

Vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimas naudojant „Google Trends“ duomenis

Nora Jasiukėnaitė, Rūta Levulienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Taikomosios matematikos institutas Naugarduko g. 24, Vilnius
nora.jasiukenaite@mif.stud.vu.lt

Santrauka. Šiame darbe į laiko eilučių prognozavimo modelius – sezoninį ARIMA modelį su egzogeniniais kintamaisiais (SARIMAX) ir atraminių vektorių regresiją (AVR) – įtraukiami internetinės paieškos sistemos „Google“ duomenys, siekiant pagerinti modelių tikslumą vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimui. Nustatyta, jog modeliai, sudaryti naudojant „Google Trends“ duomenis, pasižymi mažesnėmis paklaidomis.

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

1 Įvadas

Trumpalaikio apgyvendinimo paslaugos yra vienas svarbiausių aspektų, padedančių užtikrinti turizmo sektoriaus plėtrą – apgyvendinimo paslaugų prieinamumas ir kokybė gali lemti keliautojų sprendimus dėl kelionės maršruto bei trukmės. Be to, „Keliauk Lietuvoje“ duomenimis, 2023 m. Lietuvos piliečiai sudarė net 66 % visų turistų Lietuvoje [1]. Taigi, svarbu užtikrinti tikslingą apgyvendinimo paslaugų plėtrą bei gerinimą, o būsimų vietos turistų srautų prognozė gali būti naudojama kaip įrankis šiam tikslui.

Turistų srautų prognozavimui naudojami įvairūs metodai – tiek klasikiniai laiko eilučių [2], [3], tiek mašininio mokymosi bei dirbtinio intelekto modeliai [4], [5]. Pastarųjų metų tyrimai rodo, jog internetinių paieškos sistemų, kaip „Google“ ir „Baidu“, užklausų populiarumo duomenų įtraukimas į šiuos modelius gali žymiai pagerinti jų tikslumą [6], [7], [8]. Tačiau nėra ištirta, kaip ši metodika veikia Lietuvos kontekste, naudojant paieškos užklausas lietuvių kalba. Šio darbo tikslas – nustatyti, ar „Google Trends“ duomenų įtraukimas į klasikinius laiko eilučių bei mašininio mokymosi modelius pagerina vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimo tikslumą.

2 Duomenys

Tyrime naudojami viešai prieinami Valstybės duomenų agentūros mėnesiniai turistų skaičiaus Lietuvos apgyvendinimo įstaigose 2012-2024 m. duomenys [9]. Taip pat įtraukiami 2012-2024 m. mėnesiniai „Google Trends“ [10] duomenys užklausoms „apartamentai“, „viešbutis“, „Vilnius“, „Kaunas“ ir „lankytinos vietos“, kuriuose pateikiamas santykinis užklausos populiarumas pagal bendrą paieškų geografinėje vietovėje skaičių ir atitinkamą laikotarpį. [10] Naudotinos užklausos atrinktos iš didesnės aibės pagal koreliacijos koeficiento su turistų skaičiumi reikšmę.

Dėl situacijos specifiškumo ir įtakos prognozavimo tikslumui, COVID-19 pandemijos laikotarpio (2020 m. sausis - 2021 m. liepa) duomenys buvo laikomi išskirtimis bei koreguoti padalinus laiko eilutę pagal mėnesį ir naudojant tiesinę interpoliaciją reikšmių įrašymui.

Laiko eilutė buvo padalinta į mokymo ir testavimo aibes – paskutinių turimų, t. y. 2024, metų duomenys naudojami modelių tikslumui įvertinti.

3 Metodai ir rezultatai

Tyrime laiko eilučių prognozavimui taikyti du modeliai – sezoninis ARIMA modelis su egzogeniniais kintamaisiais (SARIMAX) [11] bei atraminių vektorių regresija (AVR) su Gauso branduoliu [12].

ARIMA yra plačiausiai taikomas klasikinis laiko eilučių analizės ir prognozavimo modelis. SARIMAX yra šio modelio atmaina, išplečianti jo galimybes įtraukiant sezoniškumo duomenyse išskyrimą – taip leidžianti modelį taikyti nestacionariems duomenims – ir naudojant išorinius kintamuosius, kurie suteikia papildomos informacijos ir pagerina modelio tikslumą [11]. SARIMAX modeliui būtina nustatyti parametrus: autoregresijos (p), integravimo (d) ir slenkančio vidurkio (q) bei sezoninių autoregresijos (P), integravimo (D) ir slenkančio vidurkio (Q) komponentų reikšmes. Parametras p nustato autoregresinių lagų skaičių, d keičia diferencijavimo skirtumų skaičių, o q nurodo slenkančio vidurkio lango dydį. Parametrai P , Q ir D yra anksčiau minėtų parametrų atitikmenys, skirti apdoroti sezoniškumu pasižyminčius duomenis [11]. Taip pat nustatomas ir stebėjimų skaičius sezone s – mėnesiniams duomenims šio parametro reikšmė lygi 12.

AVR yra prižiūrimo (angl. *supervised*) mašininio mokymosi modelio – atraminių vektorių klasifikatoriaus (AVK) – atmaina, pritaikyta regresijos uždaviniams spręsti. AVR transformuoja mokymo duomenis į aukštesnės dimensijos erdvę, kurioje duomenims yra taikomas tiesinis modelis [13]. Taikant atsitiktinių vektorių regresijos modelį, nustatomi modelio sudėtingumo pa-

parametrai – C , kuris nustato santykį tarp modelio sudėtingumo ir tikslumo, ir ϵ , kuriuo kontroliuojama leistina paklaida [13].

Šie modeliai buvo pasirinkti dėl jų populiarumo sprendžiamos problemos kontekste [14], galimybės įtraukti egzogeninius kintamuosius bei tinkamumo mažoms duomenų aibėms. Tyrimo metu buvo išbandyti ir sudėtingesni mašininio mokymosi modeliai (atsitiktinių miškų regresija [15], *XGBoost* (angl. *Extreme Gradient Boosting*) [16] ir *LightGBM* (angl. *Light Gradient Boosting Machine*) [17]), tačiau dėl riboto duomenų kiekio (turimoje duomenų aibėje yra 156 stebėjimai, pritaikius diferencijavimą ir įtraukus kintamųjų lagus skaičius sumažėjo iki 131), šių modelių veiksmingumas nebuvo optimalus ir buvo susidurta su persimokymo problema, tad jie nėra įtraukiami.

Abiejų modelių atveju išbandytos trys skirtingos kintamųjų kombinacijos: bazinis modelis (naudojant tik istorinius turistų skaičiaus duomenis); įtraukus visų užklausų „Google Trends“ duomenis; įtraukus 3 užklausų, pasižyminčių didžiausiu koreliacijos koeficientu su turistų skaičiumi, „Google Trends“ duomenis.

Siekiant padidinti modelio tikslumą ir patenkinti modelio prielaidas, SARIMAX modelių skaitiniai egzogeniniai kintamieji („Google Trends“ duomenys) paversti į stacionarius – panaikintas sezoniskumas bei tendencija naudojant diferencijavimą. Modelių parametrai buvo parinkti minimizuojant Akaičės informacinį kriterijų (AIC).

AVR modeliai sudaryti naudojant nestacionarius duomenis dėl šio modelio galimybių išskirti sudėtingus, netiesinius ryšius, sezoniskumą ir tendencijas duomenyse [13]. Modelių hiperparametrai parinkti naudojant paiešką gardelėje ir slenkančio lango kryžminę patikrą [18].

Modelių palyginimui naudota vidutinė absoliutinė procentinė paklaida (MAPE), vidutinė absoliutinė paklaida (MAE) bei šaknis iš vidutinės kvadratinės paklaidos (RMSE). Šios metrikos skaičiuojamos pagal formules:

$$MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{\hat{Y}_t - Y_t}{Y_t} \right|,$$

$$MAE = \frac{1}{N} \sum_{t=1}^N |\hat{Y}_t - Y_t|,$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (\hat{Y}_t - Y_t)^2}.$$

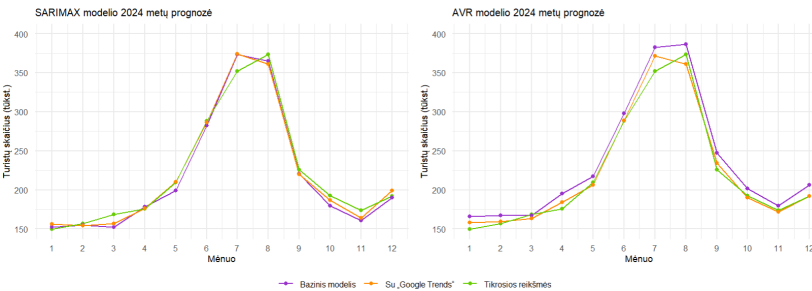
Čia N – stebėjimų skaičius testavimo aibėje, $\hat{Y}_t$ – t -ojo stebėjimo prognozuota reikšmė, Y_t – t -ojo stebėjimo tikroji reikšmė.

1 lentelėje pateiktos bazinio bei geriausio modelio su „Google Trends“ duomenimis tikslumo metrikos testinėje aibėje. Rezultatai rodo, jog tiek SARIMAX, tiek AVR modelio prognozavimo tikslumas padidėja įtraukus „Google Trends“ duomenis, tačiau AVR modelio atveju šis skirtumas yra ryškesnis. Taip pat rezultatai atskleidžia, jog AVR modelis, naudojantis „Google Trends“ duomenis, yra tiksliausias – jo absoliutinė procentinė paklaida siekia vos 2,66 %.

1 lentelė. Modelių tikslumo metrikos. Skliausteliuose nurodytas procentinis pokytis, lyginant su baziniu modeliu.

Modelis		MAPE	MAE	RMSE
SARIMAX	Bazinis	0,0394	8616	10534
	Su „Google Trends“	0,0318 (-19,3 %)	7183 (-16,6 %)	9370 (-11 %)
AVR	Bazinis	0,0614	13298	15324
	Su „Google Trends“	0,0266 (-56,7 %)	6111 (-54 %)	8154 (-46,8 %)

1 paveiksle pateikti bazinių ir „Google Trends“ duomenis naudojančių modelių prognozių ir tikrųjų turistų skaičiaus reikšmių testinėje aibėje grafikai. Grafikai patvirtina modelių tikslumo metrikų rezultatus – tiek SARIMAX, tiek AVR modelio prognozė yra arčiau tikrųjų reikšmių, kai įtraukiami „Google Trends“ duomenys. AVR atveju šis skirtumas ryškesnis – bazinis modelis žymiai pervertina turistų skaičių.



1 pav. Modelių prognozės 2024 metams.

4 Išvados

Tyrimo metu gauti rezultatai atskleidė, jog „Google Trends“ duomenys žymiai pagerina vietinių turistų skaičiaus Lietuvos apgyvendinimo įstaigose prognozavimo tikslumą. Didžiausias tikslumas pasiektas naudojant mašininio mokymosi algoritmą – atraminių vektorių regresijos modelį su internetinės paieškos užklausų populiarumo informacija. Tyrimas gali būti plečiamas įtraukiant sudėtingesnius mašininio mokymosi arba jungtinius algoritmus bei didinant internetinės paieškos užklausų aibę.

Literatūra

- [1] VŠĮ „Keliauk Lietuvoje“ (2024). Turizmo sektoriaus apžvalga. Prieiga: https://lithuania.travel/other_files/Turizmo%20sektoriaus%20ap%C5%BEvalga_Keliauk%20Lietuvoje_2024.pdf (žiūrėta 2025-04-23).
- [2] Lim, C., McAleer, M. (2001). Forecasting tourist arrivals. *Annals of Tourism Research*, 28(4), 965-977.
- [3] Cho, V. (2001). Tourism forecasting and its relationship with leading economic indicators. *Journal of Hospitality & Tourism Research*, 25(4), 399-420.
- [4] Claveria, O., Monte, E., Torra, S. (2015). Tourism Demand Forecasting with Neural Network Models: Different Ways of Treating Information. *International Journal of Tourism Research*, 17(5), 492-500.
- [5] Sun, S., Du, Z., Zhang, C., Wang, S. (2022). Improving multi-step ahead tourism demand forecasting: A strategy-driven approach. *Expert Systems with Applications*, 210, 118465.
- [6] Höpken, W., Eberle, T., Fuchs, M. et al (2019). Google Trends data for analysing tourists' online search behaviour and improving demand forecasting: the case of Åre, Sweden. *Information Technology & Tourism*, 21, 45-62.
- [7] Yang, X., Pan, B., Evans, J. A., Benfu, L. V. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, 46, 386-397.
- [8] Bokelmann, B., Lessmann, S. (2019). Spurious patterns in Google Trends data-An analysis of the effects on tourism demand forecasting in Germany. *Tourism Management*, 75, 1-12.
- [9] Valstybės duomenų agentūra (2025). Turistų skaičius apgyvendinimo įstaigose. Prieiga: <https://osp.stat.gov.lt/lt/statistiniu-rodikliu-analize?hash=21ac96a1-7229-4a43-bdbb-c1ae9175226f#/> (žiūrėta 2025-04-23).
- [10] Rogers, S. (2016). What is Google Trends data — and what does it mean? Google News Lab. Prieiga: <https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8> (žiūrėta 2025-04-23).
- [11] Alharbi, F. R., Csala, D. (2022). A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach. *Inventions*, 7(4), 94.
- [12] Ojemakinde, B. T. (2006). Support Vector Regression for Non-Stationary Time Series. *Magistro darbas. University of Tennessee*, 28-42.

- [13] Crone, S. F., Guajardo, J., Weber, R. (2006). A study on the ability of Support Vector Regression and Neural Networks to Forecast Basic Time Series Patterns. *Artificial Intelligence in Theory and Practice, IFIP 19th World Computer Congress*, 149-158.
- [14] Song, H., Qiu, R. T. R., Park, J. (2019). A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting. *Annals of Tourism Research*, 75, 338-362.
- [15] Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32.
- [16] Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- [17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30.
- [18] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4), 437-450.