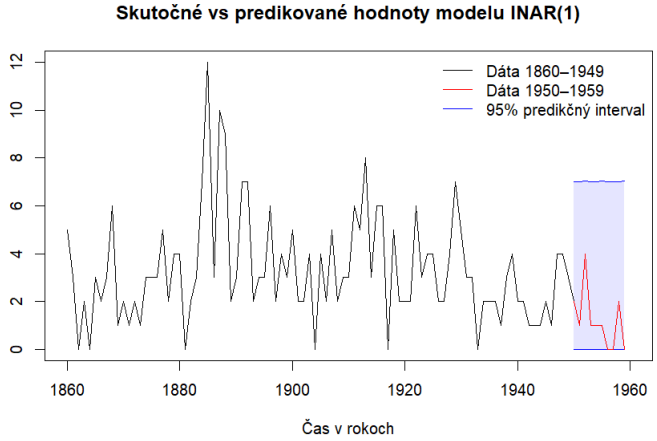
$$\begin{aligned}
h=1: \quad & \mathrm{E}[X_{t+1} \mid X_t] = \alpha X_t + \lambda \\
h=2: \quad & \mathrm{E}[X_{t+2} \mid X_t] = \mathrm{E}[\alpha \circ X_{t+1} + \varepsilon_{t+2} \mid X_t] \\
& \qquad = \alpha \mathrm{E}[X_{t+1} \mid X_t] + \lambda \\
& \qquad = \alpha(\alpha X_t + \lambda) + \lambda \\
& \qquad = \alpha^2 X_t + \alpha\lambda + \lambda \\
& \qquad \vdots \\
h: \quad & \mathrm{E}[X_{t+h} \mid X_t] = \alpha^h X_t + \alpha^{h-1}\lambda + \cdots + \lambda \\
& \qquad = \alpha^h X_t + \sum_{i=0}^{h-1} \alpha^i \lambda.
\end{aligned}$$

Využitím vzorca pre geometrický rad $\sum_{i=0}^{k} aq^i = a\dfrac{q^{k+1}-1}{q-1}$ a následným preskupením dostávame finálny tvar h krokovej predikcie pre INAR(1) model

$$\mathrm{E}[X_{t+h} \mid X_t] = \alpha^h X_t + \lambda\frac{\alpha^h - 1}{\alpha - 1} = \alpha^h \left[X_t - \frac{\lambda}{1-\alpha}\right] + \frac{\lambda}{1-\alpha}. \tag{2.56}$$

Tu ale nastáva problém, pretože (2.56) nemusí generovať celočíslené predikcie. Riešením je zobrať medián, ktorý minimalizuje $\mathrm{E}\left[\left|X_{t+h} - \hat{X}_{t+h}\right| \mid X_t\right]$.

Výpočet v softvéri R: Pri výpočte 95%-ného predikčného intervalu sme najprv odhadli parametre $\hat{\alpha}$ a $\hat{\lambda}$ pomocou funkcie `spinar_est_param`. Následne sme uložili hodnotu posledného pozorovania zo skrátenej 90 člennej vzorky. Pomocou odhadnutých parametrov modelu sme nasimulovali 1000 možných realizácií desiatich budúcich hodnôt pomocou predpisu INAR(1) modelu (2.26), pričom sme pri simulácii realizácií skutočných možných hodnôt využili funkcie `rbinom` a `rpois`. Pomocou `rbinom` sme vygenerovali prvú časť modelu $\alpha \circ X$ z binomického rozdelenia a pomocou `rpois` sme vygenerovali druhú časť ε z Poissonovho rozdelenia. Možné realizované hodnoty INAR modelu sme generovali rekurentne, kde sme vždy brali do úvahy aktuálne posledné pozorovanie. Nakoniec sme z dát vypočítali 95%-ný predikčný interval.



Obr. 13: Predikcie s predikčným intervalom pre INAR(1) model

Na základe obrázka 13 vidíme, že aj v tomto prípade sa skutočné hodnoty zmestili do predikčného intervalu.

Na záver sme porovnali všetky modely. Funkcia `predict` okrem predikčného intervalu poskytuje aj odhadnuté stredné hodnoty procesu. Pomocou týchto odhadov sme vypočítali MAE (Mean Absolute Error), ktoré je definované nasledovne

$$\text{MAE} = \frac{1}{p}\sum_{i=1}^{p}|y_i - \hat{y}_i|, \tag{2.57}$$

kde p je počet predikcií, y_i je skutočná pozorovaná hodnota a $\hat{y}_i$ je predikovaná hodnota. V prípade modelov ACP a PAR sme desať predikovaných budúcich hodnôt získali práve z funkcie `predict`.

V prípade INAR modelu knižnica `spINAR` neposkytuje žiadnu funkciu, ktorá by vedela rátať predikcie. Na výpočet predikovaných hodnôt sme využili vzorec pre podmienenú strednú hodnotu (2.29). V prvej iterácii sme využili poslednú hodnotu z 90-členného datasetu, vo zvyšných iteráciách sme postupovali už rekurzívne. Pre všetky modely sme vypočítali hodnoty MAE, ktoré uvádzame v tabuľke nižšie.

Tabuľka 4: MAE pre jednotlivé modely

Model	**MAE**
ACP(1,1)	2.2239
PAR(1)	2.2439
INAR(1)	2.2248

Z tabuľky 4 vidíme, že najnižšiu hodnotu MAE má práve model ACP(1,1). Sledovaný proces môže byť reprezentovateľný modelom ACP(1,1).

2.5 Pearsonove rezíduá

Táto podkapitola je venovaná diagnostike modelov, ktorá je dôležitá pri overovaní vhodnosti ich použitia. Predstavíme Pearsonove rezíduá a objasníme ich využitie v praxi. V tejto časti budeme čerpať najmä z [14] a [21].

Predpokladajme, že $X_1, \ldots, X_T$ je časový rad, ktorý je generovaný procesom, ktorého parameter nepoznáme. Tento model závisí od vektora parametrov $\boldsymbol{\theta}$. Označme

$$\mathrm{E}[X_t \mid X_{t-1}, X_{t-2}, \ldots; \boldsymbol{\theta}], \tag{2.58}$$

ako podmienenú strednú hodnotu a

$$\mathrm{Var}[X_t \mid X_{t-1}, X_{t-2}, \ldots; \boldsymbol{\theta}], \tag{2.59}$$

ako podmienený rozptyl, pričom obe veličiny sú funkciami parametrov $\boldsymbol{\theta}$.

Definícia 2.5.

Pearsonove rezíduá sú definované ako štandardizované veličiny

$$R_t(\boldsymbol{\theta}) = \frac{X_t - \mathrm{E}[X_t \mid X_{t-1}, X_{t-2}, \ldots; \boldsymbol{\theta}]}{\sqrt{\mathrm{Var}[X_t \mid X_{t-1}, X_{t-2}, \ldots; \boldsymbol{\theta}]}}, \tag{2.60}$$

pre $t \in \{1, \ldots, T\}$

Pearsonove rezíduá vyjadrujú, o koľko sa každé pozorovanie líši od svojej očakávanej hodnoty, pričom táto odchýlka je škálovaná príslušnou podmienenou smerodajnou

odchýlkou. V praxi nepoznáme skutočnú hodnotu parametrov $\boldsymbol{\theta}$, preto do vzorca (2.60) dosadíme odhad $\hat{\boldsymbol{\theta}}$.

Ak je fitovaný model adekvátny, očakávame, že vypočítané Pearsonove rezíduá budú mať priemer vzorky okolo 0 a rozptyl okolo 1. Autokorelačná funkcia ACF alebo PACF by mali mať hodnoty blízke 0.

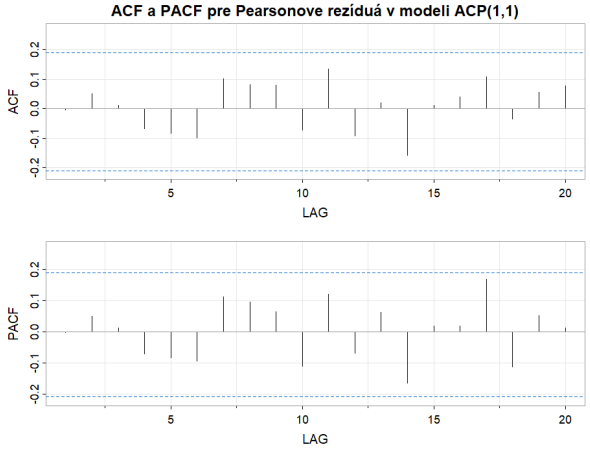
Výpočet v softvéri R: V prípade modelov ACP a PAR vieme získať Pearsonove rezíduá priamo z funkcie `residuals` nastavením parametra `type` na hodnotu *pearson.* Vstupom do tejto funkcie je výstup funkcie `tsglm`, pomocou ktorej sme odhadli parametre modelov.

V prípade INAR modelu opäť nie je k dispozícii nijaká funkcia, ktorá by nám umožnila získať hodnoty Pearsonových rezíduí priamo. Preto sme tieto hodnoty vypočítali pomocou vzorca pre Pearsonove rezíduá (2.60). Pre výpočet podmienenej strednej hodnoty sme využili vzorec (2.29) a pre výpočet podmieneného rozptylu sme využili vzorec (2.31).

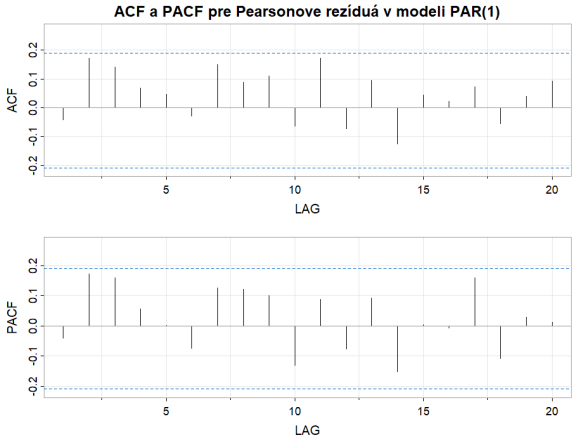
Tabuľka 5: Stredná hodnota a variancia Pearsonových rezíduí pre jednotlivé modely

Model	**Stredná hodnota**	**Variancia**
ACP(1,1)	-0.0000531	1.321
PAR(1)	-0.0011959	1.425
INAR(1)	-0.0063185	1.532

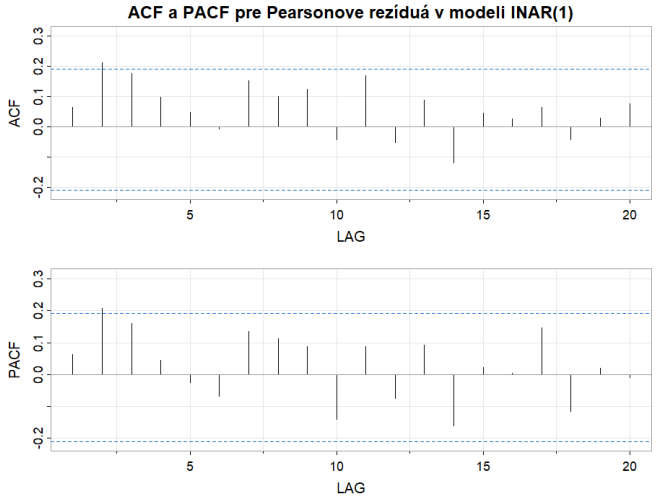
Vykreslili sme aj ACF a PACF funkcie pre Pearsonove rezíduá jednotlivých modelov.



Obr. 14: ACF a PACF Pearsonových rezíduí v modeli ACP(1,1)



Obr. 15: ACF a PACF Pearsonových rezíduí v modeli PAR(1)



Obr. 16: ACF a PACF Pearsonových rezíduí v modeli INAR(1)

Na základe strednej hodnoty a variancie Pearsonových rezíduí z tabuľky 5 a obrázkov 14, 15 a 16 ACF a PACF funkcií, má práve model ACP(1,1) najlepšie výsledky. INAR(1) nie je vhodným modelom pre tieto dáta. Už ACF a PACF naznačujú, že rezíduá sú korelované. Sledovaný proces môže byť reprezentovateľný modelom ACP(1,1).

2.6 Predikcie pre sezónne dáta

Teoretické poznatky k tejto časti budeme čerpať z [7], [8] a [14]. V tejto podkapitole sa budeme venovať práci so sezónnymi dátami v prípade časových radov počtov. Skôr než prejdeme k aplikácii na reálnych dátach v softvéri R, stručne priblížime teoretické pozadie.

Mnohé modely sa dajú zovšeobecniť pomocou logaritmickej linkovej funkcie. Tá je vhodná napr. pri práci s časovými radmi počtov, pretože zabezpečuje kladnosť λ_t. Príklad použitia logaritmickej linkovej funkcie ukážeme pre ACP(1,1) model.

Vychádzajúc z [8] a [14], model ACP(1,1) má nasledovný tvar

$$log(\lambda_t) = \omega + \alpha_1 log(X_{t-1} + 1) + \beta_1 log(\lambda_{t-1}). \tag{2.61}$$

Ak by v predpise (2.61) bolo iba $log(X_{t-1})$, tak by stabilita ACP(1,1) modelu nemusela byť zaručená v prípade $X_{t-1} = 0$.

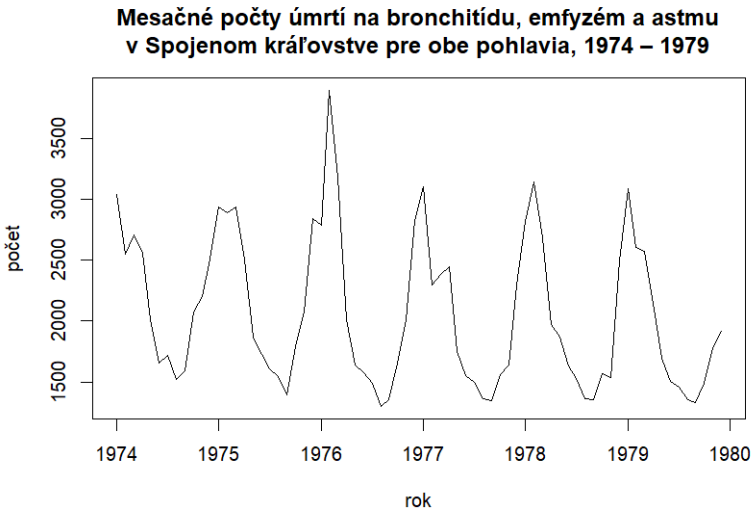
Každý model sa dá rozšíriť o tzv. kovariáty. **Kovariáty** sú vysvetľujúce, premenné obsahujúce dodatočné informácie, ktoré pomáhajú lepšie modelovať sledovaný proces. Nech $\mathbf{X}_t$ označuje časovo premenlivý r-rozmerný vektor kovariátov a $\gamma = (\gamma_1, ..., \gamma_r)^T$ označuje vektor parametrov zodpovedajúcich efektu kovariátov.

Ak do ACP(1,1) modelu s logaritmickou linkovou funkciou pridáme kovariáty, model nadobudne tvar

$$log(\lambda_t) = \omega + \alpha_1 log(X_{t-1} + 1) + \beta_1 log(\lambda_{t-1}) + \gamma^T \mathbf{X}_t, \tag{2.62}$$

pričom $t \in \mathbb{N}$, $\gamma \in \mathbb{R}$ a $\gamma^T \mathbf{X}_t > 0$.

V tejto časti sme pracovali s dátami, ktoré popisujú mesačné počty úmrtí na bronchitídu, emfyzém a astmu v Spojenom kráľovstve pre obe pohlavia pre roky 1974 až 1979. Uvedené ochorenia postihujú dýchacie cesty.



Obr. 17: Graf mesačných počtov úmrtí na ochorenia dýchacích ciest v UK pre obe pohlavia od 1974 až po rok 1979

Dáta obsahujú iba počty úmrtí v jednotlivých mesiacoch. Z obrázka 17 je vidieť sezónne správanie. Počty úmrtí začínajú rásť približne od októbra a k poklesu úmrtí prichádza približne v apríli. Na modelovanie sezónnosti sa v praxi často využívajú funkcie sínus a kosínus. Tieto funkcie vedia dobre zachytiť sezónnosť a periodickosť v dátach. K našim dátam sme tieto hodnoty dopočítali pomocou vzorcov

$$\cos\left(\frac{2\pi m}{12}\right), \sin\left(\frac{2\pi m}{12}\right), \tag{2.63}$$

kde m označuje mesiac, $m = 1, \ldots, 12$ a zároveň predpokladáme periódu jeden rok.

Následne sme opäť dáta spolu s hodnotami funkcií sínus a kosínus rozdelili na trénovaciu a testovaciu časť. Celkovo sme mali k dispozícii 72 pozorovaní, 66 pozorovaní sa použilo na odhadnutie parametrov modelu a šesť pozorovaní sa využilo na porovnanie presnosti predikcií.

Naším cieľom nebolo nájsť ten najvhodnejší model, zamerali sme sa len na 4 vybrané. Rozhodli sme sa pre modely PAR(1), ACP(1,1), ACP(1,12;1,12) a ACP(6,12;6,12). Vo všetkých modeloch boli zahrnuté kovariáty sínus a kosínus. Všetky štyri modely mali nekorelované rezíduá. Modely PAR(1) a ACP(1,1) sme vybrali ako najjednoduchšie a základné modely. V prípade modelu ACP(1,12;1,12) sa pozeráme na minulé hodnoty procesu v čase $t-1$ a $t-12$ a minulé stredné hodnoty tiež v časoch $t-1$ a $t-12$. V prípade ACP(6,12;6,12) sa pozeráme na hodnoty polroka a rok do minulosti v prípade minulých

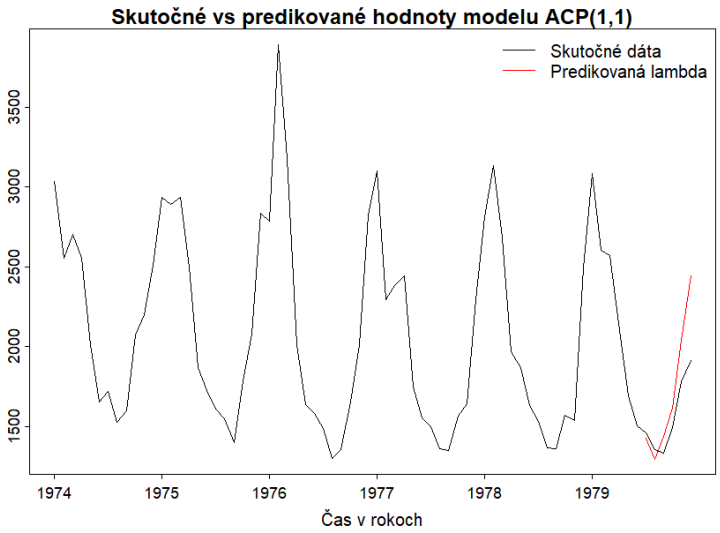
pozorovaných hodnôt a minulých stredných hodnôt.

Jednotlivé koeficienty modelov sme odhadli pomocou funkcie `tsglm`. Parameter `link` sme nastavili na hodnotu `log`, v parametri `dist` sme nastavili Poissonovo rozdelenie a v `xreg` sme dosadili 66 hodnôt prislúchajúcich funkciám sínus a kosínus z (2.63).

Tabuľka 6: Hodnoty BIC kritéria jednotlivých modelov

Model	BIC
PAR(1)	2222.604
ACP(1,1)	2100.803
ACP(1,12;1,12)	2086.104
ACP(6,12;6,12)	1983.853

Na základe najnižšieho BIC kritéria z tabuľky 6 sme pre výpočet budúcich šiestich hodnôt vybrali model ACP(6,12;6,12). Tie sme opäť rátali pomocou funckie `predict`, kde sme v parametri `xreg` dosadili šesť nových hodnôt prislúchajúcim k funkciám sínus a kosínus.



Obr. 18: Graf mesačných počtov úmrtí na ochorenia dýchacích ciest v UK pre obe pohlavia od 1974 až po rok 1979 spolu s predikovanými $\hat{\lambda}_t$, $t = 67, .., 72$.

Z obrázka 18 vidíme, že modelovanie s využitím funkcií sínus a kosínus je užitočné. Predikované $\hat{\lambda}_t$, $t = 67, .., 72$ celkom dobre zachytili sezónnosť v dátach.

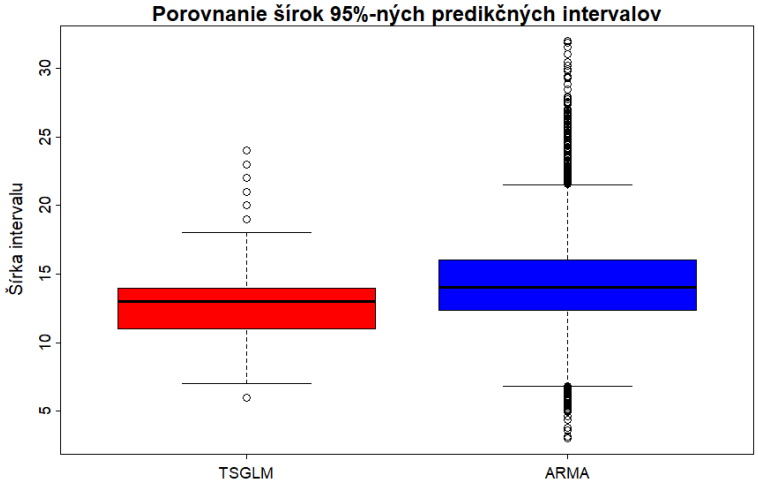
2.7 Predikcie pomocou ARMA modelu

V tejto časti bolo naším cieľom porovnať predikcie získané z ACP(1,1) modelu a predikcie z ARMA modelu pomocou logaritmickej transformácie a zistiť, či by bolo postačujúce modelovať časové rady počtov pomocou vhodnej transformáciue v rámci ARMA modelov.

Vygenerovali sme 110 dát z ACP(1,1) modelu podľa jeho predpisu s parametrami $\omega = 1.5$, $\alpha = 0.5$ a $\beta = 0.3$. Tento proces sme simulovali presne ako v podkapitole 2.3.1. Z prvých 100 dát sme odhadli parametre ACP(1,1) modelu pomocou funkcie `tsglm`, následne sme predikovali dáta na ďalších desať období dopredu pomocou funkcie `predict` a nakoniec sme vypočítali 95%-ný predikčný interval. Pre tých istých 100 dát sme urobili predikciu pomocou ARMA modelu. Tieto dáta sme predtým zlogaritmovali pomocou $log(x+1)$, pretože ACP model môže nadobúdať aj hodnoty rovné nule. Pomocou funkcie `auto.arima` sme našli model bez diferencií a pomocou funkcie `sarima.for` sme predikovali dáta na ďalších desať období dopredu. Následne sme vypočítali 95%-ný predikčný interval, ktorý sme potom odlogaritmovali. Celý tento proces sme simulovali 1000-krát. Na záver sme vyhodnotili ich priemernú šírku a presnosť.

V priemere vychádzala šírka 95%-ného predikčného intervalu pomocou ACP modelu 12.67 a šírka 95%-ného predikčného intervalu pomocou ARMA modelu 14.32. Presnosť ACP modelu bola približne 95%, zatiaľ čo v prípade ARMA modelu to bolo len približne 93%.

Následne sme vykreslili boxploty pre šírky týchto tisíc 95%-ných predikčných intervalov. Z obrázka 19 je vidieť, že TSGLM intervaly sú užšie ako ARMA, čo vidieť aj na základe mediánov. ARMA model má viac extrémne širokých alebo extrémne úzkych intervalov.



Obr. 19: Porovnanie šírok predikčných intervalov

3 Časové rady proporcií

V tejto kapitole sa budeme zaoberať časovými radmi, ktoré nadobúdajú hodnoty z intervalu $(0, 1)$. Definujeme si predpis modelu βARMA a pozrieme sa na distibúciu odhadov parametrov. Následne rozoberieme predikcie z teoretického aj praktického hľadiska, pričom výpočet budeme demonštrovať na reálnych dátach. Aj v tomto prípade porovnáme výkonnosť βARMA a ARMA modelov. V tejto kapitole sme vychádzali z [15] a [16].

3.1 Beta autoregressive moving average model - βARMA model

V tejto časti budeme čerpať najmä z [15]. Najprv priblížime základné charakteristiky beta rozdelenia.

Definícia 3.1.
Náhodná premenná X má beta rozdelenie s parametrami p, q, t.j. $X \sim \text{Beta}(p, q)$, ak má hustotu

$$f(x; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}, \tag{3.64}$$

kde $0 < x < 1$, $p > 0$, $q > 0$ a $\Gamma(.)$ je gama funkcia.

Stredná hodnota a variancia sú definované ako

$$\begin{aligned} E(X) &= \frac{p}{p+q}, \\ Var(X) &= \frac{pq}{(p+q)^2(p+q+1)}. \end{aligned} \tag{3.65}$$

Často sa využíva reparametrizácia beta rozdelenia s parametrami $\mu = \frac{p}{p+q}$ a $\phi = p+q$. Vyjadrením p a q z týchto vzťahov dostávame $p = \mu\phi$ a $q = (1-\mu)\phi$, kde $0 < \mu < 1$ a $\phi > 0$. Vzťahy (3.65) v tomto prípade prejdú do tvaru

$$\begin{aligned} E(X) &= \mu, \\ Var(X) &= \frac{\mu(1-\mu)}{1+\phi}. \end{aligned} \tag{3.66}$$

Predpokladáme, že modelovaná premenná je spojitá a nadobúda hodnoty v štandardnom intervale $(0, 1)$. Navrhovaný model je užitočný aj v situáciách, keď je modelovaná premenná obmedzená na interval (a, b), kde a a b sú známe skaláre, pričom $a < b$. V takom prípade je možné modelovať $\frac{x-a}{b-a}$ namiesto priameho modelovania x.

Ďalej predpokladajme, že kovariáty $z_t, t = 1, \ldots, n$, kde $z_t = (z_{t1}, \ldots, z_{tk})^T$ sú nenáhodné a $\beta = (\beta_1, \ldots, \beta_k)^T$ je množina neznámych lineárnych parametrov. Tu n označuje veľkosť vzorky a $k < n$.

Nech $X_t, t = 1, \ldots, n$ sú náhodné premenné a predpokladajme, že podmienené rozdelenie každej X_t, vzhľadom na predchádzajúcu informačnú množinu $\mathcal{F}_{t-1}$, t.j. najmenšiu σ-algebru, pre ktorú sú náhodné premenné $X_1, \ldots, X_{t-1}$ merateľné, má beta rozdelenie. Pre X_t je hustota vyjadrená na základe vzorca (3.64) ako

$$f(x_t \mid \mathcal{F}_{t-1}) = \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} x_t^{\mu_t\phi-1}(1-x_t)^{(1-\mu_t)\phi-1}, \tag{3.67}$$

kde $0 < x_t < 1$. Pre strednú hodnotu a varianciu platí

$$\begin{aligned} E(X_t \mid \mathcal{F}_{t-1}) &= \mu_t, \\ Var(X_t \mid \mathcal{F}_{t-1}) &= \frac{\mu_t(1-\mu_t)}{1+\phi}. \end{aligned} \tag{3.68}$$

Na základe predchádzajúcich vyjadrení vyplýva, že X_t vzhľadom na informačnú množinu $\mathcal{F}_{t-1}$ má beta rozdelenie s parametrami $\phi\mu_t$ a $(1-\mu_t)\phi$, t.j.

$$X_t \mid \mathcal{F}_{t-1} \sim \text{Beta}(\phi\mu_t, (1-\mu_t)\phi).$$

Definícia 3.2.

Model βARMA(p, q) je definovaný nasledovnými vzťahmi

$$X_t \mid \mathcal{F}_{t-1} \sim \text{Beta}(\phi\mu_t, (1-\mu_t)\phi), \tag{3.69}$$

$$g(\mu_t) = \alpha + z_t'\beta + \sum_{i=1}^{p} \varphi_i \left\{ g(X_{t-i}) - z_{t-i}'\beta \right\} + \sum_{j=1}^{q} \theta_j r_{t-j}, \tag{3.70}$$

kde stredná hodnota μ_t je transformovaná pomocou striktne monotónnej a dvakrát diferencovateľnej linkovej funkcie $g : (0, 1) \to \mathbb{R}$. Medzi najčastejšie používané linkové funkcie patria logitová, probitová a komplementárna log-log funkcia.

Parametre $p, q \in \mathbb{N}$ sú, v uvedenom poradí, rády autoregresnej a kĺzavopriemernej (moving average) zložky, φ_i $i = 1, ...p$ sú autoregresné parametere, θ_i $i = 1, ...q$ sú parametere kĺzavého priemeru a r_t je chybový člen.

Existuje niekoľko možností pre členy chyby kĺzavého priemeru r_t: napríklad chyby merané na pôvodnej mierke, t.j. $X_t - \mu_t$, na linkovej mierke, t.j. $g(X_t) - g(\mu_t)$, atď. Vyžaduje sa, aby chyba r_t bola merateľná vzhľadom na $\mathcal{F}_t$.

3.2 Odhad parametrov

Aj v prípade βARMA modelu priblížime odhad parametrov z teoretickej aj praktickej stránky. Definujeme vierohodnostnú funkciu pre tento model a pozrieme sa na distribúciu jednotlivých odhadovaných parametrov.

Odhady parametrov sa počítajú pomocou metódy maximálnej vierohodnosti. V prípade βARMA modelu má vierohodnostná funkcia tvar

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{T} \frac{\Gamma(\phi)}{\Gamma(\mu_t\phi)\Gamma((1-\mu_t)\phi)} x_t^{\mu_t\phi-1}(1-x_t)^{(1-\mu_t)\phi-1}, \tag{3.71}$$

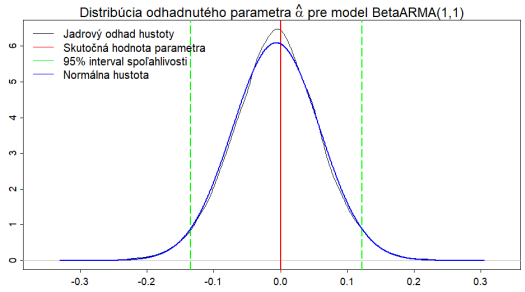
kde $\boldsymbol{\theta}$ je vektor parametrov μ_t a ϕ.

Parametre sa typicky odhadujú numericky riešením nelineárneho optimalizačného algoritmu. Podrobnosti je možné nájsť v [15].

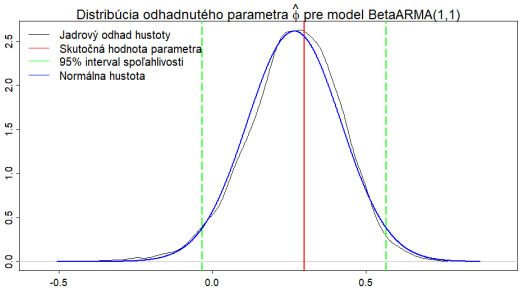
Výpočet v softvéri R: V tejto časti sme simulovali βARMA(1,1) model podľa predpisu (3.70). Simulovali sme 10000-krát sto hodnôt s nasledovnými parametrami: $p = q = 1$, $\alpha = 0$, $\varphi = 0.3$, $\theta = 0.4$ a parameter `precision` v beta rozdelení sme nastavili $\nu = 20$. Pri generovaní sme využili logitovú funkciu `logit`, ktorá transformuje dáta podľa vzťahu $log(\frac{x}{1-x})$ do $\mathbb{R}$ a inverznú transformáciu $\frac{1}{1+exp(-x)}$, ktorá mapuje hodnoty späť do intervalu $(0, 1)$. V prípade generovania kĺzavého priemeru r_t sme využili linkovú mierku $g(X_t) - g(\mu_t)$. Pomocou funkcie `btsr.fit` z balíka `BTSR` ([16]) vieme odhadnúť parametre βARMA procesu. Kedže funkcia `btsr.fit` pracuje len s hodnotami z $(0, 1)$, hodnoty nula a jeden sme ošetrili o hodnotu $eps = 10^{-6}$ tak, že k nule sme hodnotu eps pripočítali a od jednotky sme hodnotu eps odpočítali. Následne sme sa pozreli na distribúciu jednotlivých parametrov a pre každý parameter sme vypočítali jeho priemer a disperziu.

Tabuľka 7: Hodnoty priemeru a disperzie pre jednotlivé parametre βARMA(1,1) modelu

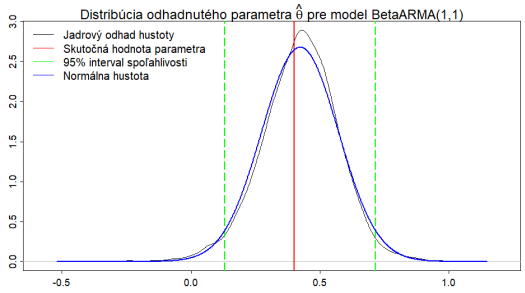
parameter	$\hat{\alpha}$	$\hat{\varphi}$	$\hat{\theta}$	$\hat{\nu}$
pôvodná hodnota	0	0.3	0.4	20
priemer	-0.00652	0.26633	0.42289	21.12282
disperzia	0.00429	0.02333	0.02215	8.93465



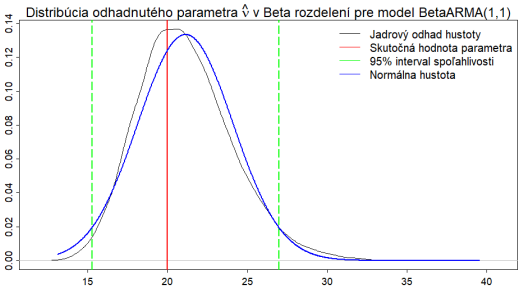
Obr. 20: Distribúcia parametra $\hat{\alpha}$ v modeli βARMA(1,1)



Obr. 21: Distribúcia parametra $\hat{\varphi}$ v modeli βARMA(1,1)



Obr. 22: Distribúcia parametra $\hat{\theta}$ v modeli βARMA(1,1)



Obr. 23: Distribúcia parametra $\hat{\nu}$ v Beta rozdelení pre model βARMA(1,1)

Na základe tabuľky 7 je vidieť, že disperzia parametra $\hat{\nu}$ vyšla vyššia než v prípade ostatných parametrov, čo môže súvisieť s tým, že tento parameter je vyjadrený v desiatkach, zatiaľ čo ostatné nadobúdajú hodnoty iba v intervale (0,1). Z obrázkov 20 - 22 je vidieť, že odhadnuté parametre sú centrované okolo nami zadaných hodnôt. Z obrázka 23 je vidieť vychýlenie odhadnutého parametra $\hat{\nu}$.

3.3 Predikcie

V tejto podkapitole ukážeme teoretický výpočet predikcií pre βARMA(1,1) model pre chyby merané na linkovej mierke, t.j. $r_{t-1} = g(X_{t-1}) - g(\mu_{t-1})$. Po dosadení r_{t-1} do predpisu modelu (3.70) dostávame upravený predpis modelu βARMA(1,1)

$$g(\mu_t) = \alpha + \varphi_1 g(x_{t-1}) + \theta_1[\, g(x_{t-1}) - g(\mu_{t-1})] \tag{3.72}$$

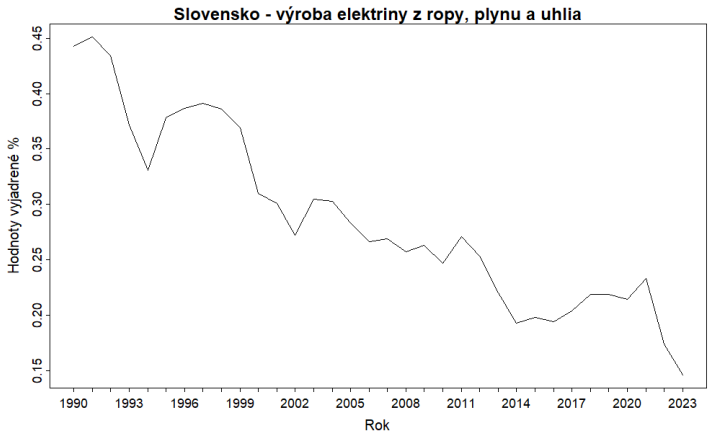
$$= \alpha + (\varphi_1 + \theta_1) g(x_{t-1}) - \theta_1 g(\mu_{t-1}). \tag{3.73}$$

Počítajme

$$
\begin{aligned}
h = 1: \quad & g(\mu_{t+1}) = \hat{\alpha} + (\hat{\varphi}_1 + \hat{\theta}_1) g(x_t) - \hat{\theta}_1 g(\mu_t) \\
h = 2: \quad & g(\mu_{t+2}) = \hat{\alpha} + (\hat{\varphi}_1 + \hat{\theta}_1) g(\hat{x}_{t+1}) - \hat{\theta}_1 g(\hat{\mu}_{t+1}) \\
& \qquad = \hat{\alpha} + (\hat{\varphi}_1 + \hat{\theta}_1) g(\hat{\mu}_{t+1}) - \hat{\theta}_1 g(\hat{\mu}_{t+1}) \\
& \qquad \vdots \\
h = h: \quad & g(\mu_{t+h}) = \hat{\alpha} + (\hat{\varphi}_1 + \hat{\theta}_1) g(\hat{x}_{t+h-1}) - \hat{\theta}_1 g(\hat{\mu}_{t+h-1}) \\
& \qquad = \hat{\alpha} + (\hat{\varphi}_1 + \hat{\theta}_1) g(\hat{\mu}_{t+h-1}) - \hat{\theta}_1 g(\hat{\mu}_{t+h-1}).
\end{aligned}
$$

V kroku $h = 1$ poznáme hodnoty x_t a μ_t, ktoré môžeme dosadiť do predikcie. V prípade ostatných parametrov dosádzame ich odhad. V prípade kroku väčšieho ako 1 môžeme odhadnúť budúce hodnoty $\hat{x}_{t+h-1}$ pomocou $\hat{\mu}_{t+h-1}$.

Výpočet v softvéri R: V tejto časti sme pracovali s dátami popisujúce výrobu elektriny z ropy, plynu a uhlia na Slovensku od roku 1990 až po rok 2023 z celkovej vyrobenej elektriny vyjadrenú v percentách. Priebeh výroby elektriny sme znázornili na obrázku 24.



Obr. 24: Výroba elektriny z ropy, plynu a uhlia na Slovensku v rokoch 1990-2023

Tieto dáta sme si rozdelili na dve časti: trénovacia časť obsahujúca 31 hodnôt a testovacia časť pozostávajúca z troch hodnôt. Kedže tieto dáta vykazujú klesajúci trend, pri modelovaní predikcií sme do modelu zahrnuli aj kovariáty v podobe času, kde tento vektor času pozostával z hodnôt od 1 až po 34. Kedže pracujeme s hodotami v intervale $(0, 1)$,

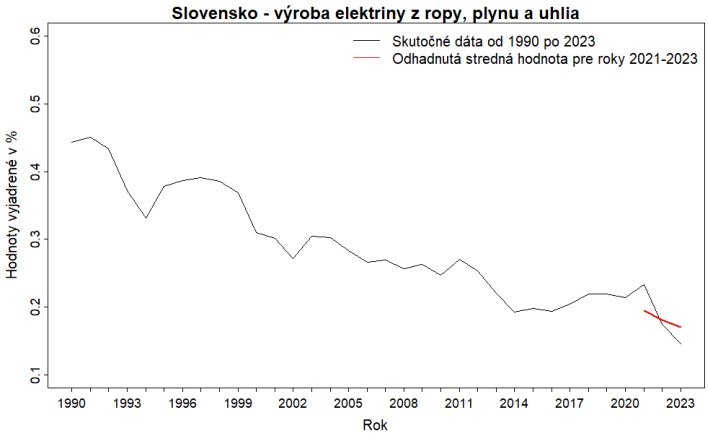
tak aj čas sme preškálovali do tohto intervalu, a taktiež sme ho rozdelili na dve časti, ktoré prislúchali k trénovacej a testovacej časti. Časový kovariát s dĺžkou 31 sme použili pri odhade modelu vo funkcii `btsr.fit` v parametri `xreg` a časový kovariát s dĺžkou tri sme použili pri výpočte predikcií vo funkcii `predict` v parametri `newdata`.

Tabuľka 8: Hodnoty BIC jednotlivých βARMA modelov

Model	**Rezíduá**	**BIC**
βARMA(1,0)	nekorelované	-141.08
βARMA(2,0)	nekorelované	-142.88
βARMA(0,1)	korelované	-
βARMA(0,2)	nekorelované	-143.23
βARMA(1,1)	nekorelované	-139.8
βARMA(2,1)	nekorelované	-139.28
βARMA(1,2)	nekorelované	-138.35
βARMA(2,2)	nekorelované	-140.17

Do úvahy sme brali osem modelov, ktoré sú popísané v tabuľke 8. Po odhade modelov sme sa pozreli na korelovanosť rezíduí, vykreslili sme si ich ACF a PACF grafy. Všetky modely mali rezíduá nekorelované, len v prípade modelu βARMA(0,1) rezíduá vykazovali korelovanosť. Následne sme vybrali model βARMA(0,2), pretože mal najnižšiu hodnotu BIC kritéria.

V ďalšom kroku sme odhadli tri budúce stredné hodnoty modelu βARMA(0,2) pre roky 2021 až 2023.



Obr. 25: Graf modelom odhadnutých hodnôt a odhadnutých budúcich stredných hodnôt

Z obrázka 25 vidíme, že predikcie zachytili pozvoľnejšie klesanie predikovaných stredných hodnôt. Na základe našich výsledkov vidíme, že trend v dátach sa dá modelovať pomocou časového kovariátu aj v prípade βARMA modelu.

3.4 Predikcie pomocou ARMA modelu

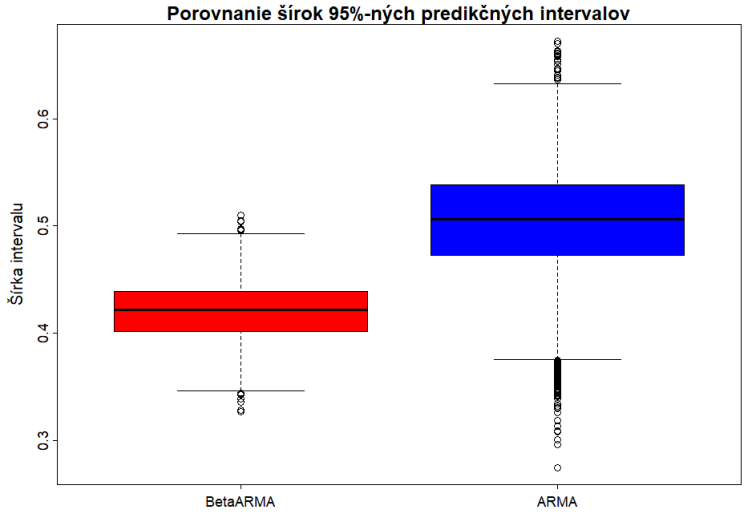
Podobne ako v podkapitole 2.7 pre nezáporné celočíslené časové rady, sme aj v prípade βARMA modelu porovnali predikcie vypočítané pomocou βARMA a ARMA modelu. Postup bol veľmi podobný ako v podkapitole 2.7.

Vygenerovali sme 110 dát z βARMA(1,1) modelu podľa jeho predpisu s parametrami $\alpha = 0$, $\varphi = 0.3$, $\theta = 0.4$ a $\nu = 20$. Proces βARMA(1,1) sme simulovali presne ako v podkapitole 3.2. Z prvých 100 dát sme odhadli parametre βARMA(1,1) modelu pomocou funkcie `btsr.fit`, následne sme predikovali strednú hodnotu na ďalších desať období dopredu pomocou funkcie `predict` a nakoniec sme vypočítali 95%-ný predikčný interval. Pre tých istých 100 dát sme urobili predikciu pomocou ARMA modelu - tieto dáta sme predtým zlogaritmovali pomocou $log(\frac{x}{1-x})$. Pomocou funkcie `auto.arima` sme našli model bez diferencií a pomocou funkcie `sarima.for` sme predikovali dáta na ďalších desať období dopredu, následne sme vypočítali 95%-ný predikčný interval, ktorý sme potom odlogaritmovali. Celý tento proces sme simulovali 1000-krát. Na záver sme vyhodnotili ich priemernú šírku a presnosť.

V priemere vychádzala šírka 95%-ného predikčného intervalu pomocou βARMA(1,1)

modelu 0.421 a šírka 95%-ného predikčného intervalu pomocou ARMA modelu 0.503. Presnosť βARMA(1,1) modelu bola približne 88%, zatiaľ čo v prípade ARMA modelu to bolo až 94%.

Následne sme si vykreslili boxploty pre šírky týchto tisíc 95%-ných predikčných intervalov. Z obrázka 26 je vidieť, že βARMA intervaly sú užšie ako ARMA, čo vidieť aj na základe mediánov, a to môže byť dôvodom, prečo vychádza presnosť βARMA modelu nižšia.



Obr. 26: Porovnanie šírok predikčných intervalov

4 Binárne časové rady

Táto kapitola bude venovaná časovým radom jednotiek a núl. Teoretické poznatky budemem čerpať z článku [10] a knihy [11]. Cieľom tejto kapitoly je definovať model pre binárne časové rady, pozrieť sa na distibúciu odhadnutých parametrov modelu a analyzovať predikcie z teoretického aj praktického hľadiska.

4.1 Binárny autoregresný model - BAR model

Pred definovaním samotného modelu pripomenieme Bernoulliho rozdelenie, inak nazývané aj alternatívne rozdelenie. Pomocou tohto rozdelenia sa modeluje jeden pokus, ktorého realizácia môže nadobudnúť dva možné výsledky - úspech s pravdepodobnosťou $P(X = 1) = p$ alebo neúspech s opačnou pravdepodobnosťou $P(X = 0) = 1 - p$, pričom $p \in (0, 1)$. Pre strednú hodnotu a rozptyl tohto rozdelenia platí

$$E(X) = p \tag{4.74}$$

$$D(X) = p(1 - p). \tag{4.75}$$

Definícia 4.1.
Nech $X_t \in \{0, 1\}$. Model BAR(p) je definovaný vzťahmi

$$X_t \mid \mathcal{F}_{t-1} \sim \text{Bernoulli}(\pi_t), \tag{4.76}$$

$$g(\pi_t) = \beta_0 + \beta_1 X_{t-1} + \cdots + \beta_p X_{t-p}, \tag{4.77}$$

kde $g(.)$ je linková funkcia, napr. logit alebo probit, β_0 je konštanta a β_i, $i = 1, ..., p$ sú autoregresné parametre. Pod pojmom π_t sa rozumie pravdepodobnosť úspechu, teda $\pi_t = P(X_t = 1 \mid \mathcal{F}_{t-1})$.

BAR model sa môže rozšíriť pridaním minulých hodnot π_t alebo vysvetľujúcich premenných, tzv. kovariátov.

4.2 Odhad parametrov

Parametre BAR(p) modelu sa odhadujú pomocou metódy maximálnej vierohodnosti. Nech $\boldsymbol{\theta} = (\beta_0, \beta_1, ..., \beta_p)^T$ je vektor parametrov. Potom pre BAR(p) model je vierohodnostná funkcia definovaná ako

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n} \pi_t^{x_t}(1-\pi_t)^{1-x_t}. \tag{4.78}$$

Výpočet v softvéri R: Simulovali sme model BAR(2) bez konštanty β_0. Simulovali sme 202 pozorovaní tohto procesu 10 000-krát. Konštantu β_0 sme zanedbali, nastavili sme ju rovnú 0, parametre β_1 sme nastavili na 0.6 a β_2 sme nastavili na 0.3. Najprv sme modelovali logit pravdepodobnosti π_t podľa predpisu (4.77), následne sme pomocou inverznej funckie $\frac{1}{1+exp(-x)}$ získali pravdepodobnosť π_t. Nakoniec sme generovali X_t z Bernoulliho rozdelenia podľa (4.76). Využili sme na to funkciu `rbinom`.

Keďže v softwéri R neexistuje balík špeciálne vytvorený pre binárne časové rady, tak je možné tieto časové rady modelovať pomocou zovšeobecnených lineárnych modelov, kde sa minulé hodnoty binárneho časového radu modelujú pomocou oneskorenia pozorovaní.

Pre odhad parametrov tohto modelu sme využili funkciu `glm` z balíku `stats`. Pred odhadom sme si vytvorili dátový rámec, ktorý pozostával z troch stĺpcov, a to X, $lag1$ a $lag2$. Stĺpec X obsahuje hodnoty časového radu, stĺpec $lag1$ obsahuje hodnoty časového radu v čase $t-1$ a $lag2$ obsahuje hodnoty časového radu v čase $t-2$. Kedže pre prvé dve pozorovania nasimulovaného časového radu chýbajú oneskorenia, simulovali sme 202 hodnôt, aby sme mali 100 pozorovaní časového radu. Z dátového rámca sme tieto prvé dva riadky odstránili.

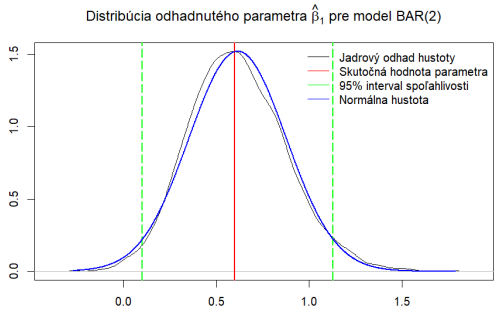
Vo funkcii `glm` sme modelovali podľa predpisu $X \sim 0 + lag1 + lag2$. Hodnota 0 v tomto predpise značí, že nechceme modelovať intercept β_0. Parameter `family` sme nastavili ako `binomial`.

V nasledujúcej tabuľke uvádzame priemerné hodnoty a disperzie odhadnutých parametrov.

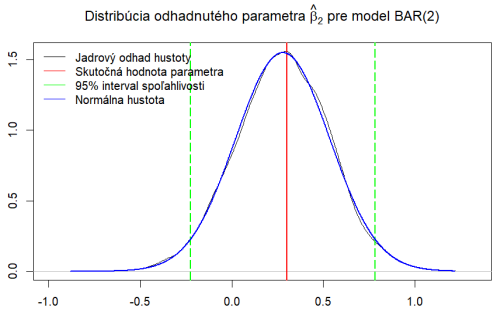
Tabuľka 9: Stredná hodnota a disperzia parametrov pre BAR(2) model

Parameter	$\hat{\beta}_1$	$\hat{\beta}_2$
pôvodná hodnota	0.6	0.3
priemer	0.616	0.279
disperzia	0.069	0.066

Vykreslili sme distribúcie jednotlivých odhadnutých parametrov, pričom sme do grafu vykreslili aj skutočné hodnoty parametrov, s ktorými sme generovali daný proces, 95%-ný interval spoľahlivosti a na porovnanie aj hustotu normálneho rozdelenia s parametrami odhadnutými z dát.



Obr. 27: Distribúcia parametra $\hat{\beta}_1$ v modeli BAR(2)



Obr. 28: Distribúcia parametra $\hat{\beta}_2$ v modeli BAR(2)

Na základe tabuľky 9 vidíme, že odhadnuté parametre v priemere vychádzajú približne rovnaké hodnotám, s ktorými sme simulovali proces BAR(2). Z grafov 27 a 28 vidíme, že distribúcie parametrov sú približne symetrické okolo stredných hodnôt.

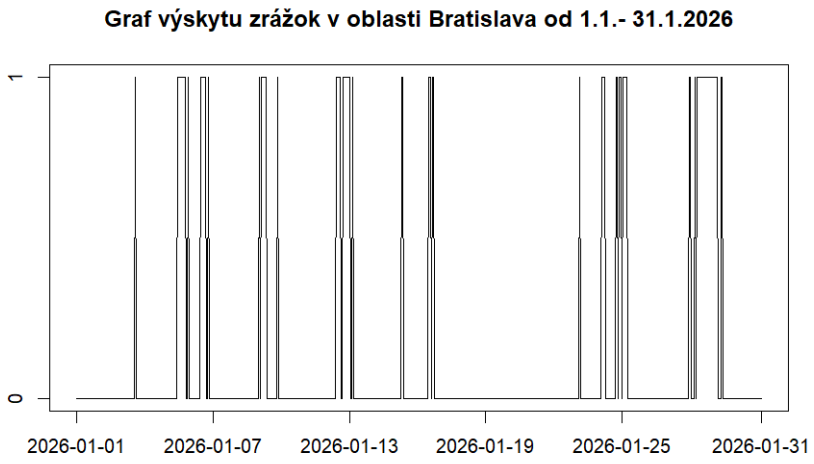
4.3 Predikcie

V tejto podkapitole si ukážeme výpočet predikcií na reálnych dátach, ale ešte pred aplikáciou si priblížime teoretický výpočet predikcií pre BAR(p) model. Počítajme

$$\begin{aligned}
h = 1: \quad g(\pi_{t+1}) &= \hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\beta}_2 x_{t-1} + \cdots + \hat{\beta}_p x_{t+1-p} \\
h = 2: \quad g(\pi_{t+2}) &= \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{t+1} + \hat{\beta}_2 x_t + \cdots + \hat{\beta}_p x_{t+2-p} \\
&= \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{t+1} + \hat{\beta}_2 x_t + \cdots + \hat{\beta}_p x_{t+2-p} \\
&\vdots \\
h = h: \quad g(\pi_{t+h}) &= \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{t+h-1} + \hat{\beta}_2 \hat{x}_{t+h-2} + \cdots + \hat{\beta}_p \hat{x}_{t+h-p} \\
&= \hat{\beta}_0 + \hat{\beta}_1 \hat{\pi}_{t+h-1} + \hat{\beta}_2 \hat{\pi}_{t+h-2} + \cdots + \hat{\beta}_p \hat{\pi}_{t+h-p}.
\end{aligned}$$

V prvom kroku dosadíme známe hodnoty x_{t+1-p} pre $p = 1, ..., p$ a beta koeficienty odhadneme. V ďalších krokoch dosadíme odhadnuté beta koeficienty a neznáme hodnoty x_{t+h-p} odhadneme pomocou $\hat{\pi}_{t+h-p}$.

Výpočet v softvéri R: Pri praktickej aplikácii BAR modelu sme zvolili meteorologické hodinové pozorovania pre Bratislavu od 1.1. po 31.1.2026. Tieto dáta obsahovali informácie o hodinových meraniach relatívnej vlhkosti vyjadrenej v percentách s označením `relative_humidity`, úhrn zrážok v mm označený ako `precipiation`, oblačnosť v percentách označená `cloud_cover` a povrchový tlak vyjadrený v hPa, označený ako `surface_pressure`. Stĺpec vyjadrujúci zrážky sme si premenili na binárne hodnoty a pomenovali sme ho `precipiation_binary`. Ak hodnoty boli väčšie ako 0 mm, označili sme ich 1, teda v daný deň padali zrážky, buď dažďové alebo snehové. Ak nepadali zrážky v daný deň, označenie bolo 0. Následne sme vykreslili priebeh výskytu zrážok, ktorý je vidieť na obrázku 29. K týmto binárnym hodnotám sme dopočítali oneskorené hodnoty ostatných premenných do času $t - 1$. K týmto jednotlivým dopočítaným premenným sme pridali príponu *lag* za názov premennej. Výpočet bol podobný ako v podkapitole 4.2.



Obr. 29: Graf výskytu zrážok počas jednotlivých dní v Bratislave od 1.1. do 31.1. 2026 v hodinových intervaloch

Následne sme dáta rozdelili na dve časti. Prvá časť, na základe ktorej sa odhadli parametre modelu, pozostávala z 671 pozorovaní a druhá časť, ktorá slúži na porovnanie predikcií s realitou, obsahovala 72 pozorovaní. Predikcie sme rátali na tri dni dopredu. Spolu sme mali k dispozícii 743 pozorovaní.

Na odhad parametrov sme využili funkciu `glm` s parametrom `family` = *binomial*. Bernoulliho rozdelenie je špeciálnym prípadom binomického rozdelenia, ak sa pokus vykoná práve raz. Odhadli sme dva modely. V prvom sme modelovali podľa predpisu

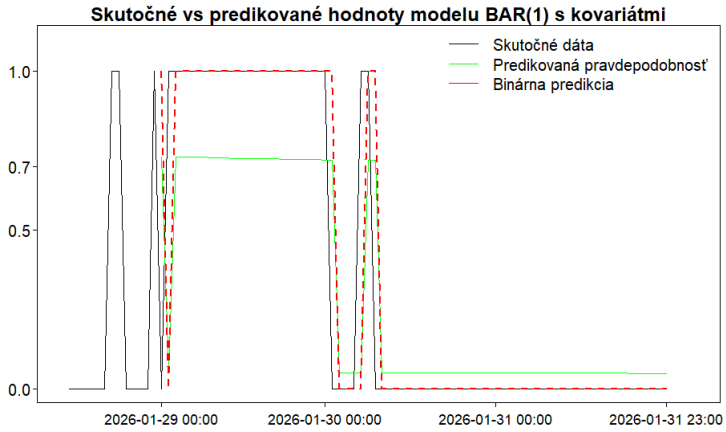
$$\begin{aligned}\text{precipitation_binary} \sim 0 &+ \text{precipitation_binary_lag} \\ &+ \text{relative_humidity_lag} \\ &+ \text{cloud_cover_lag} \\ &+ \text{surface_pressure_lag}\end{aligned}$$

V tomto modeli vyšla premenná *relative_humidity_lag* nesignifikantná. Druhý model sme teda modelovali už bez tejto premennej. Predpis druhého modelu bol nasledovný

$$\begin{aligned}\text{precipitation_binary} \sim 0 &+ \text{precipitation_binary_lag} \\ &+ \text{cloud_cover_lag} \\ &+ \text{surface_pressure_lag}.\end{aligned}$$

Predikcie na tri dni dopredu sme rátali na základe druhého modelu bez premennej *relative_humidity_lag*. Na výpočet sme využili funkciu `predict`. Do parametra `newdata` sme vložili nové hodnoty kovariátov z druhej časti datasetu bez hodnôt prislúchajúcim k *relative_humidity_lag*. V prípade binomického modelu parameter `type` = *response* rovno vypočíta predikované pravdepodobnosti.

Následne sme tieto predikované pravdepodobnosti previedli na binárne hodnoty tak, že ak odhadnutá pravdepodobnosť bola väčšia ako 0.5, tak sme jej priradili hodnotu 1, inak 0. Spolu s týmito hodnotami sme vykreslili aj pôvodné predikované pravdepodobnosti, ktoré je vidieť na obrázku 30.



Obr. 30: Graf výskytu zrážok v Bratislave spolu s predikovanými pravdepodobnosťami

Nakoniec sme jednoduchým spôsobom vyčíslili presnosť úspešných predikcií porovnaním v koľkých prípadoch predikcia zodpovedala realite. Model správne predikoval výskyt zrážok v 67 prípadoch a v piatich prípadoch sa predikcia netrafila do skutočnosti. Presnosť úspešných predikcií bola približne 93 %. K chybným predikciám došlo pri prechode medzi hodnotami 0 a 1.

Na základe našich výsledkov má zmysel modelovať binárne časové rady pomocou regresie využitím funkcie `glm`, pretože umožňuje pracovať s binomickým rozdelením a poskytuje priamo predikované pravdepodobnosti.

Záver

Hlavným cieľom tejto diplomovej práce bolo spraviť prehľad modelov časových radov pre vybrané špecifické typy dát, ilustrovať ich na reálnych dátach a prípadne porovnať porovnateľné modely. Venovali sme sa časovým radom, ktoré nie sú založené na predpoklade normálneho rozdelenia, a to časovým radom počtov, proporcií a binárnym časovým radom.

Prvá kapitola bola venovaná základným pojmom časových radov. Stručne sme definovali pojmy ako časový rad, stredná hodnota, autokovariančná funkcia, stacionarita, nestacionarita, autokorelačná a parciálna autokorelačná funkcia. Uviedli sme vlastnosti bieleho šumu a popísali sme operátor posunu. Následne sme definovali AR, MA a ARMA modely. Na záver tejto kapitoly sme predstavili Ljung-Boxov test, ktorý slúži na overenie nekorelovanosti rezíduí.

Druhá kapitola bola zameraná na časové rady počtov. Zamerali sme sa na tri modely, a to INAR(1), ACP(1,1) a PAR(1) model. Pre všetky modely sme definovali ich predpis spolu s ich vlasnosťami. V prípade INAR modelu sme bližšie popísali binomický thinning operátor, kde sme dokázali vlastnosti (2.20), (2.21) a (2.22).

Pri odhade parametrov modelu sme analyzovali ich distribúciu. Výsledky ukázali, že použité funkcie v priemere pomerne presne odhadovali nami zadané parametre, pričom menšie odchýlky sa vyskytli pri odhadovaní parametra ω v prípade modelu ACP(1,1). V podkapitole 2.4 pri aplikačnom výpočte rpedikcií sme v prípade ACP a PAR modelov porovnali dva prístupy výpočtu 95%-ných predikčných intervalov. V oboch prípadoch vychádzala priemerná šírka intervalu väčšia pri variante *quantiles*. V prípade INAR modelu sme vypočítali predikčný interval simulačne, pretože balík `spINAR` neposkytuje funkciu na výpočet predikcií.

V ďalšej podkapitole 2.5 sme definovali Pearsonove rezíduá. Na základe našich výpočtov práve model ACP dosahoval najlepšie výsledky. V podkapitole 2.6 sme sa zamerali na výpočet predikcií pre sezónne dáta. Definovali sme predpis ACP(1,1) modelu s logaritmickou linkovou funkciou spolu s kovariátmi (2.62). Prínosom tejto časti bola praktická ukážka práce so sezónnymi dátami. Na záver druhej kapitoly sme simulačne porovnali dva prístupy: predikcie pomocou funkcií pre časové rady počtov a predikcie pomocou funkcií pre ARMA modely. Výsledky vyšli približne porovnateľné, presnosť ACP modelu bola

približne 95%, zatiaľ čo v prípade ARMA modelu to bolo o niečo menej, a to 93%.

V tretej kapitole sme sa venovali časovým radom proporcií. Zamerali sme sa na βARMA model, kde sme definovali jeho predpis, a taktiež ako v prípade časových radov, sme sa pozreli na distribúciu odhadnutých parametrov. V prípade predikcií sme odvodili teoretický výpočet a ukázali sme postup modelovania pomocou βARMA modelu na dátach z reálneho sveta. Zaujímavý výsledok prinieslo simulačné porovnanie vypočtu predikcií pomocou βARMA a ARMA modelu. Predikčné intervaly pre βARMA model vychádzali v priemere užšie ako v prípade ARMA modelu, čo sa odzrkadlilo aj na presnosti. Presnosť βARMA modelu vyšla približne 88%, zatiaľ čo v prípade ARMA modelu to bolo až 94%.

V poslednej štvrtej kapitole sme sa venovali binárnym časovým radom, kde sme sa zamerali na BAR model. Hlavným prínosom tejto podkapitoly bola demonštrácia toho, že sa BAR model dá modelovať pomocou časových oneskorení využitím funkcie `glm`, ktorá slúži na odhad zovšeobecnených lineárnych modelov.

Za hlavný prínos tejto diplomovej práce považujeme komplexné spacovanie témy negaussovských časových radov od definovania modelov, popísania ich vlastností až po praktické aplikácie. Veľkým prínosom bolo spracovanie postupu v softvéri R, kde sme pri každom modeli predstavili vhodné balíky a ilustrovali použitie ich funkcií na rôznych príkladoch.

Táto práca by sa ďalej mohla rozšíriť o ďalšie typy negaussovských časových radov, prípadne o hlbšie preskúmanie jednotlivých modelov.

Zoznam použitej literatúry

[1] Al-Osh, M. A., Alzaid, A. A.: *An integer-valued pth-order autoregressive structure (INAR(p)) process*, Journal of Applied Probability, Volume 27, Issue 2, str.314-324 1990

[2] Al-Osh, M. A., Alzaid, A. A.: *First-order integer-valued autoregressive INAR(1) process*, Journal of Time Series Analysis, Wiley Online Library, 1987

[3] Balakrishna N.:*Non-Gaussian Autoregressive-Type Time Series*, Springer, 2021

[4] Brandt, P. T., Wiliams, J. T.: *A Linear Poisson Autoregressive Model: The Poisson AR(p) Model*, Political Analysis, 2001

[5] Bu R., McCabe B.: *Maximum likelihood estimation of higher-order integer-valued autoregressive processes*, Journal of Time Series Analysis, Volume 29, Issue 6, str. 973-994, 2008

[6] Faymonville, M.: *(Semi)Parametric Estimation and Bootstrapping of INAR Models*, 2024

[7] Fidino, M., Magle, S. B.: *Using Fourier series to estimate periodic patterns in dynamic occupancy models*, Ecosphere, Volume 8, Issue 9, 2017

[8] Fokianos, K., Tjøstheim, D.: *Log-linear Poisson autoregression*, Journal of multivariate analysis, Volume 102, Issue 3, str.563-578, 2011

[9] Heinen A.: *Modelling Time Series Count Data: An Autoregressive Conditional Poisson Model*, 2003

[10] Hudecová Š.: *Structural changes in autoregressive models for binary time series*, Journal of Statistical Planning and Inference, Volume 143, Issue 10, str. 1744-1752, 2013

[11] Kedem B., Fokianos K.: *Regression models for time series analysis*, John Wiley & Sons, 2005

[12] Kella O., Löpker A.: *On Binomial Thinning and Mixing*, Indagationes Mathematicae, Volume 34, Issue 5, str. 1121-1145, 2023

[13] Kirchgaessner G., Wolters J., Hassler U.: *Introduction to Modern Time Series Analysis. Second Edition* , Springer Berlin, Heidelberg, 2013

[14] Liboschik, T., Fokianos, K., Fried, R.: *tscount: An R Package for Analysis of Count Time Series Following Generalized Linear Models*, Journal of Statistical Software, Vol. 82, str.1–51, 2017

[15] Rocha A. V., Cribari-Neto F.: *Beta autoregressive moving average models*, TEST 18, str.529–545, 2009

[16] Schaedler Prass T., Pumi G.: *BTSR: Bounded Time Series Regression*, R package version 1.0.0, 2025

[17] Silva N., Pereira I., Eduarda Silva M.: *Forecasting in INAR(1) model*, Revstat - Statistical Journal 7(1), 2022, str.119-134

[18] Simarmata D. M., Novkaniza F., Widyaningsih Y.: *A time series model: First-order integer-valued autoregressive (INAR(1))*, AIP Conference Proceedings, Volume 1862, Issue 1, 2017

[19] Weiß Ch.: *An Introduction to Discrete-Valued Time Series*, 2018, str.73-94

[20] Weiß, C.H.: *Thinning operations for modeling time series of counts—a survey*, Adv Stat Anal 92, 2008, str.319–341

[21] Weiß Ch., Scherer L., Aleksandrov B.: *Checking Model Adequacy for Count Time Series by Using Pearson Residuals*, Journal of Time Series Econometrics, 2019

Príloha

Všetky kódy použité v tejto diplomovej práci je možné nájsť na stránke GitHub: https://github.com/nzifcakova/DP-Negaussovske_casove_rady.

Kódy sú pomenované podľa podkapitoly, ku ktorej sa vzťahujú. Pred použitím si prosím prečítajte `README.md`.