



VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
EKONOMETRIJOS BAKALAURO STUDIJŲ PROGRAMA

**Internetinių puslapių apsilankymų statistinė
analizė**

Statistical analysis of website visits

Baigiamasis bakalauro darbas

Autorius: Evelina Molytė
VU el. p.: evelina.molyte@mif.stud.vu.lt

Darbo vadovas: prof.dr. Vytautas Kazakevičius

Vilnius
2022

Santrauka

Šiame bakalauro baigiamajame darbe yra tiriamos Lietuvoje veikiančių 2 internetinių puslapių - kurie pažymėti kaip A ir B, apsilankymų duomenys. Darbas suskirstytas į dvi dalis - teorinę ir tiriamąją. Teorinėje dalyje trumpai aprašyti naudojami metodai. Tiriamojoje dalyje ištiriami duomenys, sudaromi GLM, ARMA ir BATS modeliai ir atliekamos jų prognozės. Tarpusavyje palyginus modelius, buvo nustatytas tiksliausias modelis, kuriuo atliktos ilgesnio laikotarpio prognozės. Techninė darbo dalis atlikta naudojantis „RStudio“ programine įranga ir paketais.

Raktiniai žodžiai: Laiko eilutės, Apibendrintasis Tiesinis Modelis, ARMA modelis, BATS modelis.

Summary

In this bachelor's thesis, the data of visits to 2 websites operating in Lithuania - which are marked as A and B - are researched. The work is divided into two parts - theory and research. The theory part briefly describes the methods that are used. The research part examines the data, creates GLM, ARMA and BATS models and forecasts. Models were compared and the most accurate model selected to make longer-term forecasts. The technical part of the work was performed using RStudio software and packages.

Keywords: Time series, GLM model, ARMA model, BATS model.

Turinys

1	Įvadas	4
2	Interneto naudojimas	5
3	Teorinė dalis	7
4	Tiriamoji dalis	13
4.1	Duomenų aprašymas	13
4.2	Pirminė duomenų analizė	13
4.3	Apibendrintasis Tiesinis Modelis	16
4.3.1	Modelio sudarymas	16
4.3.2	Modelio charakteristikos	16
4.3.3	Modelio liekanų patikrinimas	17
4.3.4	Apsilankymų prognozavimas	19
4.4	ARIMA modelis	21
4.4.1	Duomenų pavaizdavimas	21
4.4.2	Stacionarumo patikrinimas	22
4.4.3	Modelio eilės pasirinkimas	23
4.4.4	Modelių liekanų analizė	24
4.4.5	Modelių prognozės	26
4.5	BATS modelis	27
4.5.1	Modelių liekanų analizė	28
4.5.2	Modelių prognozės	29
4.6	Tinkamiausio modelio išrinkimas	31
5	Išvados	33
6	Priedai	35
6.1	A priedas	35
6.2	B priedas	40

1 Įvadas

Puslapio srauto duomenys yra svarbūs kiekvieno internetinio puslapio vystyme, ypač kai vis daugiau žmonių pasirenka naudotis internetu. Į srauto duomenis gali pakliūti įvairių rodiklių, kurie gali padėti išsiaiškinti realius vartotojų poreikius bei elgsenos ypatumus kiekviename internetiniame puslapyje. Tačiau vienas svarbiausių internetinių puslapių srauto rodiklių - yra apsilankymų skaičius (angl. *visits*), kartais dar vadinamas sesijomis (angl. *sessions*). Apsilankymų skaičius yra svarbus dėl kelių priežasčių, pirmiausia iš techninės pusės - ar puslapis atlaiko esamą sesijų kiekį ir kiek srauto pritraukia puslapis. Tai svarbu matyti esamu laiku ir prognozuoti, kad būtų galima atlikti veiksmus, kurie padėtų išvengti puslapio lūžių (angl. *page crash*), arba pritrauktų srautą į puslapį.

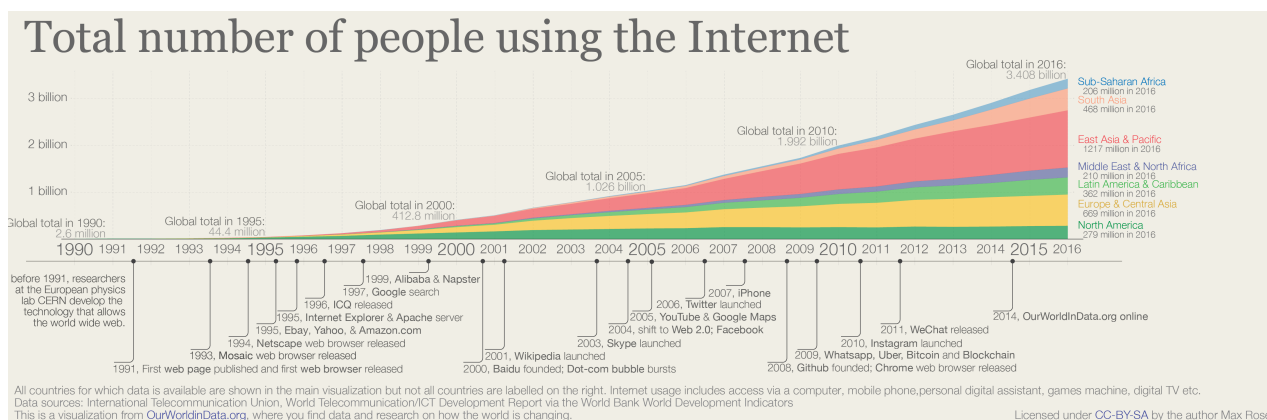
Šio baigiamojo darbo pagrindinis tikslas – surasti tinkamiausią modelį kiekvienai laiko eilutei ir atlikti prognozę metams į priekį.

Pagrindiniai uždaviniai:

- atlikti turimų duomenų analizę;
- sudaryti regresinį modelį;
- sudaryti laiko eilučių modelius;
- atlikti sudarytų modelių liekanų analizę;
- sudaryti nagrinėjamų modelių trumpalaikes prognozes ir palyginti jų tikslumą.

2 Interneto naudojimas

Interneto, kurį naudojame dabar, pradžios data yra laikoma 1983 m. sausio 1-oji. Žinoma, pradžioje interneto galimybės buvo labai ribotos, o prieigą daugiausiai turėjo universitetų bendruomenės. Nuo 1993 metų, išleidus „Mosaic“ naršyklę, prieinamumas tapo paprastesnis ir prieinamumas atsivėrė ir paprastiems žmonėms. Internetinės prekybos pradmenys buvo pradėti apie 2000-uosius metus ir pradinių verlių žlugimas padėjo išvystyti šią sferą į tai kuo naudojames dabar ([2]). Per pastaruosius beveik 30 metų, kaip matoma iš 1 paveikslėlio, internetu besinaudojančių žmonių skaičius vis augo visose nurodytose šalyse, taip pat augo ir laikas, praleidžimas internete, tam beabejo turėjo įtakos ir naujų internetinių puslapių bei socialinių tinklų vystymas ir taip pat interneto prieigos plėtojimas ([3]). Taip pat „Covid-19“ pandemijai paplitus visame pasaulyje, internetas tapo neatsiejamas nuo daugelio žmonių kasdienybės - darbas, mokslai buvo vykdomi nuotoliniu būdu, apsipirkimas ir bendravimas taip pat daugiausiai buvo galimas tik internetu. Tad interneto vartotojų skaičius šoktelėjo dar aukščiau, pirmais pandemijos metais - 2020 interneto vartotojų skaičius išaugo 10%, kaip pateikta šaltinyje [4]. Dėl šios priežasties ir verslai daugiau perėjo, labiau vystė, arba pradėjo kurtis būtent internete. Vien Lietuvoje galima surasti daugybę įvairaus pobūdžio puslapių bei programėlių, kurios palengvina kasdienybę ir yra naudojamos po kelis kartus per savaitę ar dieną, tai pavyzdžiui: maisto užsakymo, sporto namuose, bilietų įsigijimo, mokymų, technikos įsigijimo, transporto nuomos, vaizdo įrašų peržiūrų, socialinių tinklų ir kiti puslapiai. Taigi augant vartotojų skaičiui, auga ir apsilankymų skaičius įvairiuose internetiniuose puslapiuose, ir visų puslapių savininkai siekia pritraukti lankytojus būtent į savo puslapius. Stebint apsilankymų skaičiaus kitimus, galima laiku pastebėti problemas, tokias kaip puslapio lūžius, kurie atsitinka dėl per didelio srauto, tokiu atveju tinklalapio savininkas turėtų apsvarstyti galingesnio serverio įsigijimą, kad daugiau tokių problemų nepasikartotų. Taip pat esant mažam srautui, verslo vykdytojai turėtų atlikti savo puslapio analizę, pasilyginti su konkurentais - galbūt puslapis yra pasenęs (angl. *outdated*), nepatogus, nepatrauklus vartotojui, tada verta pagalvoti apie investavimą į puslapio atnaujinimą ir efektyvinimą.



1 pav.: Internetu besinaudojančių žmonių skaičius kiekvienais metais. Šaltinis [3].

3 Teorinė dalis

Skyriuje bus pateikti šiame darbe taikytų metodų ir laiko eilučių modelių trumpi aprašymai.

Laiko eilutės

Laiko eilutė - tai stebėjimų rinkinys x_t , surinktas per laiką t [4]. Laiko eilutė padeda pamatyti, kaip kintamasis keičiasi per laiką. Taip pat laiko eilučių pagalba galima prognozuoti ateities rodiklius, pasinaudojant praeities turimais duomenimis.

Apibendrintasis tiesinis modelis

Apibendrintasis tiesinis modelis (GLM) - tai tiesinis modelis, apibūdinantis priklausomo kintamojo vidurkio funkciją [7]:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K.$$

Dispersijos mažėjimo daugiklis

Dispersijos mažėjimo daugiklis (angl. *variance inflation factor*) - VIF, parodo ar yra multikolinearumo problema, tai yra ar regresoriai stipriai koreliuoja. Šis rodiklis yra apskaičiuojamas kiekvienam regresoriui ir esant didesnei nei 4 koreliacijai, sakoma, kad multikolinearumas yra. Ir atvirkščiai - kai $VIF < 4$ - multikolinearumo nėra [8].

Stjudento kriterijus

Stjudento kriterijus atskiriems regresoriams padeda nustatyti, ar atitinkamas regresorius yra statistiškai reikšmingas, t.y. jei kriterijaus p reikšmė $< 0.05 = \alpha$. Jei regresorius nėra statistiškai reikšmingas, tada jis turėtų būti pašalintas iš modelio, paliekamas modelyje gali būti tik išimtiniais atvejais [8]. Kaip matysite šiame darbe, vienas iš tokių išimtinių atvejų yra su pseudokintamaisiais, kai tarp jų pasitaikė ir statistiškai reikšmingų ir nereikšmingų, mes juos palikome, nes yra nelogiškai kai kuriuos pseudokintamuosius naikinti, kitų ne, tokiu atveju tektų pasirinkti kitą regresorių arba perkoduoti esamą.

Determinacijos koeficientas

Determinacijos koeficientas - R^2 , tai svarbiausia regresinio modelio charakteristika, kuri parodo duomenų tikimą. Šis koeficiento įgyjamos reikšmės $[0,1]$ intervale (Šaltinis [8]). Kuo koeficiento reikšmė arčiau 1, tuo modelis laikomas tinkamesnis duomenims, tačiau kai $R^2 < 0.20$, laikoma, kad modelis yra netinkamas duomenims.

Baltasis triukšmas

Pirmiausia apibrėšime balto triukšmo seką. Stacionari seka $\{Z_t\}$ vadinama balto triukšmo seka su vidurkiu 0 ir dispersija σ^2 [6], jeigu $\mathbb{E} Z_t = 0$ ir

$$\text{Cov}(Z_t, Z_s) = \begin{cases} \sigma^2, & t = s, \\ 0, & t \neq s. \end{cases}$$

Žymimas $Z_t \sim \text{BT}(0, \sigma^2)$.

Autoregresijos - slenkamojo vidurkio procesas (ARMA)

Procesas $\{X_t, t \in \mathbb{Z}\}$ vadinamas autoregresijos - slenkamojo vidurkio procesu, jeigu $\{X_t\}$ yra stacionarus ir su visais t [6]

$$X_t - \phi_1 X_{t-1} - \dots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \dots + \theta_q Z_{t-q};$$

čia $Z_t \sim \text{BT}(0, \sigma^2)$. Tokią seką žymėsime $\text{ARMA}(p, q)$.

Šią seką galima užrašyti trumpiau:

$$\phi(B)X_t = \theta(B)Z_t, t \in \mathbb{Z};$$

čia:

$$\phi(z) = 1 - \phi_1 z - \dots - \phi_p z^p,$$

$$\theta(z) = 1 + \theta_1 z + \dots + \theta_q z^q$$

ir B yra postūmio atgal operatorius, t.y. $B^j X_t = X_{t-j}$, $j = 0, 1, \dots$. Daugianariai $\phi(\cdot)$ bei $\theta(\cdot)$ vadinami atitinkamai autoregresijos bei slenkamojo vidurkio daugianariais.

Autoregresijos slenkamojo vidurkio integruotas procesas (ARIMA)

Atsitikinė seka $\{X_t, t \in \mathbb{Z}\}$ vadinama (p, d, q) eilės autoregresijos slenkamojo vidurkio integruotu procesu (čia $d = 1, 2, \dots$), jeigu skirtumai $(1 - B)^d X_t$ tenkina $\text{ARMA}(p, q)$ lygtį [6], t.y.

$$\phi(B)(1 - B)^d X_t = \theta(B)Z_t,$$

čia $Z_t \sim \text{BT}(0, \sigma^2)$. Tokį modelį žymėsime $\text{ARIMA}(p, d, q)$.

Stacionari seka

Seka $\{X_t, t \in \mathbb{Z}\}$ vadinama stacionaria plačiąja prasme, jeigu [6]:

1. $\forall t \in \mathbb{Z} \quad EX_t^2 < \infty,$
2. $\forall t \in \mathbb{Z} \quad EX_t = EX_0,$
3. $\text{Cov}(s, t) = \text{Cov}(s + h, t + h), \quad \forall s, t, h \in \mathbb{Z}.$

Išplėstinis Dickey Fullerio testas

Išplėstinis Dickey Fullerio testas (angl. *augmented Dickey–Fuller test (ADF)*) atliekamas modeliui:

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \delta_1 \Delta X_{t-1} + \dots + \delta_p \Delta X_{t-p} + Z_t,$$

čia α - konstanta, β - koeficientas, p - ankstinių skaičius ir Z_t - baltojo triukšmo procesas.

Hipotezės gali būti užrašomos taip:

$$\begin{cases} H_0 : & \gamma = 0 \quad (\text{laiko eilutė turi vienetinę šaknį}), \\ H_1 : & \gamma < 0. \end{cases}$$

Testo statistika yra

$$DF_t = \frac{\hat{\gamma}}{SE(\hat{\gamma})},$$

čia SE - standartinė koeficiento paklaida. Testo statistika lyginama su atitinkama Dickey Fullerio testo, taikomo autoregresiniam modeliui su tiesiniu trendu, kritine reikšme.

Kwiatkowski – Phillips – Schmidt – Shin testas (KPSS)

Taip pat stacionarumo testui naudojamas Kwiatkowski–Phillips–Schmidt–Shin (KPSS) kriterijus. Šis kriterijus dažnai taikomas mažoms imtims. KPSS testui naudojama lygtis [12]:

$$y_t = \beta' D_t + \mu_t + u_t,$$

$$\mu_t = \mu_{t-1} + \varepsilon_t, \quad \varepsilon \sim \text{BT}(0, \sigma^2),$$

čia D_t susideda iš deterministinių komponentų, u_t - stacionarios paklaidos, μ_t - atsitiktinis klaidžiojimas. Šio testo nulinė hipotezė H_0 teigia, kad laiko eilutė yra stacionari.

KPSS testas naudoja Lagranžo daugiklių (LM) statistiką, tikrinant nulinę hipotezę H_0 :

$$KPSS = \frac{T^{-2} \sum_{t=1}^T \hat{S}_t^2}{\hat{\lambda}^2},$$

čia $\hat{S}_t = \sum_{j=1}^t \hat{u}_j$, $\hat{u}_j$ - paklaidos, $\hat{\lambda}^2$ - paklaidų variacijos įvertis.

Autokoreliacijos funkcija (ACF)

Autokoreliacija yra kintamojo koreliacija su savimi skirtingais ankstiniais. Autokoreliacijos funkcija ACF (angl. *autocorrelation function*) yra aprašoma taip [9]:

$$r_k = \frac{c_k}{c_0} = \text{Cor}(x_t, x_{t+k}),$$

čia c_0 yra x_t autokovariacija, kai ankstinis yra lygus 0, o autokovariacijos funkcija c_k k -tajame žingsnyje yra aprašoma taip:

$$c_k = \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}).$$

Autokoreliacijos 95% pasiklovimo lygmuo apskaičiuojamas naudojantis formule:

$$-\frac{1}{n} \pm \frac{2}{\sqrt{n}},$$

čia n yra kintamųjų skaičius, naudojamas apskaičiuoti ACF.

Autokoreliacijos funkcija yra naudingas įrankis, naudojamas laiko eilučių analizei, kuri padeda atpažinti svarbius duomenų bruožus. Autokoreliacijos funkcija taip pat yra svarbus įrankis, padedant pasirinkti korektišką p ir q tvarką ARMA(p, q) modeliuose.

Dalinės autokoreliacijos funkcija (PACF)

Dalinės autokoreliacijos funkcija PACF (angl. *partial autocorrelation function*) matuoja tiesinę laiko eilučių koreliaciją x_t ir vėlavimą x_{t+k} su tiesiniu nepriklausomu $x_{t-1}, x_{t-2}, \dots, x_{t-(k-1)}$. PACF yra aprašoma taip [9]:

$$f_k = \begin{cases} \text{Cor}(x_1, x_0) = r_1 & \text{if } k = 1; \\ \text{Cor}(x_k - x_k^{k-1}, x_0 - x_0^{k-1}) & \text{if } k \geq 2. \end{cases}$$

Kaip ir autokoreliacijos funkcija, taip ir dalinės autokoreliacijos funkcija yra svarbios, norint pasirinkti korektišką p ir q tvarką ARMA(p, q) modeliuose. Pirmiausia iš autokoreliacijos ACF yra sudaromas MA(q) modelis, o iš dalinės autokoreliacijos – AR(p) modelis. Sujungus AR(p) ir MA(q) modelius gaunamas ARMA(p, q) modelis.

AIC ir BIC

AIC kriterijus - Akaike informacinis kriterijus yra apibrėžiamas taip:

$$AIC = 2k - 2\ln(L),$$

čia k - modelio parametrų skaičius, L - maksimizuota modelio tikėtinumų funkcijos reikšmė. Šis kriterijus yra linkęs parinkti didesnę eilę ir didina tikimybę įtraukti tinkamą modelį.

BIC kriterijus yra apibrėžiamas taip:

$$BIC = k \ln(n) - 2 \ln(L),$$

čia k - modelio parametrų skaičius, n - laiko eilutės stebėjimų skaičius, L - maksimizuota modelio tikėtinumų funkcijos reikšmė. Šis kriterijus yra suderintas asimptotiškai, tai prie be galo daug duomenų parenka tinkamą modelį.

Šiame darbe yra naudojamas AIC kriterijus, nes jis nėra toks griežtas ir didina tikimybę įtraukti tikrą modelį.

Ljung Box testas

Autokoreliacijai tikrinti naudojamas Ljung-Box kriterijus [11]. Hipotezė gali būti užrašyta:

$$\begin{cases} H_0 : \rho_1 = \rho_2 = \dots = \rho_k = 0 & (\text{autokoreliacijos nėra}), \\ H_1 : \rho_k \neq 0, \text{ bent vienam } k & (\text{autokoreliacija yra}). \end{cases}$$

Testo statistika yra apskaičiuojama:

$$Q = n(n+2) \sum_{k=1}^h \frac{\hat{\rho}_k^2}{n-k},$$

čia n - imties dydis, $\hat{\rho}_k$ - imties autokoreliacijos koeficientas k -tajame žingsnyje, h teste naudotų ankstinių skaičius.

BATS modelis

BATS modelis yra eksponentinio glodinimo būsenų erdvės modelis su Box-Cox transformacija, ARMA(p, q) modelio liekanomis bei trendo ir sezono komponentėmis. Šis modelis gali būti

išreiškiamas lygčių sistema [10]:

$$\begin{aligned}
y_t^{(\lambda)} &= \begin{cases} \frac{y_t^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y_t, & \lambda = 0, \end{cases} \\
y_t^{(\lambda)} &= l_{t-1} + \phi b_{t-1} + \sum_{i=1}^T s_{t-m_i}^{(i)} + d_t \\
l_t &= l_{t-1} + \phi b_{t-1} + \alpha d_t \\
b_t &= (1 - \phi)b + \phi b_{t-1} + \beta d_t \\
s_t^{(i)} &= s_{t-m_i}^{(i)} + \gamma_i d_t \\
d_t &= \sum_{i=1}^p \varphi_i d_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i} + \varepsilon_t,
\end{aligned}$$

čia $m_1, \dots, m_T$ žymi sezono periodus, l_t - vietinis lygmuo pagal periodą t , b - ilgalaikis trendas, b_t - trumpalaikis trendas pagal periodą t , $s_t^{(i)}$ nusako i -ąją sezono komponentę laiko momentu t , d_t žymi ARMA(p, q) procesą ir $Z_t \sim \text{BT}(0, \sigma^2)$, α, β, γ_i - glodinimo parametrai, ϕ - slopinimo parametras. Žymėsime tokį modelį BATS($\lambda; \{p, q\}; \phi; \{m_1, \dots, m_T\}$).

4 Tiriamoji dalis

4.1 Duomenų aprašymas

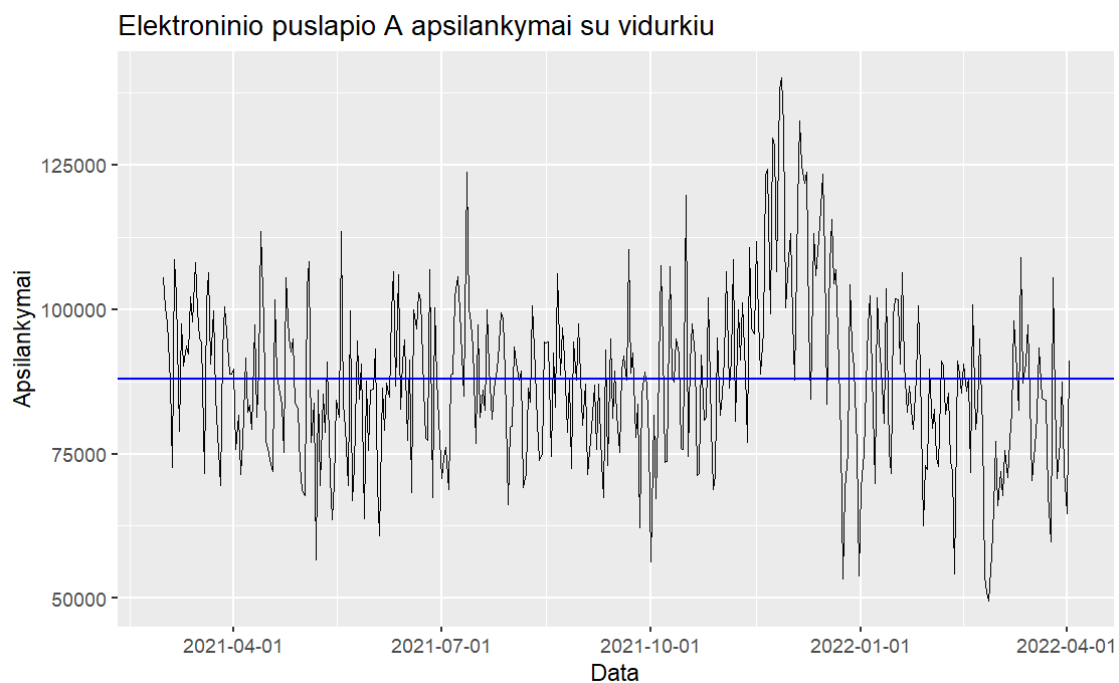
Srauto duomenys Lietuvos internetiniuose puslapiuose - internetinėse parduotuvėse - yra gauti iš „Similarweb“ puslapio. Duomenų bazė nėra laisvai prieinama bei yra mokama, tad prieiga gali naudotis tik įmonės, išsipirkusios prieigos teisę. Duomenų bazėje pateikti įvairūs rodikliai, tokie kaip: unikalių lankytojų skaičius (angl. *unique visitors*, vidutinė apsilankymo trukmė (angl. *average visit duration*, vidutinis aplankytų puslapių skaičius (angl. *average pages visited*) ir kiti rodikliai. Tačiau pagrindinis dominantis rodiklis - apsilankymų skaičius per dieną (angl. *visits*). Pasak „Similarweb“ puslapio ([1]) duomenys yra surenkami 4 būdais:

- tiesioginiu vertinimu (angl. *direct measurement*) - kai svetainės bendradarbiauja pačios ir dalinasi savo duomenimis su „Similarweb“;
- iš įrašų tinklo (angl. *contributory network*) - tai vartotojų produktų rinkinys, kuris kaupia anoniminius įrenginio elgsenos duomenis;
- iš partnerysčių (angl. *partnerships*) - pasaulinis organizacijų tinklas, kuris renka „skaitmeninius signalus“;
- ir galiausiai viešųjų duomenų ištraukimu (angl. *public data extraction*) - pažangus algoritminis variklis fiksuoja ir indeksuoja viešuosius duomenis iš įvairių svetainių ir programų.

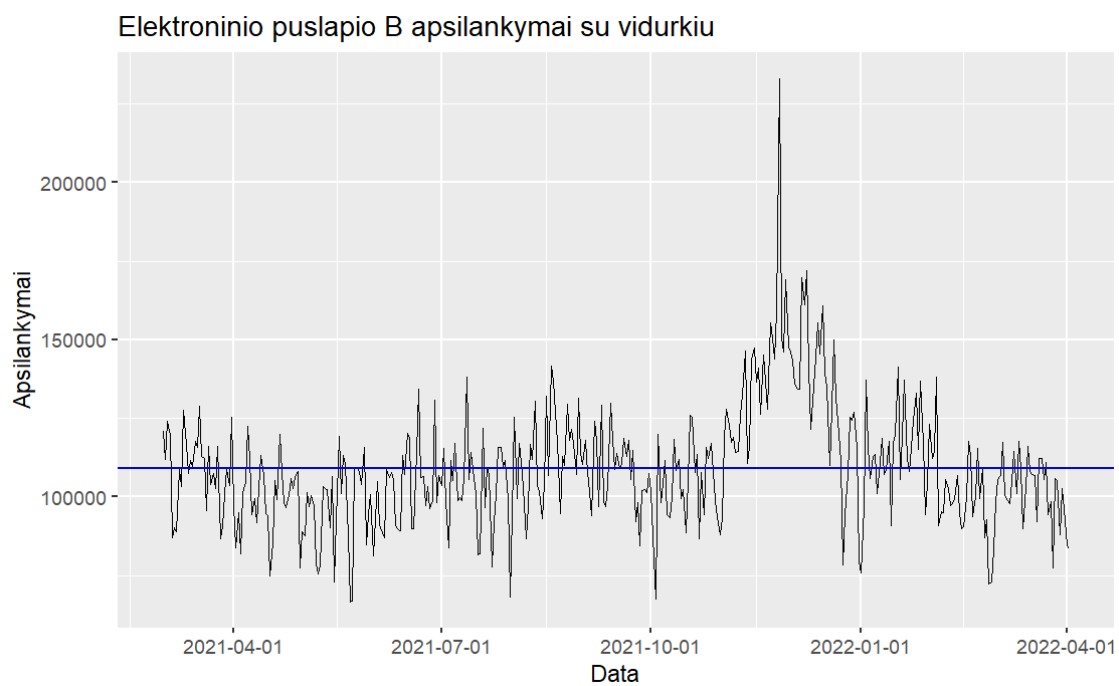
Tad žinant šią informaciją, pateiktą „Similarweb“ puslapyje, nėra galimybės nustatyti koku tikslu būdu buvo surinkti naudojamų puslapių apsilankymų duomenys, tad laikysime, kad - buvo gauti tiesioginio vertinimo būdu. Turimus apsilankymų duomenis sudaro 398 dienų stebėjimai - nuo 2021-03-01 iki 2022-04-02.

4.2 Pirminė duomenų analizė

Tyrimui naudojami dviejų Lietuvoje veikiančių internetinės prekybos puslapių apsilankymų duomenys, juos įvardinsime kaip - **puslapis A** ir **puslapis B**. Abu puslapiai prekiauja panašaus pobūdžio prekėmis ir yra konkurentai. Pirmas laiko eilučių analizės žingsnis yra - duomenų grafinis vaizdavimas [6]. Duomenų grafikas neretai gali parodyti tam tikrus duomenų pokyčius ar pasikartojimus. Pirmojo elektroninio puslapio - A apsilankymų grafikas matomas 2 pav. , o B puslapio apsilankymų grafikas - 3, mėlynos linijos grafikuose vaizduoja duomenų apsilankymų vidurkį.



2 pav.: Internetinio puslapio A apsilankymų laiko eilutė.



3 pav.: Internetinio puslapio B apsilankymų laiko eilutė.

Kaip matoma iš grafikų bendras abiejų puslapių bruožas - tai Kalėdiniu laikotarpiu padidėjęs srautas, po kurio seka vėl įprasti apsilankymų kiekiai, šis laikotarpis yra išskirtis, nes turimi tik vienerių metų duomenys, esant ilgesniam stebėjimų laikotarpiui, srauto padidėjimas Kalėdiniu laikotarpiu būtų laikomas sezoniškumu. 2022 metų vasario pabaigoje - kovo pradžioje matomas laikotarpis, kai apsilankymų srautas A ir B puslapiuose buvo labai nukritęs, tai būtų galima sieti su karo prasidėjimu Ukrainoje, kuomet daugelis Lietuvos gyventojų buvo nežinioje ir savo laiką skyrė kitoms veikloms, o ne lankymuisi internetinėse parduotuvėse. Taip pat verta paminėti, kad A puslapio apsilankymų augimas nėra matomas, jei atmestume Kalėdinį laikotarpį. Tuo tarpu B puslapio laiko eilutėje matomas nežymiai didėjantis trendas.

Statistinės charakteristikos	A puslapio, (tūkst.)	B puslapio, (tūkst.)
Mediana	87.47	107.10
Vidurkis	87.88	109.18
Dispersija	2.20×10^5	3.63×10^6
Standartinis nuokrypis	14.84	19.06
Minimumas	49.61	66.67
Maksimumas	140.10	233.03

1 lentelė: A ir B puslapių apsilankymų pagrindinės statistinės charakteristikos.

Puslapių apsilankymų vidurkiai skiriasi kiek daugiau nei 20 tūkst.: A puslapio apsilankymų vidurkis yra 87.9 tūkst. ir nuo medianos skiriasi nežymiai, tuo tarpu B puslapio apsilankymų vidurkis siekia 109 tūkst. ir nuo medianos skiriasi per 2 tūkst. Apskaičiavus kitas statistines charakteristikas, kurios pateiktos 1 lentelėje, matoma, kad A puslapyje didžiausias apsilankymų skaičius per dieną buvo - 140 tūkst., mažiausias - 49.6 tūkst., o B puslapyje daugiausiai apsilankymų buvo 233 tūkst., o mažiausias apsilankymų skaičius per dieną 66.6 tūkst. Taip pat svarbu paminėti standartinį nuokrypį, kuris parodo kaip duomenys yra išsibarstę ar nutolę nuo vidurkio, A atveju standartinis nuokrypis 14.8 tūkst., o B standartinis nuokrypis 19 tūkst., tad abiem atvejais duomenys yra plačiai išsibarstę.

Sekantis analizės žingsnis buvo suskirstyti turimą duomenų imtį į mokymosi (angl. *training*) ir testavimo (angl. *testing*) duomenis, testavimo duomenys buvo paimti paskutinio stebėto mėnesio - nuo 2022-03-03 iki 2022-04-02 imtinai, kas sudaro 31 dieną. Kaip rašoma šaltinyje [5] suskirstymas į mokymosi ir testavimo duomenis yra svarbus tuo, kad tokiu būdu yra galimybė palyginti prognozuojamas reikšmes su realiomis, taip įvertinant modelio tikslumą ir pajėgumą prognozuoti tolimesnes reikšmes.

4.3 Apibendrintasis Tiesinis Modelis

4.3.1 Modelio sudarymas

Kitas internetinių puslapių analizės žingsnis būtų Apibendrintojo tiesinio modelio sudarymas, po kurio seks šio modelio prognozių sudarymas. Sudarydami Apibendrintąjį tiesinį modelį, priklausomajam kintamajam Y - puslapių apsilankymams, parinksime regresorius. Tokiu būdu panaikinsime sezoniškumą ir globalų trendą. Sudaryto modelio liekanos bus naudojamos tolimesniems modelių parinkimams. Šiuo atveju trendo komponentė būtų dienos - nuo 1 iki 367 (nes naudojame mokymosi duomenų imtį), sezonas - tai savaitės dienos, bei išskirtinis įvykis šiuose duomenyse (nes tai 1-nerių metų duomenys) - tai Kalėdinis laikotarpis, kuris pasideda nuo 2021-11-22 iki 2021-12-21, kaip matoma iš duomenų grafiko, tuomet apsilankymų kiekis buvo neįprastai padidėjęs, po kurio vėl sekė įprastas apsilankymų kiekis. Kalėdinis laikotarpis yra faktorinis kintamasis, įgyjantys reikšmes 1, kai yra dienos nuo 2021-11-22 iki 2021-12-21, ir 0 likusiomis metų datomis. Savaitės dienos yra pseudokintamasis (angl. *dummy variable*), ir įgys reikšmes 1, kai bus atitinkama savaitės diena (pvz: antradienis) ir 0, kitomis savaitės dienomis.

Priklausomojo kintamojo Y - apsilankymų apibendrintojo tiesinio modelio formulė atrodytų taip:

$$Y(\text{apsilankymai}) = X_1(\text{diena}) + X_2(\text{Kalėdinis laikotarpis}) + X_3(\text{savaitės diena}).$$

4.3.2 Modelio charakteristikos

Sudaryto modelio koeficientai A puslapio yra pateikti 2 lentelėje. Pirmiausia A puslapio apsilankymų duomenyse pastebima, kad parametras *Diena* yra neigiamas ir statistiškai reikšmingas, nes p reikšmė mažesnė už pasikliovimo intervalą, kadangi šis kintamasis atitinka trendą, tai galima sakyti, kad trendas yra mažėjantis - mažėja po 17 apsilankymų per dieną. Matoma, kad kai yra Kalėdinis laikotarpis, tai apsilankymų padaugėja beveik 28 tūkst. Iš savaitės dienų parametrų matoma, kad visos savaitės dienos yra statistiškai reikšmingos, nes p reikšmės mažesnės už reikšmingumo lygmenį - $\alpha = 0.05$, ir matoma, kad visos reikšmės yra teigiamos, tai rodo kiek kiekvieną savaitės dieną apsilankymų yra daugiau lyginant su penktadieniu. Taigi sakome, kad A puslapio apsilankymų skaičius Y priklauso ir nuo dienos, ir Kalėdinio laikotarpio ir nuo savaitės dienų. Regresoriai neturi stiprios koreliacijos, nes visų regresorių dispersijos mažėjimo daugiklis (VIF) mažesnis už 4. Determinacijos koeficientas $R^2 = 0.44$, tad galima sakyti, kad regresoriai X_1, X_2 ir X_3 paaiškina 44% nepriklausomo kintamojo Y ir modelis yra tinkamas duomenims.

Panaši lentelė matoma ir B puslapio - 3 lentelė. Čia matoma, kad B puslapio dienos parametras yra teigiamas ir statistiškai reikšmingas, tad kasdieną apsilankymų skaičius didėja

Parametras	Koeficientas	Standartinis nuokrypis	Įvertinys	p reikšmė
Laisvasis narys	75443.66	1873.74	40.26	$< 2 \times 10^{-16}$
Diena	-17.01	5.79	-2.94	3.51×10^{-3}
Kalėdos	27929.86	2239.21	12.47	$< 2 \times 10^{-16}$
Savaitės diena pirmadienis	17267.35	2205.08	7.83	5.56×10^{-14}
Savaitės diena antradienis	20796.94	2205.06	9.43	$< 2 \times 10^{-16}$
Savaitės diena trečiadienis	17562.72	2204.75	7.97	2.20×10^{-14}
Savaitės diena ketvirtadienis	10089.55	2215.21	4.56	7.20×10^{-6}
Savaitės diena šeštadienis	18732.6	2215.21	8.46	7.09×10^{-16}
Savaitės diena sekmadienis	11771.76	2215.23	5.31	1.89×10^{-7}

2 lentelė: Puslapio A apsilankymų GLM modelio koeficientų sąrašas.

Parametras	Koeficientas	Standartinis nuokrypis	Įvertinys	p reikšmė
Laisvasis narys	96883.52	2259.8	42.87	$< 2 \times 10^{-16}$
Diena	24.63	6.98	3.52	4.73×10^{-4}
Kalėdos	39043.83	2700.57	14.46	$< 2 \times 10^{-16}$
Savaitės diena pirmadienis	15130.89	2659.42	5.69	2.65×10^{-8}
Savaitės diena antradienis	7060.77	2659.38	2.66	8.28×10^{-3}
Savaitės diena trečiadienis	12659.98	2659.01	4.76	2.80×10^{-6}
Savaitės diena ketvirtadienis	10617.39	2671.62	3.97	8.54×10^{-5}
Savaitės diena šeštadienis	-6506.44	2671.62	-2.44	0.02
Savaitės diena sekmadienis	-3271.87	2671.65	-1.23	0.22

3 lentelė: Puslapio B apsilankymų GLM modelio koeficientų sąrašas.

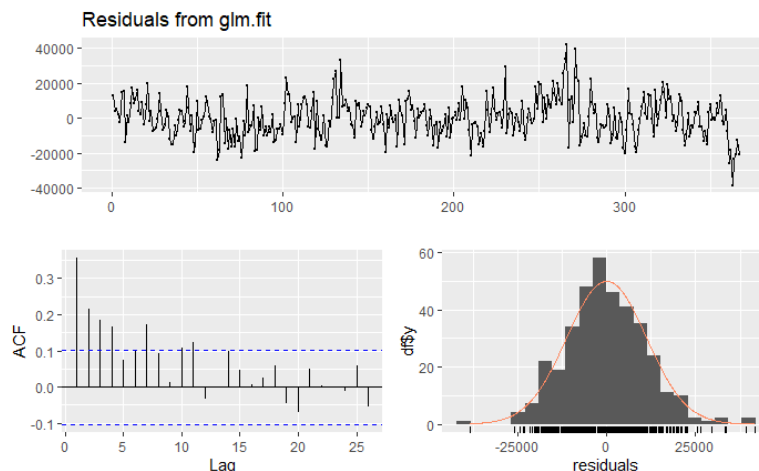
beveik 25 vienetais, o Kalėdinio laikotarpio - apsilankymų padaugėja 39 tūkst. Iš savaitės dienų parametų matoma, kad tik sekmadienis yra statistiškai nereikšmingas parametras, tai reiškia, kad sekmadienį apsilankymų skaičius yra beveik toks pat kaip ir penktadienį, tačiau šio kintamojo mes nepanaikinsime, nes panaikinti pseudokintamuosius nėra logiška. Likusios savaitės dienos yra statistiškai reikšmingos ir teigiamos, tik šeštadienis yra neigiamas, o tai reiškia, kad atitinkamai pagal ženklą apsilankymų buvo daugiau ar mažiau nei penktadienį. Taigi B puslapio apsilankymų skaičius Y taip pat priklauso nuo dienos, Kalėdinio laikotarpio ir savaitės dienų. Gautas determinacijos koeficientas $R^2 = 0.52$, šis modelis paaiškina 52% nepriklausomojo kintamojo Y ir yra tinkamas duomenims.

Taigi matome, kad abiejų internetinių puslapių sudarytų modelių charakteristikos rodo, kad modeliai yra tinkami duomenims.

4.3.3 Modelio liekanų patikrinimas

Modelio liekanų analizė padeda nustatyti ar modelis tikrai yra tinkamas duomenims. Svarbiausia yra patikrinti ar modelis yra autokoreliuotas ir ar vidurkis yra nulinis. Kai šios savybės

yra tenkinamos, liekanos vadinamos Baltuoju triukšmu. Pirmiausia pažvelkime į A puslapio liekanų pasiskirstymo, ACF grafikus ir histogramą 4 pav. Matoma, kad liekanos pasiskirsčiusios aplink 0, iš ACF grafiko matomi keli reikšmingi ankstiniai, tad liekanos atrodo autokoreliuotos. Iš histogramos matome normalaus pasiskirstymo vaizdą. Įsitikinti grafikų teikiama informacija galima atlikus testus Ljung Box testą autokoreliacijai patikrinti, ir Stjudento testą - vidurkio lygybei nuliui. Gautos p reikšmės pateiktos 4 lentėje. Taigi, kaip ir buvo nustatyta iš ACF grafiko - liekanos yra autokoreliuotos, kadangi Ljung Box testo p reikšmė mažesnė už $\alpha = 0.05$. Iš Stjudento testo gavome, kad liekanų vidurkis 0.

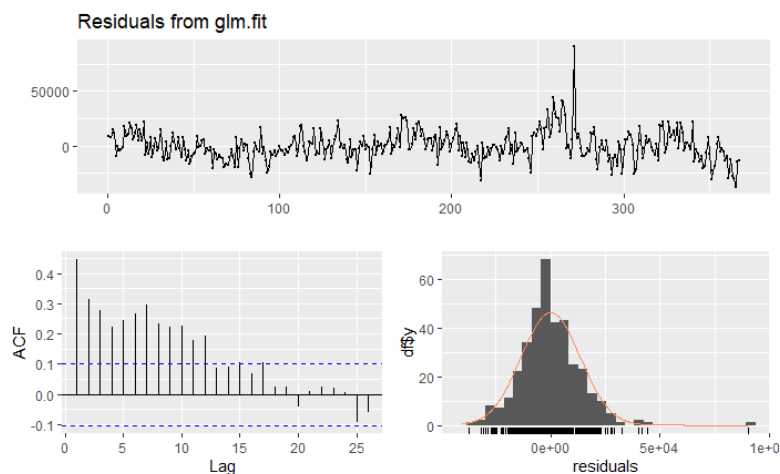


4 pav.: A puslapio GLM modelio liekanų grafikai.

	Ljung Box testas	Stjudento t-testas
A puslapio p reikšmė	8.1×10^{-12}	1
B puslapio p reikšmė	2.2×10^{-16}	1

4 lentelė: A ir B puslapių GLM modelio liekanų testų p reikšmės.

B puslapio GLM modelio liekanų grafikas 5 rodo panašų vaizdą, tad iškart žiūrime kokie testų rezultatai. Toje pačioje lentelėje - 4 pateiktos ir B puslapio testų p reikšmės. Gauta, kad ir šios liekanos yra autokoreliuotos ir turi vidurkis lygus 0.



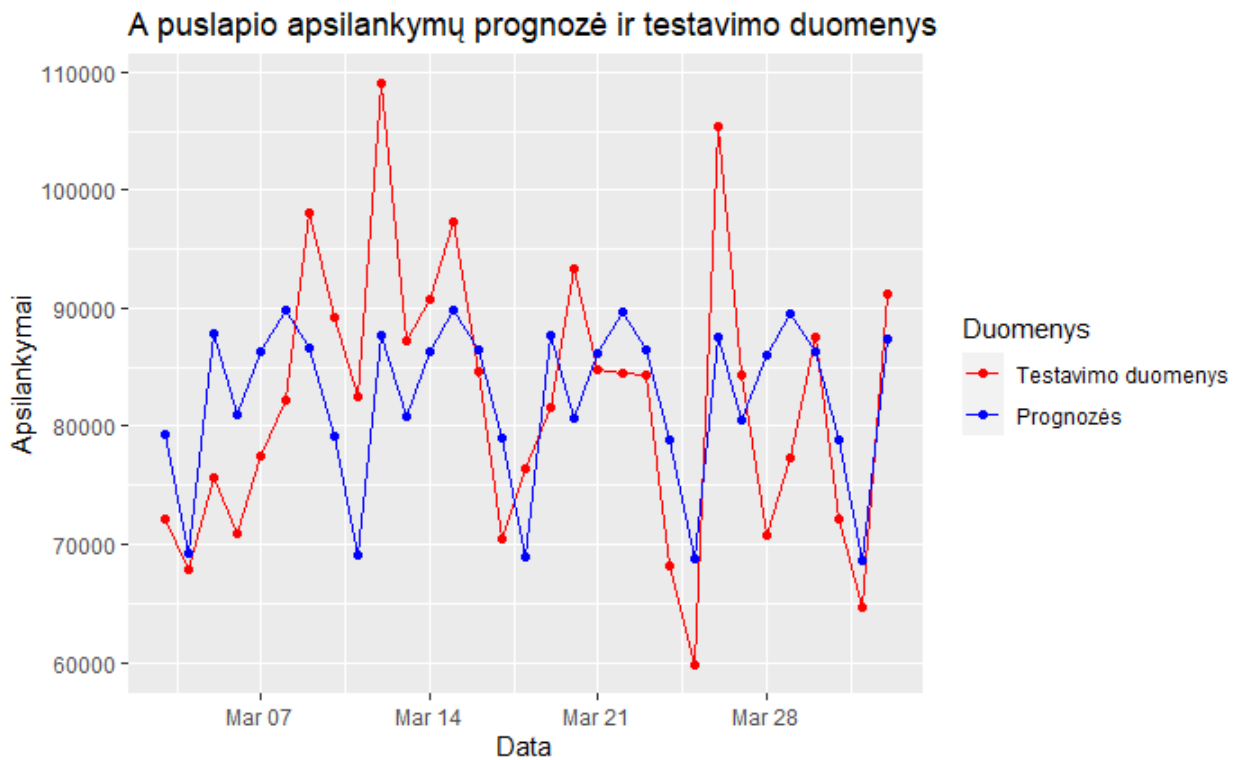
5 pav.: B puslapio GLM modelio liekanų grafikai

Taigi ir A ir B puslapių modelių liekanos yra autokoreliuotos, tad modelio liekanose dar yra naudingos informacijos. Tokiu atveju modelis nėra laikomas adekvačiu, tačiau naudodamiesi vistiek galime atlikti prognozes.

4.3.4 Apsilankymų prognozavimas

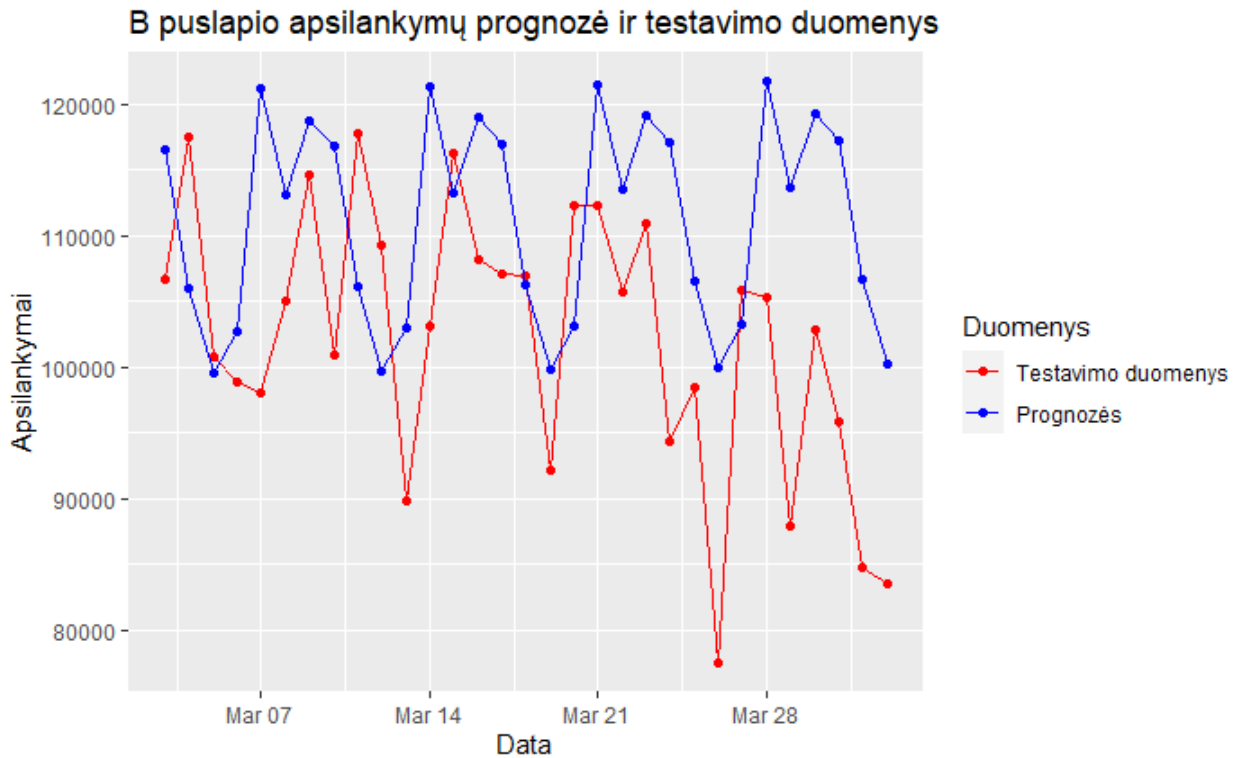
Tolimesnis žingsnis - ateities apsilankymų prognozavimas ir palyginimas su realiais duomenimis. Kadangi turime apskaičiuotus koeficientus - galime juos įsistatyti į Apibendrintojo tiesinio modelio formulę ir taip prognozuoti ateities apsilankymų reikšmes. Šiame žingsnyje jau panaudosime testavimo duomenų rinkinį, kurio dėka pamatysime kaip arti realių reikšmių yra mūsų sudarytos prognozės su Apibendrintuoju tiesiniu modeliu.

Iš A puslapio realių ir prognozuojamų apsilankymų grafiko 6 matoma, kad prognozės yra arti realybės, paskutinėmis dienomis matoma, kad reikšmės yra labai arti viena kitos. Kuo prognozuojamų reikšmių ir realių duomenų linijos yra panašesnės, tuo modelio prognozės yra tikslesnės. Vietomis realūs apsilankymai buvo didesni už prognozuojamus, kas yra natūralu, kadangi apsilankymų srautą gali paveikti ir daugelis kitų neištirtų faktorių. Šiuo laikotarpiu - pirmomis kovo savaitėmis padidėjusį srautą galima paaiškinti taip: tuo laikotarpiu Lietuvoje įsibėgėjo humanitarinės pagalbos Ukrainai misijos, kurioms Lietuvos žmonės aktyviai aukojo įvairias priemones: elektroniką, drabužius, miegmaišius, vaistinėles ir kt, taip pat panašiomis priemonėmis stengėsi apsirūpinti patys. Kai kurioms prekėms fizinėse parduotuvėse tapus neprieinamoms, internetinės parduotuvės tapo išsigelbėjimu. Kadangi šis puslapis A užsiima prekyba kai kuriomis iš paminėtų prekėmis, tai tuo laiku matomas kiek padidėjęs srautas. Matoma, kad prognozuojami apsilankymai bėgant laikui pamažu mažėja, kas yra pastebima ir realiuose duomenyse. Bendrai galima sakyti, kad apibendrintojo tiesinio modelio apsilankymų prognozės yra arti realių apsilankymų.



6 pav.: A puslapio apsilankymų GLM modelio prognozės ir realių reikšmių palyginimas.

Tuo pačiu principu nubraižytas ir B puslapio prognozių ir realių apsilankymų grafikas - 7 pav. Šiame grafike matoma, kad stebimo laikotarpio pradžioje prognozuojami ir realūs duomenys labiau sutampa. Pabaigoje matomas didelis atotrūkis, kadangi bendrai prognozė auga bėgant laikui, o tuo laikotarpiu realūs duomenys mažėja. Tačiau tikėtina, kad realus mažėjimas yra paveiktas kitų išorinių faktorių: galbūt puslapyje nėra prieinamų paklausiausių prekių. Bendrai galima sakyti, kad prognozė pradžioje buvo arti realių duomenų, tačiau matomi didesni nesutapimai pabaigoje.



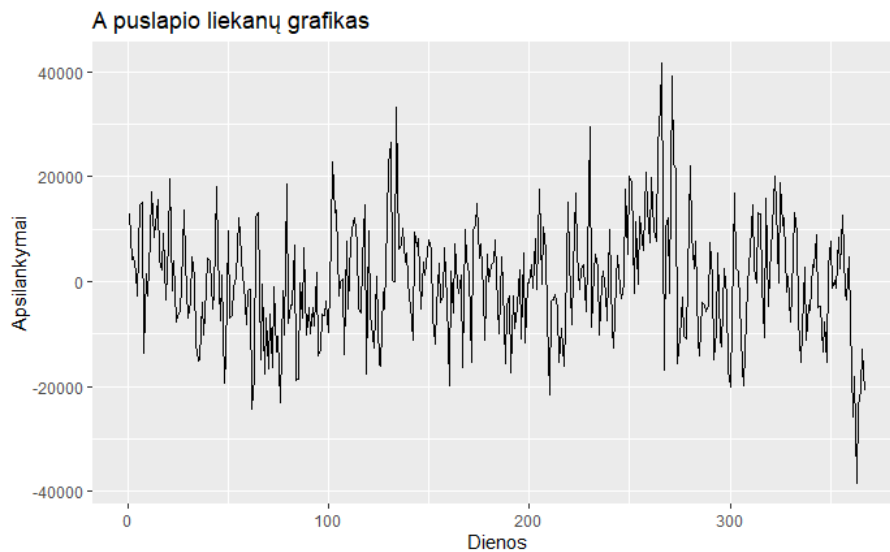
7 pav.: B puslapio apsilankymų GLM modelio prognozės ir realių reikšmių palyginimas.

Tolimesnė duomenų analizė vyks su abiejų puslapių liekanomis, kurios liko pritaikius Apibendrintąjį tiesinį modelį.

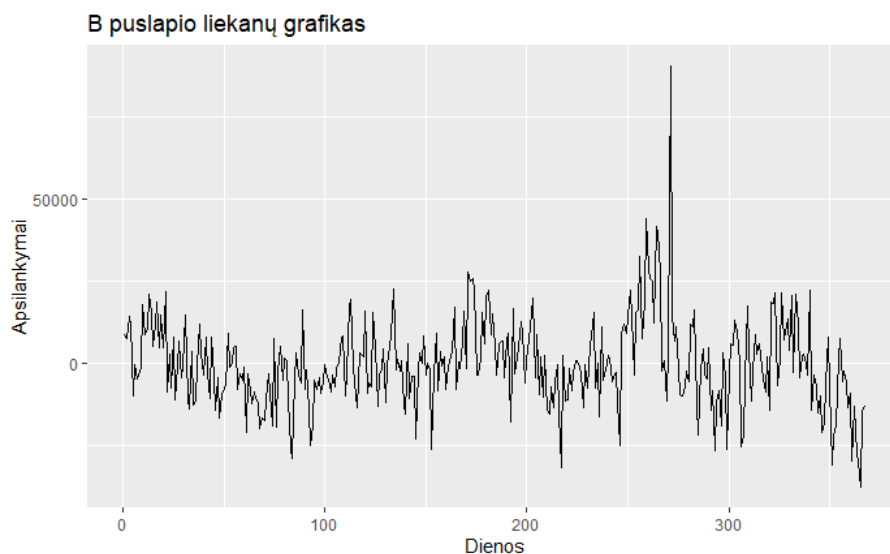
4.4 ARIMA modelis

4.4.1 Duomenų pavaizdavimas

Kadangi toliau naudojami duomenys yra GLM modelio liekanos, tai svarbu pamatyti naujų duomenų vaizdą. A puslapio liekanų laiko eilutės grafike 8 pav. matoma, kad duomenys pasiskirstę gan panšiai, nebėra išskirčių, kaip pradiniuose duomenyse. B laiko eilutėje - 9 taip pat matomas tolygesnis pasiskirstymas, tačiau duomenų stebėjimo pabaigoje pastebimas reikšmių mažėjimas.



8 pav.: A puslapio apsilankymų liekanų laiko eilutė.



9 pav.: B puslapio apsilankymų liekanų laiko eilutė.

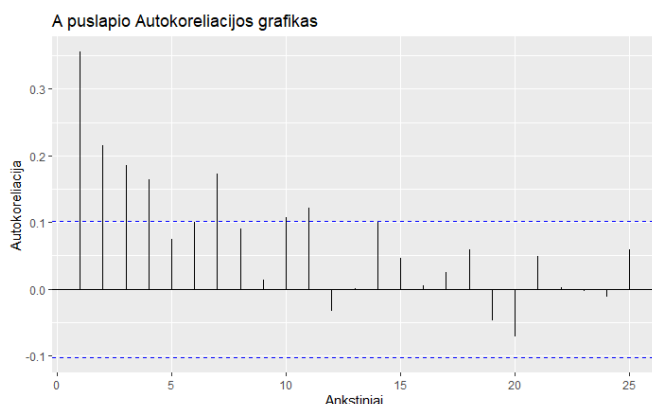
4.4.2 Stacionarumo patikrinimas

Norint sudaryti ARIMA modelį yra svarbu, kad duomenys būtų stacionarūs - nekistų per laiką, kadangi ARIMA prognozuodama reikšmės, atsižvelgia į praeities duomenis. Žvelgiant į prieš tai nubrėžtus liekanų grafikus 8 pav. ir 9 pav., matome, kad A eilutė atrodo stacionari, B eilutė taip pat atrodo stacionari, tik grafiko galas kiek skiriasi nuo pradžios. Norint įsitikinti - patikriname ar turimos liekanos yra stacionarios, tam naudosime išplėstinį Dickey Fullerio testą (ADF), kuris padeda nustatyti, ar duomenyse yra vienetinė šaknis. Patikrinus abiejų puslapių duomenis, gauta, kad A puslapio apsilankymų laiko eilutė yra stacionari. B puslapio eilutė, pasak ADF testo rezultato - ne stacionari, praktikoje nestacionarumo problemą galima

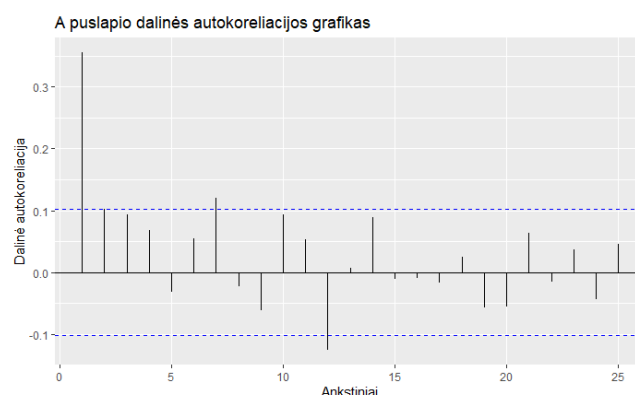
išspręsti diferencijavus duomenis. Tačiau ADF testo išvada nėra galutinė ir stacionarumą dar galima patikrinti naudojantis Kwiatkowski – Phillips – Schmidt – Shin testu (KPSS), kuriuo naudojantis jau buvo nustatyta, kad B puslapio duomenys yra stacionarus. Taigi pasirinkta neatlikti transformacijų pradiniais liekanų B eilutės duomenims ir laikyti juos stacionariais.

4.4.3 Modelio eilės pasirinkimas

Tolimesnis ARIMA modelio žingsnis - $ARIMA(p, d, q)$ tinkamiausių parametų parinkimas. Parametrai p ir q randami naudojantis Autokoreliacijos (ACF) ir Dalinės autokoreliacijos grafiais (PACF), p nustatomas pagal PACF grafiką, o q pagal ACF grafiką. Tuo tarpu d parametro reikšmę nusako diferencijavimas, kadangi abiejų puslapių eilutės nėra diferencijuotos ir neturi vienetinės šaknies, tai $d = 0$. Žvelgiame į A puslapio ACF ir PACF grafikus 10 pav., iš ACF grafiko 10a matoma, kad pradiniai ankstiniai pamažu mažėja, o PACF grafike 10b matomas reikšmingas pirmasis ankstinys, tai indikuoja AR(1) modelį, pasirinkta MA(1) eilė, nes modelis yra mišrus, tad gautas modelis - $ARIMA(1, 0, 1)$, kas yra $ARMA(1, 1)$ modelis. O `auto.arima` funkcijos pasiūlytas tinkamiausias modelis $ARIMA(1, 0, 2)$. Modelių palyginimui naudojamosi Akaike kriterijumi (AIC). Modelis, kurio AIC yra mažesnis - laikomas geresniu, šiuo atveju tai būtų `auto.arima` funkcijos pasiūlytas $ARMA(1, 2)$ modelis.



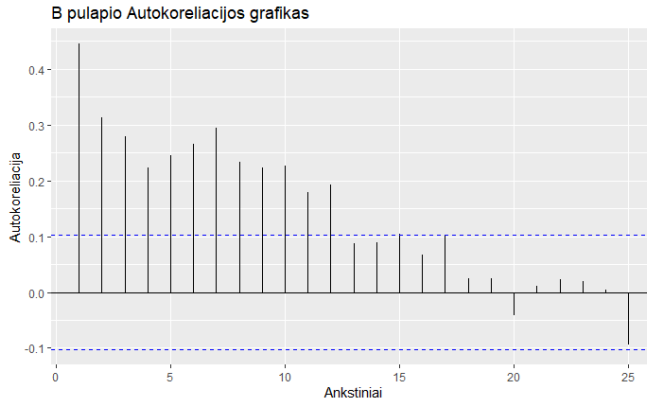
(a) A puslapio ACF grafikas



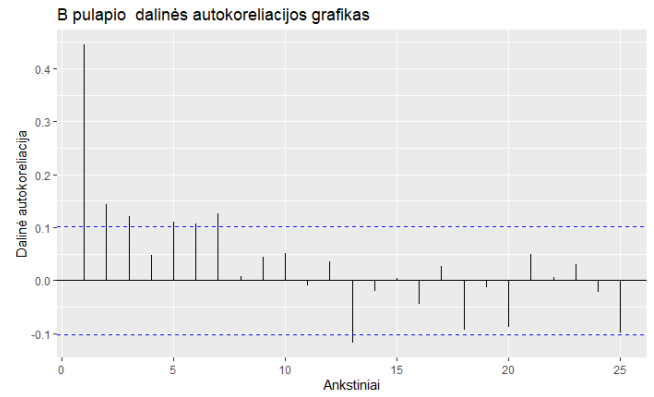
(b) A puslapio ACF grafikas

10 pav.: A puslapio laiko eilučių ACF ir PACF grafikai

Kadangi minėta, kad B eilutei tikrinsime tiek pradinius, tiek diferencijuotus duomenis, toliau žvelgiame į ACF ir PACF grafikus abiem eilutėms 11. matoma, kad ACF 11a pav. daugelis pradinių ankstinių yra reikšmingi, tačiau pamažu mažėja didėjant ankstiniams, o PACF 11b grafike matomas ryškus pirmasis ankstinys, iš esmės grafikai panašūs į A puslapio. Taigi taip pat pasirinktas $ARIMA(1, 0, 1)$ modelis, o tiksliau $ARMA(1, 1)$ modelis, tuo tarpu `auto.arima` funkcijos siūlomas tinkamiausias modelis yra $ARIMA(1, 0, 2)$, t.y. $ARMA(1, 2)$. Palyginus šių modelių AIC reikšmes gauta, kad B puslapio duomenims tinkamiausias modelis yra $ARMA(1, 2)$.



(a) B puslapio ACF grafikas



(b) B puslapio ACF grafikas

11 pav.: B puslapio laiko eilučių ACF ir PACF grafikai

	A puslapio koeficientai	B puslapio koeficientai
AR1	0.885	0.944
MA1	-0.585	-0.621
MA2	-0.116	-0.135

5 lentelė: A ir B puslapių ARMA(1, 2) modelio koeficientų sąrašas.

Abiejų puslapių laiko eilučių ARIMA modelio eilės gavosi tokios pat - ARMA(1, 2). Puslapių modelių koeficientai pateikti 5 lentelėje, kuriuos įsistatę gauname lygtis:

- A puslapio ARMA(1, 2) lygtis :

$$(1 - 0,885B)X_t = (1 - 0,585B - 0,116B^2)Z_t.$$

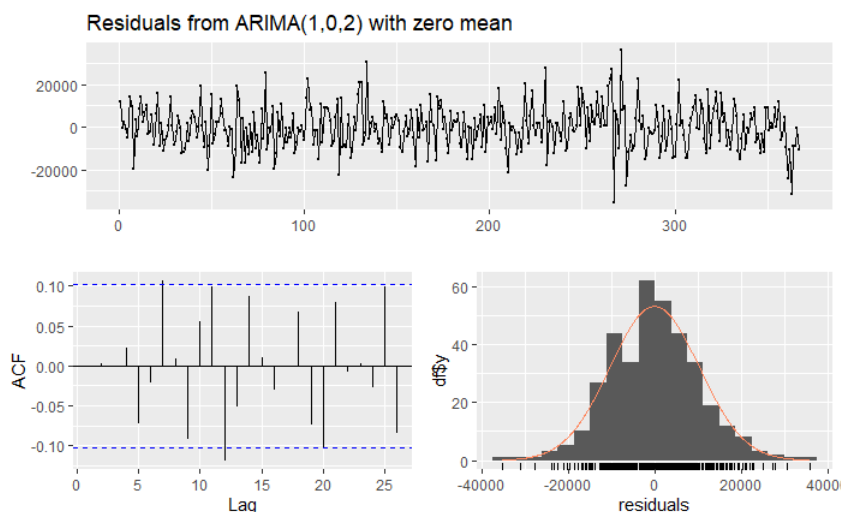
- B puslapio ARMA(1, 2) lygtis :

$$(1 - 0,944B)X_t = (1 - 0,621B - 0,135B^2)Z_t.$$

4.4.4 Modelių liekanų analizė

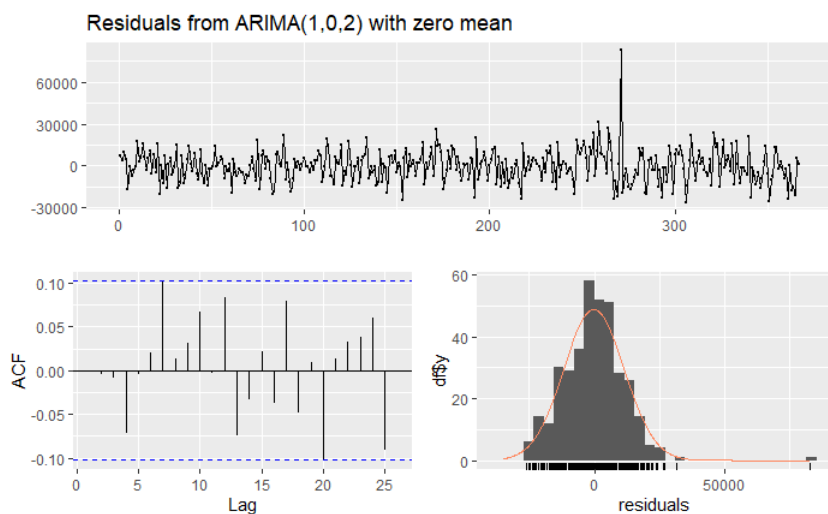
Kaip ir su GLM modeliu, taip ir su ARIMA modeliu yra svarbu atlikti liekanų analizę. Pirmausia žvelgiame į liekanų grafikus: A puslapio 12 pav. matoma, kad duomenys pasiskirstę aplink 0, tačiau sunku pasakyti ar vidurkis tikrai 0, tuo tarpu ACF grafike matomi keli ankstiniai, kurie yra reikšmingi. Iš histogramos matomas normalus liekanų pasiskirstymas. Tiksliam liekanų autokoreliacijos nustatymui toliau naudotas Ljung Box testas. O norint patikrinti ar vidurkis lygus 0 - taikytas Stjudento t- testas, gautos p reikšmės matomos 6 lentelėje. Abiem

atvejais p reikšmės didesnės už $\alpha = 0.05$, tad negalime atmesti H_0 hipotezių - nėra autokoreliuotos ir liekanų vidurkis yra nulinis. Liekanos sudaro Baltąjį triukšmą ir patvirtina prielaidą, kad A puslapiui taikytas ARMA(1,2) modelis yra tinkamas prognozėms.



12 pav.: A puslapio ARMA(1,2) modelio liekanų grafikai.

Panaši situacija matoma ir iš B puslapio modelio liekanų grafikų - 13 pav. Duomenų vidurkis atrodo arti 0. Bet ACF grafike nėra įžvelgiama reikšmingų ankstinių. Iš pasiskirstymo histogramos matyti, kad duomenys normalaus pasiskirstymo. Naudojantis tais pačiais testais, kurių rezultatai pateikti 6 lentelėje, nustatyta, kad B puslapio ARMA(1,2) modelio liekanos yra tinkamos prognozėms - liekanos neautokoreliuotos ir vidurkis lygus 0.



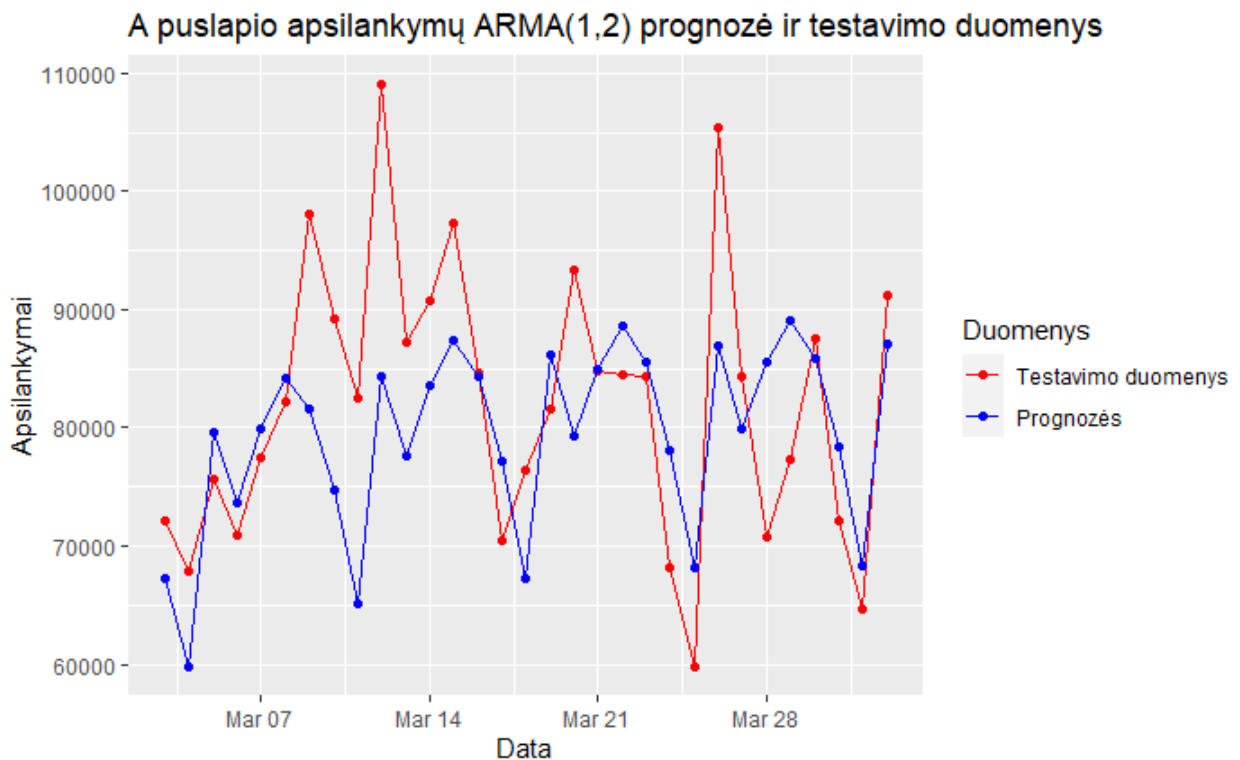
13 pav.: B puslapio ARMA(1,2) modelio liekanų grafikai.

	Ljung Box testas	Stjudento t-testas
A puslapio p reikšmė	0.14	0.837
B puslapio p reikšmė	0.318	0.763

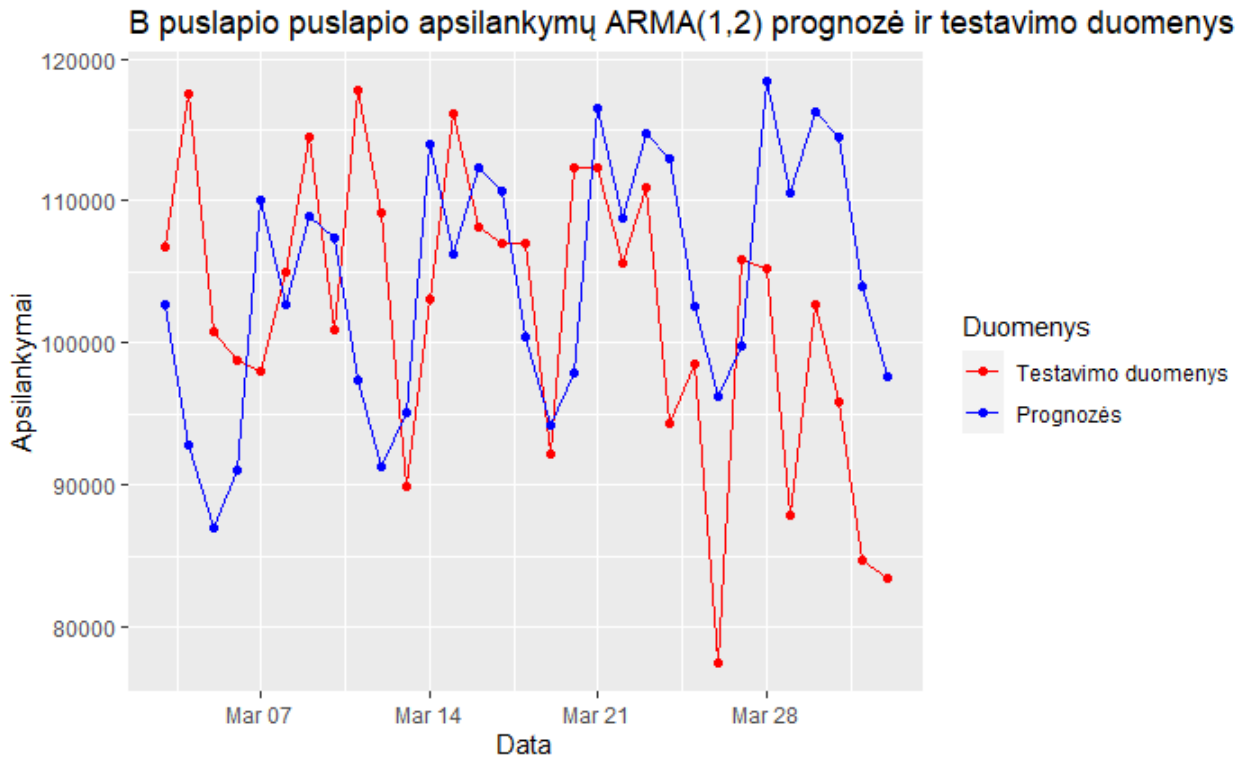
6 lentelė: A ir B puslapių modelio liekanų testų p reikšmės.

4.4.5 Modelių prognozės

Naudojant ARMA(1,2) modelį, atlikta GLM modelio liekanų prognozė mėnesiui į priekį. Kad šias prognozes būtų galima palyginti su realiais (testavimo) duomenimis - ARMA(1,2) modelio prognozės liekanoms sudėtos su GLM modelio prognozėmis. Taigi šio modelio prognozės A puslapio apsilankymams galima matyti 14 grafike. Šiame grafike matomas gan panašus vaizdas kaip GLM modelio prognozėje, prognozės vietomis yra visai šalia realių duomenų, tačiau prognozavimo 2 savaitėje matomas didesnis prognozių atotrūkis nuo realių apsilankymų, mat pradžioje prognoziuojami apsilankymai yra mažesni, o po to matomas augimas, kuris galiausiai pasiekia realias reikšmes. Bendrai galima sakyti, kad prognozuojamos reikšmės pradžioje nesutampa su realiais duomenimis, tačiau vėliau prognozės labiau atspindi realią situaciją.



14 pav.: A puslapio apsilankymų ARMA(1,2) prognozė ir realūs duomenys.



15 pav.: B puslapio apsilankymų ARMA(1,2) prognozė ir realūs duomenys.

B puslapio B puslapio apsilankymų ARMA(1,2) prognozė grafike 15 matoma kiek geresnė situacija, palyginus su to paties puslapio GLM grafiku, čia prognozės yra kiek arčiau realių duomenų, 2 - 3 savaitėmis duomenys labiausiai atitinka. Pabaigoje vis dar matomas didelis atotrūkis. Taigi bendrai prognozė yra arti realybės, tik pabaigoje matomi didesni neatitikimai.

4.5 BATS modelis

Naudojantis tais pačiais GLM modelio liekanų duomenimis, sudarytas BATS modelis. Abiejų puslapių laiko eilučių BATS modelio eilės tokios pat - BATS(1, [0, 1]). Gaunamos A ir B puslapių BATS(1, [0, 1]) lygtys:

- A puslapio lygtis :

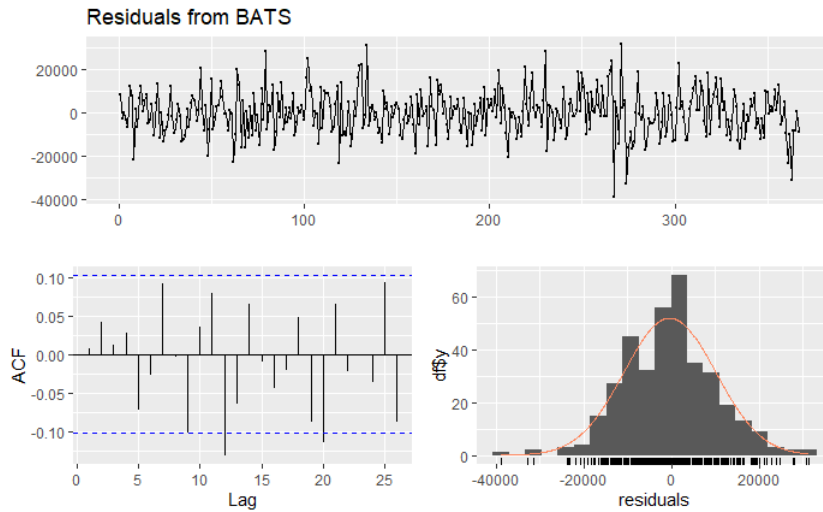
$$\begin{aligned}
 y_t^{(1)} &= y_t - 1, \\
 y_t^{(1)} &= l_{t-1} + d_t, \\
 l_t &= l_{t-1} + 0,116d_t, \\
 d_t &= 0,2\varepsilon_{t-1} + \varepsilon_t.
 \end{aligned}$$

- B puslapio lygtis :

$$\begin{aligned}y_t^{(1)} &= y_t - 1, \\y_t^{(1)} &= l_{t-1} + d_t, \\l_t &= l_{t-1} + 0,158d_t, \\d_t &= 0,182\varepsilon_{t-1} + \varepsilon_t.\end{aligned}$$

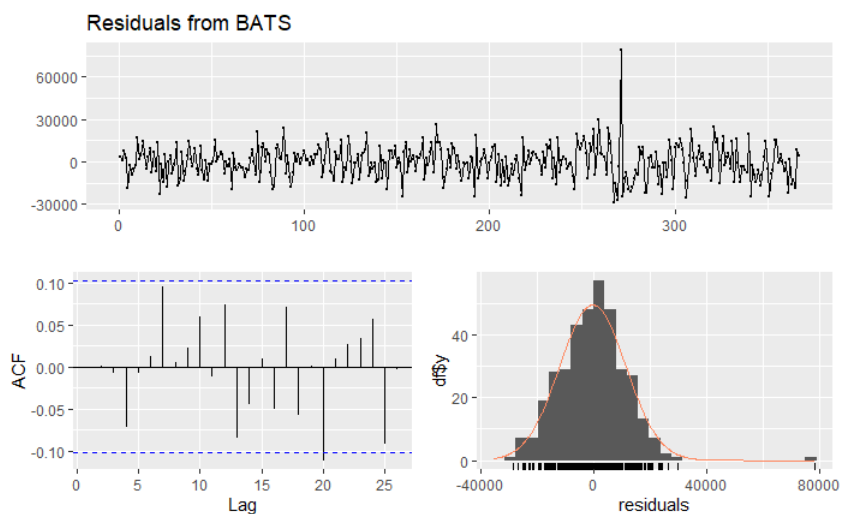
4.5.1 Modelių liekanų analizė

Atliekama BATS modelio liekanų analizė. Pirmiausia žvelgiame į liekanų grafiką - A puslapio 16 pav. matoma, kad duomenys pasiskirstę aplink 0, tačiau yra kelios labiau nutolusios reikšmės, tuo tarpu ACF grafike matomi keli ankstiniai, kurie yra reikšmingi. Tiksliam liekanų autokoreliacijos nustatymui naudotas Ljung Box testas. O vidurkio patikrinimui - taikytas Stjudento t- testas, gautos p reikšmės matomos 7 lentelėje. Abiem atvejais p reikšmės didesnės už $\alpha = 0.05$, tad negalime atmesti H_0 hipotezių - nėra autokoreliuotos ir liekanų vidurkis yra nulinis. A puslapiui taikytas BATS modelis yra tinkamas prognozėms.



16 pav.: A puslapio BATS modelio liekanų grafikai.

Panaši situacija matoma ir iš B puslapio modelio liekanų grafikų - 17 pav. Duomenų vidurkis atrodo arti 0, su matomomis labiau nutolusiomis reikšmėmis. ACF grafike išvelgiamas nutoles reikšmingas ankstinys. Iš pasiskirstymo histogramos atrodo, kad duomenys normalaus pasiskirstymo. Naudojantis tais pačiais testais, kurių rezultatai pateikti 7 lentelėje, nustatyta, kad B puslapio BATS modelio liekanos yra tinkamos prognozėms - liekanos neautokoreliuotos ir vidurkis lygus 0.



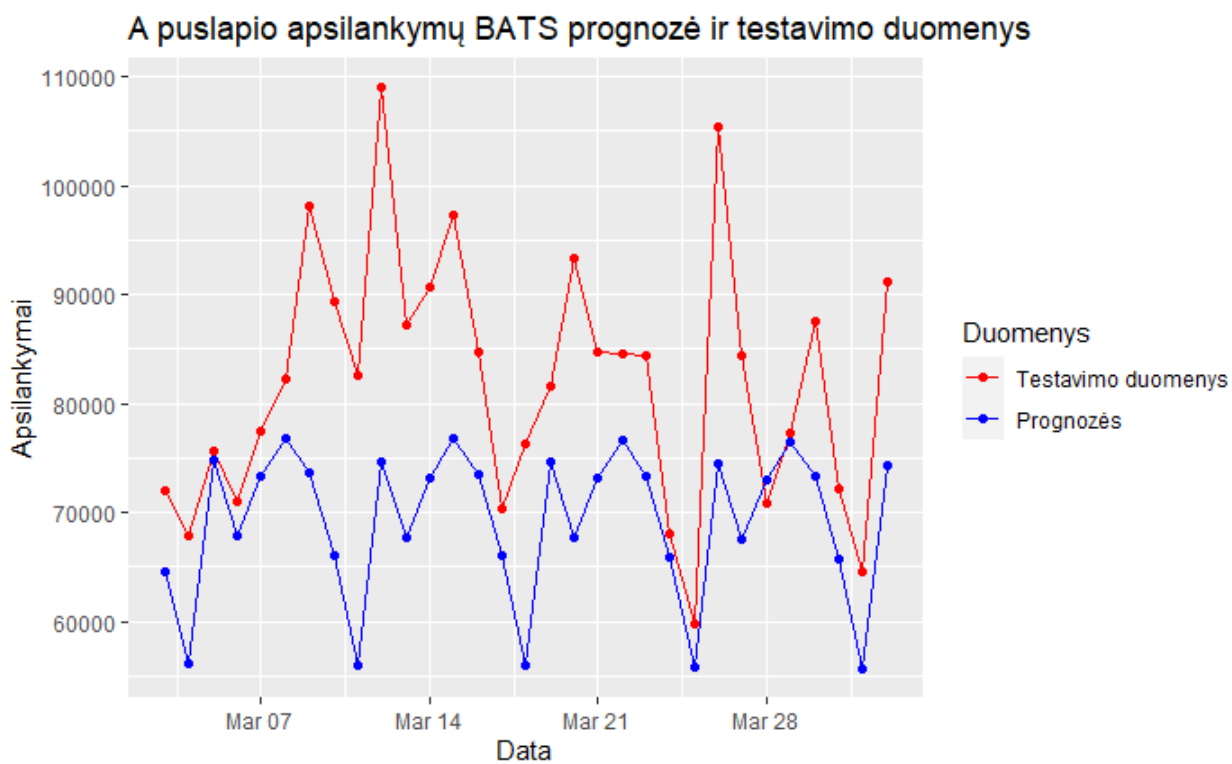
17 pav.: B puslapio BATS modelio liekanų grafikai.

	Ljung Box testas	Stjudento t-testas
A puslapio p reikšmė	0.094	0.521
B puslapio p reikšmė	0.313	0.575

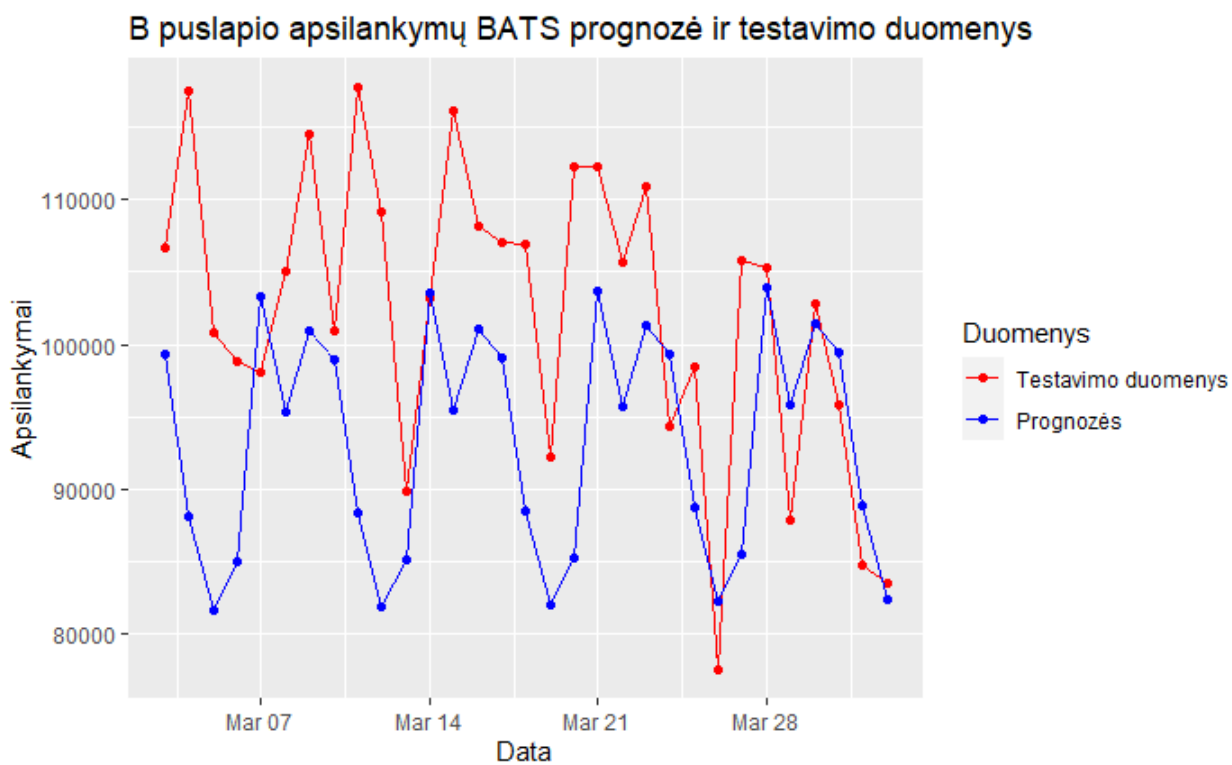
7 lentelė: A ir B puslapių BATS modelio liekanų testų p reikšmės.

4.5.2 Modelių prognozės

Kaip ir ARMA modelio atveju, taip ir čia - sudaryta prognozė GLM modelio liekanoms, tad taip pat buvo pridėtos BATS prognozių reikšmės prie GLM prognozių. Iš A puslapio BATS prognozių grafiko - 18 matoma, kad prognozės yra toliausiai realių reikšmių, lyginant su kitų modelių prognozių grafikais. Realūs apsilankymai yra žymiai didesni, nei prognozuoja BATS.



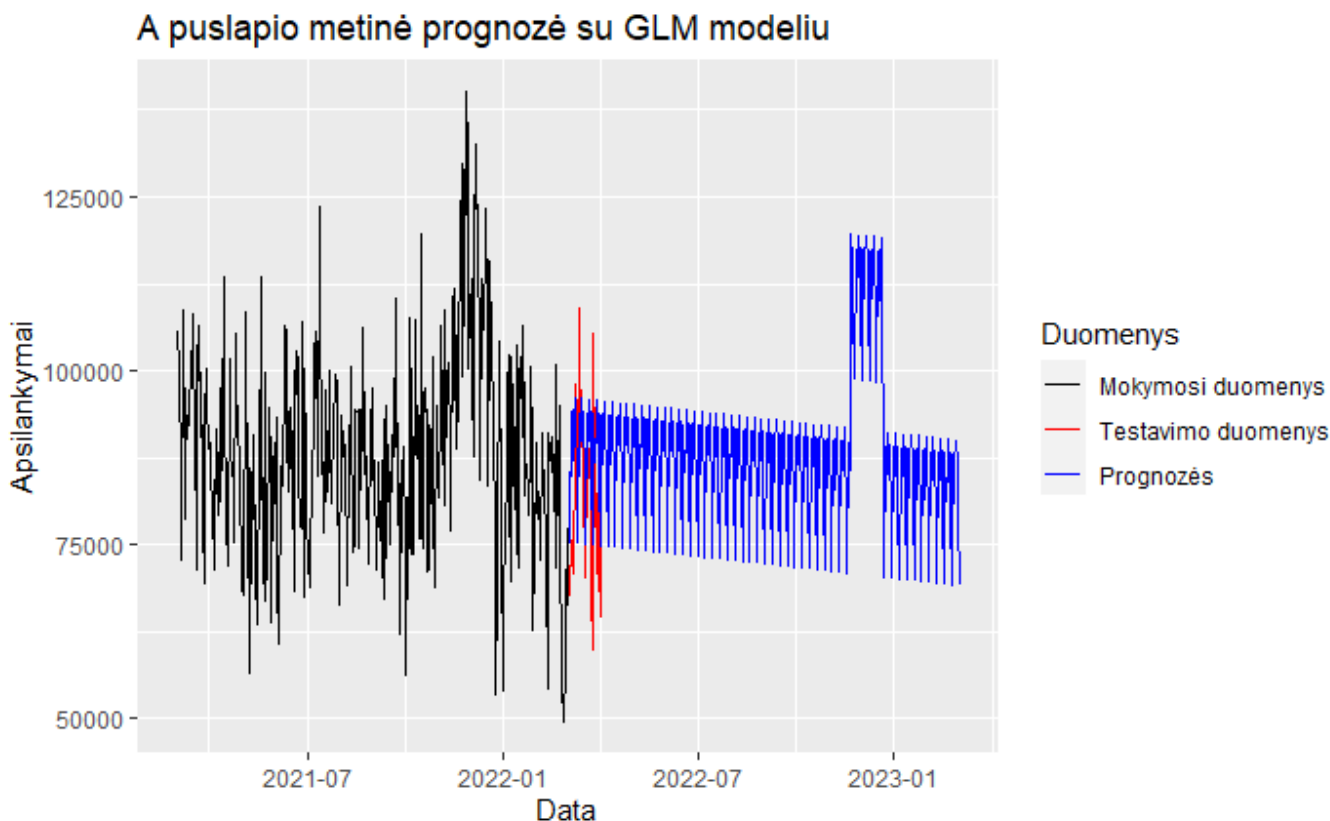
18 pav.: A puslapio apsilankymų BATS prognozė ir realūs duomenys.



19 pav.: B puslapio apsilankymų BATS prognozė ir realūs duomenys.

4.6 Tinkamiausio modelio išrinkimas

Su abiem puslapiais buvo sudarytos prognozės, naudojantis 3 skirtingais modeliais. Pirmasis - tai Apibendrintasis tiesinis modelis, kurio liekanoms buvo sudaryti likę 2 modeliai, tai ARMA ir BATS. Išrinkti geriausią modelį galima naudojantis vidutine kvadrato paklaida - RMSE ir vidutine absoliučia paklaida - MAE. Lentelėje 8 matoma puslapiui A taikytų modelių prognozių RMSE ir MAE. Matome, kad pagal RMSE geriausias modelis yra - Apibendrintasis tiesinis modelis (GLM), o pagal MAE - ARMA(1, 2). Šiuo atveju pasikliaujame RMSE rodikliu ir teigiame, kad A puslapio apsilankymų prognozavimui tinkamiausias modelis yra - Apibendrintasis tiesinis modelis (GLM). Naudojantis šiuo modeliu sudaroma ilgesnio laikotarpio (nuo 2022-03-03 iki 2023-03-03) prognozė, kuri matoma 20 pav. Iš iš turimų mokymosi duomenų (nuo 2021-03-01 iki 2022-03-02) Apibendrintasis tiesinis modelis prognozuoja A puslapio apsilankymų mažėjimą. Kalėdiniu laikotarpiu matomas srauto padidėjimas, po kurio vėl prognozuojamas apsilankymų kritimas.



20 pav.: A puslapio apsilankymų ilgesnio laikotarpio GLM modelio prognozė.

B puslapio apsilankymų duomenims taip pat buvo sudaryti tie patys 3 modeliai: Apibendrintasis tiesinis modelis, ARMA ir BATS. Sudarymo principas taip pat buvo toks pat, kaip A puslapiui. Iš šių modelių RMSE ir MAE rodiklių, kurie pateikti 9 lentelėje, matoma, kad tinkamiausias modelis tiek pagal RMSE, tiek pagal MAE yra - ARMA(1, 2). Naudojantis šiuo

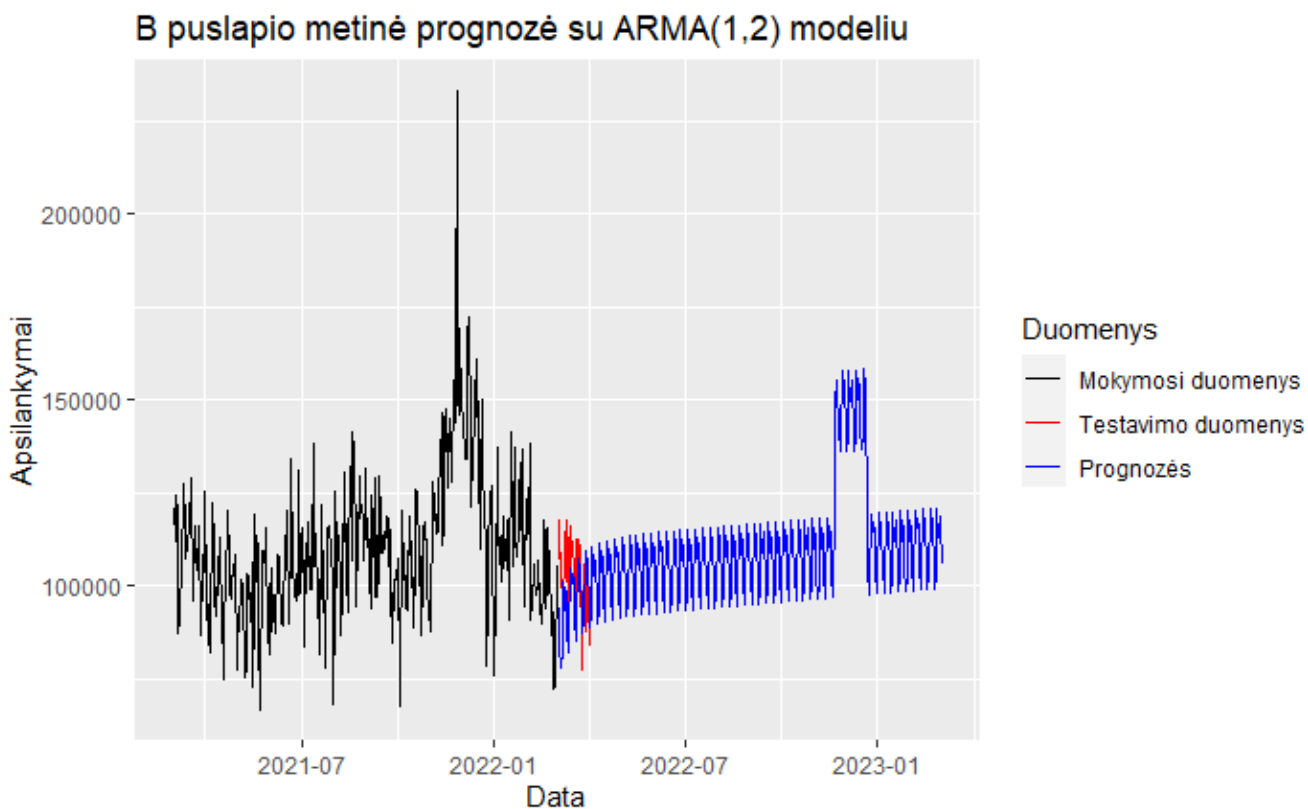
modeliu ir buvo sudaryta ilgesnio laikotarpio prognozė B puslapio apsilankymams - 21 pav. Iš šio grafiko matoma, kad ARMA(1,2) modelis prognozuoja apsilankymų augimą B internetiniame puslapyje. Prognozuojamas ir srauto augimas Kalėdiniu laikotarpiu.

Modelis	RMSE	MAE
GLM	9472.28	8129.61
ARMA(1, 2)	9981.83	7984.92
BATS(1, [0, 1])	15901.26	12941.28

8 lentelė: A puslapio modelių RMSE ir MAE reikšmės.

Modelis	RMSE	MAE
GLM	13847.8	11971.86
ARMA(1, 2)	12639.64	10711.05
BATS(1, [0, 1])	13984.75	11008.84

9 lentelė: B puslapio modelių RMSE ir MAE reikšmės.



21 pav.: B puslapio apsilankymų ilgesnio laikotarpio ARMA(1,2) modelio prognozė.

5 Išvados

Šiame darbe buvo atlikta A ir B internetinių puslapių apsilankymų statistinė analizė. Buvo sudaryta po 3 modelius A ir B internetinių puslapių apsilankymams. Apibendrintojo tiesinio modelio sudarymo metu, buvo pasirinkti regresoriai, kurie atitiko puslapių apsilankymo laiko eilučių sezonines, trendo ir Kalėdinio laikotarpio komponentes. Šio modelio liekanos toliau buvo naudojamos ARIMA ir BATS modelių ir jų prognozių sudarymui. Liekanų prognozės sudėtos su pradinių duomenų GLM prognozėmis ir taip suskaičiuotas prognozes buvo galima palyginti su realiais testavimo duomenimis. Taip pat buvo atlikta visų modelių liekanų analizė. Palyginus visus metodus tarpusavyje, A puslapio apsilankymų duomenims tinkamiausias modelis yra apibendrintasis tiesinis modelis (GLM), o B puslapio apsilankymų duomenims - ARMA(1, 2) modelis.

Galiausiai sudarytos ilgojo laikotarpio prognozės, naudojantis tinkamiausiais modeliais. A puslapiui GLM prognozuoja apsilankymų mažėjimą. Rekomendacijos puslapio savininkams būtų padaryti gilesnę puslapio analizę - galbūt puslapis nėra patogus ar patrauklus vartotojui, patyrinėti srauto atėjimo kanalus, galbūt verta investuoti daugiau resursų į vieną iš kanalų, taip pat pasilyginti su konkurentais ir t.t. Tuo tarpu B puslapiui ARMA(1, 2) modelis ilguoju laikotarpiu prognozuoja apsilankymų augimą, kuris nėra staigus ir puslapis jau yra atlaikęs ir didesnę apsilankymų srautą, tad šio puslapio savininkams rekomenduojama įsivertinti puslapio serverių pajėgumus, ir esant poreikiui įsigyti galingesnius serverius.

Literatūra

- [1] „Similarweb“ Data Methodology: <https://support.similarweb.com/hc/en-us/articles/360001631538#UUID-161d3620-5125-b5f8-2a66-8fc8e9b85a21>
- [2] „A short history of the internet“(2020): <https://www.scienceandmuseum.org.uk/objects-and-stories/short-history-internet>
- [3] M. Roser, H. Ritchie and E. Ortiz-Ospina. Internet, 2015: <https://ourworldindata.org/internet>
- [4] Internet use, 2021: <https://www.itu.int/itu-d/reports/statistics/2021/11/15/internet-use/>
- [5] R. Hyndman, G. Athanasopoulos. Forecasting: Principles and Practice, 2018 <https://otexts.com/fpp2/>
- [6] R.Leipus. Laiko eilutės, 2021.
- [7] V. Čekanavičius, G. Murauskas. Statistika ir jos taikymai III, 2009.
- [8] V. Čekanavičius, G. Murauskas. Taikomoji Regresinė analizė socialiniuose tyrimuose, 2014.
- [9] E. E. Holmes, M. D. Scheuerell, E. J. Ward. Applied time series analysis for fisheries and environmental data, 2021: <https://atsa-es.github.io/atsa-labs/>
- [10] A. M. De Livera, R. J. Hyndman, R. D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing, *Journal of the American Statistical Association*, 2011, 106(496), p. 1513-1527.
- [11] G. M. Ljung, G. E. P. Box. On a measure of lack of fit in time series models, *Biometrika*, 1978, 65, p. 297-303.
- [12] E. Zivot. Unit Root Tests, 2005: <https://faculty.washington.edu/ezivot/econ584/notes/unitroot.pdf>

6 Priedai

6.1 A priedas

A puslapio analizės su RStudio programine įranga kodas

```
#Duomenų įsikėlimas:
A_data <- read_excel("C:/Users/Evelina/Desktop/LT_1-(2021_03_01-2022_04_02) (1).xlsx",sheet =
A_data$Date <- as.Date(A_data$Date, "%Y/%m/%d")
date <- A_data$Date

#Pagrindiniai statistiniai rodikliai:
stat.desc(A_data$Visits)

#Naudojamų rodiklių išrinkimas:
A_data.fr <- data_frame(Date=A_data$Date,Visits=A_data$Visits)

#Naujo kintamojo sukūrimas - dienos nuo 1 iki 398:
A_data.fr$diena <- seq(1,nrow(A_data.fr),1)

#Naujo kintamojo sukūrimas - Kalėdų laikotarpis nuo 2021-11-22 iki 2021-12-21:
A_data.fr$kaledos <- 0
A_data.fr[(A_data.fr$Date>='2021-11-22') &
(A_data.fr$Date<='2021-12-21'),]$kaledos <- 1

#Naujo kintamojo sukūrimas - savaitės diena:
A_data.fr$sav_diena <- weekdays(date)
A_data.fr$sav_diena<-factor(A_data.fr$sav_diena)

#Duomenų pavaizdavimas grafike:
ggplot(data = A_data.fr, aes(x = Date, y = Visits))+
  geom_line(size = 0.25)+
  ggtitle("Elektroninio puslapio A apsilankymų laiko eilutė")+
  xlab("Data")+ ylab("Apsilankymai")+
  scale_x_date(date_labels = "%Y-%m-%d")+
  geom_hline(yintercept = 87883.72, col = "blue")

#Pradinių duomenų suskirstymas į mokymosi (training) ir testavimo (test) duomenis:
A_train.r <- A_data.fr%>% subset(Date <= "2022-03-02")
```

```

A_test.r <- A_data.fr %>% subset(Date >= "2022-03-03")

# Regresija ir liekanų išrinkimas
#GLM
A_fit<-glm(Visits~diena+kaledos+sav_diena,data=A_train.r)
summary(A_fit)

#multikolinearumo tikrinimas
library(performance)
check_collinearity(A_fit)

#R^2 determinacijos koeficientas
with(summary(A_fit), 1 - deviance/null.deviance)
#Hipotezės
lht(A_fit, c("sav_dienaMonday=0","sav_dienaTuesday=0",
            "sav_dienaWednesday =0","sav_dienaThursday =0",
            "sav_dienaSaturday =0","sav_dienaSunday =0"))

A_residuals <- residuals(A_fit)

#Proгноzių skaičiavimas
A_testvec <- as.vector(A_test.r)
A_test.r$Pred <-predict(A_fit, new = A_testvec)

#Tikslumas
A_acc_Reg <- accuracy(A_test.r$Pred, A_test.r$Visits)
A_acc_Reg

#modelio liekanų patikrinimas
checkresiduals(A_fit)
Box.test(A_fit$residuals, type = "Ljung-Box")
t.test(A_fit$residuals, mu = 0)

#GLM prognozės grafikas
ggplot(A_test.r,aes(Date,Visits)) +
  geom_line(data=A_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_point(data=A_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_line(data=A_test.r,aes(Date, Pred, color="Prognozės"))+

```

```

    geom_point(data=A_test.r,aes(Date, Pred, color="Prognozės"))+
    ggtitle("A puslapio apsilankymų prognozė ir testavimo duomenys" )+
    xlab("Data")+ ylab("Apsilankymai")+
    scale_color_manual(name='Duomenys',
                        breaks=c('Testavimo duomenys', 'Prognozės'),
                        values=c('Prognozės'='blue', 'Testavimo duomenys'='red'))

#ARIMA modelis
#liekanų grafiko braižymas
A_ts <- ts(A_residuals)
autoplot(A_ts) +xlab("Dienos") + ylab("Apsilankymai") + ggtitle("A puslapio liekanų grafikas")

#stacionarumo tikrinimas
adf.test(A_residuals, alternative = "stationary")
kpss.test(A_residuals)

#ACF ir PACF grafikai
library(ggfortify)
ggAcf(A_residuals, main="A puslapio Autokoreliacijos grafikas", ylab = "Autokoreliacija", xlab = "Lag")
ggPacf(A_residuals, main="A puslapio dalinės autokoreliacijos grafikas", ylab = "Dalinė autokoreliacija", xlab = "Lag")

#Modelių sudarymai
A_auto.model <- auto.arima(A_residuals,stepwise = F, approximation = F)
A_arima <- Arima(A_residuals, order=c(1,0,1)) #AIC=7832.24
coef(A_auto.model)
summary(A_auto.model)
summary(A_arima)

accuracy(A_auto.model)

#modelio liekanų patikrinimas
checkresiduals(A_auto.model)
t.test(A_auto.model$residuals, mu = 0)

#Prognozės
A_forecast <- forecast(A_auto.model, h=31,)
autoplot(A_forecast)
A_test.r$ARMA.pred <- (A_forecast$mean + A_test.r$Pred)

```

```

library(caret)
A_acc_ARIMA <- accuracy(A_test.r$ARMA.pred, A_test.r$Visits)

#Prognozės braižymas
ggplot(A_test.r,aes(Date,Visits)) +
  geom_line(data=A_test.r, aes(Date, Visits, color="Testavimo duomenys"))+
  geom_point(data=A_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_line(data=A_test.r,aes(Date, ARMA.pred, color="Prognozės"))+
  geom_point(data=A_test.r,aes(Date, ARMA.pred, color="Prognozės"))+
  ggtitle("A puslapio apsilankymų ARMA(1,2) prognozė ir testavimo duomenys" )+
  xlab("Data")+ ylab("Apsilankymai")+
  scale_color_manual(name='Duomenys',
                     breaks=c('Testavimo duomenys', 'Prognozės'),
                     values=c('Prognozės'='blue', 'Testavimo duomenys'='red'))

#BATS modelis
A_bats<-bats(A_residuals)
summary(A_bats)

#liekanų analizė
A_bats_res <-residuals(A_bats)
checkresiduals(A_bats)
t.test(A_bats_res, mu = 0)

#Prognozė
A_bts_fore<- forecast(A_bats, level=c(95), h = 31)

A_test.r$BATS_prog <- (A_bts_fore$mean + A_test.r$Pred)

#ACCURACY
A_acc_BATS <- accuracy(A_test.r$BATS_prog, A_test.r$Visits)

#Prognozės grafikas
ggplot(A_test.r,aes(Date,Visits)) +
  geom_line(data=A_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_point(data=A_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_line(data=A_test.r,aes(Date,BATS_prog , color="Prognozės"))+
  geom_point(data=A_test.r,aes(Date,BATS_prog , color="Prognozės"))+

```

```

    ggtitle("A puslapio apsilankymų BATS prognozė ir testavimo duomenys" )+
    xlab("Data")+ ylab("Apsilankymai")+
    scale_color_manual(name='Duomenys',
                        breaks=c('Testavimo duomenys', 'Prognozės'),
                        values=c('Prognozės'='blue', 'Testavimo duomenys'='red'))

#"GLM tikslumas"
A_acc_Reg

#ARIMA tikslumas
A_acc_ARIMA

#BATS tikslumas
A_acc_BATS

#Metų prognozė
#Kintamųjų x generavimas
data_seq <- seq(as.Date('2022-03-03'), as.Date('2023-03-03'), by = "1 days")

metu.prog <- data.frame(Date = data_seq)
date <- metu.prog$Date

#diena
metu.prog$diena <- seq(1,nrow(metu.prog),1)

#kaledu laikotarpis nuo 2022-11-22 iki 2022-12-21
metu.prog$kaledos <- 0
metu.prog[(metu.prog$Date>='2022-11-22') & (metu.prog$Date<='2022-12-21'),]$kaledos <- 1

# savaites diena
metu.prog$sav_diena <- weekdays(date)
metu.prog$sav_diena<-factor(metu.prog$sav_diena)

#PROGNOZAVIMAS
A_testvec <- as.vector(A_test.r)

metu.prog$Prognoze <-predict(A_fit, new = metu.prog)

```

```
#Prognozės grafikas
ggplot(A_train.r,aes(Date,Visits)) +
  geom_line(data=A_train.r, aes(Date, Visits, color="Mokymosi duomenys")) +
  geom_line(data=A_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_line(data=metu.prog,aes(Date, Prognoze, color="Prognozės"))+
  ggtitle("A puslapio metinė prognozė su GLM modeliu")+
  xlab("Data")+ ylab("Apsilankymai")+
  scale_color_manual(name='Duomenys',
                     breaks=c('Mokymosi duomenys', 'Testavimo duomenys', 'Prognozės'),
                     values=c('Prognozės'='blue', 'Testavimo duomenys'='red', 'Mokymosi duomenys'='green'))
```

6.2 B priedas

B puslapio analizės su RStudio programine įranga kodas

```
#Įmonės B duomenys
B_data <- read_excel("C:/Users/Evelina/Desktop/LT_1-(2021_03_01-2022_04_02) (1).xlsx",sheet = 1)
B_data$Date <- as.Date(B_data$Date, "%Y/%m/%d")
date <- B_data$Date

#Statistikos
stat.desc(B_data$Visits)

B_data.fr <- data_frame(Date=B_data$Date,Visits=B_data$Visits)

#dienos nuo 1 iki 398
B_data.fr$diena <- seq(1,nrow(B_data.fr),1)

#kaleėdų laikotarpis nuo 2021-11-22 iki 2021-12-21
B_data.fr$kaledos <- 0
B_data.fr[(B_data.fr$Date>='2021-11-22') & (B_data.fr$Date<='2021-12-21'),]$kaledos <- 1

#savaitės diena
B_data.fr$sav_diena <- weekdays(date)
B_data.fr$sav_diena<-factor(B_data.fr$sav_diena)

#Atskyrimas į training ir test data.
B_train.r <- B_data.fr%>% subset(Date <= "2022-03-02")
```

```

B_test.r <- B_data.fr %>% subset(Date >= "2022-03-03")

#pradinių duomenų pavaizdavimas
ggplot(data = B_data.fr, aes(x = Date, y = Visits))+
  geom_line(size = 0.25)+
  ggtitle("Elektroninio puslapio B apsilankymai su vidurkiu")+
  xlab("Data")+ ylab("Apsilankymai")+
  scale_x_date(date_labels = "%Y-%m-%d")+
  geom_hline(yintercept = 109174.8, col = "blue")

#Regresijos modelis
B_fit<-glm(Visits~diena+kaledos+sav_diena,data=B_train.r)
summary(B_fit)

#R^2 determinacijos koeficientas
with(summary(B_fit), 1 - deviance/null.deviance)

#multikolinearumo tikrinimas
library(performance)
check_collinearity(B_fit)

library(car)

#Hipotezės
lht(B_fit, c("sav_dienaMonday=0","sav_dienaTuesday=0",
            "sav_dienaWednesday =0","sav_dienaThursday =0",
            "sav_dienaSaturday =0","sav_dienaSunday =0"))

B_residuals <- residuals(B_fit)

#liekanų patikrinimas
checkresiduals(B_fit)
Box.test(B_fit$residuals, type = "Ljung-Box")
t.test(B_fit$residuals, mu = 0)

#Prognozė
B_testvec <- as.vector(B_test.r)
B_test.r$Pred_b <-predict(B_fit, new = B_testvec)

```

```

#Tikslumas
B_acc_Reg <- accuracy(B_test.r$Pred_b, B_test.r$Visits)

#Prognozės grafikas
ggplot(B_test.r,aes(Date,Visits)) +
  geom_line(data=B_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_point(data=B_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_line(data=B_test.r,aes(Date, Pred_b, color="Prognozės"))+
  geom_point(data=B_test.r,aes(Date, Pred_b, color="Prognozės"))+
  ggtitle("B puslapio apsilankymų prognozė ir testavimo duomenys" )+
  xlab("Data")+ ylab("Apsilankymai")+
  scale_color_manual(name='Duomenys',
                     breaks=c('Testavimo duomenys', 'Prognozės'),
                     values=c('Prognozės'='blue', 'Testavimo duomenys'='red'))

#Liekamų grafika
B_ts <- ts(B_residuals)
autoplot(B_ts) +xlab("Dienos") + ylab("Apsilankymai") + ggtitle("B puslapio liekamų grafikas")

#Stacionarumo patikrinimas
adf.test(B_residuals, alternative = "stationary")
kpss.test(B_residuals)

#ACF ir PACF
ggAcf(B_residuals, main="B puslapio Autokoreliacijos grafikas", ylab = "Autokoreliacija", xlab
ggPacf(B_residuals, main="B puslapio dalinės autokoreliacijos grafikas", ylab = "Dalinė autoko

#Modelio parinkimas
B_mod <- Arima(B_residuals, order=c(1,0,1))
B_auto_mod <- auto.arima(B_residuals,stepwise = F, approximation = F)
summary(B_mod)
summary(B_auto_mod)

#Liekamų analizė
checkresiduals(B_auto_mod)
t.test(B_auto_mod$residuals, mu = 0)

```

```

#Prognozės

B_forecast <- forecast(B_auto_mod, h=31)
autoplot(B_forecast)
B_test.r$ARMA.pred <- (B_forecast$mean + B_test.r$Pred_b)

B_acc_ARIMA <- accuracy(B_test.r$ARMA.pred, B_test.r$Visits)

#Prognozės grafikas
ggplot(B_test.r,aes(Date,Visits)) +
  geom_line(data=B_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_point(data=B_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_line(data=B_test.r,aes(Date, ARMA.pred, color="Prognozės"))+
  geom_point(data=B_test.r,aes(Date, ARMA.pred, color="Prognozės"))+
  ggtitle("B puslapio puslapio apsilankymų ARMA(1,2) prognozė ir testavimo duomenys" )+
  xlab("Data")+ ylab("Apsilankymai")+
  scale_color_manual(name='Duomenys',
                     breaks=c('Testavimo duomenys', 'Prognozės'),
                     values=c('Prognozės'='blue', 'Testavimo duomenys'='red'))

#BATS modelis
B_bats<-bats(B_residuals)
B_bats_res <-residuals(B_bats)

#Liekamų analizė
checkresiduals(B_bats)
t.test(B_bats_res, mu = 0)

#Prognozė
B_bts_fore<- forecast(B_bats, level=c(95), h = 31)
B_test.r$BATS_prog <- (B_bts_fore$mean + B_test.r$Pred_b)

#Tikslumas
B_acc_BATS <- accuracy(B_test.r$BATS_prog, B_test.r$Visits)

#Prognozės grafikas
ggplot(B_test.r,aes(Date,Visits)) +
  geom_line(data=B_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_point(data=B_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_line(data=B_test.r,aes(Date,BATS_prog , color="Prognozės"))+

```

```

geom_point(data=B_test.r,aes(Date,BATS_prog , color="Prognozės"))+
  ggtitle("B puslapio apsilankymų BATS prognozė ir testavimo duomenys" )+
  xlab("Data")+ ylab("Apsilankymai")+
  scale_color_manual(name='Duomenys',
                     breaks=c('Testavimo duomenys', 'Prognozės'),
                     values=c('Prognozės'='blue', 'Testavimo duomenys'='red'))

#Prognozių tikslumai
B_acc_Reg
B_acc_ARIMA
B_acc_BATS

#Metu prognoze GLM
#Kintamųjų x generavimas
data_seq <- seq(as.Date('2022-03-03'), as.Date('2023-03-03'), by = "1 days")
B_metu.prog <- data.frame(Date = data_seq)
date <- B_metu.prog$Date

#diena
B_metu.prog$diena <- seq(1,nrow(B_metu.prog),1)

#kalėdų laikotarpis nuo 2022-11-22 iki 2022-12-21
B_metu.prog$kaledos <- 0
B_metu.prog[(B_metu.prog$Date>='2022-11-22') & (B_metu.prog$Date<='2022-12-21'),]$kaledos <- 1

# savaitės diena
B_metu.prog$sav_diena <- weekdays(date)
B_metu.prog$sav_diena<-factor(B_metu.prog$sav_diena)

#PROGNOZAVIMAS
B_metu.prog$Prognoze <-predict(B_fit, new = B_metu.prog)

#Metų prognozė su ARIMA(1,2)
B_metu_forecast <- forecast(B_auto_mod, h=366)

B_metu.prog$ARMA.prog <- (B_metu_forecast$mean + B_metu.prog$Prognoze)
mean(B_metu.prog$ARMA.prog)

```

```

ggplot(B_train.r,aes(Date,Visits)) +
  geom_line(data=B_train.r, aes(Date, Visits, color="Mokymosi duomenys")) +
  geom_line(data=B_test.r, aes(Date, Visits, color="Testavimo duomenys")) +
  geom_line(data=B_metu.prog,aes(Date, ARMA.prog, color="Prognozės"))+
  ggtitle("B puslapio metinė prognozė su ARMA(1,2) modeliu")+
  xlab("Data")+ ylab("Apsilankymai")+
  scale_color_manual(name='Duomenys',
    breaks=c('Mokymosi duomenys', 'Testavimo duomenys', 'Prognozės'),
    values=c('Prognozės'='blue', 'Testavimo duomenys'='red', 'Mokymosi duomenys'='green'))

```