



VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
DUOMENŲ MOKSLO STUDIJŲ PROGRAMA

Baigiamasis bakalauro darbas

**Turistų ir nakvynių skaičiaus Lietuvos apgyvendinimo
įstaigose prognozavimas naudojant „Google Trends“
duomenis**

**Forecasting Number of Tourists and Overnight Stays in
Accommodation Businesses in Lithuania Using „Google Trends“
Data**

Nora Jasiukėnaitė

Darbo vadovas : doc. dr. Rūta Levulienė

Vilnius
2025

Santrauka

Informacija apie turistų skaičių bei jų elgseną ateityje yra svarbus įrankis, padedantis planuoti turizmo sektoriaus plėtrą ir išnaudoti jo potencialą, ypač kalbant apie apgyvendinimo įstaigas. Tyrimai rodo, jog internetinių paieškos užklausų populiarumo duomenų, kaip „Google Trends“, įtraukimas į prognostinius modelius gali pagerinti prognozių tikslumą, tačiau šios metodikos efektyvumas nėra ištirtas Lietuvos kontekste. Pagrindinis šio darbo tikslas yra ištirti, ar „Google Trends“ duomenų įtraukimas į turistų ir nakvynių skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimo modelius gali pagerinti jų tikslumą. Tikslui pasiekti, buvo lyginamas modelių su ir be „Google Trends“ duomenų prognozavimo tikslumas skirtingose apgyvendinimo įstaigų kategorijose ir pagal turistų kilmės šalis. Tyrimo eigoje išbandyti du prognozavimo modeliai: klasikinis laiko eilučių modelis SARIMAX ir mašininio mokymosi modelis – atraminių vektorių regresija. Atlikus tyrimą, nustatyta, jog „Google Trends“ duomenų įtraukimas gali stipriai pagerinti vietinių turistų skaičiaus prognozavimo tikslumą, tačiau pasižymi trūkumais prognozuojant kintamuosius užsienio turistams ir visiems keliautojams. Taip pat padaryta išvada, kad SARIMAX modelis yra pranašesnis tiek turistų, tiek nakvynių skaičiaus prognozavimui.

Raktiniai žodžiai: laiko eilutės, turizmas, mašininis mokymasis, SARIMAX, AVR, Google Trends.

Summary

Information about the number of tourists and their behaviour in the future is a crucial tool in planning the development of the tourism sector and realizing its full potential, especially when talking about accommodation businesses. Studies show that including online search query popularity data, such as „Google Trends“, in forecasting models can increase their accuracy, however, the effectiveness of this methodology has not been researched in the context of Lithuania. The primary aim of this research is to examine whether the inclusion of „Google Trends“ data in models for forecasting the number of tourists and overnight stays in accommodation businesses in Lithuania can increase their accuracy. To achieve this objective, the accuracy of models with and without „Google Trends“ data was compared across different categories of accommodation businesses and by tourists' countries of origin. Two forecasting models were tested during the study: a classical time series model SARIMAX and a machine learning model – support vector regression. The study found that the inclusion of „Google Trends“ data can significantly improve accuracy when forecasting the number of domestic tourists, but has limitations when dealing with foreign tourists and all travellers. Additionally, it was concluded that the SARIMAX model is superior in forecasting both the number of tourists and the number of overnight stays.

Keywords: time series, tourism, machine learning, SARIMAX, SVR, Google Trends.

Turinys

Santrauka	2
Summary	3
Įvadas	6
1. Literatūros apžvalga	8
2. Metodai	9
2.1. Pagrindinių komponentų analizė	9
2.2. Sezoninis ARIMA modelis su egzogeniniais kintamaisiais	9
2.3. Atraminių vektorių regresija	11
2.4. Tikslumo metrikos	12
3. Duomenys	13
3.1. „Google Trends“	13
3.2. Duomenų paruošimas ir pradinė analizė	14
4. Lietuvos turistų skaičiaus prognozavimas	21
4.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai	21
4.2. Atraminių vektorių regresijos modelių rezultatai	27
4.3. Modelių palyginimas	31
5. Visų turistų skaičiaus prognozavimas	32
5.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai	32
5.2. Atraminių vektorių regresijos modelių rezultatai	37
5.3. Modelių palyginimas	41
6. Užsienio turistų skaičiaus prognozavimas	42
6.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai	42
6.2. Atraminių vektorių regresijos modelių rezultatai	47
6.3. Modelių palyginimas	50
7. Nakvynių skaičiaus vietiniams turistams prognozavimas	52
7.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai	52
7.2. Atraminių vektorių regresijos modelių rezultatai	54
7.3. Modelių palyginimas	56
8. Nakvynių skaičiaus visiems turistams prognozavimas	57
8.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai	57
8.2. Atraminių vektorių regresijos modelių rezultatai	59

8.3. Modelių palyginimas	61
9. Nakvynių skaičiaus užsienio turistams prognozavimas	62
9.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai	62
9.2. Atraminių vektorių regresijos modelių rezultatai	63
9.3. Modelių palyginimas	65
10. Rezultatų apibendrinimas	66
Rezultatai ir išvados	68
Literatūra ir šaltiniai	69
1 priedas. Naudota programinė įranga	71
2 priedas. Nakvynių skaičiaus prognozių grafikai	72
3 priedas. Programinis kodas	86

Ivadas

Turizmo sektorius turi reikšmingą poveikį šalies ekonomikai, užimtumui ir regionų plėtrai. Šio sektoriaus sėkmė tiesiogiai veikia ir daugelį kitų sričių – keliautojai atneša pelną viešojo maitinimo, transporto ir prekybos įmonėms, o turistai iš užsienio valstybių gali būti vertingi telekomunikacijos ir ryšių bei finansinių paslaugų įstaigoms. Nors turizmo sektorius susideda iš daugelio varomųjų jėgų, apgyvendinimo įstaigos yra vienos svarbiausių, nes būtent jų prieinamumas gali lemti keliautojų sprendimus. Siekiant toliau plėsti šį sektorių ir išnaudoti visą jo potencialą, svarbu užtikrinti efektyvų ir tikslingą apgyvendinimo įstaigų darbą, o turistų skaičiaus bei jų elgsenos prognozavimas gali būti svarbus įrankis šiam tikslui. Informacija apie turistų srautus ateityje gali būti itin vertinga apgyvendinimo įstaigoms kuriant optimalias verslo plėtros strategijas, gerinant paslaugų kokybę ir prieinamumą bei užtikrinant efektyvų darbuotojų skaičių ir racionalią paslaugų kainą. Taip pat turistų skaičiaus prognozavimas gali padėti rengiant turizmo skatinimo programas, identifikuojant sektoriaus trūkumus ir galimybes.

Nors dažniausiai turistų srautų prognozavimui naudojami istoriniai keliautojų skaičiaus duomenys arba socioekonominiai rodikliai, pastaraisiais metais šioje srityje populiarėja nauja metodika – internetinės paieškos sistemų užklausų populiarumo duomenų, kaip „Google Trends“, įtraukimas į prognostinius modelius. Kelionių planavimas vis dažniau atliekamas internetinėje erdvėje, tad internetinių paieškų apimties duomenys gali būti vertingas įrankis dėl jų galimybės identifikuoti turizmo paklausos pokyčius bei keliautojų ketinimus. Vis dėlto, šios metodikos rezultatai gali stipriai priklausyti nuo tiriamos valstybės ypatumų ir tyrimui pasirinktų paieškos užklausų. Nors įvairiose šalyse atlikti tyrimai rodo, kad tokių duomenų įtraukimas gali reikšmingai pagerinti prognozių tikslumą, nėra ištirta, kaip ši metodika veikia Lietuvos kontekste, ypač naudojant paieškos užklausas lietuvių kalba.

Šio darbo tikslas – ištirti, ar „Google Trends“ duomenų įtraukimas į turistų ir nakvynių skaičių Lietuvos apgyvendinimo įstaigose prognozavimo modelius pagerina jų tikslumą. Tikslui įgyvendinti, buvo išsikelti šie uždaviniai:

1. atlikti literatūros šaltinių analizę ir identifikuoti tyrimui tinkamus metodus;
2. išrinkti paieškos užklausas „Google Trends“ duomenims;
3. atlikti pradinę duomenų analizę ir paruošti duomenis modelių parinkimui;
4. sudaryti modelius turistų ir nakvynių skaičiaus prognozavimui su ir be „Google Trends“ duomenų ir palyginti jų tikslumą;
5. tarpusavyje palyginti laiko eilučių ir mašininio mokymosi modelius, išrenkant tinkamiausią šio tyrimo kontekste.

Darbą sudaro dešimt pagrindinių skyrių. Pirmame skyriuje pateikiama su atliekamu tyrimu susijusios literatūros apžvalga. Antrame skyriuje pristatomi naudojami modeliai ir metodai. Trečias skyrius skirtas duomenų, naudojamų tyrime, pristatymui, pradinės analizės rezultatų bei duomenų paruošimo modeliams aptarimui. 4–9 skyriuose pateikiami modeliavimo rezultatai prognozuojant

turistų ir nakvynių skaičių skirtingose apgyvendinimo įstaigose. 10 skyriuje aptariami apibendrinti tyrimo rezultatai, o 11 skyrius skirtas darbo rezultatų ir įžvalgų aptarimui.

Šio darbo tema ir metodais paremtas straipsnis buvo išleistas 2025 metų konferencijos „Lietuvos magistrantų informatikos ir IT tyrimai“ darbų rinkinyje [12]. Tyrimo metodika ir rezultatai pristatyti žodiniu pranešimu šioje konferencijoje.

1. Literatūros apžvalga

Paklausos ateityje prognozavimas yra itin svarbus turizmo sektoriui priklausantiems verslams, ypač kalbant apie rizikos vertinimą [5]. Apgyvendinimo įstaigų atveju, nerezervuoti kambariai, darbuotojų trūkumas paklausiausiais laikotarpiais ar netikslingai vykdoma verslo plėtra tiesiogiai veikia šių verslų pelną [5]. Turistų srautų prognozavimas padeda kurti vertingas strategijas ir kitoms turizmo sektoriaus šakoms – transporto, pramogų ir poilsio ar maitinimo paslaugų teikėjams, kurių sėkmė taip pat tiesiogiai veikia apgyvendinimo įstaigų klestėjimą [5]. Dėl šių priežasčių, turistų skaičiaus prognozavimo metodai ir modeliai yra viena svarbiausių temų turizmo sektoriaus tyrimuose [24].

Ieškant naujų būdų, galinčių pagerinti prognozuojamų turistų srautų tikslumą, vis daugėja tyrimų, naudojančių internetinius duomenis [13]. Didžioji dalis tokių tyrimų į modelius įtraukia internetinės paieškos sistemų, kaip „Google“ ar „Baidu“, duomenis [13]. Šios metodikos pradininkais laikomi Pan, Chenguang ir Song, kurių tyrimo rezultatai parodė, jog „Google“ paieškos sistemos duomenys gali pagerinti Čarlstone, JAV esančių viešbučių užimtumo prognozavimo tikslumą naudojant ARMA modelius [19]. Nuo to laiko, tyrimai įrodė, kad internetinės paieškos sistemų duomenų įtraukimas į modelius sumažina paklaidas prognozuojant turistų srautus Vokietijoje, Švedijoje, Pietų Korėjoje, Kinijoje bei kitose valstybėse [2, 10, 11, 20, 26].

Turistų srautų prognozavimui naudojamas platus spektras modelių. Visgi, klasikiniai laiko eilučių modeliai, ypač paremti ARIMA, šioje srityje yra naudojami dažniausiai [24, 25]. Tačiau ARIMA modelis yra apribotas to, jog jis nėra pritaikytas nestacionariems (t. y. sezoninę arba tendencijos komponentę turintiems) duomenims – taikant šį modelį, laiko eilutė dažnai yra paverčiama į stacionarią, taip panaikinant sezoninę komponentę, kuri gali suteikti daug esminės informacijos, ypač prognozuojant turistų srautus [6]. Todėl sezoninis ARIMA (toliau SARIMA) modelis šioje srityje yra pranašesnis dėl jo tinkamumo sezoninę komponentę turintiems duomenims [6, 25]. Prognozuojant turistų srautus naudojant internetinės paieškos sistemų duomenis, SARIMA modelis yra išplečiamas įtraukiant egzogeninius kintamuosius [20].

Turistų skaičiaus prognozavimo uždavinys sprendžiamas ir naudojant mašininio mokymosi ir dirbtinio intelekto modelius. Šioje srityje dirbtinių neuroninių tinklų bei ansamblių modeliai pasižymi aukštu tikslumu [25], tačiau jų apmokymui reikalinga didelė duomenų aibė, tad jie yra tinkamesni aukšto dažnio (valandinių, dieninių) turistų srautų prognozavimui [21]. Tuo tarpu atraminių vektorių regresijos (AVR) modeliai mėnesinių laiko eilučių prognozavimo naudojant mašininį mokymąsi kontekste yra pranašesni dėl jų tinkamumo mažoms duomenų aibėms ir mažesnės persimokymo rizikos [8]. Šie modeliai gali būti pranašesni ir už klasikinius laiko eilučių modelius [1, 3, 22], nes nėra keliamos prielaidos duomenų skirstiniui ar stacionarumui [26]. Be to, AVR modeliai geriau identifikuoja priklausomojo kintamojo ir kovariančių sąryšius, todėl gali pasiekti geresnius prognozių rezultatus [3].

2. Metodai

2.1. Pagrindinių komponentių analizė

Pagrindinių komponentių analizė (toliau PKA) yra klasikinis statistinis metodas, naudojamas daugiamatinių duomenų dimensijos mažinimui [4]. Taikant šį metodą, didelis skaičius galimai koreliuotų kintamųjų yra transformuojamas į mažesnį skaičių nekoreliuotų kintamųjų, vadinamų pagrindinėmis komponentėmis.

Tarkime, turime duomenų matricą

$$X = \{X_1, X_2, \dots, X_m\} = \{x_{ij}, i = 1, \dots, m, j = 1, \dots, n\},$$

kur x_{ij} yra j -tojo kintamojo i -toji reikšmė. Kintamųjų reikšmėms pritaikome centravimą:

$$x'_{ij} = x_{ij} - \bar{x}_j$$

ir standartizavimą:

$$z_{ij} = \frac{x'_{ij}}{s_j},$$

čia $\bar{x}_j$ – j -tojo kintamojo aritmetinis vidurkis, s_j – j -tojo kintamojo standartinis nuokrypis. Suformuojame kovariacinę matricą $C = \{c_{kl}, k, l = 1, \dots, n\}$ su elementais:

$$c_{kl} = \frac{1}{m-1} \sum_{i=1}^m (z_{ik} - \bar{z}_k)(z_{il} - \bar{z}_l) = \frac{1}{m-1} \sum_{i=1}^m z_{ik} z_{il}. \quad (1)$$

Kadangi duomenys buvo standartizuoti ir $s_j = 1$, kovariacijos koeficiento reikšmės atitinka koreliacijos koeficiento reikšmes:

$$r_{kl} = \frac{c_{kl}}{s_k s_l} = \frac{c_{kl}}{1} = c_{kl}.$$

Tuomet apskaičiuojame kovariacinės matricos tikrinius vektorius E_k ir tikrines reikšmes λ_k , naudojant formulę

$$C E_k = \lambda_k E_k, \quad (2)$$

čia E_k yra vektorius stulpelis, C – kovariacinė matrica, λ_k – tikrinė reikšmė, randama iš charakteristinės lygties $\det(C - \lambda_k I) = 0$, kur I yra vienetinė matrica, kurios matmenys atitinka kovariacinės matricos C . Tikriniai vektoriai E_k yra surūšiuojami juos atitinkančių tikrinių reikšmių mažėjimo tvarka ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$). Gauta tikrinių vektorių matrica $A = \{E_1, E_2, \dots, E_n\}$ yra vadinama pagrindinių komponentių matrica.

2.2. Sezoninis ARIMA modelis su egzogeniniais kintamaisiais

Autoregresiniai integruoti slenkančio vidurkio modeliai (toliau ARIMA) yra vieni paprasčiausių, tačiau efektyviausių, ypač taikant mažoms duomenų aibėms, statistinių laiko eilučių modelių [15].

Tačiau šie modeliai yra pritaikyti tik vienmatėms laiko eilutėms, neturinčios sezoninės komponentės. Siekiant išspręsti šiuose trūkumus, ARIMA modelis buvo išplėstas į sezoninį ARIMA modelį su egzogeniniais kintamaisiais (toliau SARIMAX), suteikiantį galimybę apdoroti sezoniškumu pasižyminčias laiko eilutes ir įtraukti papildomas kovariantes. SARIMAX modelis gali būti išreikštas tiesinės regresijos formule [16]:

$$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \dots + \beta_k X_{k,t} + \omega_t, \quad (3)$$

čia y_t – priklausomojo kintamojo reikšmė laiko momentu t , β_0 – konstanta, $\beta_1, \dots, \beta_k$ – koeficientai egzogeniniams kintamiesiems, $X_{1,t}, \dots, X_{k,t}$ – egzogeninių kintamųjų reikšmės laiko momentu t . ω_t šioje formulėje yra liekanos, modeliuojamos kaip SARIMA(p, d, q)(P, D, Q) $_S$ procesas:

$$\omega_t = \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t, \quad (4)$$

kur ϕ_p ir Φ_P yra p eilės sezoninio ir P eilės nesezoninio autoregresinio nario svoriai, θ_q ir Θ_Q – q eilės sezoninio ir Q eilės nesezoninio slenkančio vidurkio nario svoriai, d ir D – sezoninio ir nesezoninio diferencijavimo eilė, B – atitolinimo (lagų) operatorius (angl. *backshift operator*), laiko eilutę atitolinantis per tam tikrą laikotarpį, pvz. $B^i Y_t = Y_{t-i}$. S nurodo sezono trukmę (mėnesinių duomenų atveju jo reikšmė yra 12), ε_t – liekanos, atitinkančios baltąjį triukšmą.

Sudarant SARIMAX modelį, svarbu išrinkti tinkamus parametrus p, P , atitinkančius sezoninės ir nesezoninės autoregresijos eilę, d, D , nurodančius sezoninio ir nesezoninio diferencijavimo eilę bei q ir Q , nurodančius sezoninio ir nesezoninio slenkančio vidurkio eilę. Parametrai renkami minimizuojant Akaikės informacinio kriterijaus (AIC) reikšmę.

Sudarant modelį, svarbu patikrinti prielaidas – gautos liekanos turi būti stacionarios (atitikti atsitiktinį klaidžiojimą), nepasižymėti autokoreliacija (atitikti baltąjį triukšmą). Šios prielaidos tikrinamos naudojant statistinius testus.

Pirmosios prielaidos tikrinimui naudojamas išplėstasis Dikio ir Fulerio testas (ADF), kurio hipotezės gali būti apibrėžtos taip:

$$\begin{cases} H_0 : & \text{laiko eilutė turi vienetinę šaknį;} \\ H_1 : & \text{laiko eilutė neturi vienetinės šaknies.} \end{cases}$$

ADF testo statistika skaičiuojama pagal formulę:

$$ADF = \frac{\hat{\rho}}{SE(\hat{\rho})}, \quad (5)$$

čia $\hat{\rho}$ yra koeficiento įvertis iš regresijos modelio

$$\Delta y_t = \alpha + \beta t + \rho y_{t-1} + \sum_{i=1}^p \delta \Delta y_{t-1} + \varepsilon_t.$$

Ljungo ir Bokso testo, tikrinančio antrąją prielaidą – liekanų autokoreliaciją – hipotezės apibrėžiamos taip:

$$\begin{cases} H_0 : \rho_1 = \rho_2 = \dots = \rho_m = 0; \\ H_1 : \rho_i \neq 0 \text{ bent vienam } i \in \{1, \dots, m\}. \end{cases}$$

Testo statistika skaičiuojama pagal formulę:

$$Q_{LB} = N(N+2) \sum_{i=1}^m \frac{\hat{\rho}_i^2}{N-i}, \quad (6)$$

čia N – stebėjimų skaičius, m – lagų skaičius, $\hat{\rho}_i$ – autokoreliacija i -tajame lage. Taikant šį testą, svarbu pasirinkti tinkamą lagų skaičių. Nors nėra tiksliai apibrėžto optimalaus lagų skaičiaus, naudojant $m = \ln(N)$ testas pasižymi aukšta galia [9].

Papildomai, liekanoms gali būti taikomas vienos imties t-testas, kurio hipotezės apibrėžiamos taip:

$$\begin{cases} H_0 : \mu = 0; \\ H_1 : \mu \neq 0, \end{cases}$$

t. y., kai nulinė hipotezė atmetama, liekanų vidurkis nėra lygus 0, ir modelio prognozės yra paslinktos. T-testo statistika skaičiuojama pagal formulę:

$$t = \frac{\bar{x}}{s/\sqrt{n}}, \quad (7)$$

čia $\bar{x}$ – imties aritmetinis vidurkis, s – imties standartinis nuokrypis, n – stebėjimų skaičius imtyje.

2.3. Atraminių vektorių regresija

Atraminių vektorių regresija (toliau AVR) yra prižiūrimo mašininio mokymosi algoritmas, kilęs išplėtus klasifikavimui skirtas atraminių vektorių mašinas (toliau AVM) regresijos uždaviniais spręsti. AVR tikslas – mokymo duomenų aibei $(x_1, y_1), \dots, (x_l, y_l) \subset \mathcal{X} = \mathbb{R}^d$ rasti tokią funkciją $f(x)$, kad jos paklaidos visų tikrųjų reikšmių y_l atžvilgiu būtų ne didesnės kaip ε , ir funkcija būtų kuo plokštesnė [23]. Šis uždavinys gali būti suformuluotas kaip optimizavimo uždavinys, t. y. siekiama minimizuoti

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} (\xi_i + \xi_i^*) \quad (8)$$

su apribojimais

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i, \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0. \end{cases}$$

Čia w yra optimizuojamas svorių vektorius, C – baudos parametras, nustatantis santykį tarp modelio sudėtingumo ir leistinų paklaidų, ξ_i, ξ_i^* – laisvi kintamieji, sumažinantys paklaidų apribojimus, y_i – atsakas, b – poslinkis, ε – leidžiama paklaida.

Tačiau naudojant tokią formuluotę, algoritmas yra apribotas tiesinėmis funkcijomis. Siekiant

išplėsti algoritmo panaudojimą ir atvejais, kai optimalus sprendinys nėra tiesinis, įtraukiami Lagranžo daugikliai bei naudojamos branduolio funkcijos. Tuomet optimizavimo uždavinys gali būti išreikštas taip:

maksimizuojame

$$\begin{cases} -\frac{1}{2} \sum_{i,j=1}^{\ell} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)k(x_i, x_j), \\ -\varepsilon \sum_{i=1}^{\ell} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{\ell} y_i(\alpha_i - \alpha_i^*) \end{cases} \quad (9)$$

taip, kad

$$\sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) = 0, \quad \alpha_i, \alpha_i^* \in [0, C].$$

Čia α_i, α_i^* – Lagranžo daugikliai, $k(x_i, x_j)$ – branduolio funkcija.

Branduolio funkcijos atvaizduoja įvesties erdvę į aukštesnės dimensijos požymių erdvę, kurioje yra taikomas AVR algoritmas [17, 23]. Taip apdorojant požymius, AVR gali būti taikoma netiesiniams uždaviniams. Darbe pasirinkta naudoti radialinę bazinę branduolio funkcija (dar vadinamą Gauso branduoliu), galinčią apdoroti tiek tiesinius, tiek netiesinius sąryšius. Gauso branduolio funkcija išreiškiama:

$$k(x_i, x_j) = \exp \left(-\frac{(x_i - x_j)^2}{\sigma^2} \right), \quad (10)$$

čia $\sigma > 0$ – branduolio plotis.

2.4. Tikslumo metrikos

Siekiant įvertinti prognozavimo modelių tikslumą ir palyginti juos tarpusavyje, skaičiuojamos prognozių tikslumo metrikos. Pasirinktos trys metrikos: vidutinė absoliutinė procentinė paklaida (angl. *mean absolute percentage error*) (toliau MAPE), vidutinė absoliutinė paklaida (angl. *mean absolute error*) (toliau MAE) ir šaknis iš vidutinės kvadratinės paklaidos (angl. *root mean square error*) (toliau RMSE). Metrikos skaičiuojamos naudojant formules:

$$\begin{aligned} MAPE &= \frac{1}{N} \sum_{t=1}^N \left| \frac{\hat{Y}_t - Y_t}{Y_t} \right|, \\ MAE &= \frac{1}{N} \sum_{t=1}^N |\hat{Y}_t - Y_t|, \\ RMSE &= \sqrt{\frac{1}{N} \sum_{t=1}^N (\hat{Y}_t - Y_t)^2}. \end{aligned}$$

3. Duomenys

Tyrimė naudojami duomenys – Valstybės duomenų agentūros administruojamo Oficialiosios statistikos portalo (toliau OSP) viešai prieinami mėnesiniai duomenys apie turistų ir nakvynių skaičių Lietuvos apgyvendinimo įstaigose 2012-2024 metais [27, 28]. Viso turimoje duomenų aibėje yra 156 stebėjimai. OSP pateikiami skaičiai yra išskirstyti pagal apgyvendinimo įstaigos savivaldybę, apgyvendinimo įstaigos tipą ir valstybę, iš kurios atvyko turistai. Tyrimė prognozuojami tik bendri turistų ir nakvynių skaičiai visai Lietuvos teritorijai, tad naudojami agreguoti visų savivaldybių duomenys.

Apgyvendinimo įstaigos išskirstytos į tris kategorijas:

- „Viešbučiai ir kt.“, susidedantis iš viešbučių, motelių ir svečių namų;
- „Stovyklavietės ir kt.“, susidedantis iš stovyklaviečių, poilsio namų ir namelių, kempingų ir t. t.;
- „Kitas apgyvendinimas“, susidedantis iš visų trumpojo apgyvendinimo paslaugų teikėjų, nepatenkančių į pastarąsias dvi kategorijas (poilsio namai, nakvynės namai, privatus apgyvendinimo sektorius, kaimo turizmo sodybos ir pan.).

Taip pat pateikiamas bendras skaičius visose apgyvendinimo įstaigose („Visos įstaigos“).

Turistų ir nakvynių skaičiai pateikiami kiekvienai valstybei, iš kurios keliautojas yra atvykęs, atskirai. Tyrimė naudojami agreguoti duomenys, paliekant tris kategorijas: „Lietuva“, „Užsienio šalys“ ir „Iš viso“.

3.1. „Google Trends“

„Google Trends“ – tai duomenys apie paieškos užklausų populiarumą „Google“ internetinės paieškos sistemoje. Užklausos populiarumo indeksas atitinka santykinį užklausos populiarumą pagal bendrą paieškų geografinėje vietovėje skaičių ir atitinkamą laikotarpį [7]. Indekso reikšmės pateikiamos intervale $[0; 100]$ – indekso reikšmės lygios 100 mėnesiais, kuriais dominanti užklausa buvo naudojama plačiausiai nustatytu laikotarpiu pasirinktoje geografinėje vietovėje [7]. Kadangi užklausos šaltinį galima pasirinkti tik iš sąrašo, kuriame pateiktos specifinės geografinės lokacijos (šalys, salos ir salynai) bei visas pasaulis bendrai, „Google Trends“ duomenys naudojami tik keliautojų ir nakvynių skaičiui turistams iš Lietuvos ir viso pasaulio prognozavimo modeliuose.

Paieškų, atliktų Lietuvoje, užklausoms naudota lietuvių kalba, ir jos buvo išrinktos pagal tris kriterijus:

- užklausos, atspindinčios tiriamų apgyvendinimo įstaigų kategorijas: „apartamentai“, „kempingas“, „motelis“, „nameliai“, „sanatorija“, „sodyba“, „stovykla“, „stovyklavietė“, „viešbutis“;
- užklausos, susijusios su kelionėmis bei apgyvendinimu bendrai: „apgyvendinimas“, „kelionė“, „lankytinos vietos“, „nuoma“;
- bendros užklausos apie Lietuvą: „Lietuva“, „Vilnius“, „Kaunas“.

Paieškų iš viso pasaulio užklausoms naudojama anglų kalba. Pirmoji bandyta strategija užklausių parinkimui buvo naudoti tas pačias frazes, kaip užklausoms lietuvių kalba, pridėdant Lietuvos kontekstą (pvz. „apartments in Lithuania“, „motels in Lithuania“, „accommodation in Lithuania“). Tačiau dėl užklausių specifiškumo susidurta su duomenų trūkumo problema – didžioji dalis užklausių nėra pakankamai populiarios pasaulio mastu, kad pasiektų populiarumo indeksą, aukštesnį nei 0. Kelios išbandytos užklausos („homestead in Lithuania“ (t. y. „sodyba Lietuvoje“), „motel in Lithuania“) turėjo tiek mažai duomenų, jog jie net nebuvo prieinami. Dėl šių priežasčių pasirinkta naudoti kitą strategiją – pasaulio masto užklausoms naudojamos tik frazės, bendrai susijusios su kelionėmis Lietuvoje ir į Lietuvą, ir turinčios tenkintiną stebėjimų skaičių: „camping in Lithuania“, „flights to Lithuania“, „hotels Lithuania“, „Kaunas“, „Lithuania“, „rent in Lithuania“, „things to do in Lithuania“, „travel in Lithuania“, „Vilnius“, „visit Lithuania“.

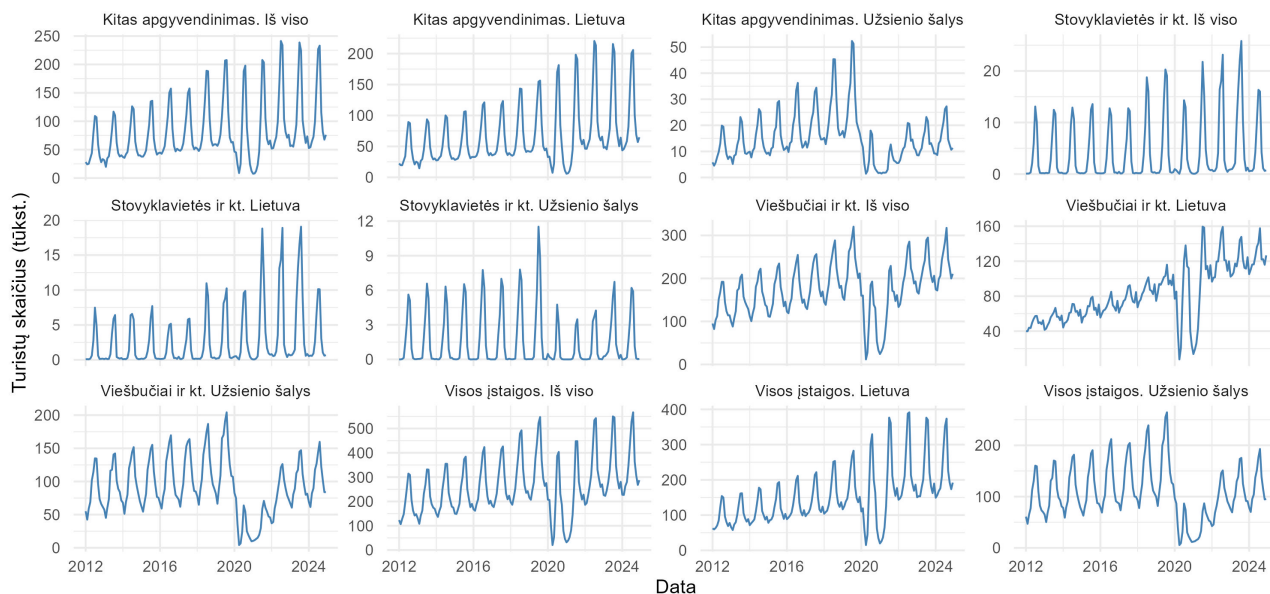
3.2. Duomenų paruošimas ir pradinė analizė

Darbe naudojami mėnesiniai 2012–2024 metų duomenys, tad laiko eilutę sudaro 156 stebėjimai. Įtraukus kintamųjų 12 mėnesių lagus, stebėjimų skaičius sumažėjo iki 144. Taip pat modelių konstravimui ir testavimui laiko eilutė buvo padalinta į mokymo ir testavimo aibes – paskutinių turimų (t. y. 2024) metų stebėjimai naudojami modelių tikslumui vertinti.

Kiekvienas iš priklausomų kintamųjų – t. y. turistų ir nakvynių skaičių – padalintas į 16 kategorijų, atitinkančių visas apgyvendinimo įstaigų kategorijų ir turistų kilmės kombinacijas. Turistų skaičių šiose kategorijose grafikai pateikti 1 paveiksle, nakvynių skaičių grafikai – 2 paveiksle. Grafikai atskleidžia, jog visi kintamieji pasižymi sezoniškumu – tiek turistų, tiek nakvynių skaičiai visose apgyvendinimo įstaigų kategorijose stipriai išauga vasaros mėnesiais. Taip pat kai kuriose apgyvendinimo įstaigų kategorijose („Kitas apgyvendinimas“, „Viešbučiai ir kt.“, „Visos įstaigos“) galima pastebėti skaičių pakilimą gruodžio mėnesiais – per didžiąsias žiemos šventes. Be to, iki COVID-19 pandemijos pradžios (t. y. 2020 metų), laiko eilutės pasižymi ryškia tendencija – turistų ir nakvynių skaičiai kasmet didėja. COVID-19 pandemijos laikotarpiu (2020-01–2021-07) daugelyje kategorijų matomas stiprus skaičių kritimas, kuris ryškiausias visose kategorijose su turistais iš užsienio šalių – būtent šiose kategorijose skaičiai vis dar nėra atsistatę į priešpandeminį lygį, kas gali būti nulemta ir 2022 metais Ukrainoje prasidėjusio karo. Iš grafikų panašu, jog mažiausiai pandemijos paveiktos apgyvendinimo įstaigos – priklausančios kategorijai „Stovyklavietės ir kt.“, likusių kategorijų skaičiai pasižymi ryškiais kritimais pandemijos laikotarpiu.

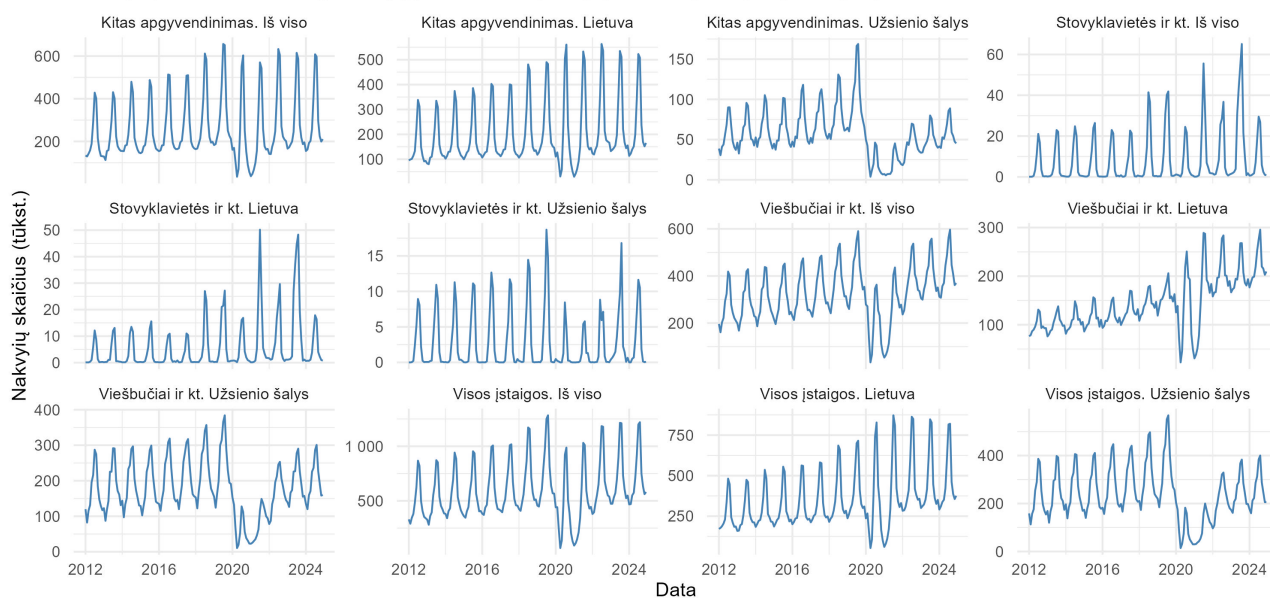
Dėl aptartų pastebėjimų ir situacijos specifiškumo pandemijos laikotarpio duomenys buvo laikomi išskirtimis ir nuspręsta juos koreguoti. Jų koregavimui naudota laiko eilučių praleistų reikšmių įrašymo paketo *imputeTS* funkcija *na_seasplit*, kuri padalina laiko eilutę pagal sezoną ir naudoja tiesinę interpoliaciją praleistų stebėjimų įrašymui [14]. Turimi duomenys yra mėnesiniai, tad laiko eilutė yra padalinama pagal mėnesį, ir reikšmės įrašomos naudojant ankstesnių ir vėlesnių metų to paties mėnesio reikšmes. Siekiant užtikrinti duomenų nuoseklumą ir logiką, funkcija naudojama tik atskirų apgyvendinimo įstaigų kategorijų (t. y. „Viešbučiai ir kt.“, „Stovyklavietės ir kt.“ ir „Kitas apgyvendinimas“) duomenų turistams iš Lietuvos ir užsienio valstybių koregavimui, o turistų ir nakvynių skaičius visose įstaigose įrašomas skaičiuojant atskirų kategorijų sumą. Turistų iš visų valstybių (t. y. „Iš viso“)

Turistų skaičius skirtingose apgyvendinimo įstaigų ir turistų tautybių kategorijose



1 pav. Turistų skaičiai visose kategorijose

Nakvyių skaičius skirtingose apgyvendinimo įstaigų ir turistų tautybių kategorijose

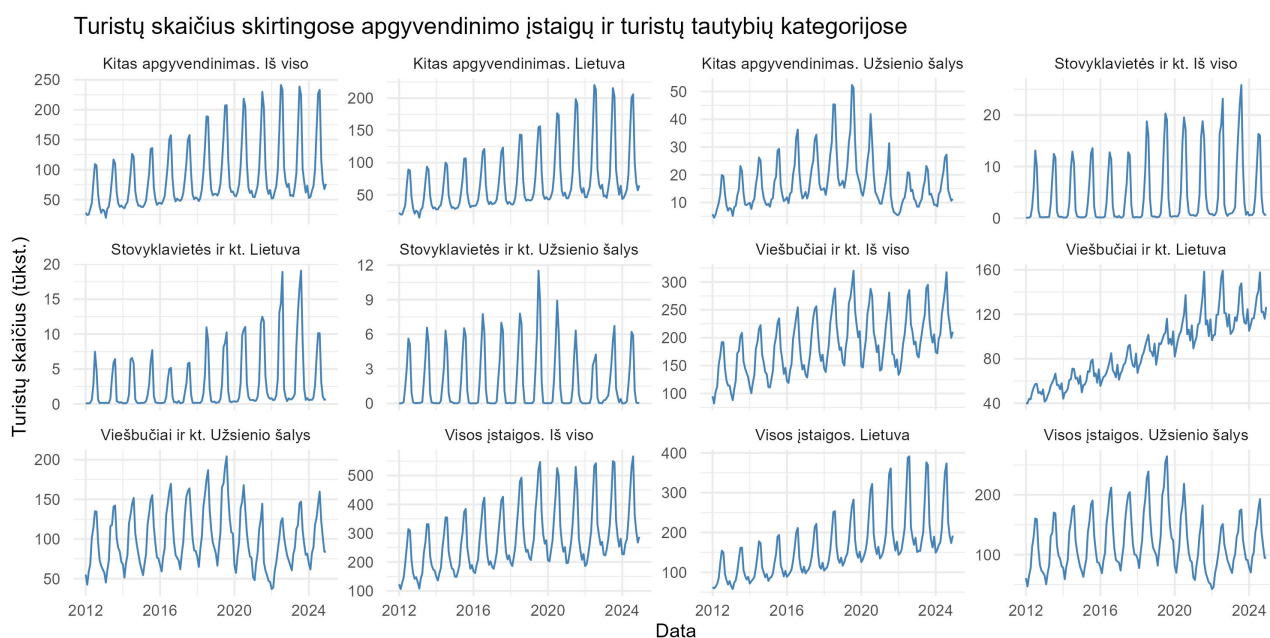


2 pav. Nakvynių skaičiai visose kategorijose

ir nakvynių skaičius atskiroms apgyvendinimo įstaigų kategorijoms užpildomas sumuojant atitinkamo kintamojo ir atitinkamos įstaigų kategorijos reikšmes turistams iš Lietuvos ir užsienio valstybių. Bendri visų turistų ir visų nakvynių visose apgyvendinimo įstaigose skaičiai koreguoti naudojant atitinkamo kintamojo visose apgyvendinimo įstaigose reikšmes kategorijose „Lietuva“ ir „užsienio šalys“ sumas.

3 ir 4 paveiksluose pateikti koreguotų turistų ir nakvynių skaičių grafikai. Koreguoti duomenys pasižymi nuoseklesne tendencija ir neturi stipriai išsiskiriančių laikotarpių. Tiesa, turistų ir nakvynių skaičiai keliautojams iš užsienio šalių vis dar pasižymi reikšmių sumažėjimu paskutiniais 4 metais, tačiau skirtumas nėra toks ryškus koregavus pandemijos laikotarpį.

Koreguotuose duomenys vis dar galima pastebėti netipinių skaičių kritimų kategorijose „Stovyklavietės ir kt. Iš viso“ bei „Stovyklavietės ir kt. Lietuva“ – paskutiniais (t. y. 2024) metais tiek turistų, tiek nakvynių skaičius šiose kategorijose yra mažas, atsižvelgiant į ankstesnių metų reikšmes. Toks netipinis reikšmių sumažėjimas gali padidinti prognozavimo paklaidas, tačiau pasirinkta šių reikšmių nekoreguoti, kad būtų išlaikytas duomenų autentiškumas.

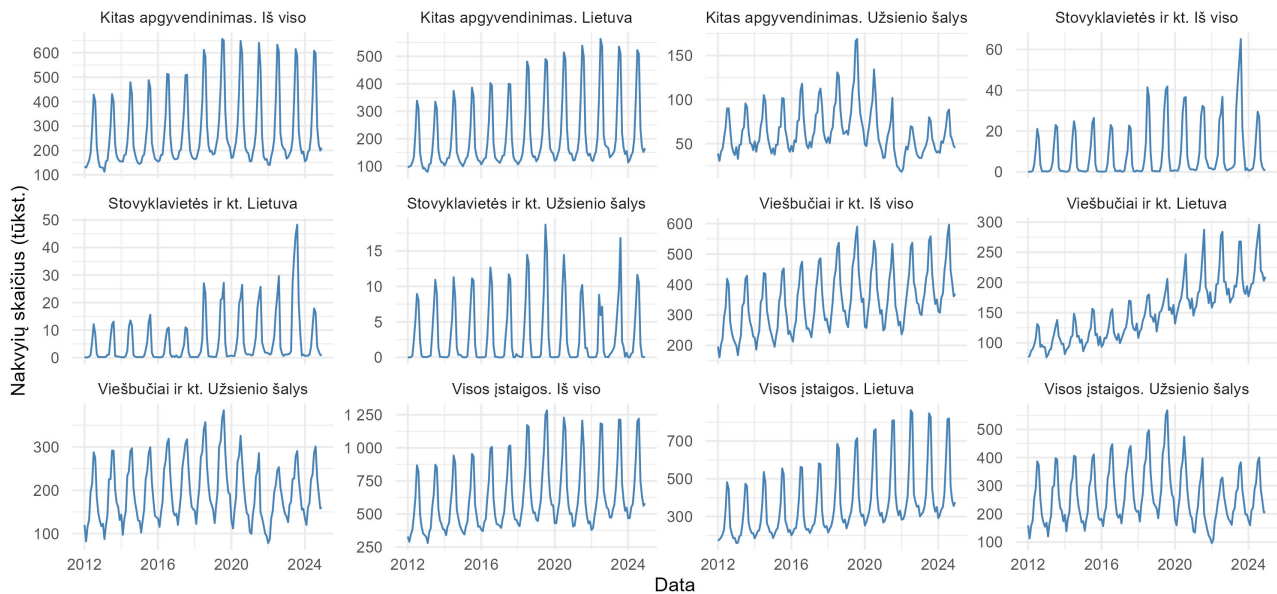


3 pav. Koreguoti turistų skaičiai visose kategorijose

Toliau analizuojami „Google Trends“ duomenys. 5 paveiksle pateikti lietuviškų užklausų duomenų grafikai, o 6 – pasaulinio masto užklausų anglų kalba duomenų grafikai. Grafikai atskleidžia, jog tendencijos yra skirtingos skirtingoms užklausoms – užklausos, kaip „Kaunas“, „Vilnius“ (tiek Lietuvos, tiek pasaulio mastu), „kelionė“, „stovyklavietė“ pasižymi ryškia kilimo tendencija, tuo tarpu užklausų „apgyvendinimas“ ir „motelis“ populiarumo indeksų reikšmės kasmet mažėja. Taip pat yra užklausų, kurių populiarumas per metus išlieka gana stabilus – tai galima pastebėti užklausose „kempingas“, „sodyba“. Be to, dauguma užklausų, ypač tų, kurioms paieška vykdyta Lietuvoje, pasižymi ryškiu sezoniškumu.

Taip pat grafikai atskleidžia, jog užklausų populiarumas nebuvo taip stipriai paveiktas pandemijos laikotarpio, tačiau vis dar galima pastebėti išsiskirčių – jos ryškiausios lietuviškose užklausose „Kaunas“, „kelionė“, „sanatorija“, „viešbutis“ ir „Vilnius“ bei pasaulinio masto užklausose „camping in

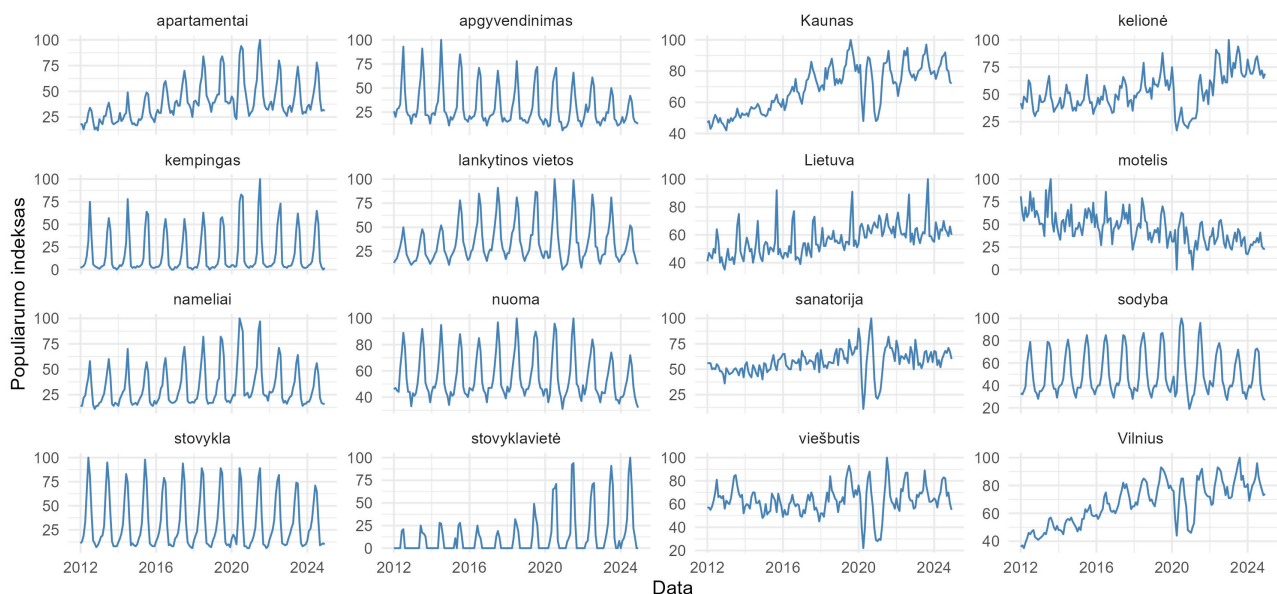
Nakvynių skaičius skirtingose apgyvendinimo įstaigų ir turistų tautybių kategorijose



4 pav. Koreguoti nakvynių skaičiai visose kategorijose

Lithuania“, „flights to Lithuania“, „Kaunas“, „things to do in Lithuania“, „Vilnius“ ir „visit Lithuania“. Šių užklausų pandemijos laikotarpio duomenis taip pat nuspręsta koreguoti.

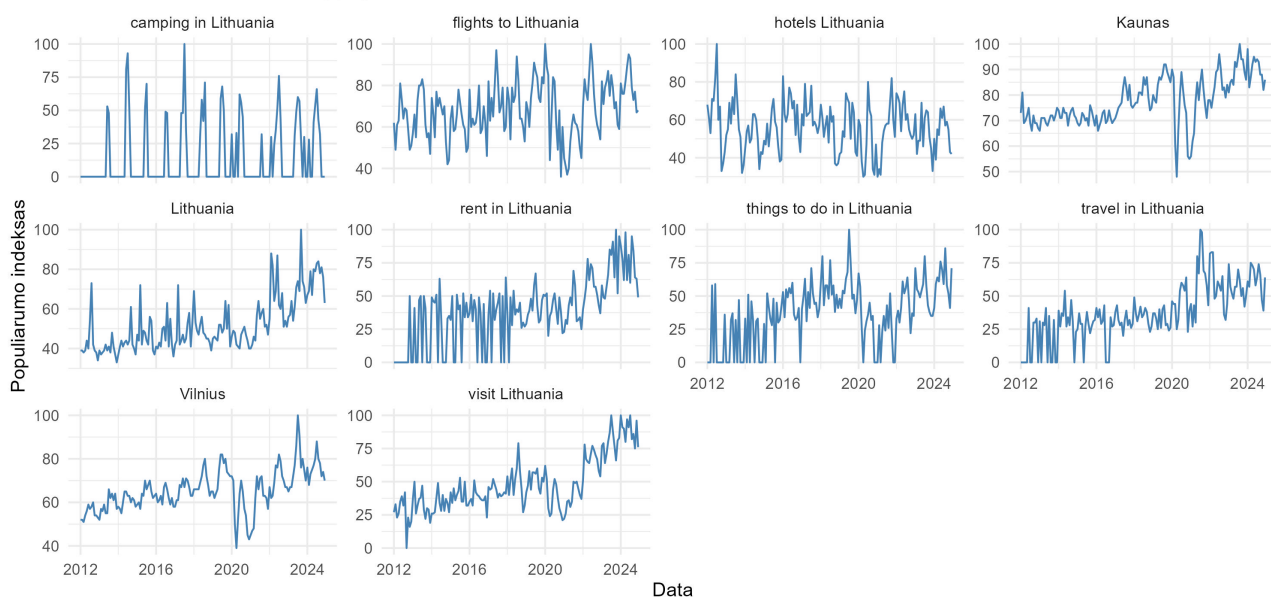
Lietuviškų užklausų populiarumas



5 pav. Lietuviškos užklausos

„Google Trends“ duomenys, pasižymintys ryškiu sezoniškumu taip pat koreguoti naudojant *na_seasplit* funkciją. Tačiau užklausų, kuriose nėra pastebimos sezoninės komponentės (t. y. „sanatorija“, „flights to Lithuania“, „Kaunas“ (pasaulio mastu), „things to do in Lithuania“ ir „visit Lithuania“), duomenys įrašyti naudojant to paties paketo funkciją *na_random*, kuria praleisti stebėjimai užpildomi atsitiktinėmis reikšmėmis [14]. Taikant šią funkciją nustatyti galimi reikšmių režiai atsižvelgiant į praleistus stebėjimus supančias reikšmes (pavyzdžiui, užklausiai „sanatorija“ nustatyti režiai

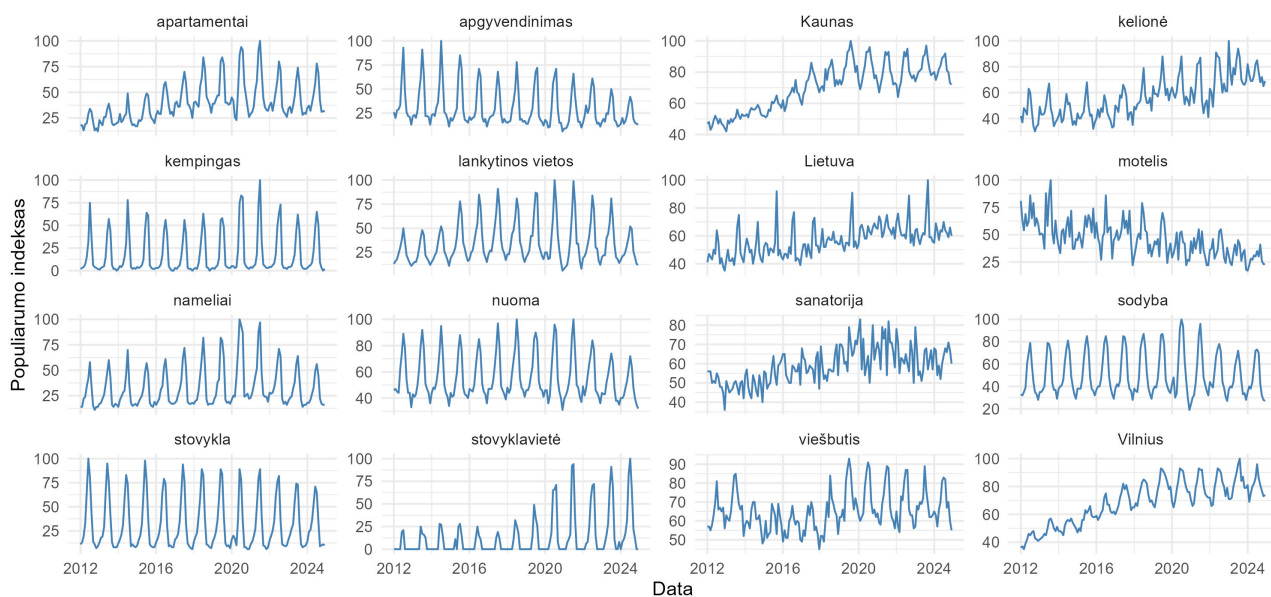
Pasaulio masto užklausų populiarumas



6 pav. Pasaulinio masto užklausos

[50; 85]), taip išvengiant išskirčių. Koreguotų „Google Trends“ duomenų grafikai pateikti 7 ir 8 paveiksluose.

Lietuviškų užklausų populiarumas

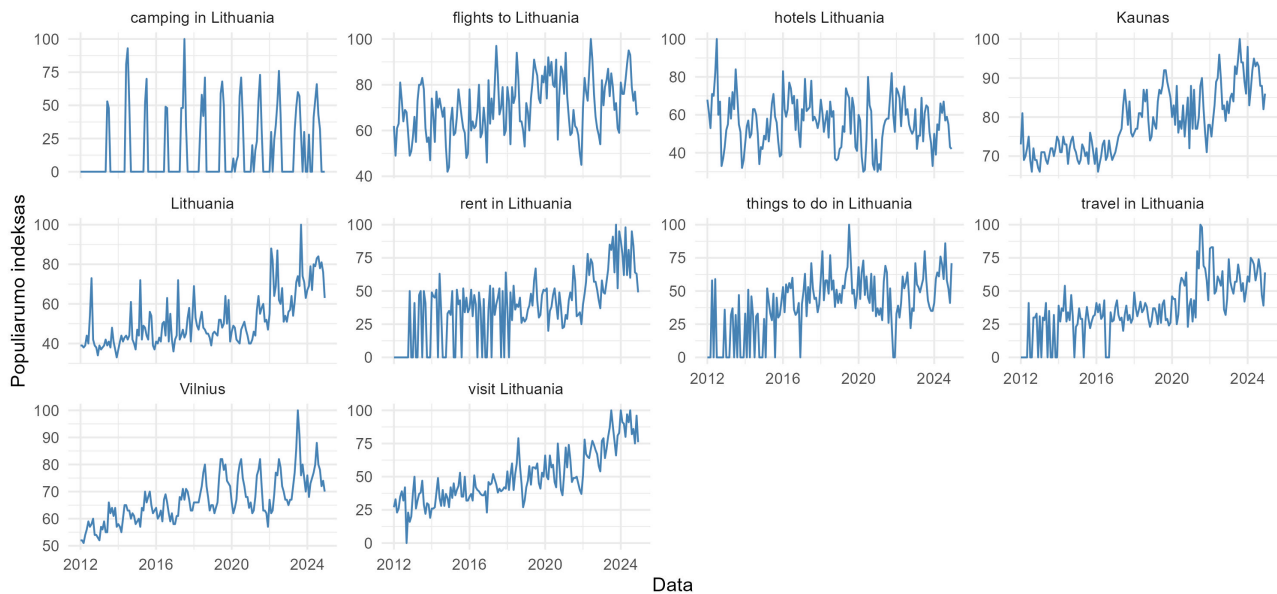


7 pav. Koreguotos lietuviškos užklausos

Taip pat tikrinti ir Pirsono koreliacijos koeficientai tarp „Google Trends“ duomenų kintamųjų. Koeficientų reikšmės skaičiuotos lietuviškos ir pasaulio masto užklausoms atskirai, nes jos naudojamos skirtinguose modeliuose. Koreliacijos koeficientų tarp lietuviškų užklausų populiarumo indekso reikšmių grafikas pateiktas 9 paveiksle. Rezultatai rodo, jog daugelis užklausų pasižymi stipria arba vidutiniška tarpusavio koreliacija, tad gali būti susidurta multikolinearumo problema, ypač sudarant atraminių vektorių regresijos modelius.

Koreliacijos koeficientų tarp pasaulinio masto užklausų reikšmių grafikas, pateiktas 10 paveik-

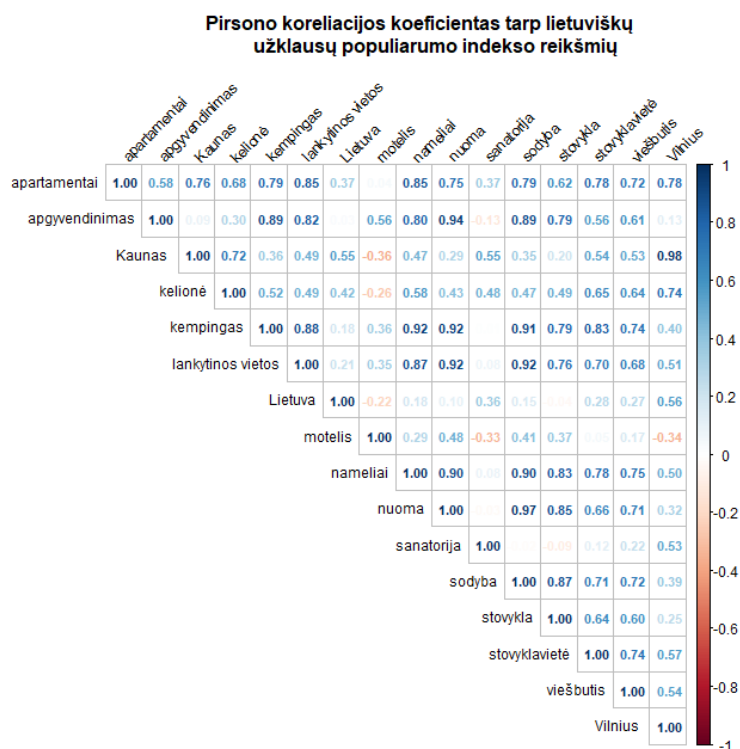
Pasaulio masto užklausų populiarumas



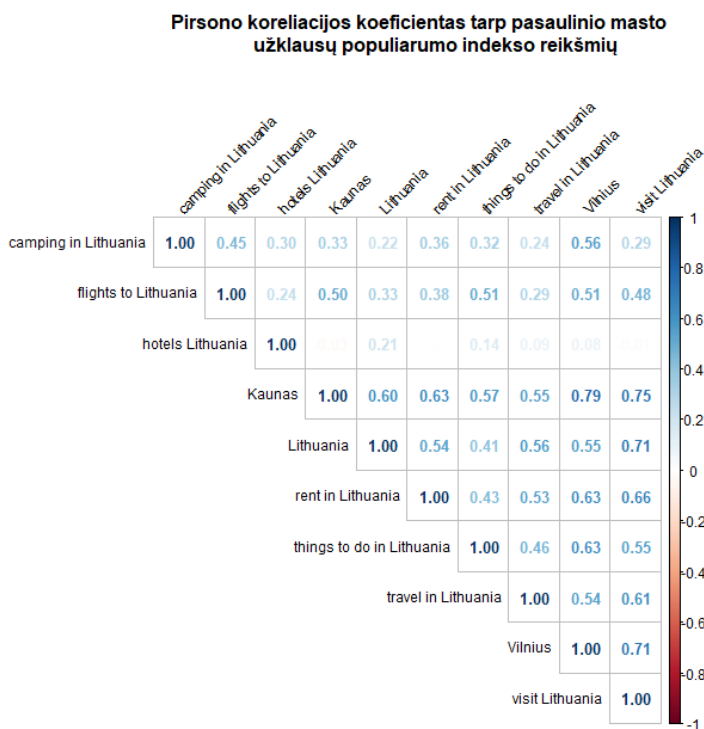
8 pav. Koreguotos pasaulinio masto užklausos

sle, atskleidžia, jog ir tarp šių užklausų koreliacijos yra gana stiprios. Stipria tarpusavio koreliacija (t. y. $|\rho| \geq 0,7$) pasižymi užklausų poros „Kaunas“ ir „Vilnius“, „Kaunas“ ir „visit Lithuania“, „Lithuania“ ir „visit Lithuania“ bei „Vilnius“ ir „visit Lithuania“. Nemaža dalis likusių porų pasižymi vidutiniškai stipria koreliacija ($|\rho| \geq 0,5$). Šie sąryšiai vėlgi gali sukelti multikolinearumą tarp kovariančių.

Dėl šių priežasčių, nuspręsta iš „Google Trends“ kintamųjų konstruoti pagrindines komponentes, taip panaikinant multikolinearumo riziką modeliuose. Pagrindinių komponentių analizė buvo atlikta kiekvienam modeliui su „Google Trends“ duomenimis atskirai. Be to, komponentės konstruotos tik mokymo aibėse. Testavimo aibėse jų reikšmės buvo apskaičiuotos pagal mokymo aibių rezultatus.



9 pav. Koreliacijos koeficientai tarp lietuviškų užklausų



10 pav. Koreliacijos koeficientai tarp pasaulio masto užklausų

4. Lietuvos turistų skaičiaus prognozavimas

Šiame skyriuje analizuojamas SARIMAX ir AVR modelių tikslumas prognozuojant vietinių turistų skaičių Lietuvos apgyvendinimo įstaigose. Modeliuose su įtrauktais „Google Trends“ duomenimis naudojamos paieškų užklausos lietuvių kalba. Modeliai apmokomi kiekvienam apgyvendinimo įstaigos tipui („Viešbučiai ir kt.“, „Stovyklavietės ir kt.“, „Kitas apgyvendinimas“ ir „Visos įstaigos“) atskirai.

4.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai

Pirmasis prognozavimui naudojamas modelis – sezoninis ARIMA modelis su egzogeniniais kintamaisiais (toliau SARIMAX). Modelių parametrus parinkti naudojama R paketo *fable* funkcija *ARIMA*, kuri automatiškai išrenka tinkamiausią modelį, minimizuojantį Akaikės informacinį kriterijų (AIC). Egzogeninių kintamųjų – „Google Trends“ duomenų – labai parinkti atsižvelgiant į koreliacijos su prognozuojamu kintamuoju koeficientą ir eksperimentiniu būdu – nuspręsta naudoti 1, 2, 11 ir 12 lagus. Siekiant pasiekti aukštesnį prognozavimo tikslumą, tiek į bazinį, tiek į „Google Trends“ duomenis naudojantį modelį įtraukiamas papildomas kokybinis kintamasis, kurio reikšmė yra 1 visiems stebėjimams nuo 2021 metų liepos mėnesio, t. y. nuo COVID-19 pandemijos pabaigos, ir 0 likusiais mėnesiais. Šį kintamąjį pasirinkta įtraukti nes, nors pagrindinio pandemijos laikotarpio duomenys buvo koreguoti, turizmo sektorius po jo vis dar nėra pilnai atsistatęs. Šis kintamasis įtraukia ir Ukrainoje vykstantį karą, kuris taip pat gali daryti stiprią įtaką turistų srautams dėl kylančio pavojaus Baltijos šalyse, ypač užsienio turistų atžvilgiu. Pagrindinių komponentų skaičius buvo renkamas pažingsniui skaičiuojant modelio tikslumą ir pagrindines modelio prielaidas tikrinančių statistinių testų (Praplėstojo Dikio ir Fulerio bei Ljungo ir Bokso) p reikšmes, taip išrenkant geriausią „Google Trends“ duomenis naudojantį modelį.

Viešbučių tipo apgyvendinimo įstaigos

Pirmiausia prognozuojami turistų skaičiai kiekvienam apgyvendinimo įstaigos tipui atskirai, pradedant nuo kategorijos „Viešbučiai ir kt.“. Ši kategorija pasižymi aukščiausiu koreliacijos koeficientu su paieškos užklausomis „Vilnius“, „Kaunas“, „apartamentai“, „sodyba“ ir „viešbutis“. Pažingsniui renkant pagrindinių komponentų skaičių, nustatyta, jog optimalus komponentų skaičius modelyje su „Google Trends“ duomenimis – 2.

Bazinio ir „Google Trends“ duomenis naudojančio modelių tinkamumas buvo tikrinamas taikant statistinius testus liekanoms. Praplėstojo Dikio ir Fulerio (toliau ADF) testo p reikšmės abiem modeliams gautos $p < 0,01$. Kadangi gautos reikšmės mažesnės už nustatytą 5 % reikšmingumo lygmenį, galima daryti išvadą, jog modelių liekanos atitinka atsitiktinį klaidžiojimą. Pritaikius Ljungo ir Bokso testą, skirtą liekanų autokoreliacijos hipotezei tikrinti, baziniam ir „Google Trends“ duomenis naudojančiam modeliui gautos p reikšmės buvo $p_{\text{bazinis}} \approx 0,084$ ir $p_{\text{su „Google Trends“}} \approx 0,785$. p reikšmė, aukštesnė už nustatytą reikšmingumo lygmenį reiškia, modelių liekanos nepasižymi autokoreliacija ir atitinka baltąjį triukšmą. Taip pat buvo tikrinamas ir liekanų vidurkis – vienos imties t -testas patvirtino, jog abiejų modelių liekanų vidurkiai yra lygūs 0 (p reikšmės $p_{\text{bazinis}} \approx 0,378$ ir

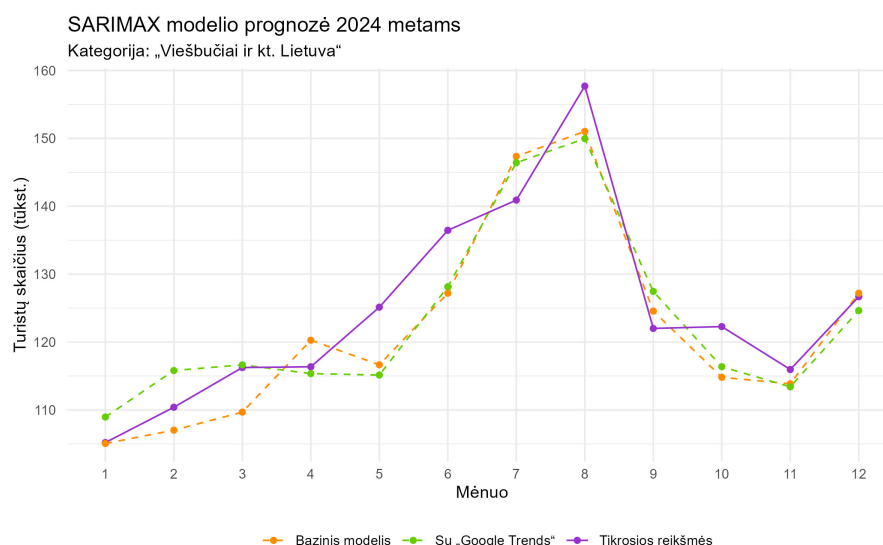
$p_{su „Google Trends“} \approx 0,37$), tad prognozės nėra paslinktos. Visų testų rezultatai parodė, jog modeliai tenkina prielaidas.

Modelių tikslumo testavimo aibėje metrikos pateiktos 1 lentelėje. Rezultatai rodo, jog abu modeliai pasižymi panašiu ir sąlyginai aukštu prognozavimo tikslumu, tačiau bazinis modelis, naudojantis tik istorinius turistų skaičiaus duomenis bei laikotarpio kintamąjį, nežymiai lenkia antrąjį modelį MAPE ir MAE metrikų atžvilgiu. Tačiau „Google Trends“ naudojančio modelio RMSE reikšmė yra šiek tiek žemesnė.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,037	4791	5631,67
Su „Google Trends“	0,038	4840	5621,21

1 lentelė. SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių viešbučiuose

Šiuos rezultatus patvirtina ir 11 paveiksle pateiktas grafikas – nei vienas iš modelių tiksliai neatspindi tikrųjų reikšmių tendencijos. Iš grafiko taip pat sunku nustatyti, kuris modelis yra tinkamesnis – bazinis modelis yra žymiai arčiau tikrųjų reikšmių sausio, rugsėjo ir gruodžio mėnesiais, tačiau „Google Trends“ duomenis naudojantis modelis pasižymi stipriai tikslesnėmis prognozėmis kovo ir balandžio mėnesiais. Likusiais mėnesiais abiejų modelių paklaidos yra panašios.



11 pav. Vietinių turistų skaičiaus viešbučių tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Stovyklaviečių tipo apgyvendinimo įstaigos

Antroji apgyvendinimo įstaigų kategorija – „Stovyklavietės ir kt.“. Šiai kategorijai atrinktos „Google“ paieškų užklausa: „kempingas“, „stovyklavietė“, „lankytinos vietos“, „apartamentai“, „nameliai“. Modelyje su „Google Trends“ duomenimis pasirinkta naudoti 5 komponentes – šis skaičius buvo optimalus tikslumo ir prielaidų tenkinimo atžvilgiu.

Modelių tinkamumo testų rezultatai parodė, jog abu modeliai atitinka prielaidas – modelių liekanos atitinka atsitiktinį klaidžiojimą (ADF testo p reikšmės $p < 0,01$ abiem modeliams), nepasižymi

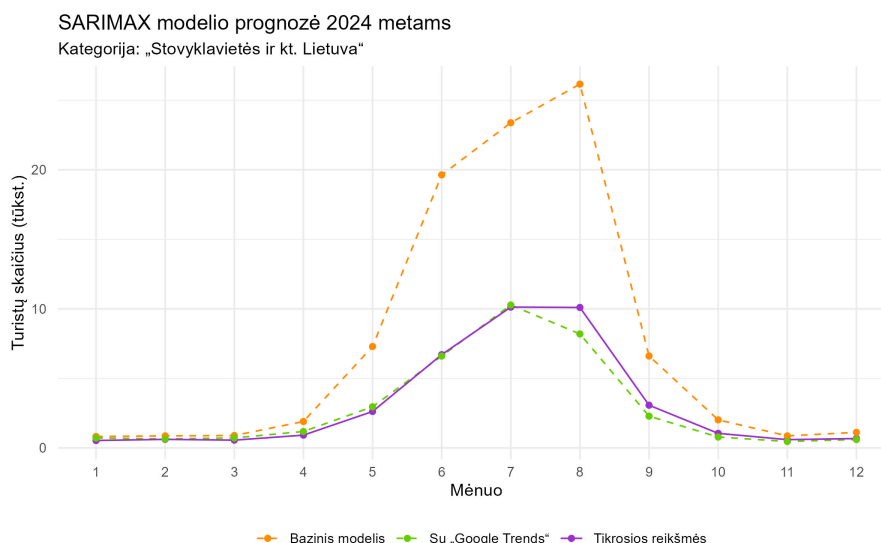
autokoreliacija (Ljungo-Bokso testo p reikšmės $p_{bazinis} \approx 0,935$ ir $p_{su „Google Trends“} \approx 0,971$) bei turi vidurkį, lygų 0 (vienos imties t-testo $p_{bazinis} \approx 0,537$ ir $p_{su „Google Trends“} \approx 0,892$).

Bazinio bei modelio, naudojančio „Google Trends“ duomenis, tikslumo metrikos, pateiktos 2 lentelėje atskleidžia, jog bazinis modelis šiai kategorijai yra netinkamas – jo vidutinė absoliutinė procentinė paklaida siekia net 103,4 %. „Google Trends“ duomenų įtraukimas prognozavimo tikslumą žymiai pagerino visų metrikų atžvilgiu, tačiau paklaidos vis dar aukštos.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	1,034	4499	7288,12
Su „Google Trends“	0,153	305	486,97

2 lentelė. SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių stovyklavietėse

Modelių prognozių grafikai, pateikti 12 paveiksle, rodo, jog bazinis modelis stipriai pervertina turistų skaičių vasaros mėnesiais, kurie yra paklausiausi, todėl ir svarbiausi, apgyvendinimo įstaigoms tiriamoje kategorijoje. Bazinis modelis vasaros sezono turistų srautus prognozuoja daugiau nei dvigubai gausesnius už tikrąsias reikšmes. Modelis su „Google Trends“ duomenimis yra žymiai arčiau tikrųjų reikšmių – prognozės stipriai nukrypsta nuo tikrovės tik rugpjūčio ir rugsėjo mėnesiais –, ką parodė ir tikslumo metrikų analizė.



12 pav. Vietinių turistų skaičiaus stovyklaviečių tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Kitas apgyvendinimas

Trečioji tirta apgyvendinimo įstaigų kategorija – „Kitas apgyvendinimas“. Šiai kategorijai atrinktos užklauskos „apartamentai“, „kempingas“, „stovyklavietė“, „lankytinos vietos“ ir „nameliai“. Į „Google Trends“ duomenis naudojantį modelį pasirinkta įtraukti tik pirmąją pagrindinę komponentę.

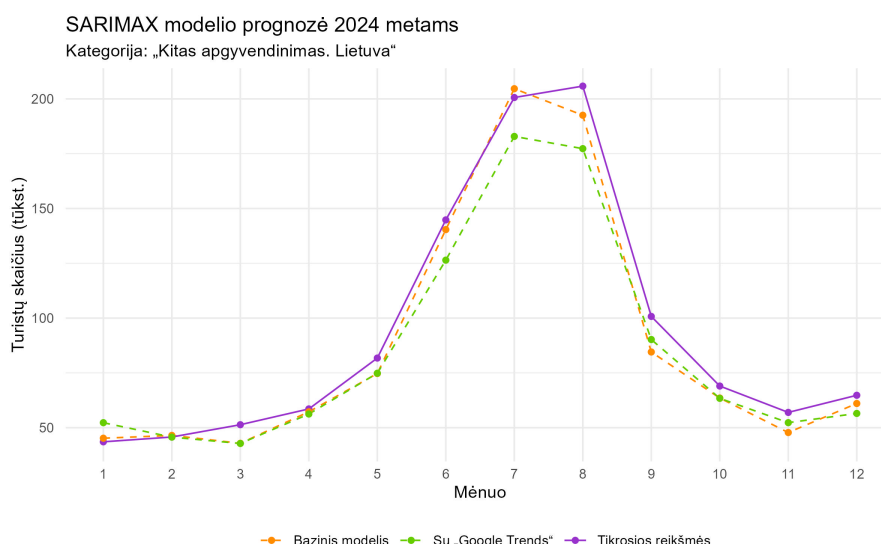
Modelių prielaidoms tikrinti naudoti statistiniai testai. ADF testo rezultatai – p reikšmės abiem modeliams gautos $p < 0,01$ – rodo, jog liekanos atitinka atsitiktinį klaidžiojimą. Pritaikius Ljungo ir Bokso testą liekanoms, baziniam modeliui gauta p reikšmė $p_{bazinis} \approx 0,894$, o modeliui su įtrauktais

„Google Trends“ duomenimis – $p_{su\ „Google\ Trends“} \approx 0,44$. Šie testo rezultatai patvirtina, jog abiejų modelių liekanos nepasižymi autokoreliacija. Galiausiai, vienos imties t-testas parodė, jog abiejų modelių liekanos turi vidurkj, lygų 0 ($p_{bazinis} \approx 0,337$ ir $p_{su\ „Google\ Trends“} \approx 0,288$). Taigi, statistiniai testai patvirtino, jog abu modeliai tenkina prielaidas.

Prognozavimo tikslumo rezultatai, pateikti 3 lentelėje rodo, jog abu modeliai pasižymi gana aukštomis paklaidomis, tačiau bazinis modelis yra žymiai tikslesnis visų metrikų atžvilgiu. 13 paveikslo grafike matoma, jog beveik visais mėnesiais bazinio modelio prognozė yra daug arčiau tikrųjų reikšmių, lyginant su antruoju modeliu. Modelis su įtrauktais „Google Trends“ užklausų populiarumo duomenimis yra linkęs gana stipriai nuvertinti turistų skaičių, didžiausias skirtumas matomas vasaros mėnesiais.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,075	6298	7802,46
Su „Google Trends“	0,103	10029	12590,44

3 lentelė. SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių kitose apgyvendinimo įstaigose



13 pav. Vietinių turistų skaičiaus kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Visos apgyvendinimo įstaigos

Bendrą turistų skaičių visose apgyvendinimo įstaigose galima prognozuoti dviem būdais: sudaryti atskirą modelį, prognozuojantį reikšmes nepriklausomai nuo anksčiau analizuotų modelių, arba sumuojant trijų apgyvendinimo įstaigų kategorijų prognozuotas reikšmes. Abi strategijos turi trūkumų. Naudojant pirmąją metodiką, tikėtina, bus pažeista duomenų logika – gautos reikšmės gali neatitikti skaičių, gautų agreguojant atskirų apgyvendinimo įstaigų kategorijų prognozes. Tačiau naudojant antrąjį būdą gali stipriai sumažėti prognozių tikslumas, ypač atsižvelgiant į gautas aukštas paklaidas stovyklaviečių ir kito apgyvendinimo kategorijose.

Pirmiausiai išbandytas pirmasis prognozavimo metodas. Užklausos, turinčios stipriausią koreliaciją su bendru vietinių turistų skaičiumi, yra „apartamentai“, „Vilnius“, „Kaunas“, „viešbutis“ ir

„sodyba“. „Google Trends“ duomenis naudojantis modelis sudarytas įtraukiant dvi pagrindines komponentes – šis komponentių skaičius buvo optimalus tikslumo metrikų ir prielaidų tikrinimo statistinių testų atžvilgiu.

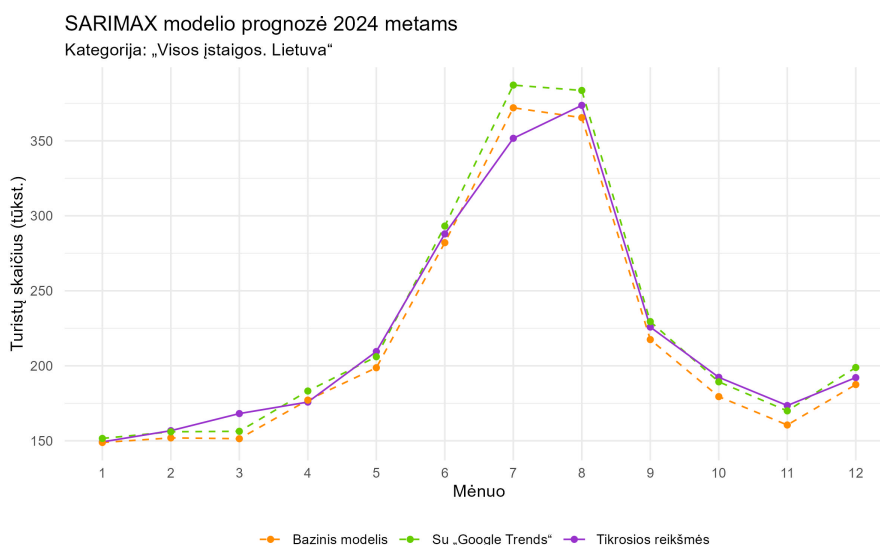
Patikrinus modelių prielaidas taikant statistinius testus buvo padaryta išvada, jog abu modeliai yra tinkami – ADF testo p reikšmės $p < 0,01$ abiem modeliams patvirtina, jog liekanos atitinka atsitiktinį klaidžiojimą, o Ljungo-Bokso testo rezultatai – $p_{\text{bazinis}} \approx 0,762$ ir $p_{\text{su „Google Trends“}} \approx 0,901$ – leidžia teigti, jog liekanos nepasižymi autokoreliacija ir atitinka baltąjį triukšmą. Pritaikius t-testą taip pat buvo patvirtinta, jog abiejų modelių liekanos turi 0 vidurkį – testo p reikšmės $p_{\text{bazinis}} \approx 0,287$ baziniam modeliui ir $p_{\text{su „Google Trends“}} \approx 0,922$ „Google Trends“ duomenis naudojančiam modeliui.

4 lentelėje pateiktos tikslumo metrikos rodo, jog modelis, naudojantis „Google Trends“ duomenis vidutiniškai yra tikslesnis – šio modelio vidutinė absoliutinė procentinė ir vidutinė absoliutinė paklaidos yra žymiai mažesnės, lyginant su baziniu modeliu. Tačiau bazinis modelis pasižymi žemesne RMSE metrikos reikšme, kas reiškia, jog šio modelio prognozės turi mažiau stiprių nukrypimų nuo tikrųjų reikšmių.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,041	8942	10659,29
Su „Google Trends“	0,031	7268	11397,09

4 lentelė. SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių visose įstaigose

14 paveiksle pavaizduotame grafike „Google Trends“ duomenis naudojantis modelis yra pastebimai arčiau tikrųjų reikšmių – ryškūs nukrypimai yra tik kovo ir liepos mėnesiais. Būtent liepos mėnesio paklaida gali paaiškinti aukštą šio modelio RMSE metrikos rezultatą – šio mėnesio bazinio modelio prognozė yra ne taip stipriai atitolusi nuo tikrosios reikšmės. Tuo tarpu bazinio modelio prognozės pasižymi didesnėmis paklaidomis gegužės, rugsėjo, spalio ir lapkričio mėnesiais, kuriais antrojo tiriamo modelio prognozuotos reikšmės yra žymiai arčiau tikrųjų.



14 pav. Vietinių turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

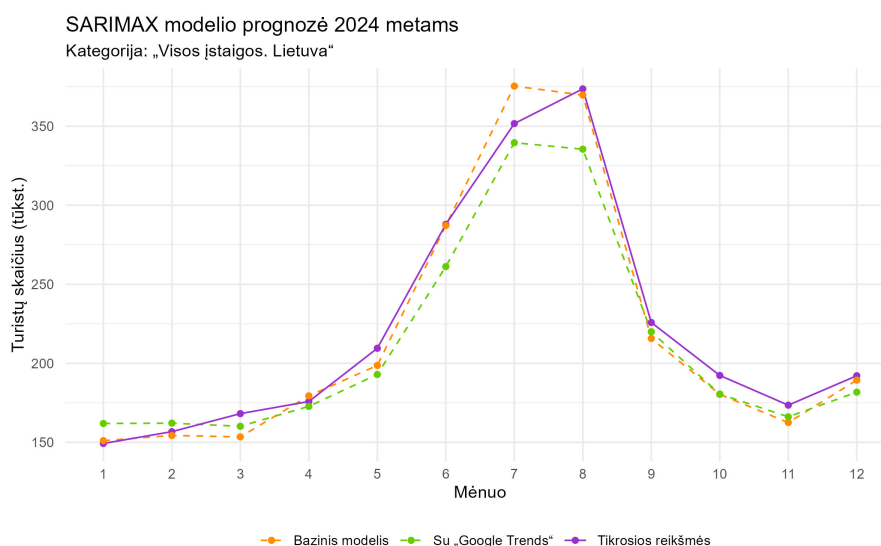
Taip pat buvo išbandyta ir antroji prognozavimo strategija. Modelių tikslumo metrikos pateik-

tos 5 lentelėje. Bazinis modelis žymiai lenkia „Google Trends“ duomenis naudojantį modelį. Be to, modelių paklaidos nėra tokios didelės, kaip tikėtasi, tad šis prognozavimo būdas yra labiau rekomenduotinas, siekiant išlaikyti duomenų nuoseklumą.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,038	8137	10447,13
Su „Google Trends“	0,056	13182	16314,73

5 lentelė. SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių visose įstaigose, naudojant ankstesnių prognozių agregavimą

Modelių prognozės, pavaizduotos 15 paveiksle, patvirtina tikslumo metrikų rezultatus – bazinio modelio prognozės yra labai artios tikrųjų reikšmių tam tikrais mėnesiais (pvz. sausį, birželį ir rugpjūtį), tačiau pasižymi didesnėmis paklaidomis kitais mėnesiais. Ta pati tendencija matoma ir modelio, naudojancio „Google Trends“ duomenis prognozėse. Šio modelio prognozuojamos reikšmės yra žymiai artimesnės tikrosioms liepos ir lapkričio mėnesiais, kuriais bazinio modelio prognozės turi dideles paklaidas. Modelių tikslumo metrikų panašumas gali būti paaiškintas šiais pastebėjimais.



15 pav. Vietinių turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

Vietinių turistų Lietuvos apgyvendinimo įstaigose prognozavimas naudojant SARIMAX modelį parodė, jog „Google Trends“ duomenų įtraukimas gali žymiai pagerinti modelio tikslumą prognozuojant turistų skaičių „Stovyklavietės ir kt.“ kategorijoje. Vienintelė kategorija, kurioje „Google Trends“ duomenis naudojantis modelis pasirodė pastebimai prasčiau, buvo „Kitas apgyvendinimas“. Prognozuojant bendrą vietinių turistų skaičių, rekomenduotina naudoti agregavimo metodą – taip išlaikomas duomenų nuoseklumas, stipriai nesumažinant rezultatų tikslumo. Naudojant šį metodą geresni rezultatai gali būti pasiekti neįtraukiant „Google Trends“ duomenų. Tačiau mažiausios paklaidos šioje kategorijoje gautos sudarant atskirą modelį bendram turistų skaičiui prognozuoti ir įtraukiant paieškos užklausų populiarumo duomenis. Ši metodika gali būti naudojama, kai prognozuojamas tik turistų skaičius visose įstaigose ir nėra grėsmės susidurti su duomenų logikos pažeidimo problema.

4.2. Atraminių vektorių regresijos modelių rezultatai

Antras prognozavimui naudotas modelis – atraminių vektorių regresija (toliau AVR). Modelių optimaliems parametrų parinkti naudota gardelės paieška ir slenkančio lango kryžminė patikra. Duomenų padalinimas kryžminei patikrai atliktas naudojant *caret* paketo funkciją *createTimeSlices*. Pasirinkta naudoti 36 mėnesių langą modelio apmokymui ir 12 mėnesių langą modelio testavimui. Parametrų paieška buvo atlikta minimizuojant šaknį iš vidutinės kvadratinės paklaidos gardelėje

$$G = \{(\sigma, C) | \sigma \in \{0,001, 0,01, 0,05, 0,06, 0,07, 0,08, 0,1, 0,15, 0,25, 0,5, 1\}, \\ C \in \{10, 50, 100, 125, 150, 200, 225, 250, 275, 300\}\}.$$

Baziniai modeliai sudaryti naudojant 1 ir 12 lagų priklausomo kintamojo duomenis. Sudarant „Google Trends“ duomenis naudojančius modelius, naudojamos tos pačios 5 užklausos, pasižyminčios stipriausiu sąryšiu su priklausomu kintamuoju, kaip ir SARIMAX atveju. Be to, pasirinkta įtraukti tik 1 ir 12 lagus, nes didesnis skaičius lagų sumažino modelių tikslumą. Pagrindinės komponentės priklausomam kintamajam ir „Google Trends“ duomenims konstruotos atskirai, taip palengvinant modelių konstravimą – į bazinį modelį įtraukiamos tik iš priklausomo kintamojo sudarytos komponentės, o į antrąjį modelį papildomai įtraukiamos ir komponentės, sukonstruotos iš paieškos užklausų populiarumo duomenų. Be to, eksperimentiniu būdu nuspręsta įtraukti visas pagrindines komponentes – taip gauti geriausi tikslumo metrikų rezultatai.

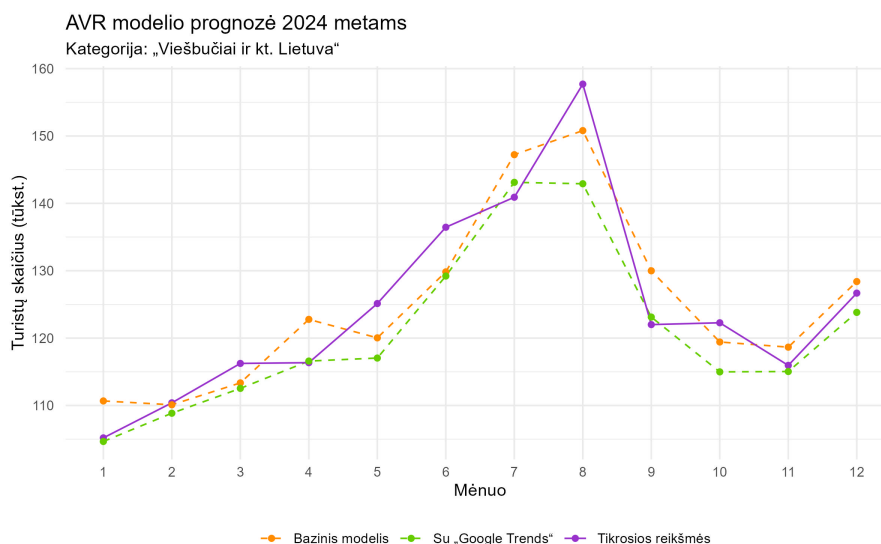
Viešbučio tipo apgyvendinimo įstaigos

Pirmoji analizuojama apgyvendinimo įstaigų kategorija – „Viešbučiai ir kt.“. Modelių tikslumo metrikų reikšmės testavimo aibėje pateiktos 6 lentelėje. Rezultatai rodo, jog „Google Trends“ duomenis naudojantis modelis yra pranašesnis už bazinį MAPE ir MAE metrikų atžvilgiu. Tačiau bazinis modelis pasižymi žemesne RMSE reikšme, kas gali reikšti, jog šio modelio prognozės yra mažiau linkę stipriai nukrypti nuo tikrųjų reikšmių.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,037	4607	5158,54
Su „Google Trends“	0,032	4211	5926,12

6 lentelė. AVR modelio tikslumas prognozuojant vietinių turistų skaičių viešbučio tipo įstaigose

Prognozių 2024 metams grafikas, pateiktas 16 paveiksle, atskleidžia, jog bazinis modelis yra labiau linkęs stipriai pervertinti turistų skaičių, ką galima pastebėti sausio, balandžio, liepos ir rugsėjo mėnesiais. Tuo tarpu modelis su įtrauktais „Google Trends“ duomenimis gegužės, rugpjūčio ir spalio mėnesiais stipriai nuvertina turistų skaičių – būtent šių mėnesių prognozėse pastebimos didelės paklaidos gali paaiškinti aukštą modelio RMSE metrikos reikšmę. Visgi, šio modelio prognozės beveik tiksliai atitinka tikrąsias reikšmes sausio, kovo, rugsėjo ir spalio mėnesiais, tuo tarpu bazinio modelio prognozės aukštu tikslumu pasižymi tik vasario mėnesį.



16 pav. Vietinių turistų skaičiaus viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Stovyklavietės tipo apgyvendinimo įstaigos

Toliau analizuojamas turistų skaičiaus prognozavimas „Stovyklavietės ir kt.“ kategorijai priklausančiose apgyvendinimo įstaigose. Prognozuojant turistų skaičių šioje kategorijoje, abu modeliai pasižymi labai aukštomis paklaidomis – bazinio modelio vidutinė absoliutinė procentinė paklaida siekia net 107,4 % (žr. 7 lentelę). Antrasis modelis pasižymi aukštesniu tikslumu visų metrikų atžvilgiu, tačiau paklaidos vis dar yra aukštos – šio modelio MAPE siekia 73,1 %. Abiejų modelių aukštos paklaidos gali būti paaiškintos tuo, jog turistų skaičius šiose apgyvendinimo įstaigose 2024 metais buvo itin žemas, lyginant su ankstesniais (t. y. 2022–2023) metais.

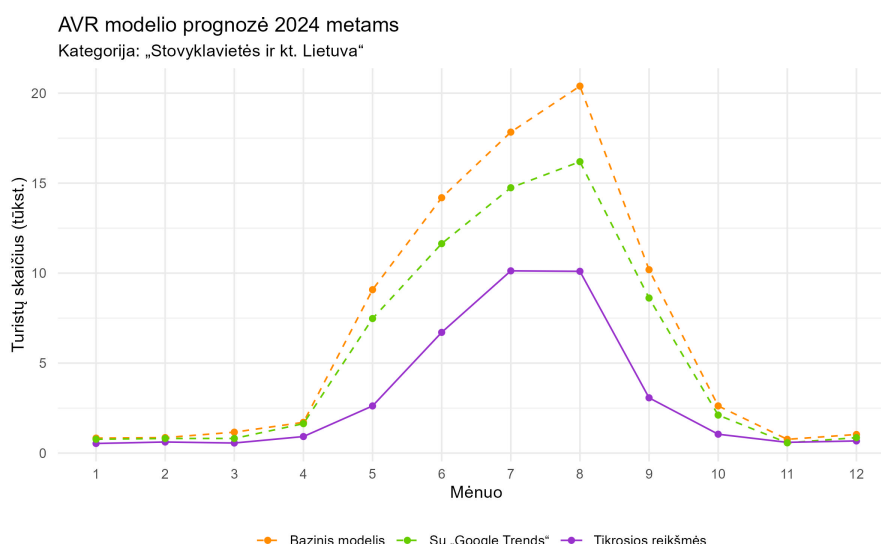
Modelis	MAPE	MAE	RMSE
Bazinis modelis	1,074	3591	5143,37
Su „Google Trends“	0,731	2393	3402,07

7 lentelė. AVR modelio tikslumas prognozuojant vietinių turistų skaičių stovyklavietės tipo įstaigose

17 paveiksle pateiktas grafikas patvirtina aptartus tikslumo metrikų rezultatus – abu modeliai yra stipriai nutolę nuo tikrųjų reikšmių paklausiausiais (t. y. gegužės–rugsėjo) mėnesiais, tačiau modelis su įtrauktais „Google Trends“ duomenimis pasižymi mažesnėmis paklaidomis. Rugsėjo mėnesį bazinio modelio prognozė yra daugiau nei du kartus didesnė už tikrąjį turistų skaičių.

Kitos apgyvendinimo įstaigos

Modelių tikslumo metrikų reikšmės prognozuojant vietinių turistų skaičių „Kitas apgyvendinimas“ kategorijai priklausančiose apgyvendinimo įstaigose pateiktos 8 lentelėje. Rezultatai rodo, jog abu modeliai pasižymi labai panašiu prognozavimo tikslumu. Bazinis modelis paklaidos vidutiniškai yra tikslesnis, ką parodo mažesnės MAPE ir MAE metrikų reikšmės. Antrasis modelis yra pranašesnis RMSE metrikos atžvilgiu. Tačiau skirtumai tarp paklaidų yra labai maži, tad šiai kategorijai gali būti naudojami abu modeliai.

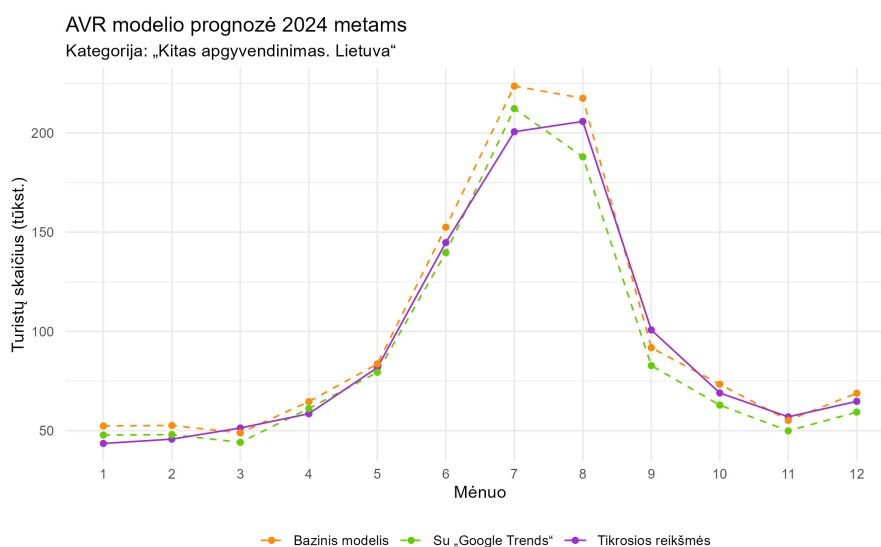


17 pav. Vietinių turistų skaičiaus stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,083	7313	9203,2
Su „Google Trends“	0,085	7486	9162,4

8 lentelė. AVR modelio tikslumas prognozuojant vietinių turistų skaičių kitose įstaigose

Iš grafiko, pateikto 18 paveiksle, taip pat sunku padaryti išvadą apie tai, kuris modelis yra tikslus. Itin ryškūs skirtumai tarp modelių paklaidų pastebimi kovo, rugpjūčio, rugsėjo ir lapkričio mėnesiais, kuriais bazinis modelis yra tikslus bei sausio, vasario ir liepos mėnesiais, kurių prognozės yra artimesnės tikriesiems turistų skaičiams naudojant „Google Trends“ duomenis. Likusiais mėnesiais abiejų modelių tikslumas yra panašus.



18 pav. Vietinių turistų skaičiaus kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Visos apgyvendinimo įstaigos

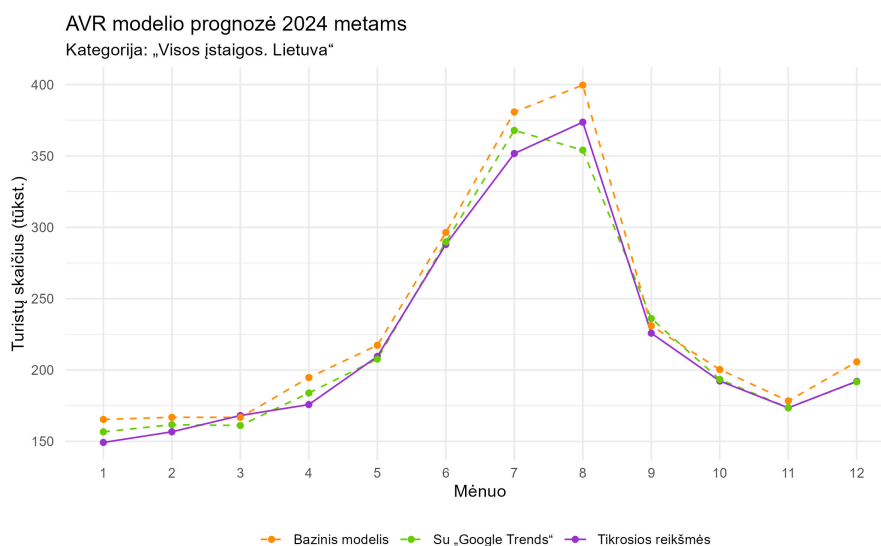
Galiausiai, prognozuojamas vietinių turistų skaičius visose apgyvendinimo įstaigose bendrai. Šiai kategorijai naudojami du prognozavimo metodai, aptarti SARIMAX modelio skyrelyje.

Pirmiausiai išbandyta konstruoti atskirus modelius turistų skaičiaus prognozavimui. Rezultatai, pateikti 19 lentelėje rodo, jog „Google Trends“ duomenis naudojantis modelis stipriai lenkia bazinį visų metrikų atžvilgiu. Bazinio modelio vidutinė absoliutinė procentinė paklaida daugiau nei dvigubai didesnė už antrojo modelio. Kitos paklaidų metrikos taip pat pasižymi dideliais skirtumais.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,058	12464	14939,39
Su „Google Trends“	0,026	6576	8926,1

9 lentelė. AVR modelio tikslumas prognozuojant vietinių turistų skaičių visose įstaigose

19 paveiksle pateiktas prognozių grafikas patvirtina tikslumo metrikų rezultatus – bazinio modelio paklaidos yra pastebimai didesnės, šis modelis yra linkęs stipriau pervertinti turistų skaičių. Gegužės, birželio, spalio, lapkričio ir gruodžio mėnesiais modelis su „Google Trends“ duomenimis beveik tiksliai atitinka tikrąsias reikšmes. Tuo tarpu bazinis modelis mažesnėmis paklaidomis pasižymi tik kovo ir rugsėjo mėnesiais, tačiau ir šiais mėnesiais skirtumai tarp modelių tikslumo nėra dideli.

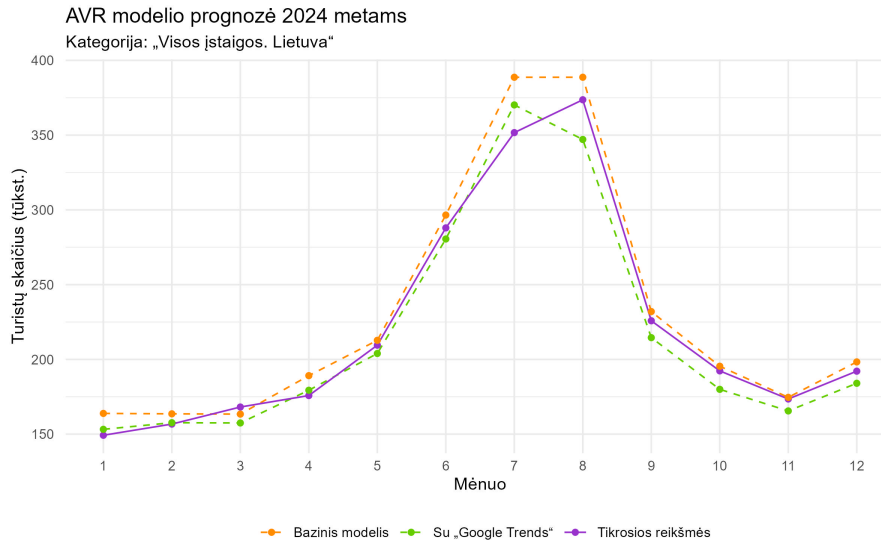


19 pav. Vietinių turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Antroji prognozavimo metodika – agreguoti ankstesnes atskirų apgyvendinimo įstaigų tipų prognozes. Tikslumo rezultatai pateikti 10 lentelėje. Naudojant šią metodiką, „Google Trends“ duomenis naudojančio modelio prognozės vėl yra tikslesnės visų metrikų atžvilgiu, tačiau šį kartą skirtumai nėra tokie dideli. Šiuos rezultatus iliustruoja 20 pateiktas grafikas – antrojo modelio prognozės yra tikslesnės visais mėnesiais, išskyrus kovą, iki rugpjūčio. Nuo rugpjūčio bazinio modelio prognozės yra arčiau tikrųjų reikšmių. Šie rezultatai paaiškina mažus skirtumus tarp modelių tikslumo metrikų reikšmių.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,043	9998	13628
Su „Google Trends“	0,041	9738	11851,72

10 lentelė. AVR modelio tikslumas prognozuojant vietinių turistų skaičių visose įstaigose, naudojant ankstesnių prognozių agregavimą



20 pav. Vietinių turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

4.3. Modelių palyginimas

Apibendrinant vietinių turistų skaičiaus prognozavimą, nei vienas iš naudotų (t. y. SARIMAX ir AVR) modelių nepasižymėjo didesniu tikslumu visose analizuotose kategorijose. Naudojant SARIMAX modelį, gaunamos mažesnės paklaidos kategorijose „Kitas apgyvendinimas“, „Stovyklavietės ir kt.“ ir „Visos įstaigos“ naudojant atskirų įstaigų kategorijų prognozių agregavimą. Tačiau AVR modelis gali duoti geresnius rezultatus prognozuojant vietinių turistų skaičių „Viešbučiai ir kt.“ kategorijoje bei sudarant modelį su „Google Trends“ duomenimis bendram turistų skaičiui visose įstaigose prognozuoti. SARIMAX modelis aukštesniu tikslumu pasižymėjo didesnėje dalyje apgyvendinimo įstaigų kategorijų ir jo prognozės visoms įstaigos naudojant duomenų logiką išlaikančią metodiką buvo tikslesnės, tad šis modelis yra labiau tinkamas vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimui.

5. Visų turistų skaičiaus prognozavimas

Šiame skyriuje analizuojamas SARIMAX ir AVR modelių tikslumas prognozuojant visų turistų skaičių Lietuvos apgyvendinimo įstaigose. Modeliuose su „Google Trends“ duomenimis naudojamos paieškos užklausos anglų kalba. Kaip ir vietinių turistų atveju, skaičiai prognozuojami kiekvienam apgyvendinimo įstaigos tipui atskirai.

5.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai

Visų turistų skaičiaus prognozavimas naudojant SARIMAX modelį yra analogiškas anksčiau pristatytai metodikai.

Viešbučio tipo apgyvendinimo įstaigos

Prognozuojant turistų skaičių apgyvendinimo įstaigose, priklausančiose kategorijai „Viešbučiai ir kt.“, modelyje su „Google Trends“ duomenimis pasirinkta naudoti užklausas „Vilnius“, „Kaunas“, „camping in Lithuania“, „things to do in Lithuania“ ir „visit Lithuania“ – šios užklausos pasižymėjo aukščiausiomis koreliacijos su priklausomu kintamuoju koeficiento reikšmėmis. Nustatytas optimalus pagrindinių komponentų skaičius – 11.

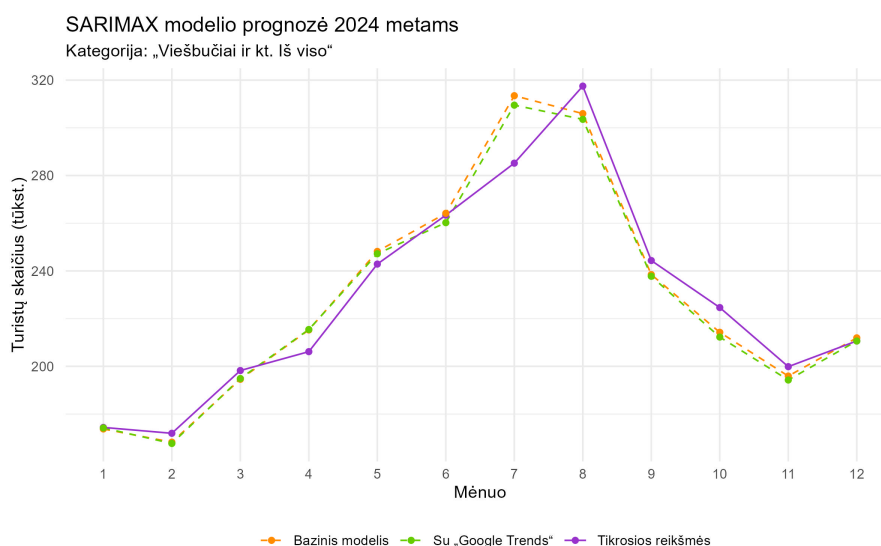
Modelių prielaidos tikrintos taikant statistinius testus. ADF testo p reikšmės abiem modeliams $p < 0,01$, tad abiejų modelių liekanos yra stacionarios. Ljungo ir Bokso testas parodė, jog liekanos nepasižymi autokoreliacija ($p \approx 0,928$ baziniam ir $p \approx 0,918$ „Google Trends“ duomenis naudojančiam modeliui). Vienos imties t-testas leidžia teigti, jog modelių prognozės nėra paslinktos – p reikšmės $p_{\text{bazinis}} \approx 0,745$ ir $p_{\text{su „Google Trends“}} \approx 0,758$. Atsižvelgiant į testų rezultatus galima daryti išvadą, jog abu modeliai tenkina prielaidas.

Modelių tikslumo metrikų reikšmės pateiktos 11 lentelėje. Modeliai pasižymi labai panašiu tikslumu – antrasis modelis pranašesnis MAPE ir RMSE metrikų atžvilgiu, tačiau skirtumas tarp modelių vidutinių absoliutinių procentinių paklaidų yra tik 0,1 %.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,029	7046	10132,3
Su „Google Trends“	0,028	7288	9824,57

11 lentelė. SARIMAX modelio tikslumas prognozuojant visų turistų skaičių viešbučio tipo įstaigose

21 paveiksle pateiktame grafike galima pastebėti, jog abiejų modelių prognozės yra labai panašios – tai gali paaiškinti ir panašumus tarp modelių tikslumo metrikų rezultatų.



21 pav. Visų turistų skaičius viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Stovyklavietės tipo apgyvendinimo įstaigos

Antroji apgyvendinimo įstaigų kategorija – „Stovyklavietės ir kt.“. Visų turistų skaičius šioje kategorijoje pasižymėjo didžiausiu koreliacijos koeficientu su paieškos užklausomis „camping in Lithuania“, „Vilnius“, „hotels Lithuania“, „flights to Lithuania“, „Kaunas“ – būtent šios užklauskos naudojamos modelyje su „Google Trends“ duomenimis. Taip pat nustatyta, jog optimalus pagrindinių komponentų skaičius, naudotinas šiame modelyje, yra 12.

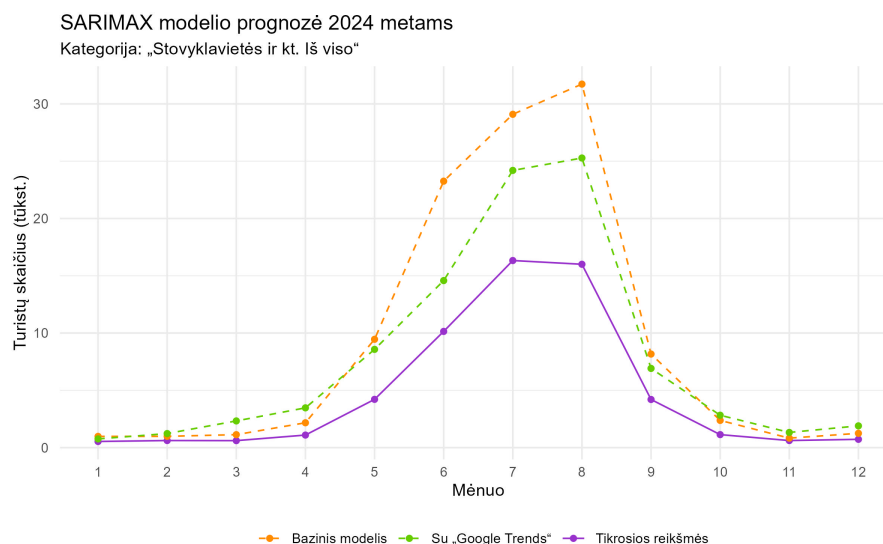
Statistiniai testai, taikyti modelių liekanoms, parodė, jog abu modeliai tenkina prielaidas – ADF testo abiejų modelių p reikšmės $p < 0,01$, Ljungo-Bokso testo rezultatai – $p_{\text{bazinis}} \approx 0,999$ ir $p_{\text{su „Google Trends“}} \approx 0,07$, t-testui gautos p reikšmės $p_{\text{bazinis}} \approx 0,674$ ir $p_{\text{su „Google Trends“}} \approx 0,888$.

12 lentelėje pateikti modelių tikslumo rezultatai rodo, jog šioje kategorijoje modelis su „Google Trends“ duomenimis pasižymi aukštesne MAPE metrikos reikšme, tačiau lenkia bazinį modelį MAE ir RMSE metrikų atžvilgiu. Visgi, abiejų modelių paklaidos yra labai aukštos.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,887	4599	7239,3
Su „Google Trends“	1,151	3096	4160,08

12 lentelė. SARIMAX modelio tikslumas prognozuojant visų turistų skaičių stovyklavietės tipo įstaigose

Prognozių grafikas, pateiktas 22 paveiksle, paaiškina gautas metrikų reikšmes – aukšta MAPE metrikos reikšmė modeliui su „Google Trends“ duomenimis gauta, tikėtina, dėl to, nes jis itin stipriai pervertina turistų skaičius kovo ir balandžio mėnesiais, kuriais tikrieji skaičiai yra labai žemi. Gegužės–rugsėjo mėnesiais šio modelio prognozės yra žymiai arčiau tikrųjų reikšmių, kas paaiškina žemesnes MAE ir RMSE metrikų reikšmes. Taigi, modelis su „Google Trends“ duomenimis yra tikslesnis prognozuojant turistų skaičius mėnesiais, kurie yra paklausiausi apgyvendinimo įstaigoms šioje kategorijoje, tačiau prognozės vis dar yra stipriai nutolę nuo tikrųjų reikšmių.



22 pav. Visų turistų skaičiaus stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Kitas apgyvendinimas

Pagal koreliacijos su visų turistų skaičiumi apgyvendinimo įstaigose kategorijoje „Kitas apgyvendinimas“ koeficientu modelyje su „Google Trends“ duomenimis pasirinkta naudoti „Google“ paieškos užklausas „camping in Lithuania“, „Vilnius“, „Kaunas“, „things to do in Lithuania“ ir „flights to Lithuania“. Be to, nustatyta, jog optimaliausia šiame modelyje naudoti vieną pagrindinę komponentę.

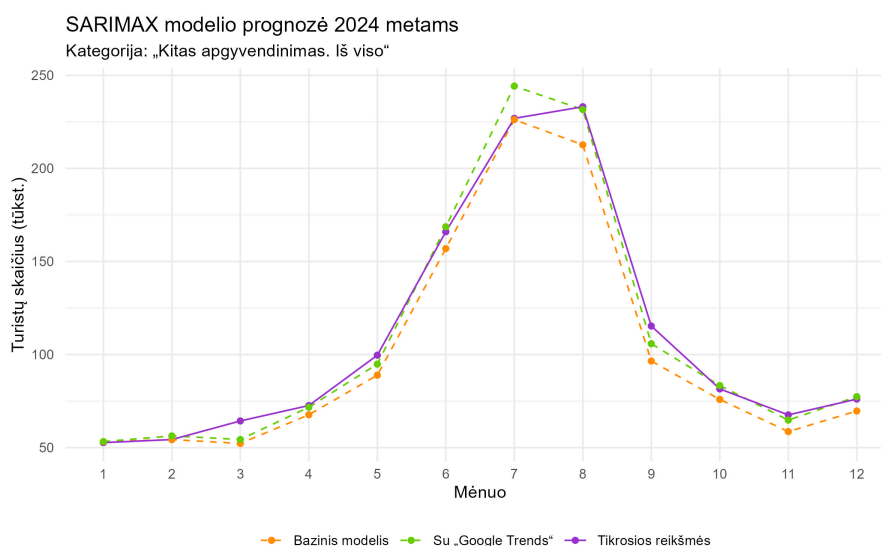
Modelių prielaidų tikrinimui naudotų statistinių testų p reikšmės leidžia daryti išvadą, jog abu modeliai tenkina prielaidas (ADF testo $p < 0,01$ abiem modeliams, Ljungo-Bokso testo p reikšmės $p_{\text{bazinis}} \approx 0,581$ ir $p_{\text{su „Google Trends“}} \approx 0,534$ ir t-testui gautos p reikšmės $p_{\text{bazinis}} \approx 0,368$ ir $p_{\text{su „Google Trends“}} \approx 0,708$).

Modelių tikslumo metrikų reikšmės, pateiktos 13 lentelėje rodo, jog „Google Trends“ duomenis naudojantis modelis šioje kategorijoje yra žymiai pranašesnis visų metrikų atžvilgiu.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,081	8188	10384,16
Su „Google Trends“	0,044	4590	6708,63

13 lentelė. SARIMAX modelio tikslumas prognozuojant visų turistų skaičių kitose įstaigose

23 paveiksle pateiktas prognozių grafikas patvirtina tikslumo metrikų rezultatus – modelis su „Google Trends“ duomenimis yra pastebimai tikslesnis. Bazinis modelis arčiau tikrųjų reikšmių yra tik liepos mėnesį, kurio antrojo modelio prognozė stipriai pervertina turistų skaičių. Kovo ir rugsėjo mėnesiais abu modeliai yra žymiai nukrypę nuo tikrųjų reikšmių, tačiau ir šiais mėnesiais „Google Trends“ duomenis naudojantis modelis yra tikslesnis.



23 pav. Visų turistų skaičius kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Visos apgyvendinimo įstaigos

Galiausiai, prognozuojamas visų turistų skaičius visose apgyvendinimo įstaigose bendrai. Kaip ir ankstesniuose skyriuose, šiai kategorijai naudojamos dvi metodikos: sudaryti atskirus modelius turistų skaičiaus prognozavimui, arba agreguoti atskirų apgyvendinimo įstaigų kategorijų prognozes.

Pirmiausiai pateikiami pirmos metodikos rezultatai. Modelyje su „Google Trends“ duomenimis naudojamos užklauskos „Vilnius“, „camping in Lithuania“, „Kaunas“, „things to do in Lithuania“, „flights to Lithuania“. Renkant optimalų pagrindinių komponentių skaičių pažingsniui, nuspręsta modelyje naudoti tik vieną komponentę.

Modelio prielaidos buvo tikrinamos naudojant statistinius testus. Pritaikius ADF testą, abiem modeliams gautos p reikšmės $p < 0,01$, tad liekanos atitinka atsitiktinį klaidžiojimą. Ljungo ir Bokso testo rezultatai ($p_{\text{bazinis}} \approx 0,655$ ir $p_{\text{su „Google Trends“}} \approx 0,661$) rodo, jog liekanos nepasižymi autokoreliacija. Vienos imties t-testo p reikšmės $p_{\text{bazinis}} \approx 0,393$ ir $p_{\text{su „Google Trends“}} \approx 0,884$ leidžia teigti, jog modelių prognozės nėra paslinktos. Testai parodė, jog abu modeliai tenkina visas prielaidas.

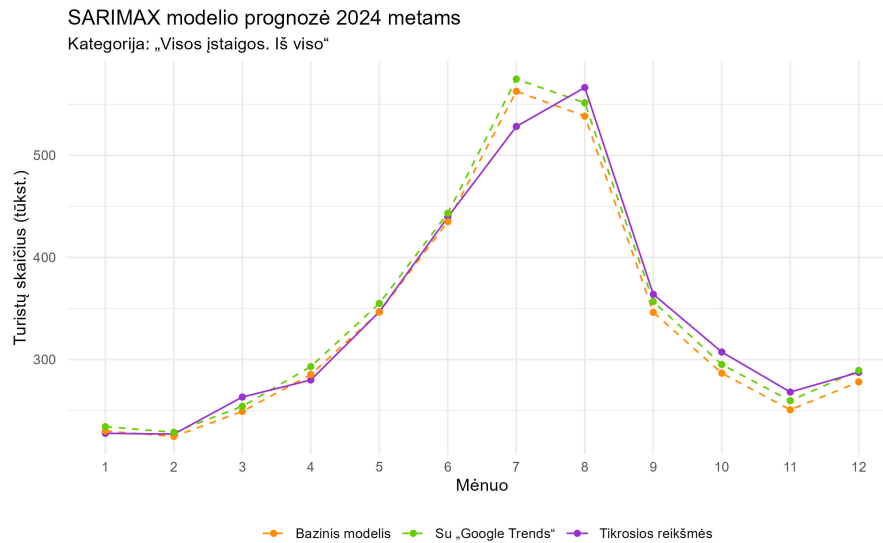
Iš modelių tikslumo metrikų reikšmių (žr. 14 lentelę) galima daryti išvadą, jog šioje kategorijoje modelis su įtrauktais „Google Trends“ duomenimis yra tikslesnis visų metrikų atžvilgiu.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,036	13053	16760,81
Su „Google Trends“	0,03	11137	15921,3

14 lentelė. SARIMAX modelio tikslumas prognozuojant visų turistų skaičių visose įstaigose

24 paveiksle pateiktas prognozių grafikas atskleidžia, jog modelio su „Google Trends“ duomenimis prognozės itin stipriai nuo tikrųjų reikšmių nukrypsta tik liepos mėnesį, tuo tarpu bazinio modelio paklaidos yra didelės visais mėnesiais, pradedant liepa. Skirtumus tarp modelių tikslumo gali lemti tai, jog mėnesiais, kai bazinis modelis yra artimesnis tikrosioms reikšmėms, antrasis modelis taip pat nepasižymi didelėmis paklaidoms (pvz. sausio, kovo, balandžio mėn.), tačiau mėnesiais, kurių antro-

jo modelio prognozės yra tikslesnės (rugpjūčio–gruodžio mėn.), bazinis modelis stipriai atitolsta nuo tikrųjų reikšmių.



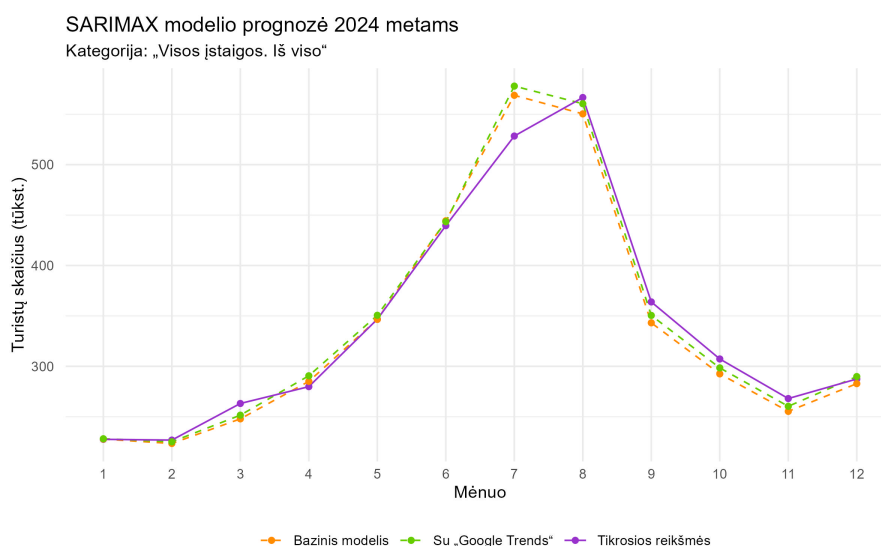
24 pav. Visų turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Taikant antrąją prognozavimo metodiką, „Google Trends“ duomenis naudojantis modelis vėl yra tikslesnis MAPE ir MAE metrikų atžvilgiu, tačiau šio modelio RMSE reikšmė yra nežymiai didesnė, lyginant su baziniu modeliu (žr. 15 lentelę).

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,031	11545	15863,41
Su „Google Trends“	0,027	10045	16065,95

15 lentelė. SARIMAX modelio tikslumas prognozuojant visų turistų skaičių visose įstaigose, naudojant ankstesnių prognozių agregavimą

Iš 25 paveiksle pateikto grafiko galima pastebėti, jog modelių prognozės yra labai panašios į gautas naudojant ankstesnę metodiką, tačiau šį kartą skirtumai tarp prognozių yra mažesni, kas, tikriausiai, ir lemia tikslumo metrikų reikšmių panašumus. Liepos mėnesį abu modeliai stipriai nutolsta nuo tikrųjų reikšmių, tačiau bazinis modelis yra šiek tiek tikslesnis – šio mėnesio paklaida ir gali lemti aukštesnę modelio su „Google Trends“ duomenimis RMSE metrikos reikšmę. Visgi, grafikas parodo, jog antrasis modelis yra tikslesnis už bazinį.



25 pav. Visų turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

Skyrelyje aptarti rezultatai rodo, jog „Google Trends“ duomenų įtraukimas į prognostinius SARIMAX modelius gali pagerinti modelių tikslumą prognozuojant visų turistų skaičių visose apgyvendinimo įstaigų kategorijose.

5.2. Atraminių vektorių regresijos modelių rezultatai

Viešbučių tipo apgyvendinimo įstaigos

Atraminių vektorių regresijos modelių tikslumo metrikų rezultatai, pateikti 16 lentelėje, rodo, jog bazinis modelis šioje kategorijoje yra žymiai tikslesnis visų metrikų atžvilgiu.

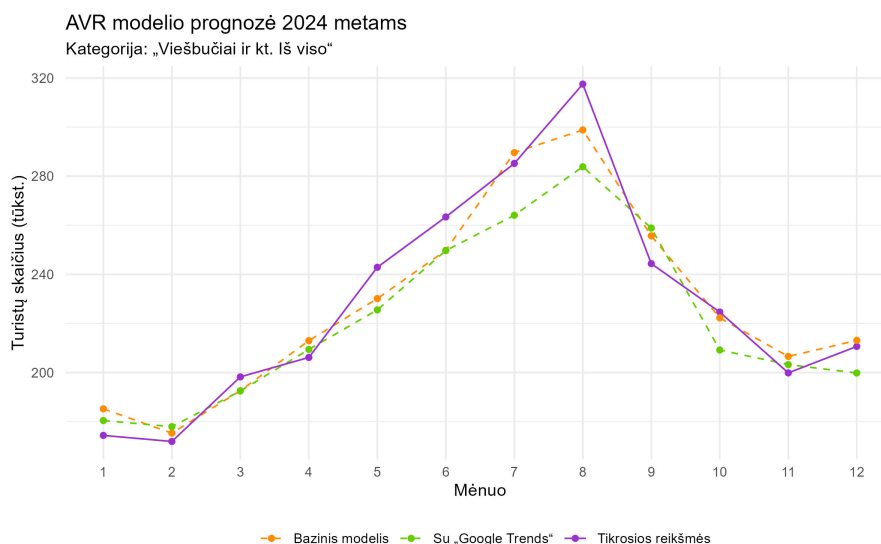
Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,035	8250	9611,44
Su „Google Trends“	0,051	12591	15171,87

16 lentelė. AVR modelio tikslumas prognozuojant visų turistų skaičių viešbučio tipo įstaigose

26 paveiksle pateiktame grafike pavaizduotos modelių prognozės atskleidžia, jog nei vienas iš modelių neatspindi tikrųjų reikšmių tendencijos. Modelis su „Google Trends“ duomenimis yra itin stipriai nutolęs nuo tikrųjų reikšmių gegužės–spalio ir gruodžio mėnesiais. Tačiau bazinis modelis, kuris pagal aptartas metrikas buvo žymiai tikslesnis, vis dar pasižymi didelėmis paklaidomis sausio, gegužės, birželio, rugpjūčio ir rugsėjo mėnesiais. Visgi, prognozių grafikas patvirtina, jog šis modelis yra tikslesnis.

Stovyklaviečių tipo apgyvendinimo įstaigos

17 lentelėje pateiktos modelių tikslumo metrikos prognozuojant visų turistų skaičių „Stovyklavietės ir kt.“ kategorijos apgyvendinimo įstaigose. Iš rezultatų galima pastebėti, jog abu modeliai



26 pav. Visų turistų skaičius viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

pasižymi labai didelėmis paklaidomis, tačiau bazinis modelis yra tikslesnis. Tačiau paklaidos vis dar yra tokios didelės, kad net ir šis modelis nebūtų tinkamas prognozėms.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	1,139	3328	4381,12
Su „Google Trends“	1,508	3719	4790,93

17 lentelė. AVR modelio tikslumas prognozuojant visų turistų skaičių stovyklavietės tipo įstaigose

27 paveiksle pateiktas grafikas patvirtina tikslumo metrikų rezultatus – abu modeliai labai stipriai pervertina turistų skaičių visu laikotarpiu. Vienintelė prognozė, kuri yra arti tikrosios reikšmės – bazinio modelio prognozė lapkričio mėnesiui. Tačiau atsižvelgiant į stiprias abiejų modelių paklaidas, šios prognozės nėra naudingos.

Kitos apgyvendinimo įstaigos

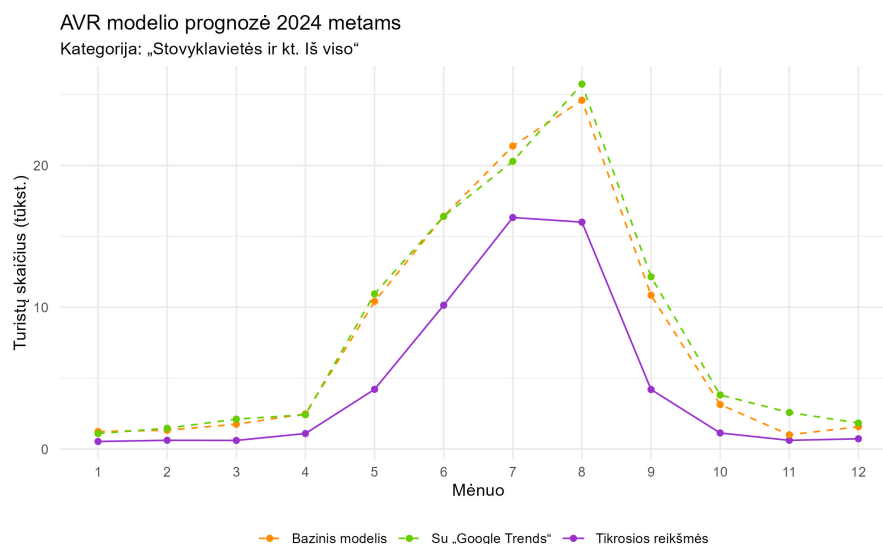
13 lentelėje pateikti tikslumo metrikų rezultatai atskleidžia, jog šioje apgyvendinimo įstaigų kategorijoje modelis su „Google Trends“ duomenimis yra pranašesnis visų metrikų atžvilgiu. Šiuos rezultatus patvirtina ir 28 paveiksle esantis prognozių grafikas, kuriame „Google Trends“ duomenis naudojantis modelis yra pastebimai tikslesnis.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,058	5372	7017,87
Su „Google Trends“	0,042	3616	4144,83

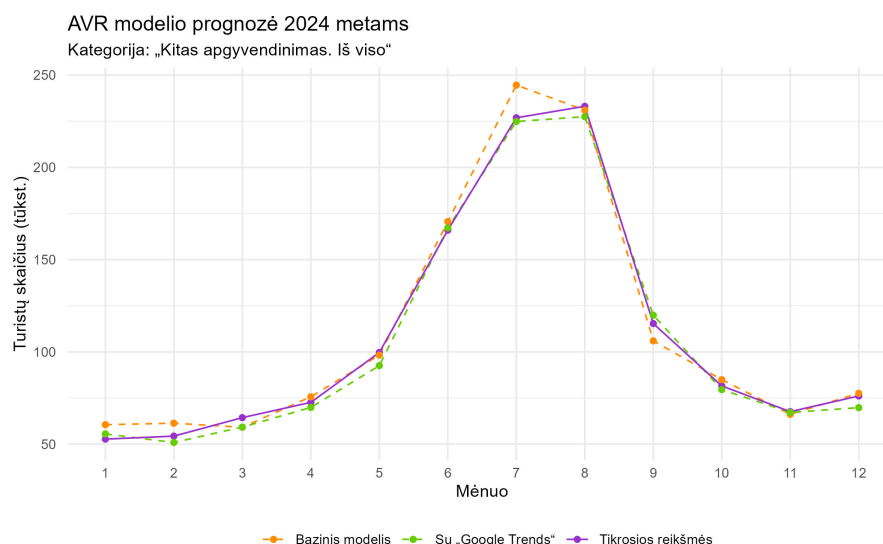
18 lentelė. AVR modelio tikslumas prognozuojant visų turistų skaičių kitose įstaigose

Visos apgyvendinimo įstaigos

Visų turistų skaičius visose Lietuvos apgyvendinimo įstaigose prognozuojamas naudojant dvi anksčiau pristatytas metodikas. Pirmiausiai analizuojamas tikslumas sudarant atskirus modelius tu-



27 pav. Visų turistų skaičiaus stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



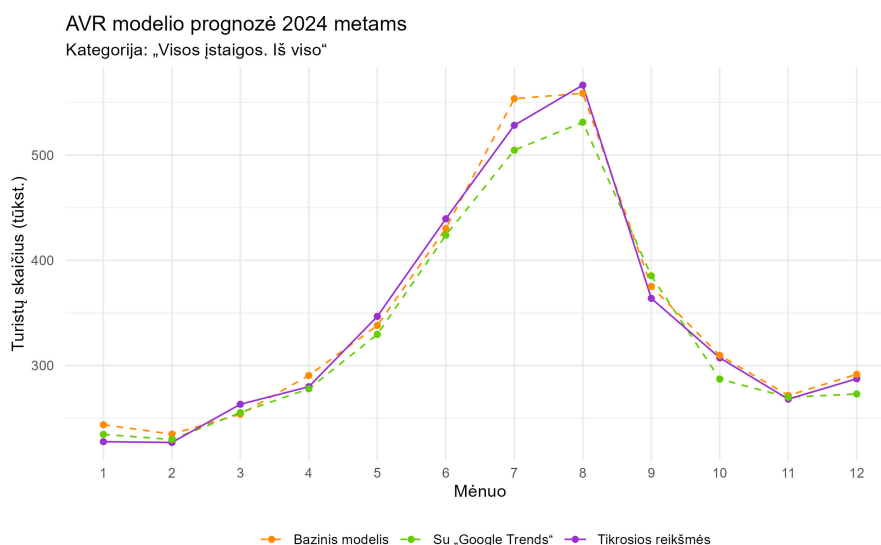
28 pav. Visų turistų skaičiaus kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

ristų skaičiaus prognozavimui, kurių rezultatai pateikti 19 lentelėje. Tikslumo metrikų reikšmės atskleidžia, jog bazinis modelis naudojant šią metodiką yra žymiai tikslesnis visų metrikų atžvilgiu. Skirtumas tarp modelių tikslumo gali būti pastebėtas ir 29 paveiksle pateiktame grafike. Nors sausio, vasario ir balandžio mėnesiais antrojo modelio prognozės yra arčiau tikrųjų reikšmių, jis pasižymi aukštomis paklaidomis liepos, rugpjūčio, rugsėjo, spalio ir gruodžio mėnesiais, tad bazinis modelis yra pastebimai tikslesnis.

Antrosios metodikos – prognozių sudarymo agreguojant atskirų apgyvendinimo įstaigų kategorijų prognozes – tikslumo metrikų rezultatai pateikti 20 lentelėje. Iš rezultatų galima daryti išvadą, jog ir naudojant šią metodiką bazinis modelis pasižymi žemesnėmis paklaidomis visų metrikų atžvilgiu, lyginant su „Google Trends“ duomenis naudojančiu modeliu.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,029	9700	11364,33
Su „Google Trends“	0,038	14105	17188,58

19 lentelė. AVR modelio tikslumas prognozuojant visų turistų skaičių visose įstaigose

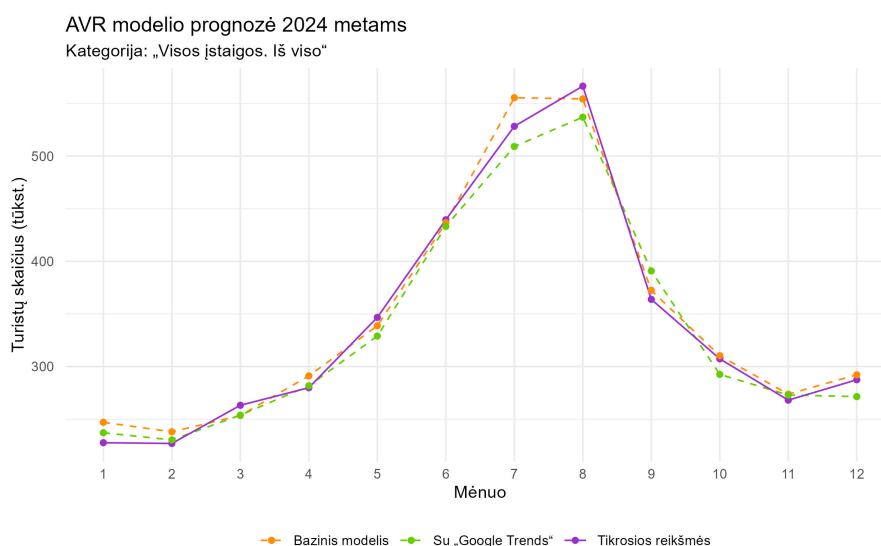


29 pav. Visų turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,032	10291	12291,55
Su „Google Trends“	0,037	13290	15846,68

20 lentelė. AVR modelio tikslumas prognozuojant visų turistų skaičių visose įstaigose, naudojant anksčiau prognozių agregavimą

30 paveiksle pateiktame prognozių grafike galima pastebėti, jog rezultatai yra panašūs anksčiau aptartiems. Tačiau naudojant šią metodiką, nežymiai sumažėja modelio su įtrauktais „Google Trends“ duomenimis prognozių paklaidos birželio, liepos, rugsėjo ir spalio mėnesiais. Visgi, bazinio modelio prognozės šiais mėnesiais taip pat kiek labiau priartėjo prie tikrųjų reikšmių. Taigi, grafikas patvirtina, jog naudojant ir šią metodiką, bazinis modelis yra tikslesnis.



30 pav. Visų turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

Visų turistų skaičiaus prognozavimo naudojant AVR modelius rezultatai rodo, jog „Google Trends“ duomenys nepagerina modelių tikslumo visose apgyvendinimo įstaigų kategorijose. Užklausų populiarumo duomenų įtraukimas į modelį jo paklaidas sumažino tik kategorijoje „Kitas apgyvendinimas“. Taigi, prognozuojant visų turistų skaičių atraminių vektorių regresijos atveju, galima gauti tikslesnius ir patikimesnius rezultatus naudojant tik istorinius turistų skaičiaus duomenis.

5.3. Modelių palyginimas

Modelių tikslumo prognozuojant visų turistų skaičių Lietuvos apgyvendinimo įstaigose rezultatai parodė, jog nei vienas iš modelių nėra tikslesnis visose tirtose kategorijose. Naudojant AVR modelį, galima gauti kiek tikslesnes prognozes kategorijose „Kitas apgyvendinimas“ ir sudarant modelį be „Google Trends“ duomenų visų turistų skaičiaus visose įstaigose prognozavimui. Tačiau šis modelis pasižymėjo išskirtinai didelėmis paklaidomis kategorijoje „Stovyklavietės ir kt.“ (žemiausia MAPE reikšmė siekė net 113,9 %), kurioje bazinio SARIMAX modelio prognozės buvo žymiai tikslesnės (vidutinė absoliutinė procentinė paklaida 88,7 %). Taip pat SARIMAX modelis pasižymėjo aukštesniu tikslumu likusiose apgyvendinimo įstaigų kategorijose, tad ir prognozuojant visų turistų skaičių, šis modelis yra tinkamesnis.

6. Užsienio turistų skaičiaus prognozavimas

Užsienio turistų skaičiaus prognozavimui nebuvo sudaromi atskiri modeliai, nes pasaulinio masto „Google Trends“ duomenys įtraukia ir paieškas, vykdytas Lietuvoje, o užklausų populiarumo duomenys atskiroms šalims yra sunkiai prieinami – kiekvienos šalies duomenis reikėtų rinkti rankiniu būdu, kadangi tinklapio duomenų kopijavimas (angl. *web scraping*) šiuo metu nėra leidžiamas ar prieinamas „Google Trends“ tinklapyje. Taip pat „Google Trends“ duomenims iš atskirų valstybių būtų optimalu naudoti užklausas valstybine arba plačiausiai dominančioje šalyje naudojama kalba, tačiau ši metodika yra už atliekamo tyrimo ribų. Be to, kiekvienais metais šalių, iš kurių į Lietuvą atvyksta turistai, aibė keičiasi, ir yra sunku nustatyti, kurių valstybių duomenys yra aktualūs.

Dėl šių priežasčių pasirinkta užsienio turistų skaičių prognozuoti skaičiuojant skirtumą tarp visų turistų ir vietinių turistų skaičiaus tiriamoje kategorijoje. Taikant šią metodiką, išlaikomas ir duomenų nuoseklumas.

6.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai

Užsienio turistų skaičiaus prognozavimui nėra konstruojami nauji modeliai, tad SARIMAX modelių prielaidos šiame skyriuje nėra tikrinamos.

Viešbučio tipo apgyvendinimo įstaigos

Užsienio turistų skaičiaus apgyvendinimo įstaigose, priklausančiose kategorijai „Viešbučiai ir kt.“, prognozavimo rezultatai pateikti 21 lentelėje. Tikslumo metrikų reikšmės rodo, jog bazinis modelis yra žymiai tikslesnis už antrąjį modelį.

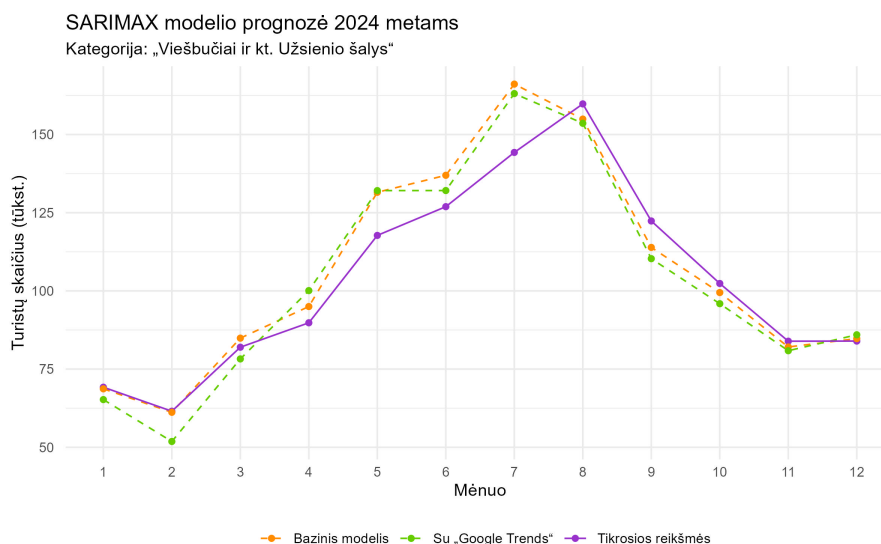
Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,051	6122	8715,27
Su „Google Trends“	0,077	7978	9371,65

21 lentelė. SARIMAX modelio tikslumas prognozuojant užsienio turistų skaičių viešbučio tipo įstaigose

31 paveiksle pateiktas prognozių grafikas. Abu modeliai kovo–liepos ir rugsėjo mėnesį pasižymi gana aukštomis paklaidomis. Tačiau bazinio modelio prognozės yra artimesnės tikroms reikšmėms kovo, balandžio ir rugpjūčio–gruodžio mėnesiais bei beveik tiksliai atitinka tikrąsias reikšmes sausio ir vasario mėnesiais, tad šis modelis yra pastebimai tikslesnis prognozuojant užsienio turistų skaičių viešbučio tipo apgyvendinimo įstaigose.

Stovyklavietės tipo apgyvendinimo įstaigos

Toliau analizuojamas modelių tikslumas prognozuojant užsienio turistų skaičių „Stovyklavietės ir kt.“ kategorijos apgyvendinimo įstaigose. Tikslumo metrikų rezultatai (žr. 22 lentelę) rodo, jog modelis su „Google Trends“ duomenimis šioje kategorijoje yra netinkamas – jo paklaidos visų metrikų atžvilgiu yra labai aukštos, o vidutinės absoliutinės procentinės paklaidos reikšmė siekia net 2524,5 %. Tiesa, bazinio modelio MAPE reikšmė taip pat yra labai aukšta, tačiau kaip jau buvo pastebėta,



31 pav. Užsienio turistų skaičiaus viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

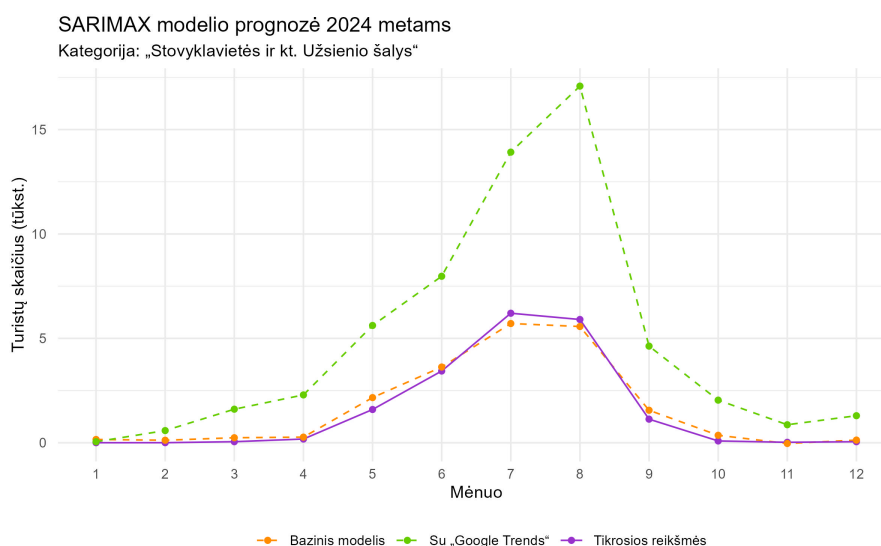
tokios reikšmės gali būti nulemtos mažų turistų skaičių žiemos mėnesiais, kurie nėra tokie svarbūs apgyvendinimo įstaigoms šioje kategorijoje, tad „Stovyklavietės ir kt.“ kategorijoje ši metrika gali būti šiek tiek apgaulinga. Bazinio modelio MAE ir RMSE metrikų reikšmės yra sąlyginai žemos, tad šis modelis yra tinkamesnis.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	7,756	248	297,68
Su „Google Trends“	25,245	2885	3892,69

22 lentelė. SARIMAX modelio tikslumas prognozuojant užsienio turistų skaičių stovyklavietės tipo įstaigose

32 paveiksle pateiktas prognozių grafikas patvirtina tikslumo metrikų rezultatus – bazinis modelis yra žymiai tikslesnis visais mėnesiais. Modelis su „Google Trends“ duomenimis itin stipriai pervertina turistų skaičių, o didžiausi nukrypimai pastebimi liepos ir rugpjūčio mėnesiais, kurie apgyvendinimo įstaigoms šioje kategorijoje yra svarbiausi ir paklausiausi. Bazinis modelis taip pat nukrypsta nuo tikrųjų reikšmių, tačiau paklaidos nėra didelės. Šis modelis pervertina turistų skaičių sausio, vasario, kovo ir spalio mėnesiais, kuriais tikrosios reikšmės yra itin žemos, kas galėtų paaiškinti aukštą MAPE metrikos reikšmę.

Ankstesni rezultatai parodė, kad bazinis modelis yra žymiai mažiau tikslus prognozuojant tiek lietuvių, tiek visų turistų skaičių stovyklavietės tipo įstaigose – abiem atvejais šis modelis labai stipriai pervertino turistų skaičių. Tačiau skaičiuojant skirtumą pervertintas vietinių turistų skaičius atsvėrė pervertinimą, gautą prognozuojant visų turistų skaičių, tad gautos prognozės užsienio turistams pasižymi sąlyginai mažomis paklaidomis. Tuo tarpu modelis su „Google Trends“ duomenimis pasižymėjo aukštu tikslumu prognozuojant vietinių turistų skaičių šioje kategorijoje, tačiau pervertino visų turistų skaičių. Skaičiuojant prognozes užsienio turistams gautos itin aukštos paklaidos atskleidžia šios prognozavimo metodikos trūkumą – ji labiau baudžia modelius, kurių viena iš prognozių buvo labai tiksli, nei modelius, kurių abi prognozės buvo panašiai stipriai nukrypę nuo tikrųjų reikšmių.



32 pav. Užsienio turistų skaičiaus stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Kitos apgyvendinimo įstaigos

23 lentelėje pateikti modelių tikslumo metrikų rezultatai kategorijai „Kitas apgyvendinimas“. Iš metrikų reikšmių galima daryti išvadą, jog šioje kategorijoje bazinis modelis vėl yra žymiai tikslus, tačiau jo paklaidos vis dar yra gana aukštos. Modelio su „Google Trends“ duomenimis tikslumo rezultatai leidžia teigti, jog šis modelis yra netinkamas užsienio turistų skaičiaus prognozavimui.

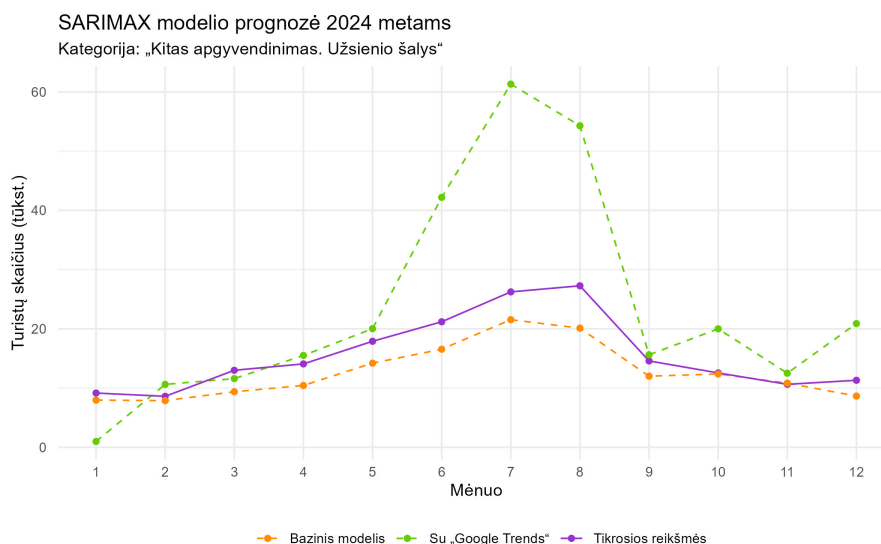
Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,172	2914	3539,51
Su „Google Trends“	0,422	7324	10432,78

23 lentelė. SARIMAX modelio tikslumas prognozuojant užsienio turistų skaičių kitose įstaigose

Iš prognozių grafiko, pateikto 33 paveiksle, galima daryti tas pačias išvadas – modelio su „Google Trends“ duomenimis prognozės visiškai neatitinka tikrųjų reikšmių sausio, birželio, liepos, rugpjūčio, spalio ir gruodžio mėnesiais. Bazinio modelio prognozių paklaidos yra žymiai mažesnės, ką parodė ir tikslumo metrikų analizė, tačiau šis modelis yra linkęs gana stipriai nuvertinti turistų skaičių kovo–rugsėjo bei gruodžio mėnesiais.

Visos apgyvendinimo įstaigos

Galiausiai, prognozuojamas užsienio turistų skaičius visose apgyvendinimo įstaigose. Šiai daliai naudojamos dvi metodikos: skaičiuoti skirtumą tarp prognozių, gautų konstruojant atskirus modelius turistų iš Lietuvos ir visų šalių skaičiaus prognozavimui (t. y. pirmosios metodikos vietinių ir visų turistų skaičiaus prognozavimui visose įstaigose rezultatų), ir skaičiuoti skirtumą tarp prognozių, gautų agreguojant atskirų kategorijų prognozes turistams iš Lietuvos ir viso pasaulio (t. y. naudojant antrąją metodiką vietinių ir visų turistų skaičiaus prognozavimui visose įstaigose). Taip pat galima ir trečia metodika – agreguoti anksčiau gautas užsienio turistų skaičiaus prognozes atskiroms apgyvendinimo



33 pav. Užsienio turistų skaičiaus kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

įstaigų kategorijoms, tačiau taikant šį metodą gaunami rezultatai, analogiški antrajai metodikai, tad jis nebus naudojamas.

Pirmosios metodikos tikslumo rezultatai, pateikti 24 lentelėje, rodo, jog ir prognozuojant užsienio turistų skaičių visose apgyvendinimo įstaigose modelis su „Google Trends“ duomenimis pasižymi aukštesnėmis paklaidomis visų metrikų atžvilgiu. Šiuos rezultatus patvirtina ir 34 paveiksle pateiktas grafikas – nors abiejų modelių prognozių tendencijos yra panašios, bazinis modelis yra žymiai tikslesnis kovo–birželio mėnesiais.

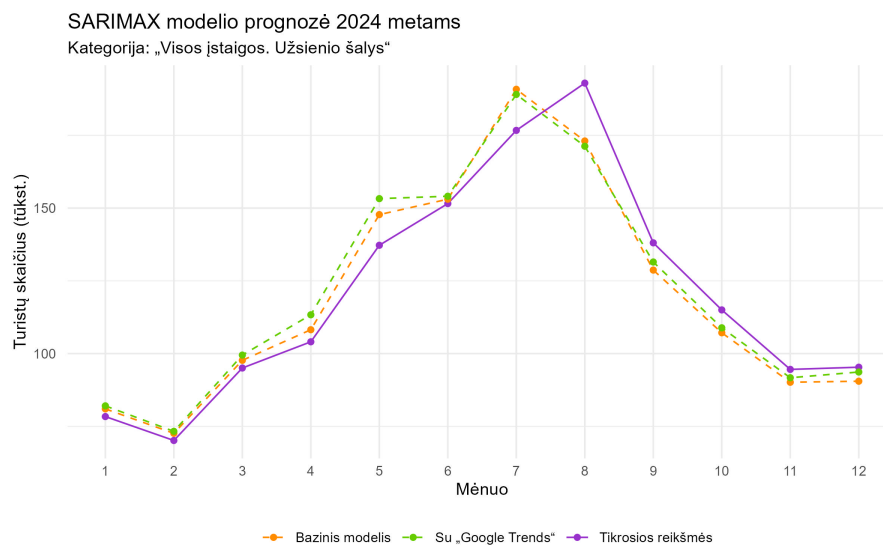
Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,053	7031	8837,97
Su „Google Trends“	0,058	7532	9608,37

24 lentelė. SARIMAX modelio tikslumas prognozuojant užsienio turistų skaičių visose įstaigose

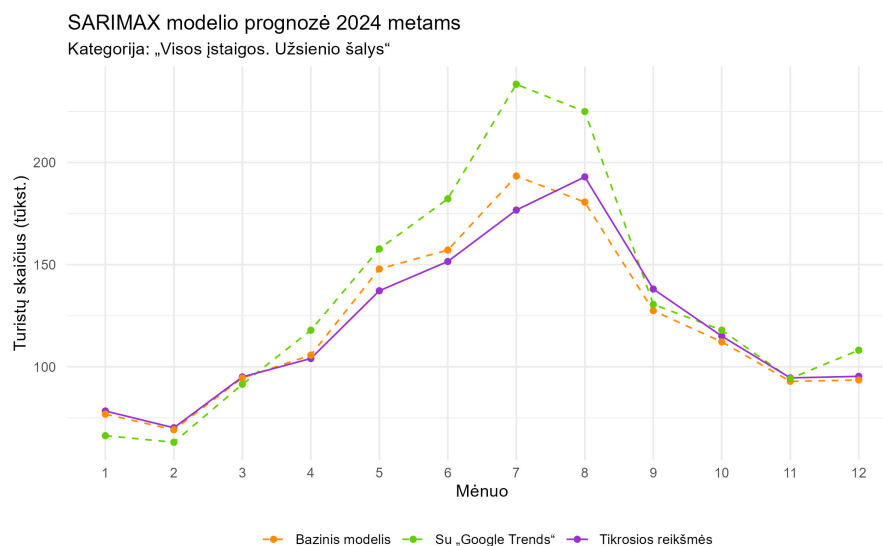
Antrosios metodikos modelių tikslumo rezultatai pateikti 25 lentelėje. Taikant šią metodiką, skirtumai tarp modelių yra didesni, tačiau bazinis modelis vis tiek yra žymiai tikslesnis visų metrikų atžvilgiu. Didesni skirtumai tarp modelių tikslumo pastebimi ir 35 paveiksle pateiktame prognozių grafike. Modelio su „Google Trends“ duomenimis prognozės pasižymi aukštomis paklaidomis visais mėnesiais, išskyrus kovą, spalį ir lapkritį. Šis modelis itin stipriai pervertina užsienio turistų skaičių balandžio–rugpjūčio mėnesiais. Bazinio modelio prognozės yra žymiai arčiau tikrųjų reikšmių, nors ir šis modelis pasižymi gana aukštomis paklaidomis gegužės–rugsėjo mėnesiais.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,039	5573	7680,04
Su „Google Trends“	0,114	14742	19064,54

25 lentelė. SARIMAX modelio tikslumas prognozuojant užsienio turistų skaičių visose įstaigose, naudojant atskirų kategorijų prognozių agregavimą



34 pav. Užsienio turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



35 pav. Užsienio turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant atskirų kategorijų prognozių agregavimą

Užsienio turistų skaičiaus prognozavimo, naudojant SARIMAX modelį, rezultatai atskleidė, jog „Google Trends“ duomenys nepagerino tikslumo nei vienoje iš apgyvendinimo įstaigų kategorijų. Tokie rezultatai gali būti nulemti pasaulinio masto paieškų užklausų populiarumo duomenų trūkumo – dėl šios priežasties pasirinktos užklausos gali būti netinkamos sprendžiamam uždaviniui ir suteikti modeliams klaidingą informaciją.

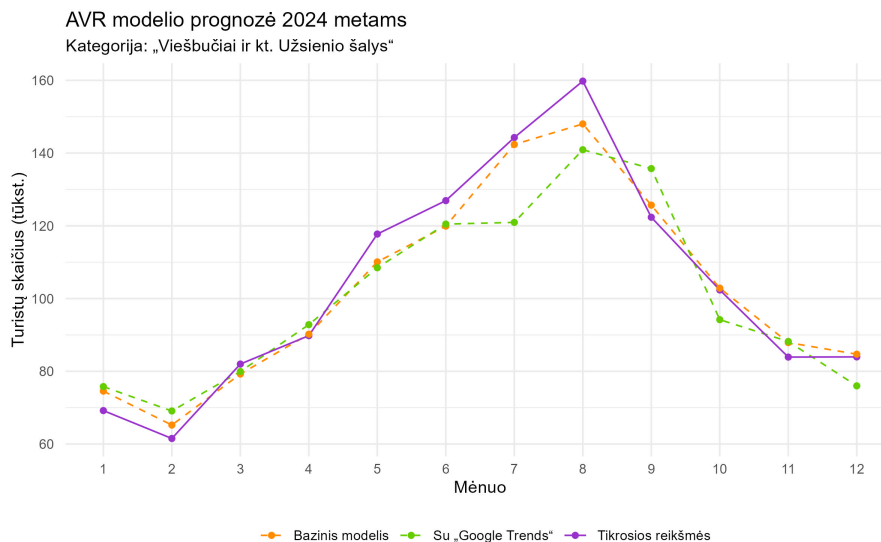
6.2. Atraminių vektorių regresijos modelių rezultatai

Viešbučio tipo apgyvendinimo įstaigos

Modelių tikslumo metrikų reikšmės apgyvendinimo įstaigų kategorijai „Viešbučiai ir kt.“, pateiktos 26 lentelėje rodo, jog bazinis modelis yra žymiai tikslesnis visų metrikų atžvilgiu. Šiuose rezultatus patvirtina ir 36 paveiksle pateiktas grafikas, iš kurio galima pastebėti, jog bazinis modelis kiek tiksliau atspindi tikrųjų reikšmių tendenciją, tačiau gana stipriai nuvertina turistų skaičių gegužės, birželio ir rugpjūčio mėnesiais. Visgi, šis modelis yra tikslesnis, atsižvelgiant į tai, jog antrojo modelio prognozės vasario, liepos, rugpjūčio, spalio ir gruodžio mėnesiais visiškai neatitinka tikrųjų reikšmių.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,039	4092	5226,88
Su „Google Trends“	0,085	9246	11065,57

26 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių viešbučio tipo įstaigose



36 pav. Užsienio turistų skaičiaus viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Stovyklavietės tipo apgyvendinimo įstaigos

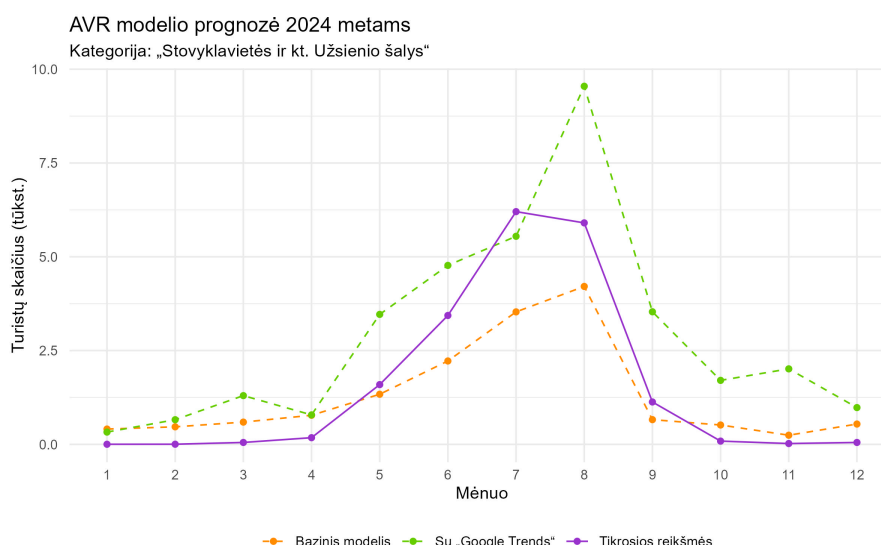
Užsienio turistų skaičiaus apgyvendinimo įstaigose, priklausančiose kategorijai „Stovyklavietės ir kt.“, prognozių tikslumo metrikos pateiktos 27 lentelėje. Rezultatai atskleidžia, jog abu modeliai

šioje kategorijoje pasižymi itin aukštomis MAPE metrikos reikšmėmis. Abiejų modelių MAE ir RMSE metrikos taip pat yra sąlyginai aukštos, tad prognozės yra nepatikimos. Visgi, bazinis modelis šioje kategorijoje pasirodė kiek geriau visų metrikų atžvilgiu. Taip pat analizės metu pastebėtas ir šios prognozių skaičiavimo metodikos trūkumas – modeliai, kurių tik viena iš prognozių pasižymi aukštomis paklaidomis, yra baudžiami labiau, nei modeliai, kurių abi prognozės yra labai netikslios.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	24,26	788	1051,79
Su „Google Trends“	35,924	1440	1699,8

27 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių stovyklavietės tipo įstaigose

Prognozių grafikas, pateiktas 37 paveiksle, patvirtina tikslumo metrikų rezultatus – abu modeliai yra stipriai nukrypę nuo tikrųjų reikšmių. Bazinis modelis kiek geriau atskleidžia tikrųjų reikšmių tendenciją, tačiau yra linkęs stipriai nuvertinti turistų skaičių vasaros mėnesiais, ir žymiai jį pervertinti likusiu laikotarpiu. Modelio su „Google Trends“ duomenimis prognozės visiškai neatitinka tikrųjų reikšmių, ko buvo galima tikėtis iš šio modelio tikslumo metrikų rezultatų.



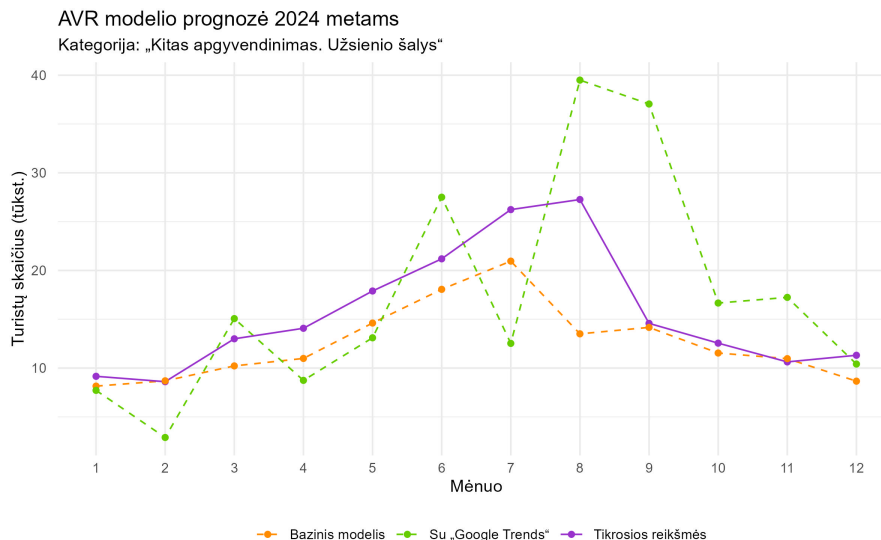
37 pav. Užsienio turistų skaičiaus stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Kitos apgyvendinimo įstaigos

Toliau analizuojamas modelių tikslumas prognozuojant užsienio turistų skaičių kategorijoje „Kitas apgyvendinimas“. Modelių tikslumo metrikų rezultatai rodo, jog ir šioje kategorijoje bazinis modelis yra žymiai tikslesnis visų metrikų atžvilgiu (žr. 28 lentelę). Šie rezultatai patvirtinti ir 38 paveiksle pateikto grafiko – modelio su „Google Trends“ duomenimis prognozės vėl visiškai neatitinka tikrųjų reikšmių. Vieninteliai mėnesiai, kurių šio modelio prognozės pasižymi mažomis paklaidomis, yra sausis ir gruodis. Tuo tarpu bazinis modelis yra pastebimai tikslesnis, nors šio modelio prognozės yra gerokai mažesnės už tikrąsias reikšmes kovo–rugpjūčio bei gruodžio mėnesiais.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,164	3068	4692,89
Su „Google Trends“	0,456	7141	9282,99

28 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių kitose įstaigose



38 pav. Užsienio turistų skaičiaus kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

Visos apgyvendinimo įstaigos

Galiausiai, analizuojamas modelių tikslumas prognozuojant užsienio turistų skaičių visose apgyvendinimo įstaigose. Šiai kategorijai naudojamos dvi metodikos, pristatytos ankstesniame skyrelyje. Modelių tikslumo metrikų rezultatai pirmajai ir antrajai metodikoms pateikti 29 ir 30 lentelėse atitinkamai. Abiejų metodikų rezultatai rodo, jog bazinis modelis ir šioje kategorijoje yra žymiai pranašesnis visų metrikų atžvilgiu. Naudojant antrąją metodiką, kuri yra tinkamesnė dėl jos galimybės išlaikyti duomenų nuoseklumą, galima gauti tikslesnius rezultatus naudojant bazinį modelį, lyginant su pirmąja.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,07	9401	13010,13
Su „Google Trends“	0,09	12520	16490,75

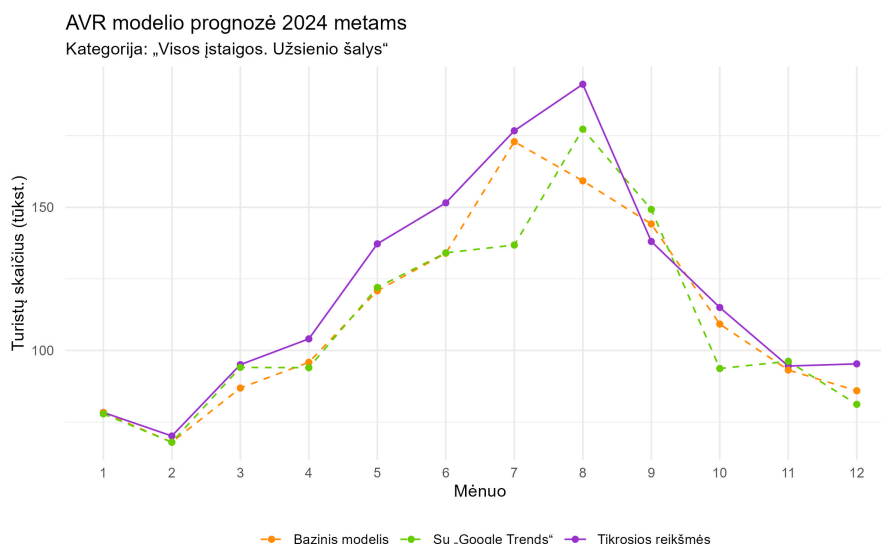
29 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių visose įstaigose

39 paveiksle pateiktas prognozių grafikas atskleidžia, jog taikant pirmąją metodiką bazinis modelis yra pastebimai tikslesnis. Tiesa, modelio su „Google Trends“ duomenimis kai kurių mėnesių (kovo ir rugpjūčio) prognozės yra žymiai arčiau tikrųjų reikšmių, tačiau šis pasižymi itin didelėmis paklaidomis liepos, rugsėjo, spalio ir gruodžio mėnesiais.

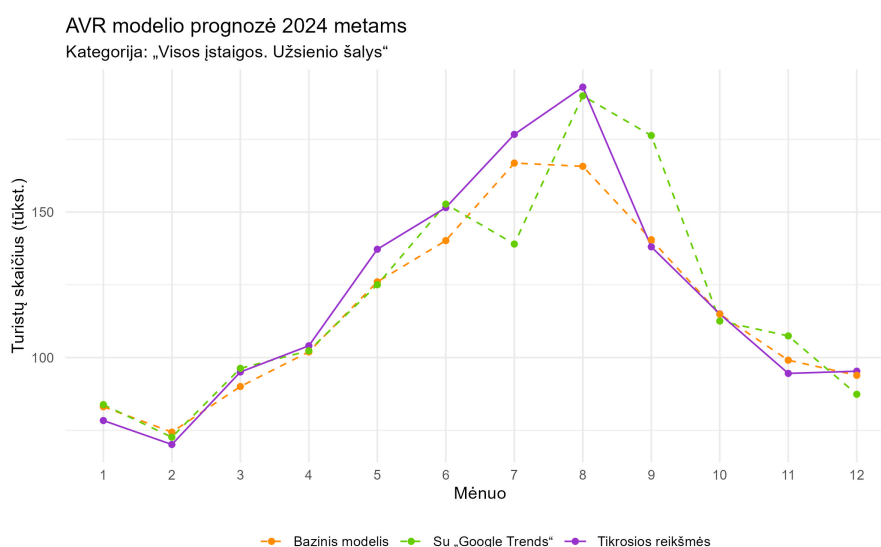
Panašius rezultatus galima matyti ir antrosios metodikos prognozių grafike, pateiktame 40 paveiksle. „Google Trends“ naudojantis modelis yra arčiau tikrųjų reikšmių vasario, kovo ir rugpjūčio mėnesiais, tačiau itin didelės šio modelio paklaidos liepos, rugsėjo, lapkričio ir gruodžio mėnesiais leidžia daryti išvadą, jog bazinis modelis yra tikslesnis.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,052	7017	9962,58
Su „Google Trends“	0,082	10544	16629,78

30 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių visose įstaigose, naudojant atskirų kategorijų prognozių agregavimą



39 pav. Užsienio turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



40 pav. Užsienio turistų skaičiaus visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant atskirų kategorijų prognozių agregavimą

6.3. Modelių palyginimas

Užsienio turistų prognozavimo naudojant SARIMAX ir AVR modelius tikslumo rezultatai atskleidė, jog AVR modelis yra pranašesnis prognozuojant turistų skaičių apgyvendinimo įstaigose, priklausančiose kategorijoms „Viešbučiai ir kt.“ ir „Kitas apgyvendinimas“. Tuo tarpu naudojant SARIMAX modelius galima pasiekti aukštesnį prognozių tikslumą kategorijoje „Stovyklavietės ir kt.“ ir progno-

zuojant užsienio turistų skaičių visose apgyvendinimo įstaigose. Atsižvelgiant į šiuos pastebėjimus, sunku išrinkti modelį, kuris bendrai būtų tinkamesnis.

7. Nakvynių skaičiaus vietiniams turistams prognozavimas

Šiame skyriuje analizuojamas SARIMAX ir AVR modelių tikslumas prognozuojant nakvynių skaičių vietiniams turistams skirtingose Lietuvos apgyvendinimo įstaigų kategorijose. Modeliai konstruojami naudojant metodiką, analogišką taikytai tiriant modelių tinkamumą prognozuojant vietinių turistų skaičių. Šiame bei sekančiuose dviejuose skyriuose prognozių grafikai nėra pateikiami. Juos galima rasti 2 priede (žr. priedą).

7.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai

Viešbučio tipo apgyvendinimo įstaigos

Pirmoji analizuojama apgyvendinimo įstaigų kategorija – „Viešbučiai ir kt.“ Šioje kategorijoje vietinių turistų nakvynių skaičius pasižymėjo didžiausia koreliacija su užklausomis „Vilnius“, „Kaunas“, „apartamentai“, „sodyba“ ir „viešbutis“. Taikant pažingsninį pagrindinių komponentų skaičiaus parinkimą nustatyta, jog optimalus komponentų skaičius yra 14.

Modelių prielaidoms tikrinti taikyti statistiniai testai parodė, jog abu modeliai atitinka prielaidas: ADF testo p reikšmės abiem modeliams $p < 0,01$, Ljungo-Bokso testo p reikšmės $p_{\text{bazinis}} \approx$

31 lentelėje pateikti modelių tikslumo metrikų rezultatai rodo, jog modelis su „Google Trends“ duomenimis šioje kategorijoje pasižymi prastesniu prognozavimo tikslumu visų metrikų atžvilgiu. Bazinis modelis yra kiek tikslesnis, tačiau ir jo paklaidos yra sąlyginai aukštos.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,091	20233	22274,37
Su „Google Trends“	0,098	21786	24075,81

31 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams viešbučio tipo įstaigose

Stovyklavietės tipo apgyvendinimo įstaigos

Toliau analizuojamas modelių tikslumas prognozuojant nakvynių skaičių vietiniams turistams apgyvendinimo įstaigų kategorijoje „Stovyklavietės ir kt.“. Šios kategorijos priklausomas kintamasis pasižymėjo stipriausia koreliacija su paieškos užklausomis „stovyklavietė“, „kempingas“, „lankytinos vietos“, „apartamentai“ ir „nameliai“. Nustatytas optimalus pagrindinių komponentų skaičius modeliui su „Google Trends“ duomenimis – 15.

Modelių prielaidos tikrintos taikant statistinius testus. Pritaikius ADF testą, abiem modeliams gautos p reikšmės $p < 0,01$, tad liekanos yra stacionarios. Ljungo ir Bokso testui gautos p reikšmės $p_{\text{bazinis}} \approx 0,935$ ir $p_{\text{su „Google Trends“}} \approx 0,36$ laidžia teigti, jog liekanos atitinka baltąjį triukšmą. Tuo tarpu t-testo p reikšmės $p_{\text{bazinis}} \approx 0,537$ ir $p_{\text{su „Google Trends“}} \approx 0,921$ rodo, jog modelių prognozės nėra paslinktos. Testų rezultatai leidžia daryti išvadą, jog abu modeliai tenkina visas prielaidas.

Modelių tikslumo metrikų rezultatai testinėje aibėje, pateikti 32 lentelėje atskleidžia, kad bazinis modelis yra pranašesnis MAPE metrikos atžvilgiu, tačiau „Google Trends“ duomenis naudojantis

modelis pasižymi žemesnėmis MAE ir RMSE metrikų reikšmėmis. Šie rezultatai, tikėtina, gali būti paaiškinti tuo, kad antrasis modelis stipriai pervertina nakvynių skaičių mėnesiais, kai tikrosios reikšmės yra labai mažos, tačiau pasižymi didesniu tikslumu mėnesiais, kurie yra paklausiausi įstaigoms šioje kategorijoje, kas buvo pastebėta prognozuojant visų turistų skaičių „Stovyklavietės ir kt.“ kategorijoje taikant SARIMAX modelį. Dėl šios priežasties, antrasis modelis galėtų būti tinkamesnis, tačiau abiejų modelių paklaidos yra aukštos.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,461	2667	4255,27
Su „Google Trends“	0,683	2109	3020,79

32 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams stovyklavietės tipo įstaigose

Kitos apgyvendinimo įstaigos

Trečioji apgyvendinimo įstaigų kategorija – „Kitas apgyvendinimas“. Nustatyta, jog nakvynių skaičius vietiniams turistams šioje kategorijoje pasižymi aukščiausiu koreliacijos koeficientu su užklauseriais „kempingas“, „lankytinos vietos“, „nuoma“, „apartamentai“ ir „sodyba“. Taip pat nustatyta, jog optimaliausia modelyje su „Google Trends“ duomenimis naudoti vieną pagrindinę komponentę.

Modelių liekanoms taikyti statistiniai testai, tikrinantys modelio prielaidas. Panaudojus ADF testą liekanų stacionarumui, gautos p reikšmės abiem modeliams $p < 0,01$. Ljungo-Bokso testas parodė, jog liekanos nepasižymi autokoreliacija ($p_{\text{bazinis}} \approx 0,613$ ir $p_{\text{su „Google Trends“}} \approx 0,699$). Gautos t-testo p reikšmės $p_{\text{bazinis}} \approx 0,778$ ir $p_{\text{su „Google Trends“}} \approx 0,814$ leidžia teigti, jog modelių prognozės nėra paslinktos. Taigi, abu modeliai tenkina prielaidas.

Modelių tikslumo metrikų reikšmės, pateiktos 33 lentelėje atskleidžia, kad modelis su „Google Trends“ duomenimis šioje kategorijoje pasižymi žymiai didesnėmis prognozavimo paklaidomis visų metrikų atžvilgiu.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,046	10559	13404,54
Su „Google Trends“	0,064	15080	19596,86

33 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams kitose apgyvendinimo įstaigose

Visos apgyvendinimo įstaigos

Nakvynių skaičių visose apgyvendinimo įstaigose prognozavimui taikomos tos pačios dvi metodikos, kaip ir prognozuojant turistų skaičių.

Taikant pirmąją metodiką vietinių turistų nakvynių skaičius pasižymėjo didžiausia koreliacija su paieškos užklauseriais „apartamentai“, „kempingas“, „stovyklavietė“, „lankytinos vietos“, „nameliai“. Nustatyta, jog optimalus pagrindinių komponentių skaičius, naudotinas modelyje su „Google Trends“ duomenimis, yra 1.

Modelio prielaidų testai parodė, jog abu modeliai tenkina visas prielaidas – ADF testo rezultatai (p reikšmės abiemis modeliams $p < 0,01$) ir Ljungo-Bokso testui gautos p reikšmės $p_{\text{bazinis}} \approx 0,613$ ir $p_{\text{su „Google Trends“}} \approx 0,699$ leidžia teigti, jog liekanos yra stacionarios ir nepasižymi autokoreliacija. T-testo rezultatai ($p_{\text{bazinis}} \approx 0,613$ ir $p_{\text{su „Google Trends“}} \approx 0,699$) rodo, jog abiejų modelių prognozės yra nepaslinktos.

Pirmosios metodikos rezultatai, pateikti 34 lentelėje rodo, jog modelis su „Google Trends“ duomenimis yra šiek tiek pranašesnis MAPE metrikos atžvilgiu, tačiau pasižymi aukštesnėmis MAE ir RMSE metrikų reikšmėmis. Atsižvelgiant į šiuos rezultatus, galima teigti, jog bazinis modelis yra tinkamesnis.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,032	14653	17897,85
Su „Google Trends“	0,027	16417	27550,46

34 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams visose įstaigose

Antrosios metodikos rezultatai pateikti 35 lentelėje. Iš tikslumo metrikų reikšmių galima pastebėti, jog taikant šią metodiką, kuri yra labiau rekomenduotina dėl jos galimybių išlaikyti logiką duomenyse, modelis, naudojantis „Google Trends“ duomenis, pasižymi žemesnėmis paklaidomis visų metrikų atžvilgiu, tad yra tinkamesnis.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,035	15147	17871,68
Su „Google Trends“	0,031	13890	15885,53

35 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams visose įstaigose, naudojant ankstesnių prognozių agregavimą

Apibendrinant nakvynių skaičiaus prognozavimą vietiniams turistams naudojant SARIMAX modelius, galima teigti, jog šioje srityje „Google Trends“ duomenys modelių tikslumo nepagerino pakankamai, kad būtų galima rekomenduoti jų įtraukimą. Užklausų populiarumo duomenų įtraukimas į modelį tik nežymiai paaukštino prognozavimo tikslumą „Stovyklavietės ir kt.“ kategorijoje bei prognozuojant nakvynių skaičių visose įstaigose, naudojant ankstesnių prognozių agregavimą.

7.2. Atraminių vektorių regresijos modelių rezultatai

Viešbučio tipo apgyvendinimo įstaigos

Atraminių vektorių regresijos modelių tikslumo metrikų rezultatai prognozuojant nakvynių skaičių vietiniams turistams viešbučio tipo apgyvendinimo įstaigose pateikti 36 lentelėje. Modelis su „Google Trends“ duomenimis pasižymi labai aukštomis visų metrikų reikšmėmis, lyginant su baziniu modeliu. Bazinio modelio vidutinė absoliutinė procentinė paklaida yra tik 3 %, tad jo prognozės yra gana tikslios ir patikimos.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,03	6429	7602,61
Su „Google Trends“	0,083	19081	22274,18

36 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams viešbučio tipo įstaigose

Stovyklavietės tipo apgyvendinimo įstaigos

Prognozavimo metrikų rezultatai, pateikti 37 lentelėje rodo, jog šioje kategorijoje modelis su „Google Trends“ duomenimis yra žymiai pranašesnis visų metrikų atžvilgiu. Visgi, paklaidos yra labai aukštos, tad nei vieno iš modelių prognozės nėra artimos tikrosioms reikšmėms.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	1,857	9950	14436,34
Su „Google Trends“	1,072	5284	7595,23

37 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams stovyklavietės tipo įstaigose

Kitos apgyvendinimo įstaigos

Toliau analizuojamas AVR modelių tikslumas prognozuojant nakvynių skaičių vietiniams turistams įstaigose, priklausančiose kategorijai „Kitas apgyvendinimas“. Modelių tikslumo metrikų rezultatai atskleidžia, jog ir šioje kategorijoje modelis su įtrauktais „Google Trends“ duomenimis pasižymi mažesnėmis paklaidomis visų metrikų atžvilgiu (žr. 38 lentelę). Taigi, užklausų populiarumo duomenų įtraukimas į modelį šioje apgyvendinimo įstaigų kategorijoje gali žymiai pagerinti prognozių tikslumą.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,069	12891	15324,76
Su „Google Trends“	0,061	10734	12123,83

38 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams kitose įstaigose

Visos apgyvendinimo įstaigos

Nakvynių skaičiaus vietiniams keliautojams prognozavimui visose apgyvendinimo įstaigose naudojant AVR modelius naudojamos anksčiau pristatytos metodikos. Modelių tikslumo rezultatai taikant pirmąją ir antrąją metodiką pateikti 39 ir 40 lentelėse atitinkamai. Šie rezultatai leidžia daryti išvadą, jog „Google Trends“ duomenų įtraukimas į modelį pagerina prognozavimo tikslumą naudojant abi metodikas. Tiesa, pirmosios metodikos atveju antrojo modelio MAPE reikšmė yra nežymiai didesnė, tačiau šis modelis stipriai lenkia bazinį MAE ir RMSE metrikų atžvilgiu, tad yra pranašesnis.

Skyrelyje pristatyti rezultatai parodė, jog „Google Trends“ duomenys yra vertingi prognozuojant nakvynių skaičių vietiniams turistams naudojant AVR modelius. Vienintelė kategorija, kurioje užklausų populiarumo duomenis naudojantis modelis pasirodė prasčiau, buvo „Viešbučiai ir kt.“. Tiesa, AVR

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,041	19975	26364,66
Su „Google Trends“	0,043	17517	21098,62

39 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams visose įstaigose

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,045	20867	26175,75
Su „Google Trends“	0,04	18390	22982,01

40 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių vietiniams turistams visose įstaigose, naudojant ankstesnių prognozių agregavimą

modelis pasižymi itin aukštomis paklaidomis prognozuojant nakvynių skaičių „Stovyklavietės ir kt.“ kategorijos apgyvendinimo įstaigose, tad jo rezultatai gali būti nepatikimi.

7.3. Modelių palyginimas

Šiame skyriuje atlikta modelių tikslumo analizė atskleidė, jog SARIMAX modelis gali būti tinkamesnis nakvynių skaičiaus vietiniams turistams prognozavimui – su šiuo modeliu gautos tikslesnės prognozės visoms apgyvendinimo įstaigų kategorijoms, išskyrus „Viešbučiai ir kt.“.

8. Nakvynių skaičiaus visiems turistams prognozavimas

Šiame skyriuje pateikiama SARIMAX ir AVR modelių tikslumo analizė prognozuojant nakvynių skaičių visiems turistams skirtingose apgyvendinimo įstaigų kategorijose. Abejų modelių atveju naudojama metodika, pristatyta analizuojant jų tikslumą turistų skaičiaus prognozavimui.

8.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai

Viešbučio tipo apgyvendinimo įstaigos

Pirmoji apgyvendinimo įstaigų kategorija, kuriai prognozuojamas nakvynių skaičius visiems turistams – „Viešbučiai ir kt.“. Paieškos užklausos, pasižyminčios aukščiausia koreliacija su priklausomu kintamuoju ir naudotinos modelyje su „Google Trends“ duomenimis: „camping in Lithuania“, „Vilnius“, „Kaunas“, „flights to Lithuania“, „things to do in Lithuania“. Be to, nustatyta, jog šiame modelyje optimalus pagrindinių komponentių skaičius yra 1.

Modelių liekanoms taikytų statistinių testų rezultatai leidžia teigti, jog abu modeliai tenkina prielaidas: ADF testo p reikšmės abiem modeliams $p < 0,01$, Ljungo ir Bokso testo p reikšmės – $p_{\text{bazinis}} \approx 0,954$ ir $p_{\text{su „Google Trends“}} \approx 0,969$, vienos imties t-testo p reikšmės $p_{\text{bazinis}} \approx 0,832$ ir $p_{\text{su „Google Trends“}} \approx 0,829$

Modelių tikslumo metrikų rezultatai, pateikti 41 lentelėje rodo, jog bazinis modelis yra pranašesnis visų tikslumo metrikų atžvilgiu, tad šis modelis yra tinkamesnis prognozuojant nakvynių skaičių šioje kategorijoje.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,034	14688	17916,68
Su „Google Trends“	0,037	16136	19828,26

41 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių visiems turistams viešbučio tipo apgyvendinimo įstaigose

Stovyklavietės tipo apgyvendinimo įstaigos

Antroji įstaigų kategorija – „Stovyklavietės ir kt.“. Nakvynių skaičius visiems turistams apgyvendinimo įstaigose šioje kategorijoje turi stipriausią sąryšį su užklausomis „camping in Lithuania“, „Vilnius“, „Kaunas“, „flights to Lithuania“ ir „hotels Lithuania“, tad šių užklausų lagai ir yra naudojami sudarant pagrindines komponentes. Pažingsniui renkant optimalų komponentių skaičių modeliui su „Google Trends“ duomenimis, nustatyta, jog tinkamiausia yra įtraukti tik vieną pagrindinę komponentę.

Modelių prielaidoms tikrinti taikyti statistiniai testai jų liekanoms. ADF testo metu gautos p reikšmės abiem modeliams $p < 0,01$, tad liekanos yra stacionarios. Ljungo-Bokso testo rezultatai ($p_{\text{bazinis}} \approx 0,796$ ir $p_{\text{su „Google Trends“}} \approx 0,888$) rodo, kad liekanos nepasižymi autokoreliacija ir atitinka baltąjį triukšmą. Pritaikius vienos imties t-testą, gautos p reikšmės $p_{\text{bazinis}} \approx 0,758$ ir

$p_{su, „Google Trends“} \approx 0,566$ – modelių prognozės nėra paslinktos. Visi testai parodė, jog modeliai tenkina prielaidas.

42 lentelėje pateikti rezultatai rodo, jog ir šioje kategorijoje modelis su „Google Trends“ duomenimis pasižymi aukštesnėmis paklaidomis, tad tinkamesnis yra bazinis modelis. Visgi, abiejų modelių paklaidos yra labai aukštos, tad nei vieno iš modelių prognozės nėra patikimos.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	1,695	13502	20804,04
Su „Google Trends“	1,805	15490	24423,54

42 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių visiems turistams stovyklavietės tipo apgyvendinimo įstaigose

Kitos apgyvendinimo įstaigos

Toliau analizuojamas SARIMAX modelių tikslumas prognozuojant nakvynių skaičių visiems turistams „Kitas apgyvendinimas“ kategorijos apgyvendinimo įstaigose. Šioje kategorijoje prognozuojamas kintamasis pasižymėjo aukščiausiu koreliacijos koeficientu su paieškos užklauskomis „camping in Lithuania“, „Vilnius“, „Kaunas“, „flights to Lithuania“ ir „things to do in Lithuania“ – būtent šių užklausų duomenys naudojami modelyje su „Google Trends“. Šiame modelyje naudojama viena pagrindinė komponentė, nes šis skaičius buvo optimaliausias.

Modelio prielaidų tikrinimui taikyti statistiniai testai leidžia daryti išvadą, jog abu modeliai tenkina visas prielaidas: ADF testo p reikšmės $p < 0,01$ abiem modeliams, pritaikius Ljungo ir Bokso testą gautos p reikšmės $p_{bazinis} \approx 0,483$ ir $p_{su, „Google Trends“} \approx 0,667$, o t-testui gauti rezultatai – $p_{bazinis} \approx 0,138$ ir $p_{su, „Google Trends“} \approx 0,329$.

Modelių tikslumo metrikų rezultatai, pateikti 43 rodo, jog modelis su „Google Trends“ duomenimis pasižymi aukštesnėmis paklaidomis visų metrikų atžvilgiu. „Viešbučiai ir kt.“ kategorijos apgyvendinimo įstaigose visų turistų rezervuojamų nakvynių skaičiaus prognozavimui tinkamesnis yra bazinis modelis.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,03	6688	8180,93
Su „Google Trends“	0,04	10613	11703,7

43 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių visiems turistams kitose apgyvendinimo įstaigose

Visos apgyvendinimo įstaigos

Galiausiai, analizuojamas modelių tikslumas kategorijoje „Visos įstaigos“. Kaip ir ankstesniuose skyriuose, šioje dalyje naudojamos dvi metodikos.

Pirmosios metodikos atveju konstruojami nauji modeliai nakvynių skaičiaus prognozei. Modelyje su „Google Trends“ duomenimis naudojama viena pagrindinė komponentė, sukonstruota iš

paieškos užklausų „camping in Lithuania“, „Vilnius“, „Kaunas“, „flights to Lithuania“ ir „things to do in Lithuania“ populiarumo duomenų lagų.

Modelių liekanoms taikyti statistiniai testai prielaidų tikrinimui. ADF testo p reikšmės abiem modeliams $p < 0,01$ – liekanos atitinka atsitiktinį klaidžiojimą. Ljungo ir Bokso testo rezultatai parodė, jog liekanos nepasižymi autokoreliacija ($p_{\text{bazinis}} \approx 0,613$ ir $p_{\text{su „Google Trends“}} \approx 0,699$). Pritaikius vienos imties t-testą, rezultatai patvirtino, jog modelių prognozės nėra paslinktos ($p_{\text{bazinis}} \approx 0,778$ ir $p_{\text{su „Google Trends“}} \approx 0,814$). Taigi, abu modeliai tenkina prielaidas.

Modelių tikslumo rezultatai atskleidžia, jog taikant pirmąją metodiką, bazinis modelis vidutiniškai yra pranašesnis – jo MAPE ir MAE metrikų reikšmės yra žemesnės (žr. 44 lentelę). Tačiau modelis su „Google Trends“ duomenimis pasižymi mažesne RMSE metrikos reikšme, kas gali reikšti, jog šio modelio prognozės yra mažiau linkę itin stipriai nukrypti nuo tikrųjų reikšmių.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,03	21917	29434,89
Su „Google Trends“	0,033	22145	27024,4

44 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių visiems turistams visose apgyvendinimo įstaigose

Naudojant antrąją metodiką, gauti tikslumo rezultatai, pateikti 45 lentelėje. Iš tikslumo metrikų galima pastebėti, jog modelis su „Google Trends“ duomenimis pasižymi žemesnėmis MAE ir RMSE reikšmėmis, tačiau labai nežymiai nusilenkia baziniam modeliui MAPE metrikos atžvilgiu. Visgi, skirtumas tarp šios metrikos reikšmių yra labai mažas, tad antrasis modelis yra tinkamesnis naudojant šią metodiką.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,029	21266	27281,27
Su „Google Trends“	0,03	20280	25795,98

45 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių visiems turistams visose apgyvendinimo įstaigose, naudojant ankstesnių prognozių agregavimą

SARIMAX modelių tikslumo prognozuojant nakvynių skaičių visiems turistams skirtingose apgyvendinimo įstaigų kategorijose rezultatai atskleidė, jog „Google Trends“ duomenys nesuteikia pakankamai papildomos informacijos, kad galėtų pagerinti prognozavimo tikslumą. Modelis su įtrauktais užklausų populiarumo duomenimis pasirodė tik nežymiai geriau prognozuojant nakvynių skaičių visose apgyvendinimo įstaigose, agreguojant atskirų įstaigų kategorijų prognozes.

8.2. Atraminių vektorių regresijos modelių rezultatai

Toliau pateikiami AVR modelių su ir be „Google Trends“ duomenų nakvynių skaičiaus visiems turistams skirtingose apgyvendinimo įstaigų kategorijose prognozavimui.

Viešbučio tipo apgyvendinimo įstaigos

Pirmiausia analizuojamas AVR modelių tikslumas kategorijoje „Viešbučiai ir kt.“. 46 lentelėje pateikti tikslumo metrikų rezultatai atskleidžia, jog šioje kategorijoje bazinis modelis pasižymi žymiai mažesnėmis paklaidomis visų metrikų atžvilgiu.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,031	13508	16229,09
Su „Google Trends“	0,052	24250	30482,23

46 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių visiems turistams viešbučio tipo įstaigose

Stovyklavietės tipo apgyvendinimo įstaigos

Toliau aptariami rezultatai, gauti prognozuojant nakvynių skaičių visiems turistams apgyvendinimo įstaigose, priklausančiose kategorijai „Stovyklavietės ir kt.“. Tiek bazinis, tiek modelis su „Google Trends“ duomenimis pasižymi aukštomis paklaidomis, tačiau antrasis modelis yra šiek tiek tikslesnis visų metrikų atžvilgiu. Visgi, atsižvelgiant į aukštas metrikų reikšmes, nei vieno iš modelių prognozės nėra patikimos.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	1,927	10576	14646,75
Su „Google Trends“	1,911	9408	12796,25

47 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių visiems turistams stovyklavietės tipo įstaigose

Kitos apgyvendinimo įstaigos

Trečioji apgyvendinimo įstaigų kategorija – „Kitas apgyvendinimas“. Nakvynių skaičiaus visiems turistams prognozavimo tikslumo rezultatai pateikti 48 lentelėje. Iš metrikų reikšmių galima daryti išvadą, jog šioje kategorijoje bazinis modelis yra gerokai tikslesnis. Šio modelio vidutinė absoliutinė procentinė paklaida yra tik 3,5 %, tad prognozės yra labai artimos tikrosioms reikšmės.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,035	7316.00	9061,27
Su „Google Trends“	0,061	20309	28787,82

48 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių visiems turistams kitose įstaigose.

Visos apgyvendinimo įstaigos

Modelių tikslumo rezultatai naudojant anksčiau pristatytą pirmąją ir antrąją metodiką prognozių skaičiavimui visoms apgyvendinimo įstaigoms pateikti 49 ir 50 lentelėse atitinkamai. Abiejų me-

todikų atveju tikslumo metrikų reikšmės rodo, jog bazinis modelis yra žymiai geresnis bei itin aukštu tikslumu, ypač MAPE metrikos atžvilgiu.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,026	15893	20839,41
Su „Google Trends“	0,047	37114	50973,12

49 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių visiems turistams visose įstaigose

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,025	15050	19453,3
Su „Google Trends“	0,048	38583	51173,43

50 lentelė. AVR modelio tikslumas prognozuojant nakvynių skaičių visiems turistams visose įstaigose, naudojant ankstesnių prognozių agregavimą

Apibendrinat skyrelį, galima teigti, jog „Google Trends“ duomenys nepagerina AVR modelių tikslumo prognozuojant nakvynių skaičių visiems turistams. Net ir kategorijoje, kurioje modelis, naudojantis šiuos duomenis, pasirodė geriau (t. y. „Stovyklavietės ir kt.“), vis tiek gautos labai aukštos paklaidos, tad prognozės nėra patikimos.

8.3. Modelių palyginimas

Skyriuje pristatyti SARIMAX ir AVR modelių tikslumo rezultatai parodė, jog naudojant bazinius SARIMAX modelius galima pasiekti aukštesnį prognozių tikslumą apgyvendinimo įstaigų kategorijose „Stovyklavietės ir kt.“ bei „Kitas apgyvendinimas“. Likusiose kategorijose AVR modeliai pasirodė ne-žymiai geriau. Taigi, yra sunku nustatyti, kuris iš modelių yra tinkamesnis nakvynių skaičiaus visiems turistams prognozavimui.

9. Nakvynių skaičiaus užsienio turistams prognozavimas

Galiausiai, analizuojamas SARIMAX ir AVR modelių su ir be „Google Trends“ duomenų tikslumas prognozuojant nakvynių skaičių užsienio turistams Lietuvos apgyvendinimo įstaigose. Kaip ir kituose skyriuose, prognozės sudaromos kiekvienai apgyvendinimo įstaigų kategorijai atskirai. Šiame skyriuje naudojama ta pati metodika, kaip ir prognozuojant užsienio turistų skaičių, t. y. nesudaromi nauji modeliai, o prognozės gaunamos skaičiuojant skirtumą tarp ankstesnių prognozių.

9.1. Sezoninių ARIMA modelių su egzogeniniais kintamaisiais rezultatai

Pirmiausia analizuojamas prognozių, gautų skaičiuojant skirtumą tarp anksčiau gautų SARIMAX modelio prognozių nakvynių skaičiui vietiniams ir visiems turistams. Kaip ir skyriuje apie užsienio turistų skaičių prognozavimą, SARIMAX modelių prielaidos nėra tikrinamos.

Viešbučio tipo apgyvendinimo įstaigos

Pirmoji apgyvendinimo įstaigų kategorija – „Viešbučiai ir kt.“. Suskaičiavus skirtumą tarp anksčiau sudarytų nakvynių skaičiaus prognozių Lietuvos ir visiems turistams, gauti tikslumo rezultatai, pateikti 51 lentelėje. Tikslumo metrikų reikšmės rodo, jog bazinis modelis yra šiek tiek tikslesnis MAPE ir MAE metrikų atžvilgiu, tačiau antrasis modelis pasižymi žemesne RMSE metrikos reikšme. Taigi, yra sunku nustatyti, kuris modelis yra tikslesnis bendrai, nes skirtumai tarp paklaidų nėra dideli. Atsižvelgiant į vidutinę absoliutinę procentinę paklaidą, bazinis modelis yra pranašesnis.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,059	11535	15215,92
Su „Google Trends“	0,064	11580	14734,6

51 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių užsienio turistams viešbučio tipo apgyvendinimo įstaigose

Stovyklavietės tipo apgyvendinimo įstaigos

Toliau pateikta modelių tikslumo analizė prognozuojant nakvynių skaičių užsienio turistams „Stovyklavietės ir kt.“ kategorijos įstaigose. 52 lentelėje pateikti modelių tikslumo rezultatai rodo, jog abu modeliai pasižymi labai aukštomis paklaidomis. Modelis su „Google Trends“ duomenimis yra tikslesnis MAPE metrikos atžvilgiu, tačiau bazinis modelis pranašesnis MAE ir RMSE metrikų atžvilgiu. Visgi, abiejų modelių visų metrikų reikšmės tokios aukštos, jog nei vieno iš modelių prognozės nėra patikimos.

Kitos apgyvendinimo įstaigos

Trečioji apgyvendinimo įstaigų kategorija – „Viešbučiai ir kt.“. Modelių tikslumo metrikų šioje kategorijoje rezultatai atskleidžia, jog bazinis modelis yra žymiai tikslesnis visų metrikų atžvilgiu.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	53,588	10910	16623,7
Su „Google Trends“	32,983	15218	25769,2

52 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių užsienio turistams stovyklavietės tipo apgyvendinimo įstaigose

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,14	8629	10471,53
Su „Google Trends“	0,359	22399	26117,1

53 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių užsienio turistams kitose apgyvendinimo įstaigose

Visos apgyvendinimo įstaigos

Nakvynių skaičiaus užsienio turistams visose įstaigose prognozavimui naudojamos dvi metodikos, pristatytos prognozuojant užsienio turistų skaičių toje pačioje kategorijoje. Pirmosios metodikos tikslumo metrikų rezultatai pateikti 54 lentelėje. Taikant šią metodiką, bazinis modelis yra žymiai tikslus visų metrikų atžvilgiu. Tačiau antroji metodika duoda kitokius rezultatus – modelis su „Google Trends“ duomenimis šį kartą yra žymiai pranašesnis. Taikant antrąją metodiką ir įtraukus „Google Trends“ duomenis gaunamos tiksliausios prognozės šioje kategorijoje.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,059	15585	18610,52
Su „Google Trends“	0,087	22767	29491,21

54 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių užsienio turistams visose apgyvendinimo įstaigose

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,056	15601	20557,83
Su „Google Trends“	0,053	13238	15065,88

55 lentelė. SARIMAX modelio tikslumas prognozuojant nakvynių skaičių užsienio turistams visose apgyvendinimo įstaigose, naudojant atskirų įstaigų prognozių agregavimą

Skyrelyje aptarti rezultatai parodė, jog „Google Trends“ duomenys nepagerino prognozių tikslumo nei vienoje iš tirtų apgyvendinimo įstaigų kategorijų.

9.2. Atraminių vektorių regresijos modelių rezultatai

Šis skyrelis skirtas AVR modelių su ir be „Google Trends“ duomenų tikslumo prognozuojant nakvynių skaičių užsienio turistams skirtingose apgyvendinimo įstaigų kategorijose analizei. Naudojama metodika, analogiška pristatytai ankstesniuose skyriuose apie nakvynių skaičiaus prognozavimą.

Viešbučio tipo apgyvendinimo įstaigos

Pirmiausia tirtas prognozių tikslumas apgyvendinimo įstaigoms, priklausančioms kategorijai „Viešbučiai ir kt.“. Tikslumo metrikų rezultatai, pateikti 56 lentelėje, rodo, jog šioje kategorijoje „Google Trends“ duomenys stipriai sumažino prognozių tikslumą visų metrikų atžvilgiu.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,039	4092	5226,88
Su „Google Trends“	0,085	9246	11065,57

56 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių viešbučio tipo įstaigose

Stovyklavietės tipo apgyvendinimo įstaigos

Toliau analizuojamas prognozių tikslumas kategorijoje „Stovyklavietės ir kt.“. Modelių tikslumo rezultatai atskleidžia, kad ir šioje kategorijoje bazinio modelio prognozės yra žymiai arčiau tikrųjų reikšmių (žr. 57 lentelę). Šio modelio MAPE reikšmė yra labai aukšta, tačiau, kaip jau buvo aptarta ankstesniuose skyriuose, tai gali būti nulemta itin žemų nakvynių skaičių ne vasaros mėnesiais.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	24,26	788	1051,79
Su „Google Trends“	35,924	1440	1699,8

57 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių stovyklavietės tipo įstaigose

Kitos apgyvendinimo įstaigos

Trečioji apgyvendinimo įstaigų kategorija – „Kitas apgyvendinimas“. Prognozių tikslumo metrikų reikšmės, pateiktos 58 lentelėje leidžia daryti išvadą, jog „Google Trends“ duomenys nesuteikia vertingos informacijos modeliams. Bazinio modelio prognozės ir šioje kategorijoje yra tikslesnės.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,164	3068	4692,89
Su „Google Trends“	0,456	7141	9282,99

58 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių kitose įstaigose

Visos apgyvendinimo įstaigos

Galiausiai, analizuojamas modelių tikslumas prognozuojant nakvynių skaičių užsienio turistams visose apgyvendinimo įstaigose. Šioje kategorijoje naudojamos dvi anksčiau pristatytos metodikos. Pirmosios ir antrosios metodikų tikslumo rezultatai pateikti 59 ir 60 lentelėse atitinkamai. Abiejų metodikų rezultatai rodo, jog bazinis modelis yra žymiai tikslesnis visų metrikų atžvilgiu.

Apibendrinant AVR modelių tikslumą prognozuojant nakvynių skaičių užsienio turistams, galima daryti išvadą, jog „Google Trends“ duomenys šiuo atveju teikia modelius klaidinančią informaciją ir padidina paklaidas. Bazinių modelių prognozės šiame kontekste yra patikimesnės.

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,07	9401	13010,13
Su „Google Trends“	0,092	12520	16490,75

59 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių visose įstaigose

Modelis	MAPE	MAE	RMSE
Bazinis modelis	0,052	7017	9962,58
Su „Google Trends“	0,082	10544	16629,78

60 lentelė. AVR modelio tikslumas prognozuojant užsienio turistų skaičių visose įstaigose, naudojant atskirų apgyvendinimo kategorijų prognozių agregavimą

9.3. Modelių palyginimas

Skyriuje pristatyti SARIMAX ir AVR modelių rezultatai atskleidė, jog nei vienas iš modelių nėra tikslesnis visose apgyvendinimo įstaigų kategorijose. Prognozuojant nakvynių skaičių užsienio turistams įstaigų kategorijose „Viešbučiai ir kt.“ bei „Stovyklavietės ir kt.“, galima pasiekti aukštesnį tikslumą naudojant AVR modelius. Kitose apgyvendinimo įstaigų kategorijose geriau pasirodė SARIMAX modelis.

10. Rezultatų apibendrinimas

Siekiant apibendrinti visus analizuotus rezultatus, palyginti SARIMAX ir AVR modelių tikslumą ir „Google Trends“ duomenų vertingumą tiriamam uždaviniui, buvo apskaičiuotos vidutinės MAPE metrikos reikšmės modeliams su ir be „Google Trends“ duomenų prognozuojant turistų ir nakvynių skaičius skirtingoms keliautojų tautybių kategorijoms. Ši metrika pasirinkta palyginimui, nes ji yra nepriklausoma nuo prognozuojamo kintamojo skalės.

Modelių tikslumo, prognozuojant turistų ir nakvynių skaičių vietiniams keliautojams, apibendrinti rezultatai pateikti 61 lentelėje. Iš vidutinių MAPE metrikos reikšmių galima daryti išvadą, jog „Google Trends“ duomenys žymiai pagerino modelių tikslumą trijuose iš keturių tirtų scenarijų. Be to, tiek turistų, tiek nakvynių skaičiaus prognozavimui yra tinkamesnis SARIMAX modelis – jo vidutinės MAPE reikšmės yra mažesnės.

	Modelis	Turistų skaičiaus prognozavimas	Nakvynių skaičiaus prognozavimas
SARIMAX	Bazinis	0,245	0,133
	Su „Google Trends“	0,076	0,183
AVR	Bazinis	0,259	0,408
	Su „Google Trends“	0,183	0,26

61 lentelė. Vidutinės MAPE metrikų reikšmės modeliams, prognozuojantiems turistų ir nakvynių skaičių vietiniams keliautojams

Vidutinis modelių tikslumas prognozuojant turistų ir nakvynių skaičių visiems keliautojams pateiktas 62 lentelėje. Rezultatai rodo, jog šiuo atveju „Google Trends“ duomenys nepagerino modelių tikslumo, išskyrus AVR modelio turistų skaičiaus prognozes. Tokie rezultatai, tikėtina, gaunami, nes dėl duomenų trūkumo, nepavyko išrinkti tinkamų paieškos užklausų, o turimi duomenys galimai yra nepakankamai informatyvūs. Tikslumo rezultatai atskleidžia, kad SARIMAX modeliai yra tikslesni prognozuojant nakvynių skaičių visiems keliautojams, tačiau AVR modeliai su „Google Trends“ duomenimis yra pranašesni turistų skaičiaus prognozavimui.

	Modelis	Turistų skaičiaus prognozavimas	Nakvynių skaičiaus prognozavimas
SARIMAX	Bazinis	0,213	0,364
	Su „Google Trends“	0,256	0,389
AVR	Bazinis	0,335	0,409
	Su „Google Trends“	0,183	0,424

62 lentelė. Vidutinės MAPE metrikų reikšmės modeliams, prognozuojantiems turistų ir nakvynių skaičių visiems keliautojams

Galiausiai, apibendrinamas modelių tikslumas prognozuojant turistų ir nakvynių skaičių užsienio keliautojams. 63 lentelėje pateikti rezultatai rodo, kad „Google Trends“ duomenys modelių tikslumą pagerino tik viename kontekste – naudojant SARIMAX modelius nakvynių prognozavimui. Visais kitais atvejais šie duomenys stipriai sumažino prognozių tikslumą. Tokie rezultatai taip pat, manytina, yra nulemti netinkamų paieškos užklausų duomenų. Be to, modelių tikslumas ir vėl išsiskiria, priklausomai nuo konteksto – naudojant SARIMAX modelius, galima gauti patikimesnes užsienio tu-

ristų skaičiaus prognozes, tačiau AVR modeliai yra pranašesni, kai prognozuojamas nakvynių skaičius užsienio turistams.

	Modelis	Turistų skaičiaus prognozavimas	Nakvynių skaičiaus prognozavimas
SARIMAX	Bazinis	1,614	10,78
	Su „Google Trends“	5,183	6,709
AVR	Bazinis	4,917	4,918
	Su „Google Trends“	7,327	7,322

63 lentelė. Vidutinės MAPE metrikų reikšmės modeliams, prognozuojantiems turistų ir nakvynių skaičių keliautojams iš užsienio

Rezultatai ir išvados

Darbe atliktas nuoseklus tyrimas, siekiant nustatyti, ar „Google Trends“ duomenys gali pagerinti modelių tikslumą prognozuojant turistų ir nakvynių skaičių Lietuvos apgyvendinimo įstaigose.

Pirmiausiai, išrinkti tinkami ir vertingą informaciją galintys suteikti internetinės paieškos sistemos „Google“ užklausų raktažodžiai.

Atlikus pradinę analizę, pastebėta COVID-19 pandemijos įtaka turistų ir nakvynių skaičiui, bei šio laikotarpio duomenys pakoreguoti. Taip pat rastos reikšmingos koreliacijos tarp kovariančių, tad tolesniuose tyrimo etapuose nuspręsta naudoti pagrindinių komponentų analizę.

Tiriant „Google Trends“ duomenų efektyvumą, panaudotas klasikinis laiko eilučių – sezoninis ARIMA su egzogeniniais kintamaisiais – modelis ir mašininio mokymosi srities siūloma atraminių vektorių regresija.

Pritaikius šiuos modelius turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimui, pastebėta, jog tikslesnės prognozės gali būti gautos naudojant SARIMAX modelį. Be to, analizė parodė, kad „Google Trends“ duomenų įtraukimas į modelius gali stipriai pagerinti jų tikslumą prognozuojant vietinių turistų skaičių. Tokie rezultatai, tikėtina, yra nulemti didesnės ir reprezentatyvesnės lietuviškų užklausų duomenų aibės.

Modelių tikslumo prognozuojant nakvynių skaičių apgyvendinimo įstaigose analizės metu pastebėta, jog ir šiame kontekste SARIMAX modelis yra pranašesnis, o „Google Trends“ duomenys modelių tikslumą pagerina tik tam tikrose situacijose.

Tyrimas parodė, jog internetinių užklausų populiarumo informacija gali būti vertinga vietinių turistų kontekste, tačiau pasižymi trūkumais prognozuojant kintamuosius užsienio turistams ir visiems keliautojams. „Google Trends“ duomenų trūkumas pasaulio masto užklausoms apie Lietuvą galėtų būti išspręstas išplečiant tyrimą – renkant informaciją apie užklausų populiarumą atskirai kiekvienai valstybei, iš kurios į Lietuvos apgyvendinimo įstaigas atvyksta turistai. Tyrimas galėtų būti plečiamas ir įtraukiant kitų internetinių šaltinių, pavyzdžiui, socialinių medijų, duomenis.

Literatūra ir šaltiniai

- [1] M. Akin. „A novel approach to model selection in tourism demand modeling“. Iš: *Tourism Management* 48 (2015), puslapiai 64–72.
- [2] S. L. B. Bokelmann. „Spurious patterns in Google Trends data-An analysis of the effects on tourism demand forecasting in Germany“. Iš: *Tourism Management* 75 (2019), puslapiai 1–12.
- [3] K.-Y. Chen, C.-H. Wang. „Support vector regression with genetic algorithms in forecasting tourism demand“. Iš: *Tourism Management* 28.1 (2007), puslapiai 215–226.
- [4] G. Dzemyda, O. Kurasova, J. Žilinskas. *Daugiamačių duomenų vizualizavimo metodai*. Mokslo aidai, 2008.
- [5] D. Frechtling. *Forecasting Tourism Demand: Methods and Strategies*. 5-6 p. London, United Kindgom: Taylor & Francis Group, 2001.
- [6] C. Goh, R. Law. „Modeling and forecasting tourism demand for arrivals with stochastic non-stationary seasonality and intervention“. Iš: *Tourism Management* 23.5 (2002), puslapiai 499–510.
- [7] Google News Lab. *What is Google Trends data — and what does it mean?* URL: <https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8> (žiūrėta 2025-05-16).
- [8] N. Guo, W. Gui, W. Chen, X. Tian, W. Qiu, Z. Tian, X. Zhang. „Using improved support vector regression to predict the transmitted energy consumption data by distributed wireless sensor network“. Iš: *Wireless Com Network* 2020.1 (2020).
- [9] H. Hassani, M. R. Yeganegi. „Selecting optimal lag order in Ljung–Box test“. Iš: *Physica A: Statistical Mechanics and its Applications* 541 (2020).
- [10] W. Höpken, T. Eberle, M. Fuchs. „Google Trends data for analysing tourists’ online search behaviour and improving demand forecasting: the case of Åre, Sweden“. Iš: *Information Technology & Tourism* 21 (2019), puslapiai 45–62.
- [11] X. Yang, B. Pan, J. A. Evans, L. V. Benfu. „Forecasting Chinese tourist volume with search engine data“. Iš: *Tourism Management* 46 (2015), puslapiai 386–397.
- [12] N. Jasiukėnaitė, R. Levulienė. „Vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimas naudojant „Google Trends“ duomenis“. Iš: *Konferencijos „Lietuvos magistrantų informatikos ir IT tyrimai“ darbai. 2025*. Vilnius, Lietuva: Vilnius University Open Series, 2025, puslapiai 64–69.
- [13] X. Li, R. Law, G. Xie, S. Wang. „Review of tourism forecasting research with internet data“. Iš: *Tourism Management* 83 (2021).
- [14] S. Moritz, T. Bartz-Beielstein. „imputeTS: Time Series Missing Value Imputation in R“. Iš: *The R Journal* 9.1 (2017), puslapiai 207–218.
- [15] A. Nielsen. *Practical Time Series Analysis*. O’Reilly Media, Inc., 2019.

- [16] C. Nontapa, C. Kesamoon, N. Kaewhawong, P. Intrapaiboon. „A New Hybrid Forecasting Using Decomposition Method with SARIMAX Model and Artificial Neural Network“. Iš: *International Journal of Mathematics and Computer Science* 16.4 (2021).
- [17] B. T. Ojemakinde. „Support Vector Regression for Non-Stationary Time Series“. Magistro darbas. University of Tennessee, 2006, puslapiai 28–42.
- [18] OpenAI. *ChatGPT*. 2024. URL: <https://www.openai.com/> (žiūrėta 2025-05-22).
- [19] B. Pan, D. C. Wu, H. Song. „Forecasting hotel room demand using search engine data“. Iš: *Journal of Hospitality and Tourism Technology* 3.3 (2012), puslapiai 196–210.
- [20] S. Park, J. Lee, W. Song. „Short-term forecasting of Japanese tourist inflow to South Korea using Google trends data“. Iš: *Journal of Travel & Tourism Marketing* 34.3 (2017), puslapiai 357–368.
- [21] B. Peng, H. Song, G. I. Crouch. „A meta-analysis of international tourism demand forecasting and implications for practice“. Iš: *Tourism Management* 45 (2014), puslapiai 181–193.
- [22] M.-L. Shen, H.-H. Liu, Y.-H. Lien, C.-F. Lee, C.-H. Yang. „Hybrid Approach for Forecasting Tourist Arrivals“. Iš: *Proceedings of the 2019 8th International Conference on Software and Computer Applications*. New York, NY, USA: Association for Computing Machinery, 2019, puslapiai 392–396.
- [23] A. J. Smola, B. Schölkopf. „A tutorial on support vector regression“. Iš: *Statistics and Computing* 14.3 (2004), puslapiai 199–222.
- [24] H. Song, G. Li. „Tourism demand modelling and forecasting—A review of recent research“. Iš: *Tourism Management* 29.2 (2008), puslapiai 203–220.
- [25] H. Song, R. T. Qiu, J. Park. „A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting“. Iš: *Annals of Tourism Research* 75 (2019), puslapiai 338–362.
- [26] S. Sun, Y. Wei, K.-L. Tsui, S. Wang. „Forecasting tourist arrivals with machine learning and internet search index“. Iš: *Tourism Management* 70 (2019), puslapiai 1–10.
- [27] Valstybės duomenų agentūra. *Nakvynių skaičius apgyvendinimo įstaigose*. 2025. URL: <https://osp.stat.gov.lt/lt/statistiniu-rodikliu-analize?hash=3a4d407a-4c23-40d6-b9f5-767d7dba4ced> (žiūrėta 2025-05-22).
- [28] Valstybės duomenų agentūra. *Turistų skaičius apgyvendinimo įstaigose*. 2025. URL: https://osp.stat.gov.lt/lt/statistiniu-rodikliu-analize?hash=21ac96a1-7229-4a43-bdbb-c1ae9175226f# (žiūrėta 2025-05-22).

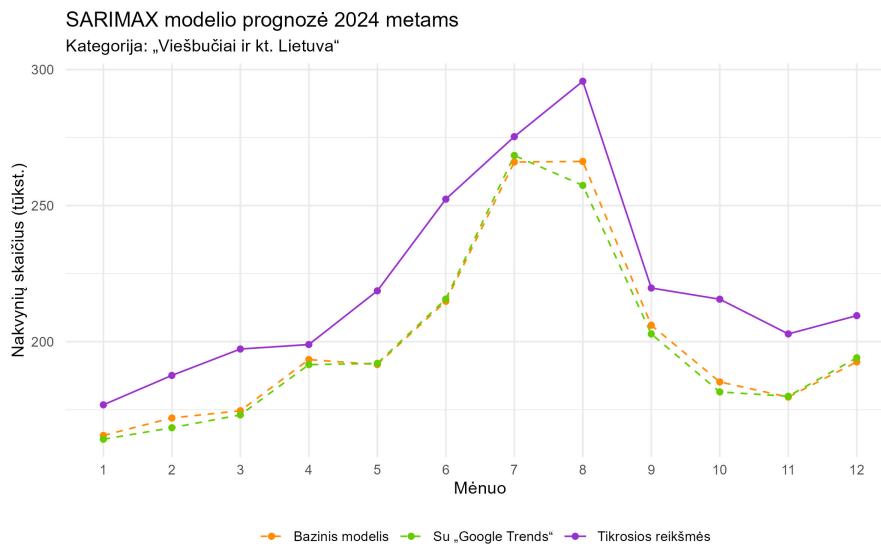
1 priedas. Naudota programinė įranga

Duomenų apdorojimas, pradinė duomenų analizė, prognostinių modelių sudarymas ir grafikų braižymas buvo atliktas naudojant R-4.4.2 programą ir jos paketus. Susidūrus su sudėtingomis klaidomis kode, jų taisymui buvo naudotas *ChatGPT-4* kalbos modelis [18]. Šio modelio siūlymai naudoti „Google Trends“ duomenų nuskaitymui (t. y. sujungiant atskirus duomenų failus į vieną lentelę) ir paruošiant laiko eilutes grafikų braižymui.

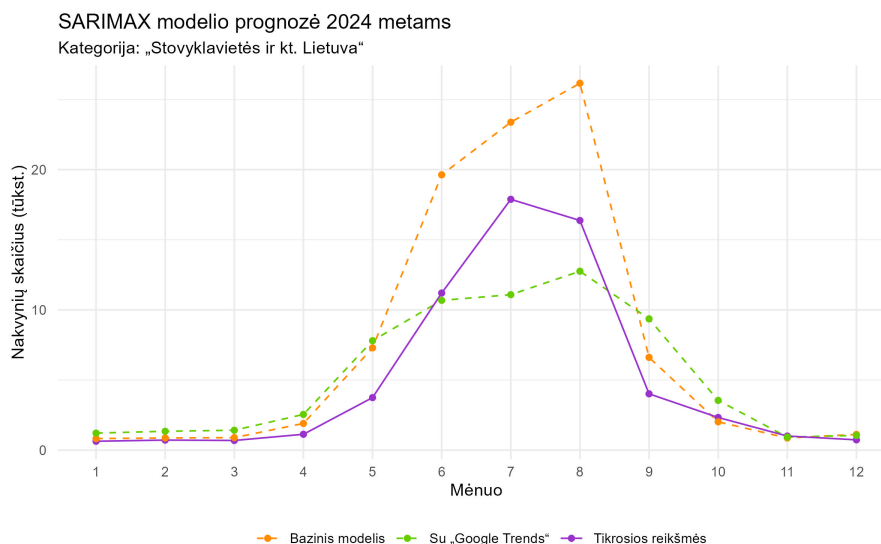
2 priedas. Nakvynių skaičiaus prognozių grafikai

Nakvynių skaičiaus vietiniams turistams prognozavimas

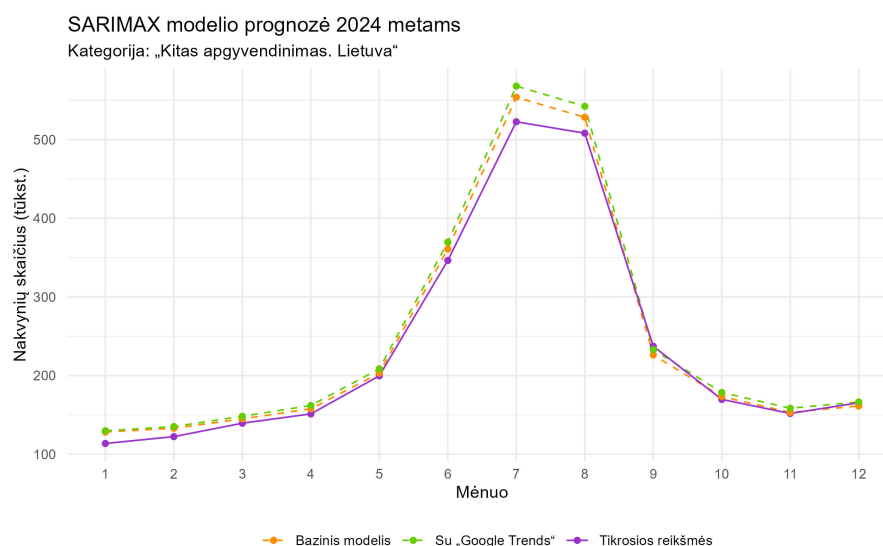
SARIMAX



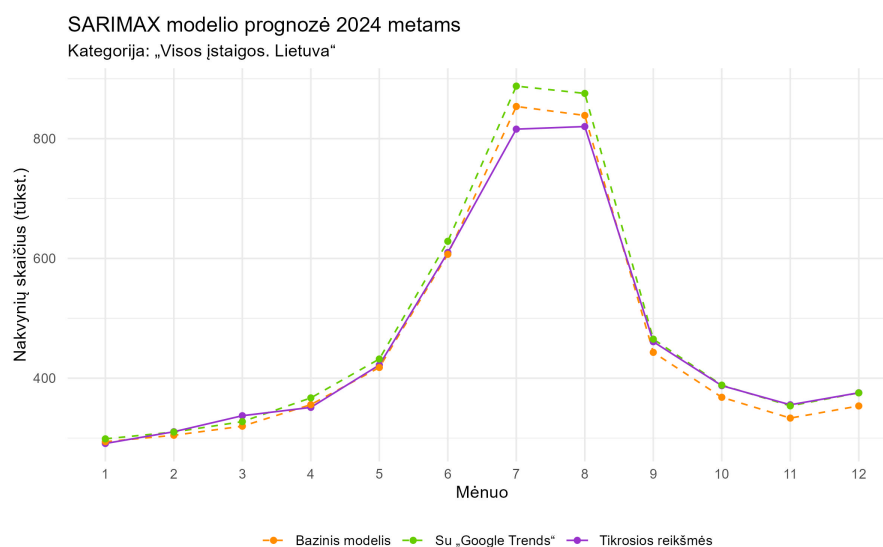
41 pav. Nakvynių skaičiaus vietiniams turistams viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



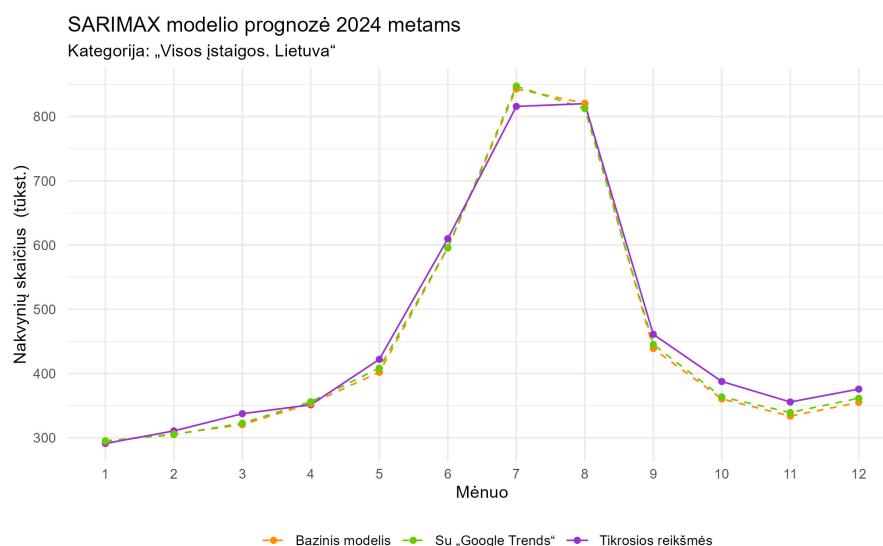
42 pav. Nakvynių skaičiaus vietiniams turistams stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



43 pav. Nakvynių skaičiaus vietiniams turistams kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

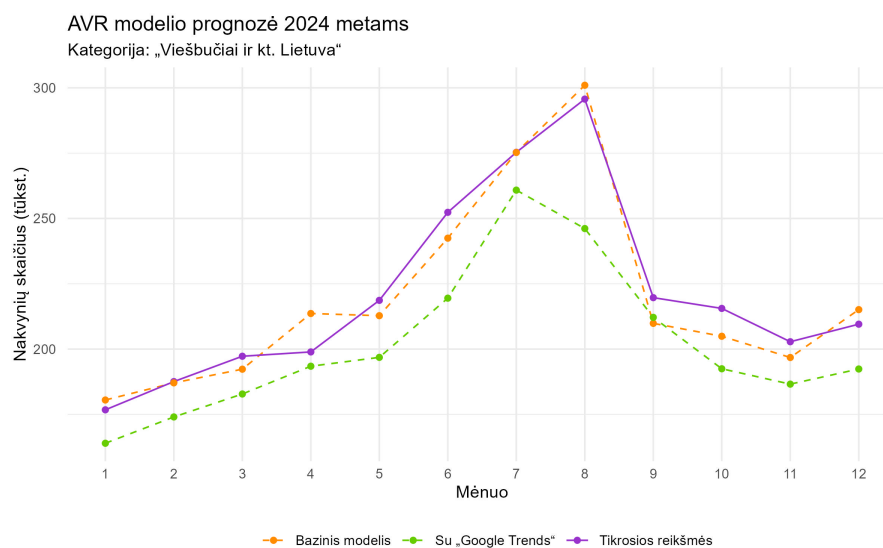


44 pav. Nakvynių skaičiaus vietiniams turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

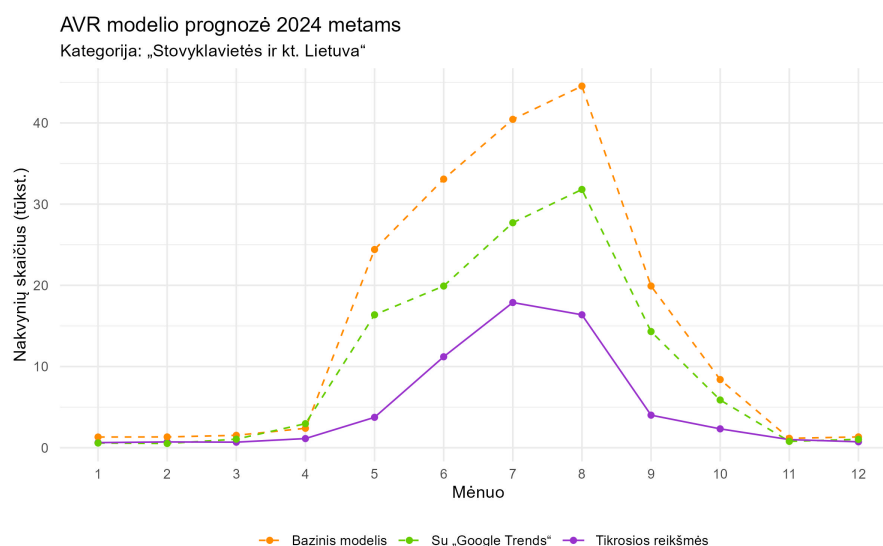


45 pav. Nakvynių skaičiaus vietiniams turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

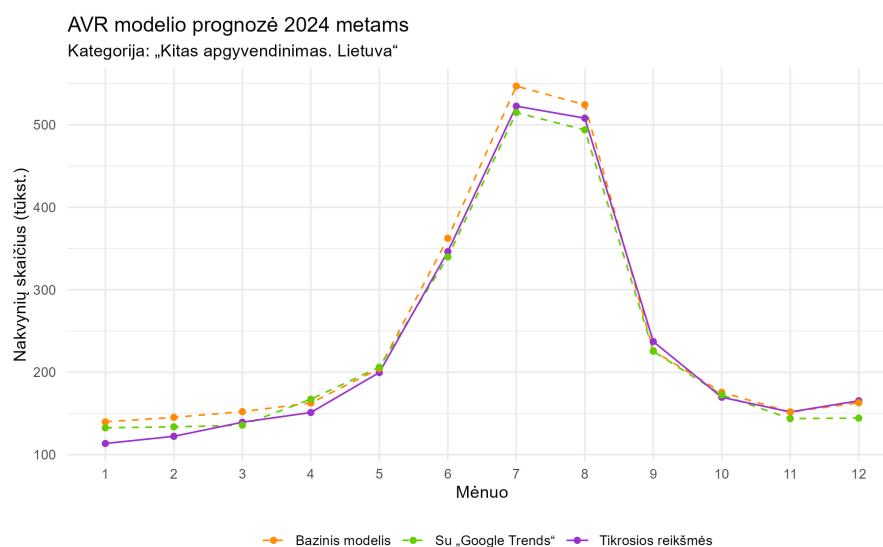
Atraminių vektorių regresija



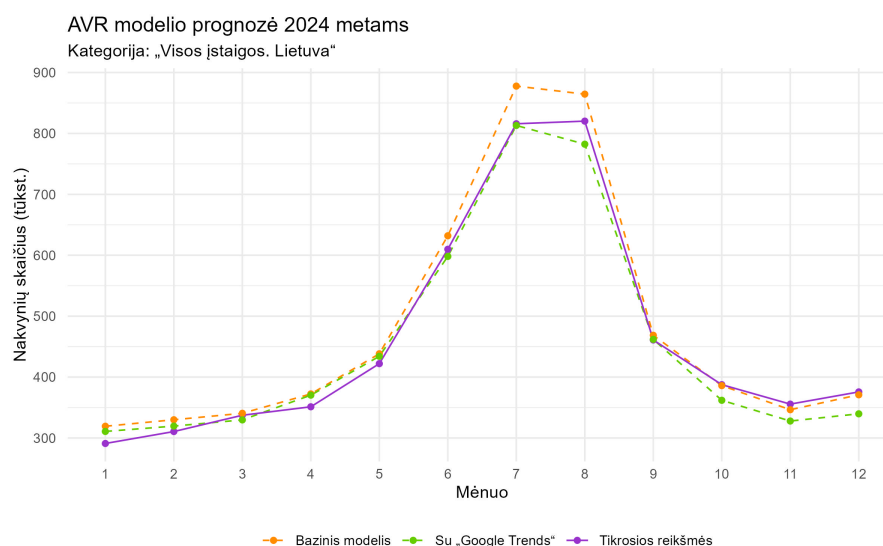
46 pav. Nakvynių skaičiaus vietiniams turistams viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



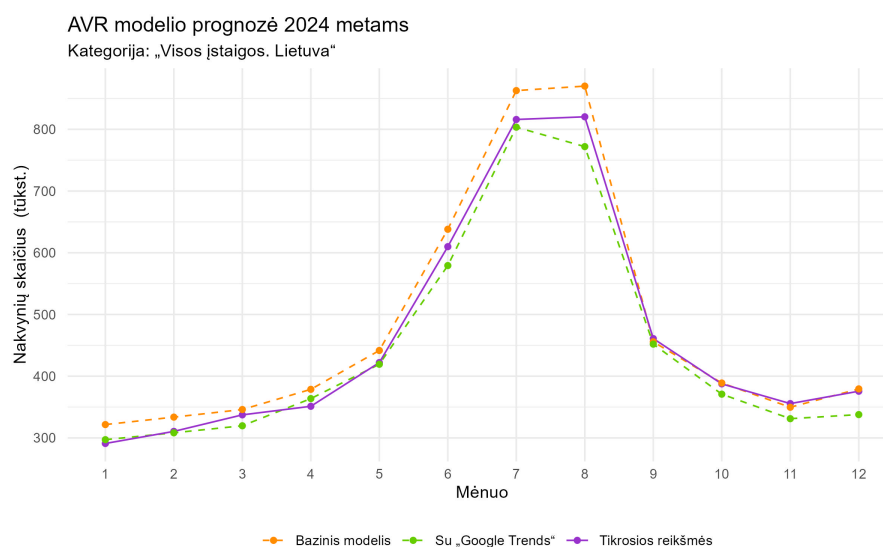
47 pav. Nakvynių skaičiaus vietiniams turistams stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



48 pav. Nakvynių skaičiaus vietiniams turistams kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



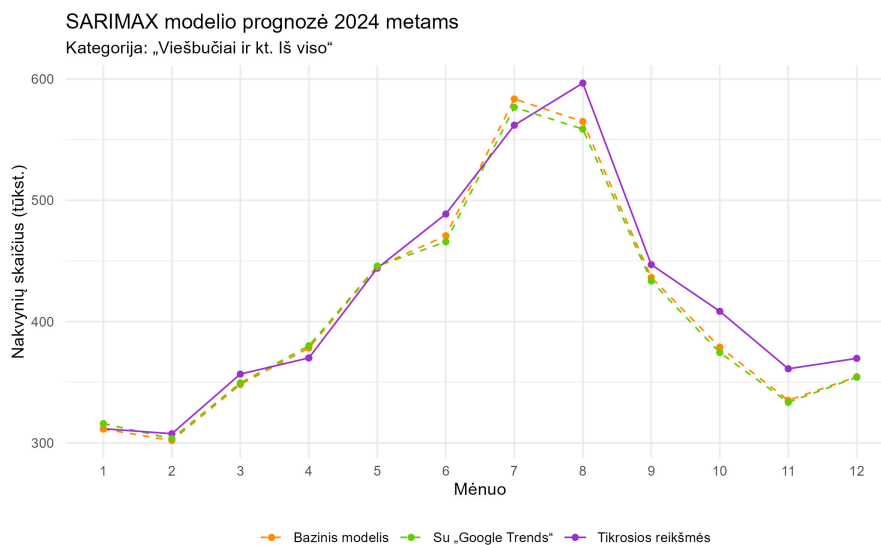
49 pav. Nakvynių skaičiaus vietiniams turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



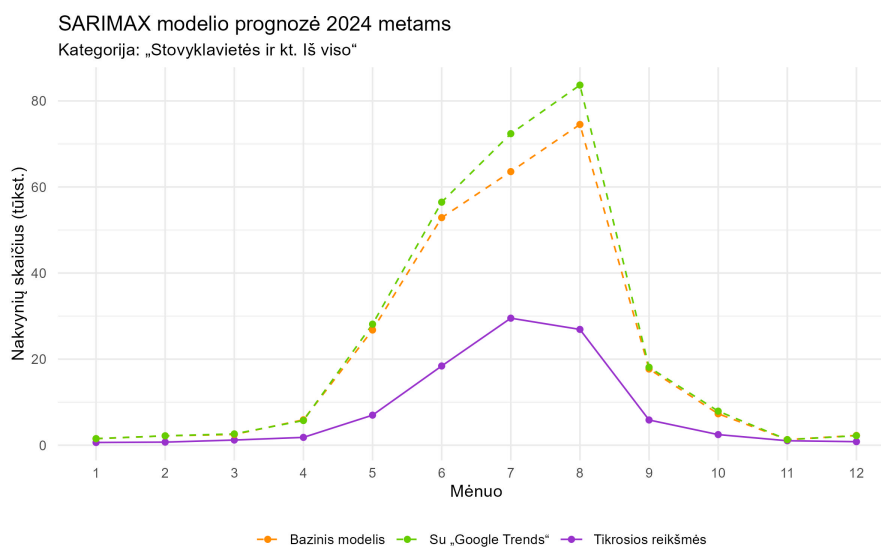
50 pav. Nakvynių skaičiaus vietiniams turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

Nakvynių skaičiaus visiems turistams prognozavimas

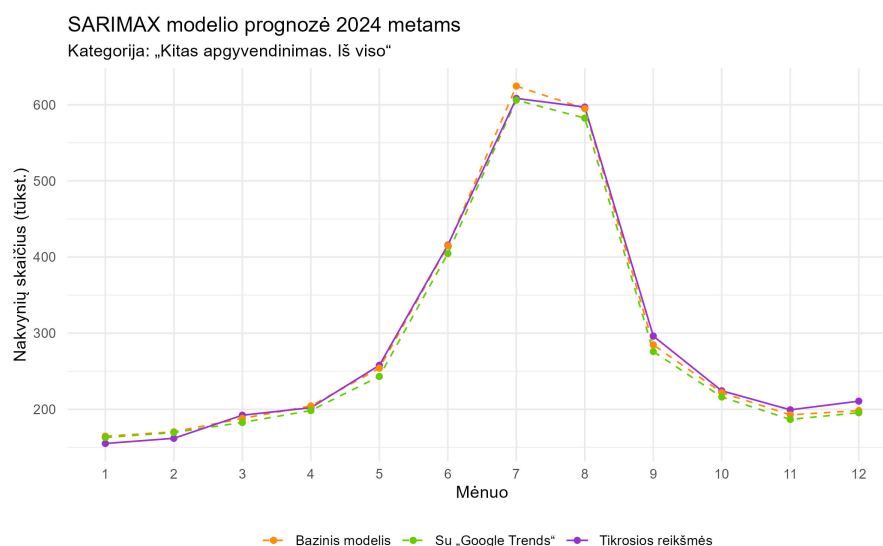
SARIMAX



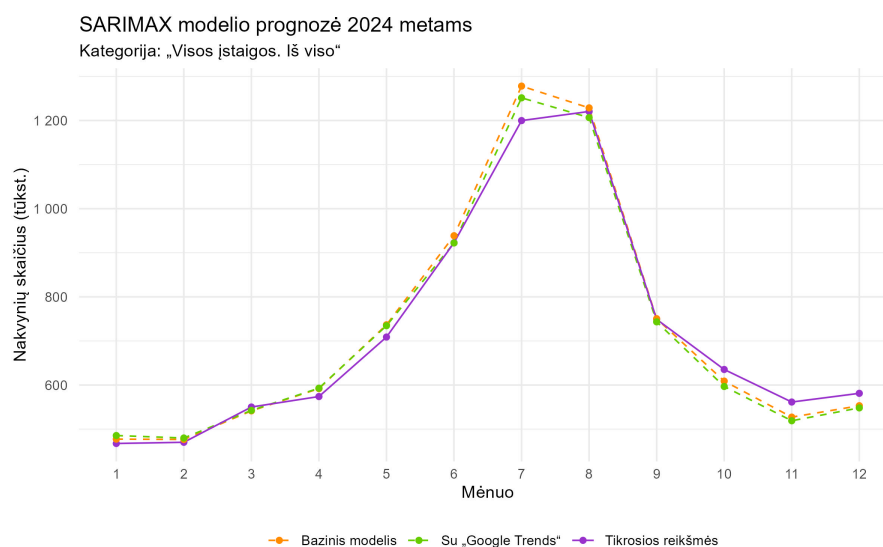
51 pav. Nakvynių skaičiaus visiems turistams viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



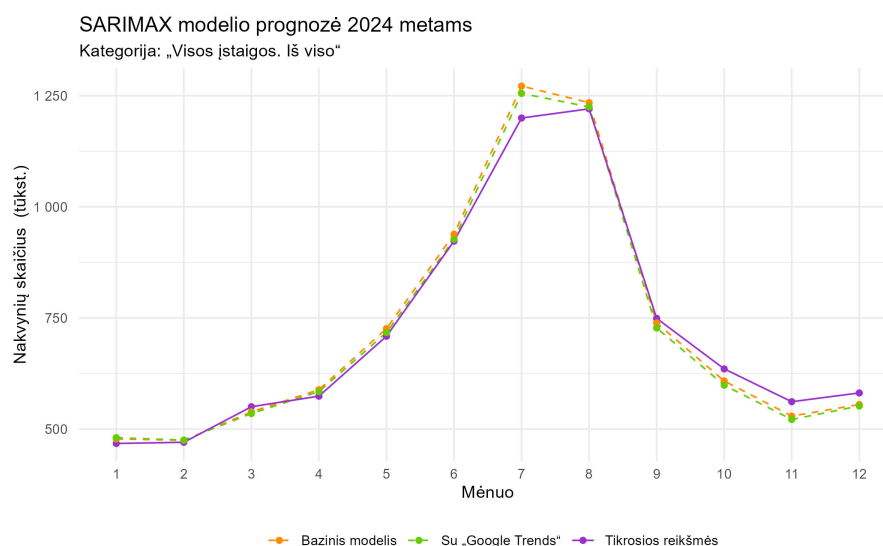
52 pav. Nakvynių skaičiaus visiems turistams stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



53 pav. Nakvynių skaičiaus visiems turistams kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

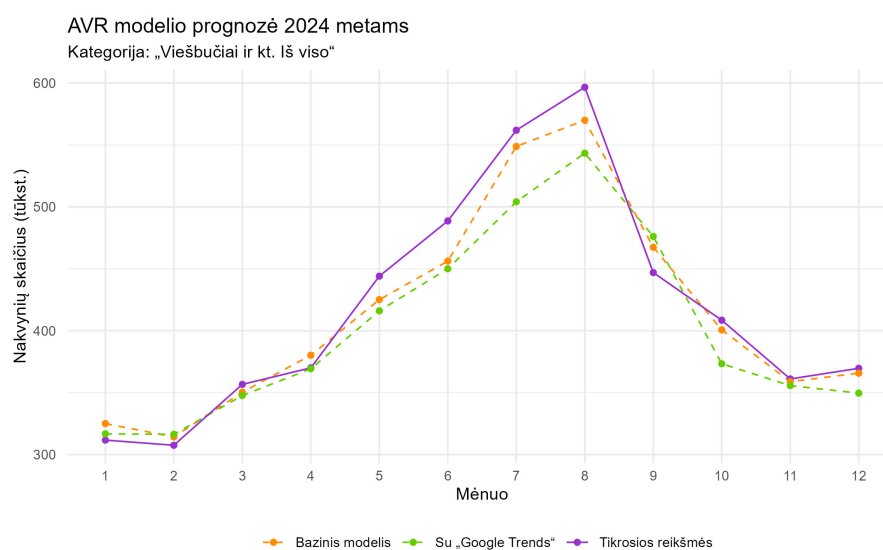


54 pav. Nakvynių skaičiaus visiems turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

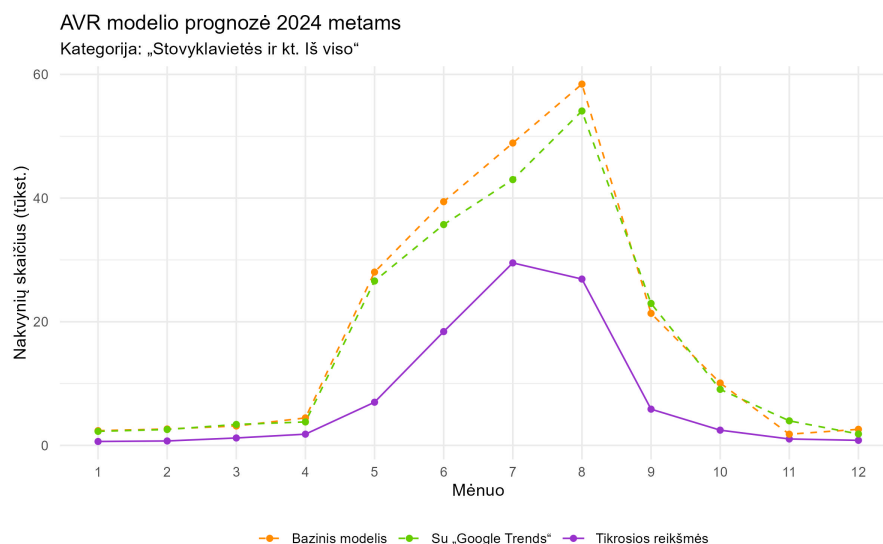


55 pav. Nakvynių skaičiaus visiems turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

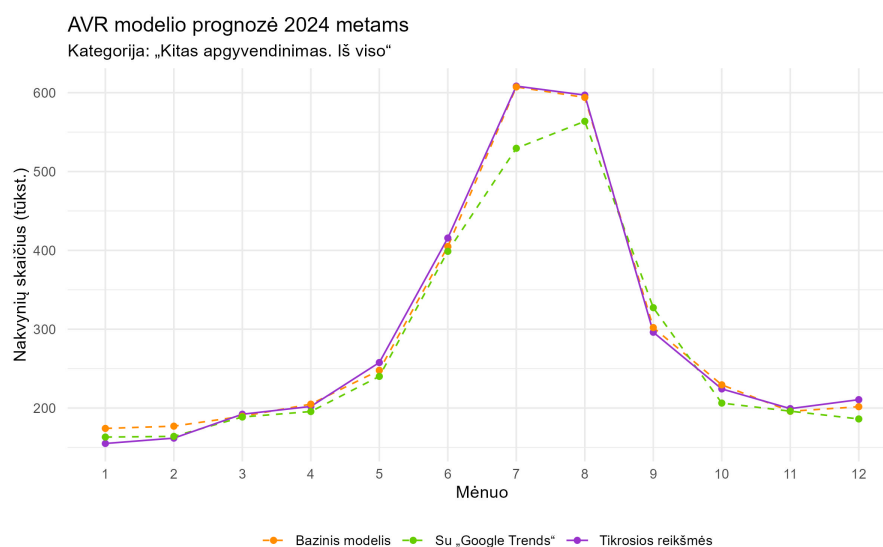
Atraminių vektorių regresija



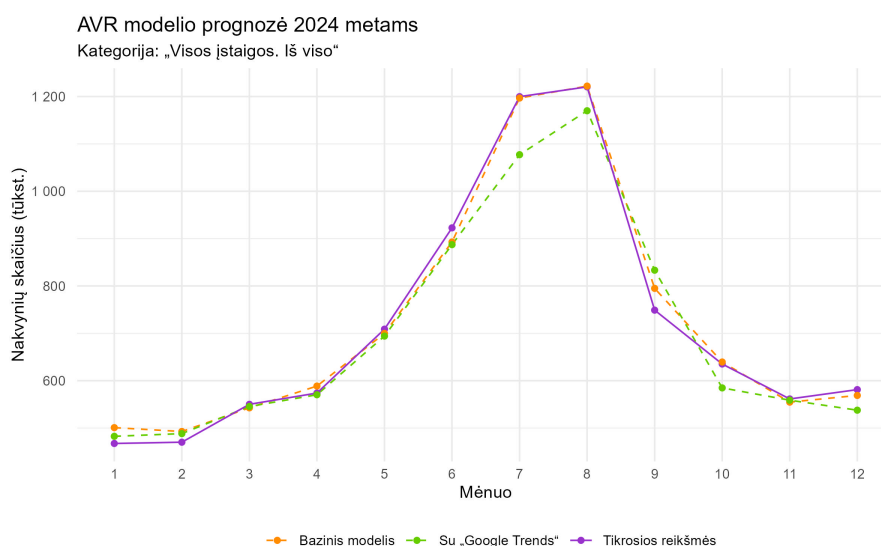
56 pav. Nakvynių skaičiaus visiems turistams viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



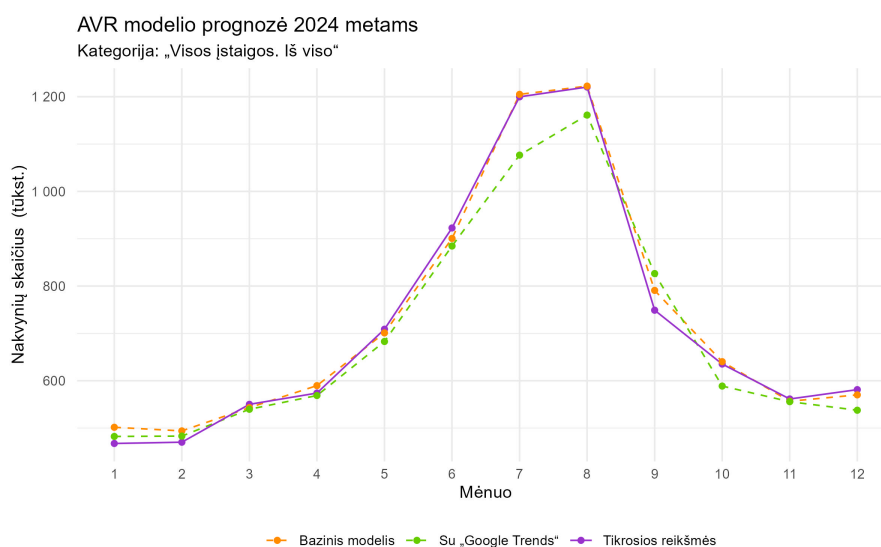
57 pav. Nakvynių skaičiaus visiems turistams stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



58 pav. Nakvynių skaičiaus visiems turistams kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



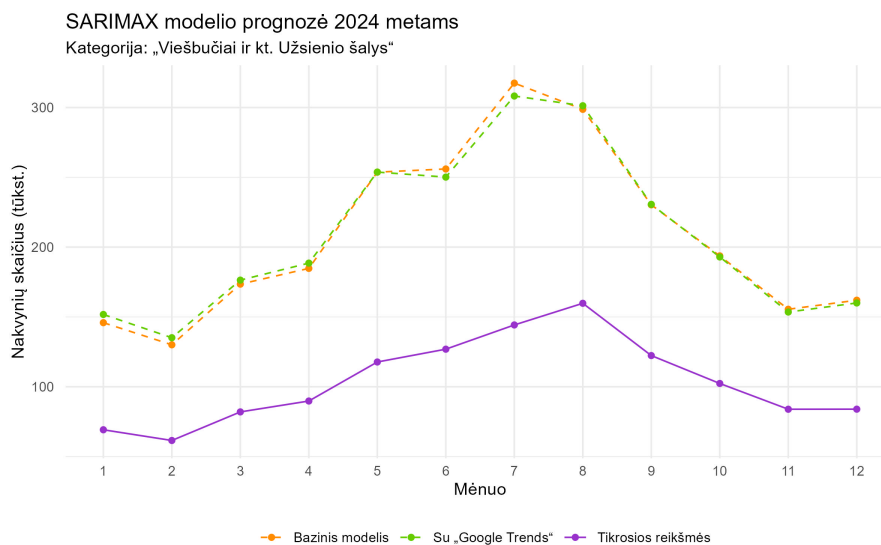
59 pav. Nakvynių skaičiaus visiems turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



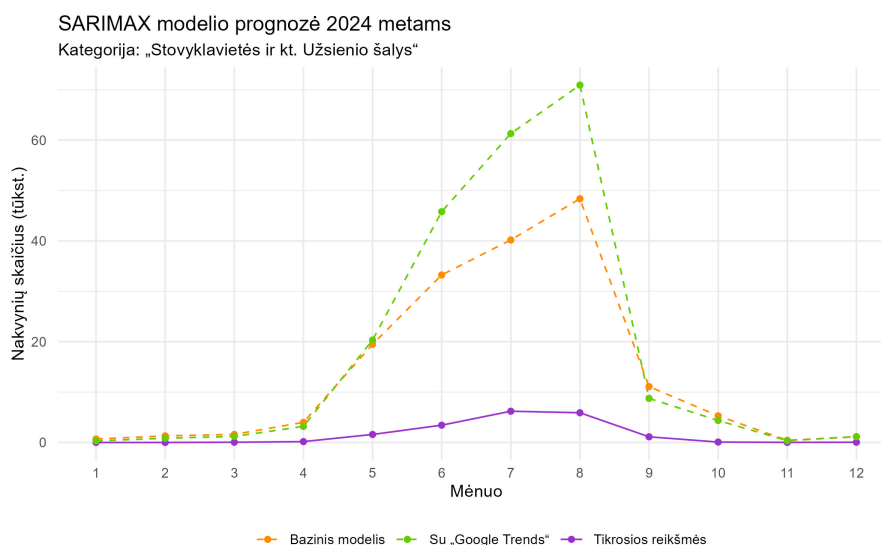
60 pav. Nakvynių skaičiaus visiems turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

Nakvynių skaičiaus užsienio turistams prognozavimas

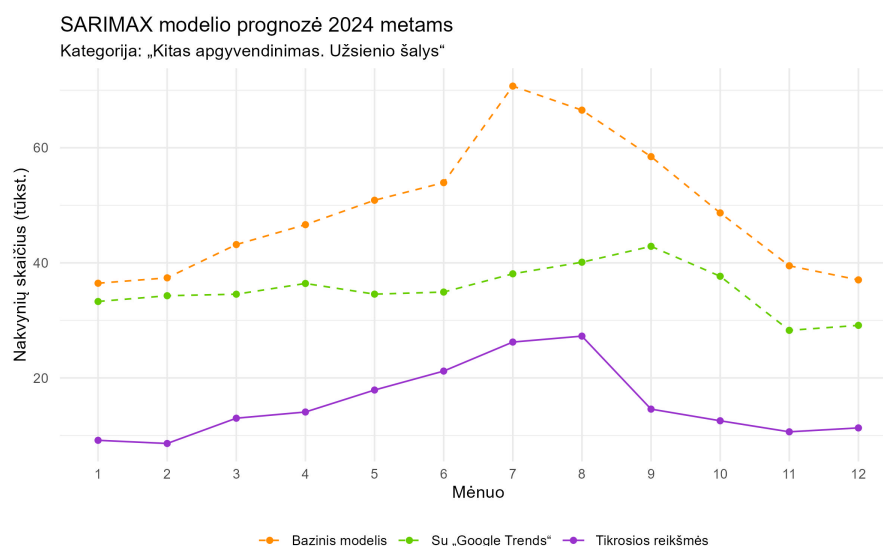
SARIMAX



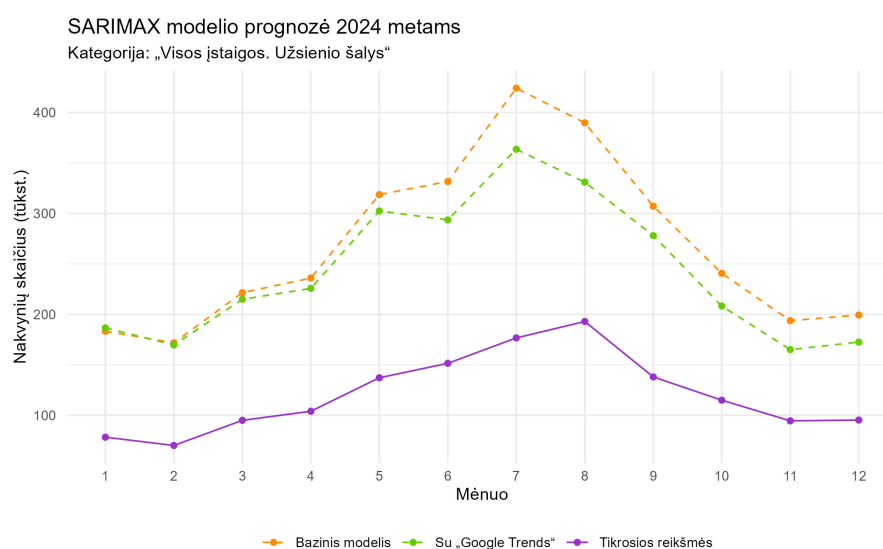
61 pav. Nakvynių skaičiaus užsienio turistams viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



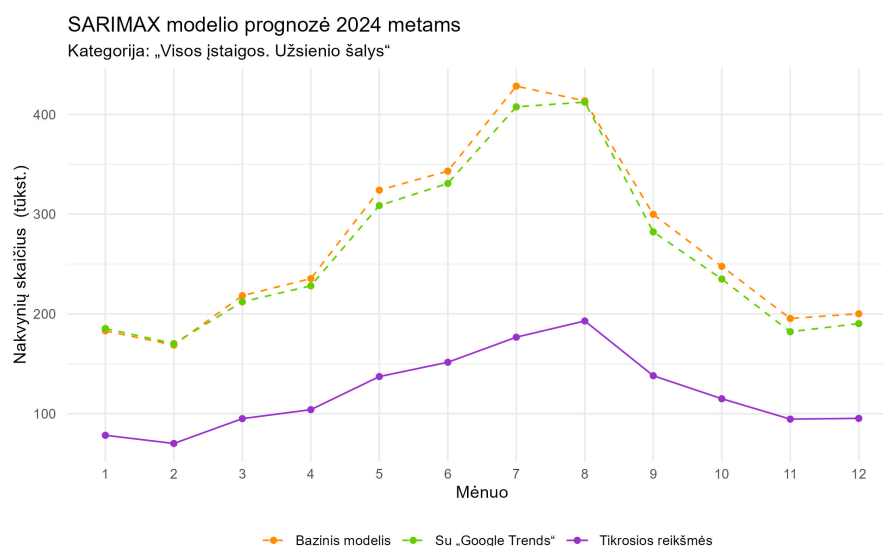
62 pav. Nakvynių skaičiaus užsienio turistams stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



63 pav. Nakvynių skaičiaus užsienio turistams kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

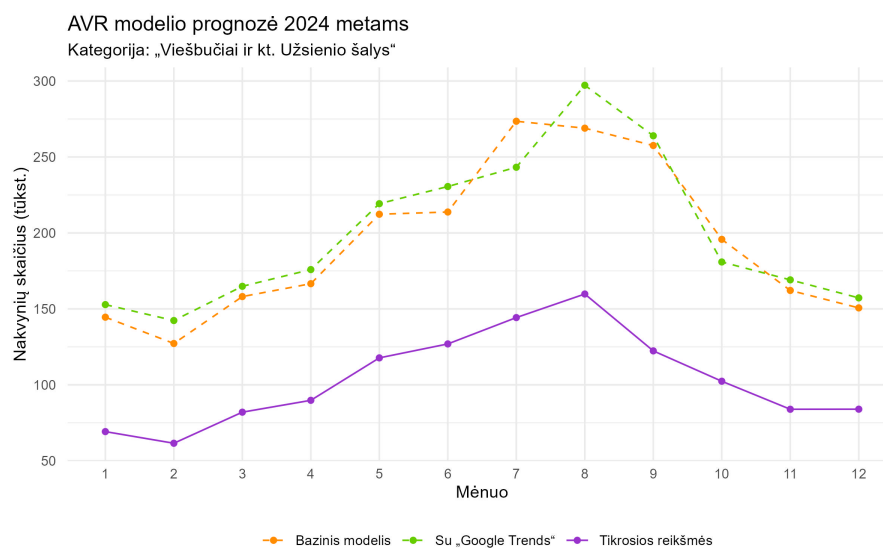


64 pav. Nakvynių skaičiaus užsienio turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams

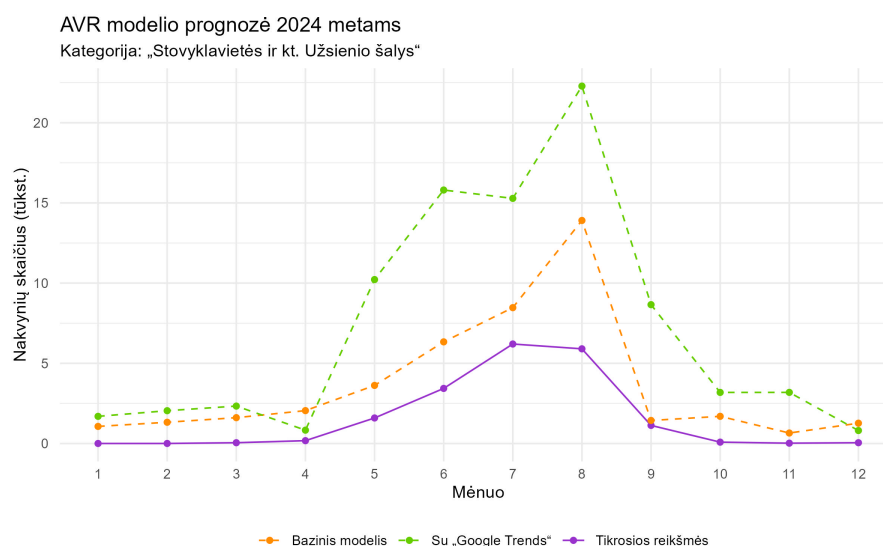


65 pav. Nakvynių skaičiaus užsienio turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

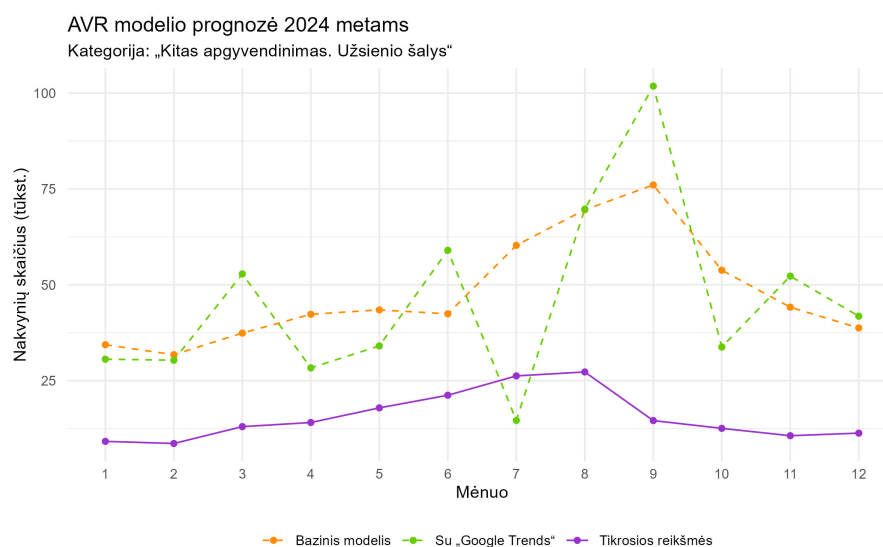
Atraminių vektorių regresija



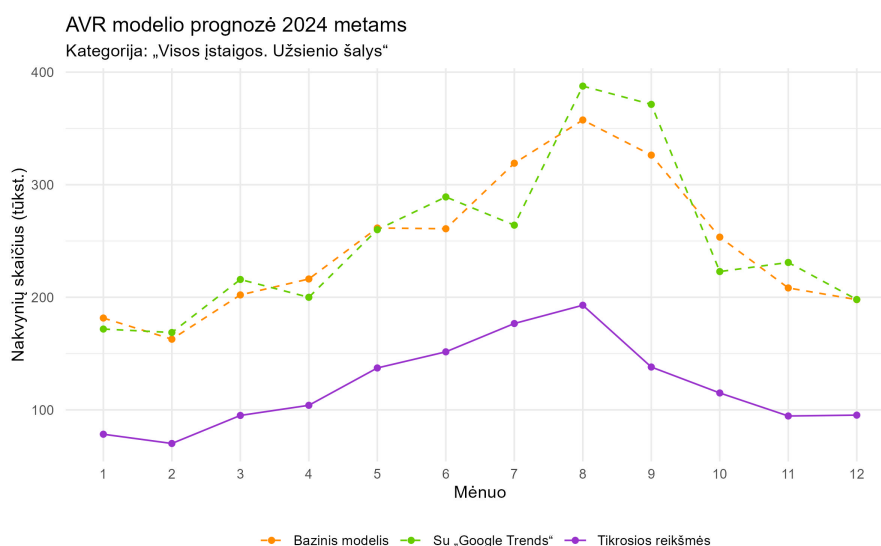
66 pav. Nakvynių skaičiaus užsienio turistams viešbučio tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



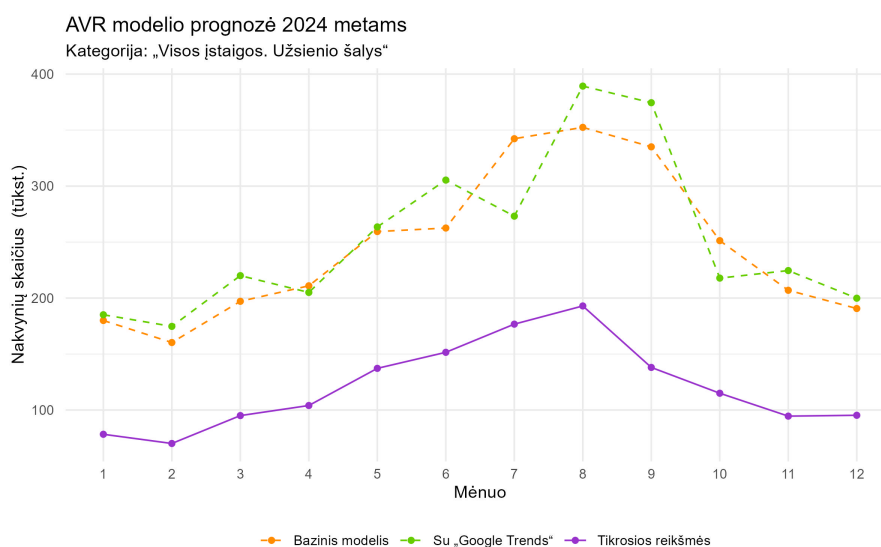
67 pav. Nakvynių skaičiaus užsienio turistams stovyklavietės tipo Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



68 pav. Nakvynių skaičiaus užsienio turistams kitose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



69 pav. Nakvynių skaičiaus užsienio turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams



70 pav. Nakvynių skaičiaus užsienio turistams visose Lietuvos apgyvendinimo įstaigose prognozė 2024 metams, naudojant ankstesnių prognozių agregavimą

3 priedas. Programinis kodas

```
library(readxl)
library(tsibble)
library(forecast)
library(gridExtra)
library(tidyverse)
library(tools)
library(corrplot)
library(tseries)
```

```

library(fpp3)
library(imputeTS)
library(nortest)
library(seasonal)
library(lmtest)
library(ggpubr)
library(e1071)
library(caret)
library(reshape2)
library(FactoMineR)
library(xtable)
library(scales)

# Duomenų įkėlimas ir tvarkymas-----

start <- yearmonth("2012-01")
end <- yearmonth("2024-12")

# Funkcija turistų skaičiaus ir nakvynių skaičiaus duomenų įkėlimui
import_data <- function(file_name){
  tourism_data <- read_excel(file_name, .name_repair = "minimal")

  colnames(tourism_data)[1] <- "Mėnuo"

  tourism_data$Mėnuo <- gsub("M", "-", tourism_data$Mėnuo)
  tourism_data$Mėnuo <- as_date(tourism_data$Mėnuo, format = '%Y-%m')

  tourism_data <- tourism_data %>%
    mutate(across(where(is.character), as.numeric)) %>%
    replace(is.na(.), 0) %>%
    filter(between(tourism_data$Mėnuo, start, end))

  tourism_data <- sort_by(tourism_data, tourism_data$Mėnuo)

  return(tourism_data)
}

# Funkcija Google Trends duomenų įkėlimui ir sujungimui
import_GT<- function(data, wd){
  setwd(wd)

```

```

csv_list <- list.files(wd, include.dirs = FALSE, recursive = TRUE,
                      pattern = "*.csv", full.names = T)

GT_data <- Reduce(function(df1, df2) full_join(df1, df2, by = "Mėnuo"),
                 lapply(csv_list, function(i) {
                   df <- read.csv(i, comment.char = "#", skip = 2)
                   colnames(df) <- c("Mėnuo",
                                     paste0("Paieškos_",
                                             file_path_sans_ext(basename(i))))
                   return(df)
                 }))

setwd("~/Bakalauras")

GT_data$Mėnuo <- as_date(GT_data$Mėnuo, format = '%Y-%m')

GT_data <- GT_data %>%
  filter(between(GT_data$Mėnuo, start, end)) |>
  mutate(Mėnuo = yearmonth(Mėnuo)) |>
  as_tsibble(index = Mėnuo)

data_full <- full_join(data, GT_data, by = "Mėnuo")

colnames(data_full)[c(3, 4, 5, 7, 8, 9, 11, 12, 13)] <-
  c("Viešbučiai ir kt. Iš viso", "Kitas apgyvendinimas. Iš viso",
    "Stovyklavietės ir kt. Iš viso", "Viešbučiai ir kt. Lietuva",
    "Kitas apgyvendinimas. Lietuva", "Stovyklavietės ir kt. Lietuva",
    "Viešbučiai ir kt. Užsienio šalys", "Kitas apgyvendinimas. Užsienio šalys",
    "Stovyklavietės ir kt. Užsienio šalys")

return(data_full)
}

# Apsibrėžiame pandemijos laikotarpius
pandemic_start <- yearmonth("2020-01")
pandemic_gap_st <- yearmonth("2020-06")
pandemic_gap_end <- yearmonth("2020-11")
pandemic_end <- yearmonth("2021-07")

```

```

data_prep <- function(data){

  #Pasiverčiame duomenis į laiko eilutę
  ts <- data |>
    mutate(Mėnuo = yearmonth(Mėnuo)) |>
    as_tsibble(index = Mėnuo)

  rm(data)

  # Įsivedame pandemijos kintamąjį
  ts <- ts |>
    mutate(pandemija = case_when(
      Mėnuo < pandemic_start ~ 0,
      pandemic_start <= Mėnuo & Mėnuo < pandemic_gap_st ~ 1,
      Mėnuo >= pandemic_gap_st & Mėnuo < pandemic_gap_end ~ 2,
      Mėnuo >= pandemic_gap_end & Mėnuo < pandemic_end ~ 1,
      Mėnuo >= pandemic_end ~ 3)) %>%
    mutate(pandemija = factor(pandemija)) %>%
    mutate(laikotarpis =
      factor(ifelse(pandemija == 3, 1, 0)))

  return(ts)
}

# Įkeliamė duomenis ir juos paruošiamė analizei
tourist_num <- import_data("data-table.xlsx")
tourist_num_GT <- import_GT(tourist_num, "~/Bakalauras/Lietuva")
tourist_num_full <- import_GT(tourist_num_GT, "~/Bakalauras/Visos šalys")
ts <- data_prep(tourist_num_full)

overnight_stays <- import_data("Nakvynių skaičius.xlsx")
overnight_stays_GT <- import_GT(overnight_stays, "~/Bakalauras/Lietuva")
overnight_stays_full <- import_GT(overnight_stays_GT, "~/Bakalauras/Visos šalys")
ts_os <- data_prep(overnight_stays_full)

GT_cols <- colnames(ts)[14:39]
LT_GT_cols <- colnames(ts)[14:29]
W_GT_cols <- colnames(ts)[30:39]

```

```

# Priklausomų kintamųjų grafikai
plot_var <- function(long_data, var, file_name){
  plot <- ggplot(long_data, aes(x = as_date(Mėnuo, format = "%Y-%m"), y = value)) +
    geom_line(color = "steelblue") +
    facet_wrap(~ series, scales = "free_y") +
    scale_x_date(date_labels = "%Y",
                 breaks = seq(from = as.Date("2012-01-01"),
                              to = as.Date("2024-12-01"),
                              by = "4 year")) +
    scale_y_continuous(labels = unit_format(unit = "", scale = 1e-3)) +
    labs(title = paste(var,
                       "skirtingose apgyvendinimo įstaigų ir turistų tautybių kategorijose"),
         x = "Data", y = paste(var, "(tūkst.)")) +
    theme_minimal()

  print(plot)

  ggsave(paste(file_name, ".jpg"), plot, width = 10, height = 5, units = "in",
         dpi = 300)
}

```

```

# Google Trends duomenų grafikai
plot_GT <- function(long_data, area, file_name){
  plot <- ggplot(long_data, aes(x = as_date(Mėnuo), y = value)) +
    geom_line(color = "steelblue") +
    facet_wrap(~ series, scales = "free_y") +
    scale_x_date(date_labels = "%Y",
                 breaks = seq(from = as.Date("2012-01-01"),
                              to = as.Date("2024-12-01"),
                              by = "4 year")) +
    labs(title = paste(area, "užklausų populiarumas"),
         x = "Data", y = "Populiarumo indeksas") +
    theme_minimal()

  print(plot)

  ggsave(paste(file_name, ".jpg"), plot, width = 10, height = 5, units = "in",
         dpi = 300)
}

```



```

}

LT_GT_long <- ts %>%
  pivot_longer(cols = LT_GT_cols,
               names_to = "series",
               values_to = "value") %>%
  mutate(series = str_remove(series, "^Paieškos_")) %>%
  mutate(series = str_remove(series, "\\(LT\\)"))

plot_GT(LT_GT_long, "Lietuviškų", "LT_GT_nekoreguoti")

W_GT_long <- ts %>%
  pivot_longer(cols = W_GT_cols,
               names_to = "series",
               values_to = "value") %>%
  mutate(series = str_remove(series, "^Paieškos_")) %>%
  mutate(series = str_replace(series, "lithuania", "Lithuania"))

plot_GT(W_GT_long, "Pasaulio masto", "W_GT_nekoreguoti")

tourist_num_long <- ts %>%
  pivot_longer(cols = -c(starts_with("Paieškos"), "Mėnuo", "pandemija",
                           "laikotarpis"),
               names_to = "series",
               values_to = "value")

plot_var(tourist_num_long, "Turistų skaičius", "Turistų_sk_nekoreguoti")

overnight_stays_long <- ts_os %>%
  pivot_longer(cols = -c(starts_with("Paieškos"), "Mėnuo", "pandemija",
                           "laikotarpis"),
               names_to = "series",
               values_to = "value")

plot_var(overnight_stays_long, "Nakvyių skaičius", "Nakvynių_sk_nekoreguoti")

# Pandemijos laikotarpio duomenų taisymas/įrašymas-----

```

```

correct_pandemic <- function(ts_data){

# Paverčiame pandemijos laikotarpį į praleistus stebėjimus
ts_imputed <- ts_data %>% mutate_at(
  c("Viešbučiai ir kt. Iš viso", "Kitas apgyvendinimas. Iš viso",
    "Stovyklavietės ir kt. Iš viso", "Viešbučiai ir kt. Lietuva",
    "Kitas apgyvendinimas. Lietuva", "Stovyklavietės ir kt. Lietuva",
    "Viešbučiai ir kt. Užsienio šalys", "Kitas apgyvendinimas. Užsienio šalys",
    "Stovyklavietės ir kt. Užsienio šalys"),
  ~ifelse(between(Mėnuo, pandemic_start, pandemic_end), NA, .x))

# Įrašome reikšmes praleistiems stebėjimams naudojant na_seasplit()
ts_imputed <- ts_imputed %>%
  mutate_at(
    c("Viešbučiai ir kt. Lietuva", "Kitas apgyvendinimas. Lietuva",
      "Stovyklavietės ir kt. Lietuva", "Viešbučiai ir kt. Užsienio šalys",
      "Kitas apgyvendinimas. Užsienio šalys", "Stovyklavietės ir kt. Užsienio šalys"),
    function(x) as.numeric(round(na_seasplit(x, find_frequency = TRUE), 0)))

# Įrašome praleistus stebėjimus agregaciniams kintamiesiems
ts_imputed <- ts_imputed %>%
  mutate_at(c("Visos įstaigos. Lietuva"),
    ~ifelse(between(Mėnuo, pandemic_start, pandemic_end),
      rowSums(ts_imputed[, c(7,8,9)]), .x)) %>%
  mutate_at(c("Visos įstaigos. Užsienio šalys"),
    ~ifelse(between(Mėnuo, pandemic_start, pandemic_end),
      rowSums(ts_imputed[, c(11,12,13)]), .x)) %>%
  mutate_at(c("Viešbučiai ir kt. Iš viso"),
    ~ifelse(between(Mėnuo, pandemic_start, pandemic_end),
      rowSums(ts_imputed[, c(7,11)]), .x)) %>%
  mutate_at(c("Kitas apgyvendinimas. Iš viso"),
    ~ifelse(between(Mėnuo, pandemic_start, pandemic_end),
      rowSums(ts_imputed[, c(8,12)]), .x)) %>%
  mutate_at(c("Stovyklavietės ir kt. Iš viso"),
    ~ifelse(between(Mėnuo, pandemic_start, pandemic_end),
      rowSums(ts_imputed[, c(9,13)]), .x))
ts_imputed <- ts_imputed %>%
  mutate_at(c("Visos įstaigos. Iš viso"),
    ~ifelse(between(Mėnuo, pandemic_start, pandemic_end),
      rowSums(ts_imputed[, c(6,10)]), .x))

```

```

# Atrenkame "Google Trends" užklaudas, kurių duomenyse yra ryškių išskirčių
# pandemijos laikotarpiu
GT_outliers <- colnames(ts_imputed)[c(16, 17, 21, 24, 28, 29, 30, 31, 33, 36, 38, 39)]

# "Google Trends" duomenis, pasižyminčius ryškiu sezoniškumu, įrašome naudojant
# na_seasplit, be sezoniškumo - naudojant na_random su nustatytais režiais
set.seed(1)
ts_imputed <- ts_imputed %>%
  mutate(across(all_of(GT_outliers),
    ~ifelse(between(Mėnuo, pandemic_start, pandemic_end), NA, .x))) %>%
  mutate(across(all_of(GT_outliers[c(1,2,5,6,7,11)]),
    ~ as.numeric(round(na_seasplit(.x, find_frequency = TRUE), 0)))) %>%
  mutate(Paieškos_motelis =
    as.numeric(round(na_random(
      Paieškos_motelis, lower_bound = 20, upper_bound = 55), 0))) %>%
  mutate(Paieškos_sanatorija =
    as.numeric(round(na_random(
      Paieškos_sanatorija, lower_bound = 50, upper_bound = 85), 0))) %>%
  mutate(`Paieškos_flights to lithuania` =
    as.numeric(round(na_random(
      `Paieškos_flights to lithuania`, lower_bound = 55,
      upper_bound = 100), 0))) %>%
  mutate(`Paieškos_Kaunas` =
    as.numeric(round(na_random(
      `Paieškos_Kaunas`, lower_bound = 70, upper_bound = 90), 0))) %>%
  mutate(`Paieškos_things to do in lithuania` =
    as.numeric(round(na_random(
      `Paieškos_things to do in lithuania`, lower_bound = 25,
      upper_bound = 75), 0))) %>%
  mutate(`Paieškos_visit lithuania` =
    as.numeric(round(na_random(
      `Paieškos_visit lithuania`, lower_bound = 30, upper_bound = 75), 0)))
return(ts_imputed)
}

ts_imputed_covid <- correct_pandemic(ts)
ts_os_imputed <- correct_pandemic(ts_os)

# Koreguotų duomenų grafikai

```

```

tourist_num_long_corrected <- ts_imputed_covid %>%
  pivot_longer(cols = -c(starts_with("Paieškos"), "Mėnuo", "pandemija",
    "laikotarpis"),
    names_to = "series",
    values_to = "value")

plot_var(tourist_num_long_corrected, "Turistų skaičius", "Turistų_sk_koreguoti")

overnight_stays_long_corrected <- ts_os_imputed %>%
  pivot_longer(cols = -c(starts_with("Paieškos"), "Mėnuo", "pandemija",
    "laikotarpis"),
    names_to = "series",
    values_to = "value")

plot_var(overnight_stays_long_corrected, "Nakvyių skaičius", "Nakvynių_sk_koreguoti")

LT_GT_long_corrected <- ts_imputed_covid %>%
  pivot_longer(cols = LT_GT_cols,
    names_to = "series",
    values_to = "value") %>%
  mutate(series = str_remove(series, "^Paieškos_")) %>%
  mutate(series = str_remove(series, "\\(LT\\)"))

plot_GT(LT_GT_long_corrected, "Lietuviškų", "LT_GT_koreguoti")

W_GT_long_corrected <- ts_imputed_covid %>%
  pivot_longer(cols = W_GT_cols,
    names_to = "series",
    values_to = "value") %>%
  mutate(series = str_remove(series, "^Paieškos_")) %>%
  mutate(series = str_replace(series, "lithuania", "Lithuania"))

plot_GT(W_GT_long_corrected, "Pasaulio masto", "W_GT_koreguoti")

# Koreliacijos koeficientų grafikai
cor_LT_GT <- cor(ts_imputed_covid[LT_GT_cols])

cor_LT_GT_names <- colnames(cor_LT_GT) %>%
  str_remove("^Paieškos_") %>%

```

```

    str_remove("\\(LT\\)")
colnames(cor_LT_GT) <- cor_LT_GT_names
rownames(cor_LT_GT) <- cor_LT_GT_names

png("LT_GT_corrplot.png", width = 700, height = 700)
corrplot(cor_LT_GT, type = "upper", tl.col = "black", tl.srt = 45, number.cex = 0.95,
         method = "number", cl.cex = 1, cl.ratio = 0.1)
title("Pirsono koreliacijos koeficientas tarp lietuviškų
      užklausų populiarumo indekso reikšmių",
      line = 1, cex.main = 1.3)

dev.off()

cor_W_GT <- cor(ts_imputed_covid[W_GT_cols])
cor_W_GT_names <- colnames(cor_W_GT) %>%
  str_remove("^Paieškos_") %>%
  str_replace("lithuania", "Lithuania")
colnames(cor_W_GT) <- cor_W_GT_names
rownames(cor_W_GT) <- cor_W_GT_names

png("W_GT_corrplot.png", width = 700, height = 700)

corrplot(cor_W_GT, type = "upper", tl.col = "black", tl.srt = 45, number.cex = 1.1,
         method = "number", cl.cex = 1, cl.ratio = 0.1)
title("Pirsono koreliacijos koeficientas tarp pasaulinio masto
      užklausų populiarumo indekso reikšmių",
      line = 1, cex.main = 1.3)

dev.off()

# Funkcija koreliacijų koeficientų tarp priklausomo kintamojo ir "Google Trends"
# užklausų ir jų lagų skaičiavimui
corr_coef <- function(gt_cols, df, var) {
  corr_df <- data.frame(Užklausa = character(),
                       stringsAsFactors = FALSE)

  names(corr_df) <- c("Užklausa")

  # Skaičiuojame koreliacijos koeficientus su kiekviena užklausa,
  # bei su lag 1-12 laiko eilutėmis

```

```

for(i in gt_cols){
  cortest <-
    cor.test(y = scale(df[var]),
             x = scale(df[i]))
  cor <- round(cortest$estimate, 3)
  new_row <- data.frame(Užklausa = i, `Kor. koef.` = cor)

  for(l in 1:12){
    cortest_lag <-
      cor.test(y = scale(df[var]),
               x = lag(scale(df[i]),1))
    cor_lag <- round(cortest_lag$estimate, 3)
    new_row[paste0("Lag ",l," kor. koef.")] <- cor_lag
  }
  corr_df <- rbind(corr_df, new_row)
}
corr_df <- corr_df[order(
  corr_df$`Kor..koef.` , decreasing = TRUE),]
}

# Top 5 užklausių ir jų lagų koreliacijos koeficiento su priklausomu kintamuoju
# grafikai
plot_corr <- function(gt_cols, data, var, var_name){
  corr <- corr_coef(gt_cols, data, var)

  GT_top5 <- corr$Užklausa[1:5]
  top5 <- corr[1:5,]

  top5_long <- top5 %>%
    pivot_longer(cols = starts_with("Lag"), names_to = "Lag",
                  values_to = "Correlation") %>%
    mutate(Lag = as.numeric(gsub("[^0-9]", "", Lag)))

  p <- ggplot(top5_long, aes(x = Lag, y = Correlation, fill = Užklausa)) +
    geom_bar(stat = "identity", position = "dodge") +
    geom_hline(yintercept = 0.7, linetype = "dashed", color = "black") +
    theme_minimal() +
    scale_x_continuous(breaks = seq(1, 12, 1)) +
    scale_y_continuous(breaks = seq(-0.5, 1, 0.1)) +
    labs(title = paste0("Koreliacijos koeficientas tarp ",var_name,

```

```

        " ir užklausių su lagais"),
    x = "Lagas (mėnesiai)",
    y = "Koreliacijos koeficientas",
    fill = "Užklausa")
print(p)
return(GT_top5)
}

# SARIMAX modeliai -----

# Funkcija lagų pridėjimui į duomenis
add_lags <- function(data, var, GT_top5){
  ts_lag <- data %>%
    select(c(Mėnuo, pandemija, laikotarpis, var, all_of(GT_top5))) %>%
    mutate(across(c(var, all_of(GT_top5)),
      list(Lag1 = ~ lag(., 1),
           Lag2 = ~ lag(., 2),
           Lag11 = ~ lag(., 11),
           Lag12 = ~ lag(., 12)),
      .names = "{.col}. {.fn}")) %>%
    drop_na()
  return(ts_lag)
}

# Pagrindinių komponentų analizės funkcija
PCA <- function(data, var){
  train_data <- data %>% filter(Mėnuo < yearmonth("2024-01"))
  test_data <- data %>% filter(Mėnuo >= yearmonth("2024-01"))
  vars <- train_data %>%
    as.data.frame() %>%
    select(contains("Lag") & contains("Paieškos"))

  train_pca <- prcomp(vars, scale. = TRUE, center = TRUE)
  train_PC <- as.data.frame(train_pca$x)

  # Testinėje aibėje pagrindines komponentes prognozuojame, kad nebūtų
  # mokymo duomenų nutekėjimo
  test_PC <- as.data.frame(predict(train_pca, newdata = test_data))
  test_data <- test_data %>% select(c(Mėnuo, var, laikotarpis))
  test <- bind_cols(test_data, test_PC)

```

```

train_data <- train_data %>% select(c(Mėnuo, var, laikotarpis))
train <- bind_cols(train_data, train_PC)

return(list(train, test))
}

# Funkcija optimaliam pagrindinių komponentų skaičiaus parinkimui SARIMAX modeliams
find_best_model <- function(var, train_data, test_data){
  predictors <- names(train_data)[grepl("^PC", names(train_data))]
  best_metrics <- c(Inf, Inf, Inf)
  assumptions <- 0
  for (i in length(predictors):1){
    print(i)
    formula <- as.formula(paste(paste0("log(`",var,"`) ~ laikotarpis +"),
                                paste(predictors[1:i], collapse = " + ")))

    sarimax <- train_data |>
      model(ARIMA(formula))

    pred <- predict(sarimax, new_data = test_data)
    acc <- forecast::accuracy(pred, test_data)[,c(7,5,4)]

    resid <- resid(sarimax)$resid
    adf_p <- adf.test(resid)$p.value
    lb_p <- Box.test(resid, lag = log(length(resid)), type = "Ljung-Box")$p.value

    assum <- ifelse(adf_p < 0.05 & lb_p > 0.05, 2, 0)

    if(acc$MAPE < best_metrics[1] & acc$MAE < best_metrics[2] &
       acc$RMSE < best_metrics[3] & assum > assumptions){
      best_metrics <- c(acc$MAPE, acc$MAE, acc$RMSE)
      assumptions <- assum - 1
      best_model <- sarimax
      pc_num <- i
    }
  }
  print(pc_num)
  return(best_model)
}

```



```

# Funkcija SARIMAX modelių tikslumui apskaičiuoti
get_ARIMA_results <- function(base_model, GT_model, test_data, var, y_name){
  predictions_df <- data.frame(Mėnuo = 1:12)

  fcst_base <- forecast(base_model, new_data = test_data)
  predictions_df[["Bazinis modelis"]] <- fcst_base$.mean
  fcst_GT <- forecast(GT_model, new_data = test_data)
  predictions_df[["Su „Google Trends“"]] <- fcst_GT$.mean
  predictions_df[["Tikrosios reikšmės"]] <- test_data[[var]]

  base_acc <- forecast::accuracy(fcst_base, test_data)
  GT_acc <- forecast::accuracy(fcst_GT, test_data)

  results <- bind_rows(base_acc, GT_acc)[,c(1,7,5,4)] %>%
    rename(Modelis = ".model") %>%
    mutate(Modelis = c("Bazinis modelis", "Su „Google Trends“")) %>%
    mutate(MAE = round(MAE, 0)) %>%
    mutate(MAPE = round(MAPE/100, 3))

  plot_results(predictions_df, var, y_name)

  return(list(results, predictions_df))
}

# Funkcija SARIMAX modelių prognozėms vizualizuoti
plot_results <- function(predictions, var, y_name){
  predictions_long <- tibble(
    index = predictions$Mėnuo,
    "Tikrosios reikšmės" = predictions$`Tikrosios reikšmės`,
    "Bazinis modelis" = predictions$`Bazinis modelis`,
    "Su „Google Trends“" = predictions$`Su „Google Trends“`
  ) %>%
    pivot_longer(cols = c(`Tikrosios reikšmės`, `Bazinis modelis`,
                          `Su „Google Trends“`),
                 names_to = "series", values_to = "value")

  range <- range(predictions_long$value)
  max = max(range)
  min = min(range)

```

```

plot <- ggplot(predictions_long, aes(x = index, y = value, color = series,
                                   linetype = series)) +

  geom_line() +
  geom_point() +
  scale_color_manual(values = c("Tikrosios reikšmės" = "darkorchid",
                                "Bazinis modelis" = "darkorange",
                                "Su „Google Trends“" = "chartreuse3")) +
  scale_linetype_manual(values = c("Tikrosios reikšmės" = "solid",
                                    "Bazinis modelis" = "dashed",
                                    "Su „Google Trends“" = "dashed")) +

  ylim(min, max) +
  scale_x_continuous(breaks = seq(1, 12, by = 1)) +
  scale_y_continuous(labels = unit_format(unit = "", scale = 1e-3)) +
  theme_minimal() +
  theme(panel.grid.minor.x = element_blank()) +
  labs(title = "SARIMAX modelio prognozė 2024 metams",
        subtitle = paste0("Kategorija: „", var, "\""),
        x = "Mėnuo", y = paste0(y_name, " (tūkst.)"),
        color = "", linetype = "") +
  theme(legend.position = 'bottom')

print(plot)
ggsave(paste0("SARIMAX. ", var, y_name, ".jpg"), plot,
        width = 8, height = 5, units = "in", dpi = 300)
}

# Funkcija SARIMAX modelių prielaidoms tikrinti
test_assumptions <- function(base_model, GT_model){
  base_resid <- resid(base_model)$resid
  GT_resid <- resid(GT_model)$resid
  print(adf.test(base_resid))
  print(adf.test(GT_resid))
  print(Box.test(base_resid, lag = log(length(base_resid)), type = "Ljung-
Box"))
  print(Box.test(GT_resid, lag = log(length(GT_resid)), type = "Ljung-Box"))
  print(t.test(base_resid))
  print(t.test(GT_resid))
}

# Funkcija apskaičiuoti prognozes, naudojant agregavimą

```

```

get_aggregative_preds <- function(hotels_pred, camps_pred, other_pred, test_data,
                                var){
  preds <- data.frame(Mėnuo = hotels_pred$Mėnuo,
                    `Bazinis modelis` = hotels_pred$`Bazinis modelis` +
                      camps_pred$`Bazinis modelis` + other_pred$`Bazinis modelis`,
                    `Su „Google Trends“` = hotels_pred$`Su „Google Trends“` +
                      camps_pred$`Su „Google Trends“` +
                      other_pred$`Su „Google Trends“`,
                    `Tikrosios reikšmės` = test_data[var])

  colnames(preds) <- c("Mėnuo", "Bazinis modelis", "Su „Google Trends\"",
                      "Tikrosios reikšmės")

  return(preds)
}

# Funkcija apskaičiuoti prognozes užsienio turistams
get_foreign_predictions <- function(W_preds, LT_preds, var){
  F_pred <- data.frame(Mėnuo = W_preds$Mėnuo,
                    "Bazinis modelis" = W_preds$`Bazinis modelis` -
                      LT_preds$`Bazinis modelis`,
                    "Su „Google Trends\"" = W_preds$`Su „Google Trends“` -
                      LT_preds$`Su „Google Trends“`,
                    "Tikrosios reikšmės" = test_F[var])
  colnames(F_pred) <- c("Mėnuo", "Bazinis modelis", "Su „Google Trends\"",
                      "Tikrosios reikšmės")

  return(F_pred)
}

# Funkcijos tikslumo metrikų reikšmėms skaičiuoti
MAPE_fun <- function(test_data, var, pred){
  round(mean(abs(test_data[[var]] - pred) /
            test_data[[var]]), 3)
}

RMSE_fun <- function(test_data, var, pred){
  round(sqrt(mean((test_data[[var]] - pred)^2)), 2)
}

```

```

MAE_fun <- function(test_data, var, pred){
  round(mean(abs(test_data[[var]] - pred)), 0)
}

# Funkcija modelių tikslumui apskaičiuoti
calculate_accuracy <- function(test_data, var, predictions){
  accuracy <- data.frame(Modelis = c("Bazinis modelis", "Su „Google Trends“"),
    MAPE =
      c(MAPE_fun(test_data, var, predictions$`Bazinis modelis`),
        MAPE_fun(test_data, var, predictions$`Su „Google Trends“`)),
    MAE =
      c(MAE_fun(test_data, var, predictions$`Bazinis modelis`),
        MAE_fun(test_data, var, predictions$`Su „Google Trends“`)),
    RMSE =
      c(RMSE_fun(test_data, var, predictions$`Bazinis modelis`),
        RMSE_fun(test_data, var, predictions$`Su „Google Trends“`))

  return(accuracy)
}

# SARIMAX modelių sudarymas-----
# Pateiktas kodas tik vietinių turistų skaičiaus prognozavimui. Visų turistų
# skaičiaus ir nakvynių skaičiaus prognozavimas atliekamas naudojant
# analogišką kodą.

# Užklausų su didžiausia koreliacija išskyrimas kiekvienoje kategorijoje
top5_LT <- plot_corr(LT_GT_cols, ts_imputed_covid,
  "Visos įstaigos. Lietuva", "lietuvių turistų skaičiaus")
top5_LT_hotels <- plot_corr(LT_GT_cols, ts_imputed_covid,
  "Viešbučiai ir kt. Lietuva",
  "lietuvių turistų skaičiaus")
top5_LT_camp <- plot_corr(LT_GT_cols, ts_imputed_covid,
  "Stovyklavietės ir kt. Lietuva",
  "lietuvių turistų skaičiaus")
top5_LT_other <- plot_corr(LT_GT_cols, ts_imputed_covid,
  "Kitas apgyvendinimas. Lietuva",
  "lietuvių turistų skaičiaus")

# Išrinktų užklausų lagų įtraukimas į duomenis
ts_lagged_LT <- add_lags(ts_imputed_covid, "Visos įstaigos. Lietuva", top5_LT)

```

```

ts_lagged_LT_camps <- add_lags(ts_imputed_covid,
                              "Stovyklavietės ir kt. Lietuva", top5_LT_camp)
ts_lagged_LT_hotels <- add_lags(ts_imputed_covid,
                              "Viešbučiai ir kt. Lietuva", top5_LT_hotels)
ts_lagged_LT_other <- add_lags(ts_imputed_covid,
                              "Kitas apgyvendinimas. Lietuva", top5_LT_other)

# Sudarome mokymo ir testines aibes kiekvienai kategorijai atskirai
train_LT <- PCA(ts_lagged_LT, "Visos įstaigos. Lietuva")[[1]]
test_LT <- PCA(ts_lagged_LT, "Visos įstaigos. Lietuva")[[2]]

train_LT_camps <- PCA(ts_lagged_LT_camps, "Stovyklavietės ir kt. Lietuva")[[1]]
test_LT_camps <- PCA(ts_lagged_LT_camps, "Stovyklavietės ir kt. Lietuva")[[2]]

train_LT_hotels <- PCA(ts_lagged_LT_hotels, "Viešbučiai ir kt. Lietuva")[[1]]
test_LT_hotels <- PCA(ts_lagged_LT_hotels, "Viešbučiai ir kt. Lietuva")[[2]]

train_LT_other <- PCA(ts_lagged_LT_other, "Kitas apgyvendinimas. Lietuva")[[1]]
test_LT_other <- PCA(ts_lagged_LT_other, "Kitas apgyvendinimas. Lietuva")[[2]]

# VIETINIAI TURISTAI-----

# Viešbučiai ir kt.
arimax_LT_hotels <- find_best_model("Viešbučiai ir kt. Lietuva", train_LT_hotels,
                                   test_LT_hotels)

arima_LT_hotels <- train_LT_hotels |>
  model(ARIMA(log(`Viešbučiai ir kt. Lietuva`) ~ laikotarpis))

test_assumptions(arima_LT_hotels, arimax_LT_hotels)

LT_hotels_results <- get_ARIMA_results(arima_LT_hotels, arimax_LT_hotels,
                                     test_LT_hotels, "Viešbučiai ir kt. Lietuva",
                                     "Turistų skaičius")
LT_hotels_test_res <- LT_hotels_results[[1]]
LT_hotels_pred <- LT_hotels_results[[2]]
xtable(LT_hotels_test_res,
       caption = "SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių
       viešbučio tipo įstaigose.",
       label = "tab:s-v-lt")

```

```

# Stovyklavietės ir kt.
arimax_LT_camps <- find_best_model("Stovyklavietės ir kt. Lietuva",
                                   train_LT_camps, test_LT_camps)
arima_LT_camps <- train_LT_camps |>
  model(ARIMA(log(`Stovyklavietės ir kt. Lietuva`) ~ laikotarpis))

test_assumptions(arima_LT_camps, arimax_LT_camps)

LT_camps_results <- get_ARIMA_results(arima_LT_camps, arimax_LT_camps,
                                       test_LT_camps, "Stovyklavietės ir kt. Lietuva",
                                       "Turistų skaičius")

LT_camps_test_res <- LT_camps_results[[1]]
LT_camps_pred <- LT_camps_results[[2]]
xtable(LT_camps_test_res,
       caption = "SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių
stovyklavietės tipo įstaigose.",
       label = "tab:s-s-lt")

# Kitas apgyvendinimas
arimax_LT_other <- find_best_model("Kitas apgyvendinimas. Lietuva", train_LT_other,
                                   test_LT_other)
arima_LT_other <- train_LT_other |>
  model(ARIMA(log(`Kitas apgyvendinimas. Lietuva`) ~ laikotarpis ))

test_assumptions(arima_LT_other, arimax_LT_other)

LT_other_results <- get_ARIMA_results(arima_LT_other, arimax_LT_other,
                                       test_LT_other, "Kitas apgyvendinimas. Lietuva",
                                       "Turistų skaičius")

LT_other_test_res <- LT_other_results[[1]]
LT_other_pred <- LT_other_results[[2]]
xtable(LT_other_test_res,
       caption = "SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių
kitose apgyvendinimo įstaigose.",
       label = "tab:s-k-lt")

# Visos įstaigos

```

```

arimax_LT <- find_best_model("Visos įstaigos. Lietuva", train_LT, test_LT)

arima_LT <- train_LT |>
  model(ARIMA(log(`Visos įstaigos. Lietuva`) ~ laikotarpis))
test_assumptions(arima_LT, arimax_LT)

LT_results <- get_ARIMA_results(arima_LT, arimax_LT, test_LT,
                                "Visos įstaigos. Lietuva", "Turistų skaičius")

LT_test_res <- LT_results[[1]]
LT_pred <- LT_results[[2]]
xtable(LT_test_res,
       caption = "SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių
visose įstaigose.",
       label = "tab:s-vi-lt")

# Visos įstaigos, agregacinis metodas
LT_all_pred <- get_aggregative_preds(LT_hotels_pred, LT_camps_pred,
                                     LT_other_pred, test_LT,
                                     "Visos įstaigos. Lietuva")

plot_results(LT_all_pred, "Visos įstaigos. Lietuva", "Turistų skaičius")
LT_all_res <- calculate_accuracy(test_LT, "Visos įstaigos. Lietuva", LT_all_pred)
xtable(LT_all_res,
       caption = "SARIMAX modelio tikslumas prognozuojant vietinių turistų skaičių
visose įstaigose, naudojant ankstesnių prognozių agregavimą.",
       label = "tab:s-vi-lt2")

# Atraminių vektorių regresija -----

# Funkcija lagų įtraukimui į duomenis
add_lags_SVR <- function(data, var, GT_top5){
  ts_lag <- data %>%
    select(c(Mėnuo, pandemija, laikotarpis, var, all_of(GT_top5))) %>%
    mutate(across(c(var, all_of(GT_top5)),
                  list(Lag1 = ~ lag(., 1),
                       Lag12 = ~ lag(., 12)),
                  .names = "{.col}. {.fn}")) %>%
    drop_na()

```

```

    return(ts_lag)
}
# Funkcija pagrindinems komponentams sudaryti AVR modeliams.
# PK sudaromos priklausomam kintamajam ir GT duomenims atskirai
PCA_SVR <- function(data, var){
  train_data <- data %>% filter(Mėnuo < yearmonth("2024-01"))
  test_data <- data %>% filter(Mėnuo >= yearmonth("2024-01"))

  GT_vars <- train_data %>%
    as.data.frame() %>%
    select(contains("Lag") & contains("Paieškos"))
  vars <- train_data %>%
    as.data.frame() %>%
    select(contains("Lag") & !contains("Paieškos"))

  train_pca_GT <- prcomp(GT_vars, scale. = TRUE, center = TRUE)
  train_pca_var <- prcomp(vars, scale. = TRUE, center = TRUE)

  train_GT_PC <- as.data.frame(train_pca_GT$x)
  train_var_PC <- as.data.frame(train_pca_var$x)

  colnames(train_GT_PC) <- paste0("GT_PC", seq_len(ncol(train_GT_PC)))
  colnames(train_var_PC) <- paste0("Var_PC", seq_len(ncol(train_var_PC)))

  test_GT_PC <- as.data.frame(predict(train_pca_GT, newdata = test_data))
  test_var_PC <- as.data.frame(predict(train_pca_var, newdata = test_data))

  colnames(test_GT_PC) <- paste0("GT_PC", seq_len(ncol(test_GT_PC)))
  colnames(test_var_PC) <- paste0("Var_PC", seq_len(ncol(test_var_PC)))

  test_data <- test_data %>% select(c(var))
  test <- bind_cols(test_data, test_GT_PC, test_var_PC)
  train_data <- train_data %>% select(c(var))
  train <- bind_cols(train_data, train_GT_PC, train_var_PC)

  return(list(train, test))
}

# Formulų sudarymas AVR modeliams
SVR_formulas <- function(train_data, var){

```



```

predictors_GT <- names(train_data)[grepl("^GT_PC", names(train_data))]
predictors_var <- names(train_data)[grepl("^Var_PC", names(train_data))]
formula_GT <- as.formula(paste(paste0("`",var,"` ~ "), paste(predictors_GT,
collapse = " + "), "+", paste(predictors_var, collapse = " + ")))

formula_var <- as.formula(paste(paste0("`",var,"` ~ "),
paste(predictors_var, collapse = " + ")))

return(list(formula_var, formula_GT))
}

# Gardelė optimaliems hiperparametrams parinkti
svrGrid <- expand.grid(
  sigma = c(0.001, 0.01, 0.05, 0.06, 0.07, 0.08, 0.1, 0.15, 0.25, 0.5, 1.0),
  C = c(10, 50, 100, 125, 150, 200, 225, 250, 275, 300))

# Duomenų padalinimas kryžminei validacijai
train_control <- trainControl(
  method = "timeslice",
  initialWindow = 36,
  horizon = 12,
  fixedWindow = TRUE)

# Funkcija AVR modelių tikslumo metrikoms ir prognozėms gauti
get_SVR_results <- function(train_control, grid, train_data, test_data, var){
  results_test <- data.frame(Modelis = character(),
                             MAPE = numeric(),
                             MAE = numeric(),
                             MSE = numeric(),
                             stringsAsFactors = FALSE)

  results_train <- data.frame(Modelis = character(),
                              MAPE = numeric(),
                              MAE = numeric(),
                              MSE = numeric(),
                              stringsAsFactors = FALSE)
  predictions_df <- data.frame(Mėnuo = 1:12)

  formulas <- SVR_formulas(train_data, var)
  base_model <- train(formulas[[1]], data = train_data, method = 'svmRadial',

```

```

        trControl = train_control, tuneGrid = grid)
base_pred <- predict(base_model, newdata = test_data)
base_pred_train <- predict(base_model)
mape <- MAPE_fun(test_data, var, base_pred)
mae <- MAE_fun(test_data, var, base_pred)
rmse <- RMSE_fun(test_data, var, base_pred)

mape_train <- MAPE_fun(train_data, var, base_pred_train)
mae_train <- MAE_fun(train_data, var, base_pred_train)
rmse_train <- RMSE_fun(train_data, var, base_pred_train)

new_rows_test <- data.frame(Modelis = "Bazinis modelis",
                             MAPE = mape,
                             MAE = mae,
                             RMSE = rmse)

new_rows_train <- data.frame(Modelis = "Bazinis modelis",
                             MAPE = mape_train,
                             MAE = mae_train,
                             RMSE = rmse_train)

predictions_df[["Bazinis modelis"]] <- base_pred

results_test <- rbind(results_test, new_rows_test)
results_train <- rbind(results_train, new_rows_train)

GT_model <- train(formulas[[2]], data = train_data, method = 'svmRadial',
                  trControl = train_control, tuneGrid = grid)
GT_pred <- predict(GT_model, newdata = test_data)
GT_pred_train <- predict(GT_model)
mape <- MAPE_fun(test_data, var, GT_pred)
mae <- MAE_fun(test_data, var, GT_pred)
rmse <- RMSE_fun(test_data, var, GT_pred)

mape_train <- MAPE_fun(train_data, var, GT_pred_train)
mae_train <- MAE_fun(train_data, var, GT_pred_train)
rmse_train <- RMSE_fun(train_data, var, GT_pred_train)

new_rows_test <- data.frame(Modelis = "Su „Google Trends\"",
                             MAPE = mape,

```

```

        MAE = mae,
        RMSE = rmse)

new_rows_train <- data.frame(Modelis = "Su „Google Trends\"",
                             MAPE = mape_train,
                             MAE = mae_train,
                             RMSE = rmse_train)

predictions_df[["Su „Google Trends"]] <- GT_pred
predictions_df[["Tikrosios reikšmės"]] <- test_data[[var]]

results_test <- rbind(results_test, new_rows_test)
results_train <- rbind(results_train, new_rows_train)

return(list(results_test, results_train, predictions_df))

}

# AVR modelių prognozių grafikai
plot_SVR_results <- function(predictions, var, y_name){
  predictions_long <- tibble(
    index = predictions$Mėnuo,
    "Tikrosios reikšmės" = predictions$`Tikrosios reikšmės`,
    "Bazinis modelis" = predictions$`Bazinis modelis`,
    "Su „Google Trends\"" = predictions$`Su „Google Trends`
  ) %>%
    pivot_longer(cols = c(`Tikrosios reikšmės`, `Bazinis modelis`, `Su „Google Trends`),
                 names_to = "series", values_to = "value")

  range <- range(predictions_long$value)
  max = max(range)
  min = min(range)

  plot <- ggplot(predictions_long, aes(x = index, y = value, color = series,
                                     linetype = series)) +
    geom_line() +
    geom_point() +
    scale_color_manual(values = c("Tikrosios reikšmės" = "darkorchid",

```

```

        "Bazinis modelis" = "darkorange",
        "Su „Google Trends“" = "chartreuse3")) +
scale_linetype_manual(values = c("Tikrosios reikšmės" = "solid",
        "Bazinis modelis" = "dashed",
        "Su „Google Trends“" = "dashed")) +

ylim(min, max) +
scale_x_continuous(breaks = seq(1, 12, by = 1)) +
scale_y_continuous(labels = unit_format(unit = "", scale = 1e-3)) +
theme_minimal() +
theme(panel.grid.minor.x = element_blank())+
labs(title = "AVR modelio prognozė 2024 metams",
      subtitle = paste0("Kategorija: „",var,"""),
      x = "Mėnuo", y = paste0(y_name, ' (tūkst.)'),
      color = "", linetype = "") +
theme(legend.position = 'bottom')

print(plot)
ggsave(paste0("AVR. ", var, y_name,".jpg"), plot,
        width = 8, height = 5, units = "in", dpi = 300)
}

# Pridedame top 5 užklausų lagus (naudojamos tos pačios užklausos, kaip ir
# SARIMAX atveju)
ts_lagged_LT_SVR <- add_lags_SVR(ts_imputed_covid, "Visos įstaigos. Lietuva",top5_LT)
ts_lagged_LT_hotels_SVR <- add_lags_SVR(ts_imputed_covid,
        "Viešbučiai ir kt. Lietuva",top5_LT_hotels)
ts_lagged_LT_camp_SVR <- add_lags_SVR(ts_imputed_covid,
        "Stovyklavietės ir kt. Lietuva",top5_LT_camp)
ts_lagged_LT_other_SVR <- add_lags_SVR(ts_imputed_covid,
        "Kitas apgyvendinimas. Lietuva",top5_LT_other)

# Sudarome mokymo ir testinės aibės kiekvienai kategorijai atskirai
train_LT_SVR <- PCA_SVR(ts_lagged_LT_SVR, "Visos įstaigos. Lietuva")[[1]]
test_LT_SVR <- PCA_SVR(ts_lagged_LT_SVR, "Visos įstaigos. Lietuva")[[2]]

train_LT_hotels_SVR <- PCA_SVR(ts_lagged_LT_hotels_SVR,
        "Viešbučiai ir kt. Lietuva")[[1]]
test_LT_hotels_SVR <- PCA_SVR(ts_lagged_LT_hotels_SVR,
        "Viešbučiai ir kt. Lietuva")[[2]]

```

```

train_LT_camp_SVR <- PCA_SVR(ts_lagged_LT_camp_SVR,
                             "Stovyklavietės ir kt. Lietuva")[[1]]
test_LT_camp_SVR <- PCA_SVR(ts_lagged_LT_camp_SVR,
                             "Stovyklavietės ir kt. Lietuva")[[2]]

train_LT_other_SVR <- PCA_SVR(ts_lagged_LT_other_SVR,
                              "Kitas apgyvendinimas. Lietuva")[[1]]
test_LT_other_SVR <- PCA_SVR(ts_lagged_LT_other_SVR,
                              "Kitas apgyvendinimas. Lietuva")[[2]]

# VIETINIAI TURISTAI-----
# Pateikiamas tik modelių vietinių turistų skaičiaus prognozavimui sudarymas.
# Kiti modeliai sudaromi naudojant analogišką kodą.

# Viešbučiai ir kt.
SVR_LT_hotels_res <- get_SVR_results(train_control, svrGrid, train_LT_hotels_SVR,
                                     test_LT_hotels_SVR, "Viešbučiai ir kt. Lietuva")

SVR_LT_hotels_test_res <- SVR_LT_hotels_res[[1]]
SVR_LT_hotels_preds <- SVR_LT_hotels_res[[3]]

xtable(SVR_LT_hotels_test_res,
       caption = "AVR modelio tikslumas prognozuojant vietinių turistų skaičių
       viešbučio tipo įstaigose.",
       label = "tab:a-v-lt")

plot_SVR_results(SVR_LT_hotels_preds, "Viešbučiai ir kt. Lietuva",
                 "Turistų skaičius")

# Stovyklavietės ir kt.
SVR_LT_camp_res <- get_SVR_results(train_control, svrGrid, train_LT_camp_SVR,
                                   test_LT_camp_SVR, "Stovyklavietės ir kt. Lietuva")

SVR_LT_camp_test_res <- SVR_LT_camp_res[[1]]
SVR_LT_camp_preds <- SVR_LT_camp_res[[3]]

xtable(SVR_LT_camp_test_res,
       caption = "AVR modelio tikslumas prognozuojant vietinių turistų skaičių
       stovyklavietės tipo įstaigose.",
       label = "tab:a-s-lt")

```

```

plot_SVR_results(SVR_LT_camp_preds, "Stovyklavietės ir kt. Lietuva",
                  "Turistų skaičius")

# Kitas apgyvendinimas
SVR_LT_other_res <- get_SVR_results(train_control, svrGrid, train_LT_other_SVR,
                                    test_LT_other_SVR, "Kitas apgyvendinimas. Lietuva")

SVR_LT_other_test_res <- SVR_LT_other_res[[1]]
SVR_LT_other_preds <- SVR_LT_other_res[[3]]

xtable(SVR_LT_other_test_res,
       caption = "AVR modelio tikslumas prognozuojant vietinių turistų skaičių
       kitose įstaigose.",
       label = "tab:a-k-lt")

plot_SVR_results(SVR_LT_other_preds, "Kitas apgyvendinimas. Lietuva",
                  "Turistų skaičius")

# Visos įstaigos
SVR_LT_res <- get_SVR_results(train_control, svrGrid, train_LT_SVR, test_LT_SVR,
                              "Visos įstaigos. Lietuva")

SVR_LT_test_res <- SVR_LT_res[[1]]
SVR_LT_preds <- SVR_LT_res[[3]]

xtable(SVR_LT_test_res,
       caption = "AVR modelio tikslumas prognozuojant vietinių turistų skaičių
       visose įstaigose.",
       label = "tab:a-vi-lt")

plot_SVR_results(SVR_LT_preds, "Visos įstaigos. Lietuva", "Turistų skaičius")

# Visos įstaigos. Agregacinis metodas
SVR_LT_all_pred <- get_aggregative_preds(SVR_LT_hotels_preds, SVR_LT_camp_preds,
                                         SVR_LT_other_preds, test_LT_SVR,
                                         "Visos įstaigos. Lietuva")

SVR_LT_all_res <- calculate_accuracy(test_LT_SVR, "Visos įstaigos. Lietuva",
                                     SVR_LT_all_pred)

```

```
xtable(SVR_LT_all_res,  
       caption = "AVR modelio tikslumas prognozuojant vietinių turistų skaičių  
       visose įstaigose, naudojant ankstesnių prognozių agregavimą.",  
       label = "tab:a-vi-lt2")  
  
plot_SVR_results(SVR_LT_all_pred, "Visos įstaigos. Lietuva", "Turistų skaičius")
```