Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Anže Schwarzmann

Analiza javnega mestnega potniškega prometa z napovedovanjem časovnih vrst

magistrsko delo

magistrski program druge stopnje
računalništvo in informatika

Mentor: doc. dr. Matej Guid

Ljubljana, 2018

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
ANN	artificial neural network	umetna nevronska mreža
APC	automatic passenger counters	samodejni števec potnikov
AR	autoregression model	avtoregresijski model
ARIMA	autoregressive integrated moving average	avtoregresijski integriran model drsečih sredin
ARIMAX	autoregressive integrated moving average with explanatory variable	avtoregresijski integriran model drsečih sredin s pojasnjevalnimi spremenljivkami
ARSO	Slovenian environment agency	Agencija Republike Slovenije za okolje
AVL	automatic vehicle location	samodejni lokalizator vozil
CSV	comma-separated values	vrednosti, ločene z vejicami
GPS	global positioning system	globalni sistem pozicioniranja
JPP	public passenger transport	javni potniški promet
k-NN	k-nearest neighbors	k-najbližjih sosedov
LPP	Ljubljana passenger transport	Ljubljanski potniški promet
MAE	mean absolute error	povprečna absolutna napaka
MOL	City of Ljubljana	Mestna občina Ljubljana
P + R	park and ride	parkiraj in pelji
PCA	principal component regression	analiza glavnih komponent
RME	relative mean error	relativna srednja napaka
RMSE	root-mean-square error	korenski srednji kvadrat napake
SVM	support vector machine	metoda podpornih vektorjev
VAR	vector autoregression	vektorski avtoregresijski model

Povzetek

Naslov: Analiza javnega mestnega potniškega prometa z napovedovanjem časovnih vrst

Javni potniški promet je pomemben sestavni del vsakega večjega mesta, saj omogoča preprosto in med drugim tudi okolju bolj prijazno potovanje po mestu. Pričakujemo lahko, da se bo s povečevanjem števila prebivalstva povečevalo tudi povpraševanje po prevozih z javnim potniškim prometom. Še zlasti pri uporabi mestnih avtobusov se med potniki kot ena izmed najpomembnejših pomanjkljivosti šteje nepredvidljivost časa vožnje. Z namenom, da bi pripomogli k izboljšanju kakovosti javnega potniškega prometa, smo s pomočjo izbranih metod za napovedovanje časovnih vrst analizirali, kako dobro lahko napovedujemo število potnikov na postajališčih in čas vožnje avtobusa med dvema zaporednima postajališčema. Zanimala nas je tudi morebitna povezava med napovedno točnostjo ter nepredvidljivostjo števila potnikov in časa vožnje avtobusov. Razpolagali smo s podatki o prihodih avtobusov na postajališča in o številu potnikov na posameznih postajališčih po posameznih urah v mestu Ljubljana.

V naši raziskavi smo uporabili več različnih metod za napovedovanje časovnih vrst. Analizirali smo tudi vpliv vremenskih dejavnikov na točnost napovedi. Pri napovedovanju časa vožnje avtobusa, še posebej pa pri napovedovanju števila potnikov na postajališčih, smo pokazali, da lahko dobro poznavanje vremenskih razmer v bližnji prihodnosti, denimo v naslednjih urah, pomembno vpliva na izboljšanje točnosti napovedi. To smo pokazali s pomočjo primerjave metode ARIMAX, ki pri napovedih uporablja tudi po-

jasnjevalne spremenljivke, z ostalimi klasičnimi metodami za napovedovanje časovnih vrst. Za predstavnike ostalih metod smo izbrali: (1) AR – avtoregresijski model, (2) ARIMA – avtoregresijski integriran model drsečih sredin in (3) VAR – vektorski avtoregresijski model. Metoda ARIMAX, edina od naštetih metod pri napovedovanju, predvideva, da poznamo prihodnje vrednosti nekaterih spremenljivk, ki sicer niso predmet napovedi. V eksperimentih se je izkazalo, da v primeru, ko se za pojasnjevalne spremenljivke uporabljajo vremenski podatki, metoda ARIMAX vodi do občutno in hkrati tudi konsistentno boljših napovedi.

Po pričakovanjih se je izkazalo, da so napovedne točnosti boljše v poletnih mesecih in ob koncih tedna, ko je število potnikov manjše in ko so časi vožnje avtobusov praviloma krajši. Enako velja tudi za postajališča, ki niso v samem središču mesta. Pokazali smo, da število potnikov na postajališču nima pomembnega vpliva na izboljšanje točnosti napovedi časa vožnje avtobusa med dvema zaporednima postajališčema. Eksperimentalni rezultati tudi dajejo slutiti, da povečan promet sam po sebi ne pomeni nujno večje nepredvidljivosti trajanja vožnje.

Ključne besede

javni potniški promet, avtobusi, čas trajanja vožnje, število potnikov, zunanji dejavniki, vremenski podatki, znanost o podatkih, napovedovanje časovnih vrst, ARIMAX

Abstract

Title: Analysis of Public Transport with Time Series Forecasting

Public transport is a vital part of each major city, since it enables simple and environmentally-friendly transportation throughout the city. It is expected that the rising number of inhabitants will increase the demand for public transport. When using the city buses, the passengers especially see the unpredictable arrival times as one of the largest flaws. In order to improve the quality of public transport, time series forecasting methods have been chosen to analyse how accurately we can foretell the number of passengers at a bus stop and the bus travel times between two bus stops. We were also interested in finding a possible connection between the forecasted accuracy and unpredictability of the number of passengers and bus travel times. We had data of bus arrival times to the bus stops and number of passengers at bus stops at different hours in Ljubljana.

Different time series forecasting methods have been used in our research. We also analysed the influence of various weather conditions on the accuracy of our prediction. It has been indicated that when predicting the bus driving period or especially the number of passengers at bus stops, good knowledge of the weather conditions in the near future, e. g. in next hours, can significantly influence the improvement of prediction accurate. This was presented using ARIMAX method that uses explanatory variables together with other classical methods for time series forecasting. The following were chosen as representatives of other classical methods: (1) AR – Autoregressive model, (2) ARIMA – Autoregressive integrated moving average model, and (3) VAR

– Vector auto-regressive model. ARIMAX method is the only of the listed prediction methods that assumes that we know the future values of some of the variables that are, however, not the subject of our prediction. The experiments have shown that in cases when we use weather data for explanatory variables, the ARIMAX method leads to the better and more consistent predictions.

As expected, the predicted accuracies were better in summer months and at weekends, when the number of passengers is smaller and when bus driving periods are shorter. The same is true for the bus stops that are not located in the city center. We have proved that the number of passengers at a bus stop does not have a particular influence on improving the accuracy of predicting bus travel time between two bus stops. The experimental results indicate that the increased traffic does not necessarily mean more unpredictable driving periods.

Keywords

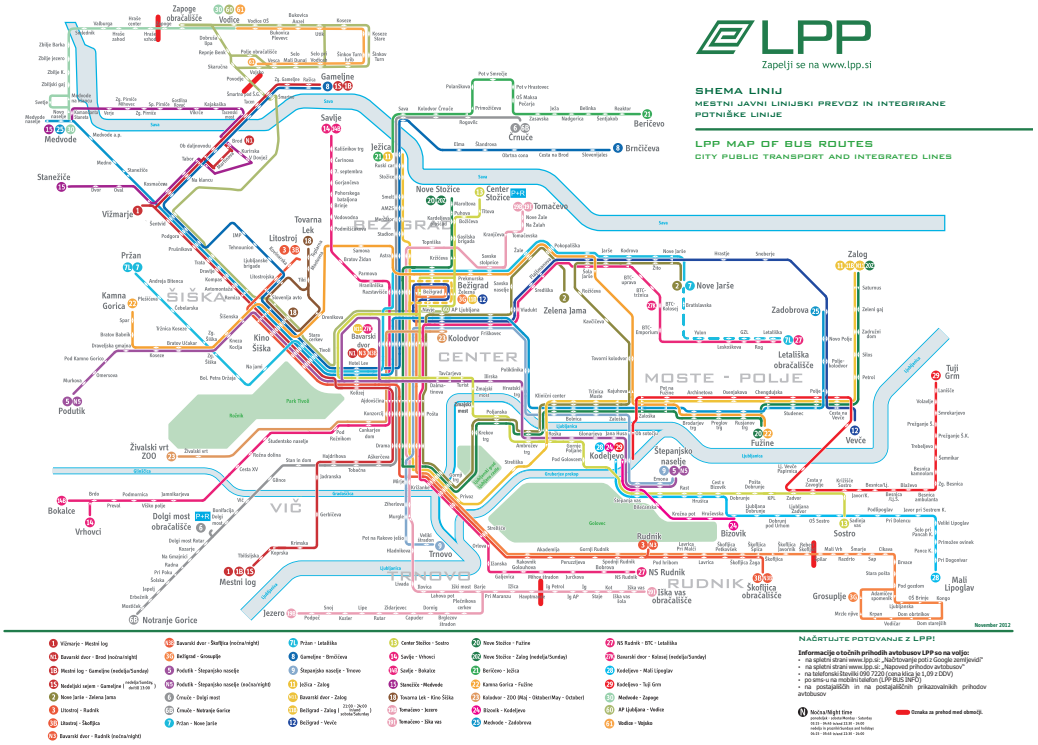
public transport, buses, travel time, number of passengers, external factors, weather data, data science, time series forecasting, ARIMAX

Poglavje 1

Uvod

S povečevanjem števila prebivalstva v mestu in okolici se z zasičenostjo mesta z avtomobili povečuje onesnaževanje okolja. Onesnaževanje okolja niso le toplogredni plini in izpusti trdih delcev, temveč tudi hrup, poraba neobnovljivih goriv ter prometna infrastruktura. Vsako leto se povečuje število dnevnih migracij prebivalcev iz okolice mesta, ki se vozijo na delo v mesto in po dnevnih opravkih, zaradi česar, še zlasti zjutraj in popoldne, nastajajo prometni zastoji v središču mesta in na mestnih vpadnicah.

Ena izmed rešitev je uporaba javnega potniškega prometa (v nadaljevanju JPP), saj bi se s tem zmanjšala prometna obremenjenost osebnih vozil v samem mestu. Pri uporabi JPP-ja je pomembno, da je potnik še zmeraj deležen udobja in preprostosti kot pri lastnem prevozu. V večjih mestih za prebivalce iz okolice mesta gradijo parkirišča na obrobju mesta, ki omogočajo parkiranje in nadaljno vožnjo z avtobusom, tako imenovani sistem parkiraj in pelji (P + R). Z uporabo sistema parkiraj in pelji ne bi bilo potrebe po iskanju parkirišča, ki jih v središču mest po navadi ni veliko oziroma so zasedena, ter strošku plačevanja parkirnine. JPP uporabljajo tudi drugi prebivalci mesta in okolice, kot so osnovnošolci, dijaki, študentje, delovno aktivni, starejše osebe, bolniki, invalidi ter drugi, ki nimajo možnosti uporabe drugih transportnih sredstev. Zanje je še posebej pomembno, da uporabljajo zanesljiv in ugoden transportni sistem, saj so od njega odvisni pri premagovanju vsakodnevnih



Slika 1.1: Shema avtobusnih linij v Ljubljani leta 2012 [1].

poti.

Ne glede na organizacijo in sodoben informacijski sistem je razumljivo, da prihaja do težav. JPP je prostorsko in časovno omejen na vnaprej določene linije in časovne presledke, ki se ne posodabljajo glede na povpraševanje potnikov. Prihod avtobusa na postajališče ni zmeraj točno ob času, ki je napisan na voznem redu posamezne linije. Potnik s tem pridobi slabo izkušnjo uporabe, saj je čas trajanja vožnje avtobusa med postajališči nepredvidljiv in s tem prav tako prihod avtobusa na postajališče. Ob določenih delih dneva prihaja do prezasedenosti avtobusov, saj so časovni intervali med avtobusi predolgi oziroma avtobus ni primerne velikosti, kar je posledica neprilagajanja potrebam potnikov.

Določene pomanjkljivosti za predvidljivejše potovanje so v Mestni občini Ljubljana (v nadaljevanju MOL) v sodelovanju s podjetjem Ljubljanski pot-

niški promet (v nadaljevanju LPP), rešili z uvedbo posebnih cestnih pasov, namenjenih za avtobuse na lokalnih vpadnicah v središče mesta.

V magistrskem delu smo se osredotočili na mesto Ljubljana. Podatke smo pridobili v podjetju LPP in jih uporabili za analizo javnega mestnega potniškega prometa. Vsa pridobljena avtobusna postajališča in avtobusne linije vidimo na Sliki 1.1. Za analizo smo izbrali množico postajališč. Na izbranih postajališčih smo napovedovali število potnikov na postajališču in čas vožnje avtobusa med dvema zaporednima postajališčema. Za napovedovanje smo uporabili metode za napovedovanje časovnih vrst.

1.1 Motivacija

V magistrskem delu smo analizirali, kako kakovostno je mogoče napovedovati število potnikov na postajališčih in čas vožnje avtobusa med dvema zaporednima postajališčema. V ta namen smo med seboj primerjali različne metode za napovedovanje časovnih vrst. Analizirali smo, kako dobro z uporabo metod za napovedovanje časovnih vrst lahko napovedujemo. Za napovedovanje smo uporabili podatke o JPP-ju v Ljubljani. Poleg podatkov o JPP-ju smo vključili še podatke o vremenu, pridobljene s spletne strani Agencije Republike Slovenije za okolje (v nadaljevanju ARSO), ter preverili njihov vpliv na napovedno točnost števila potnikov na postajališču in časa vožnje avtobusa med dvema zaporednima postajališčema.

Zanima nas, ali lahko z metodami in pridobljenimi podatki ugotovimo nepredvidljivost števila potnikov določenega avtobusnega postajališča in časa vožnje avtobusa med dvema zaporednima postajališčema na določenem postajališču oziroma avtobusni liniji v različnih mesecih, dnevih v tednu in urah v dnevu. Težava pri JPP-ju je nepredvidljivost časa vožnje avtobusa in s tem prihod avtobusa na postajališče. Urniki prihodov avtobusov na postajališče so vnaprej znani in čas vožnje pri praznem cestišču tudi poznamo, vendar v praksi prihaja do nepredvidljivih dogodkov, ki privedejo do zamud. Čas trajanja vožnje se spreminja glede na del dneva, ob različnih dnevih v tednu

in ob različnih mesecih. Zelo pogosto, še zlasti za potnike, ni glavna težava čas vožnje, ampak nepredvidljivost časa vožnje in s tem povezan prihod avtobusa na postajališče. Vendar je to nepredvidljivost težko ocenjevati oziroma določiti. Predpostavili smo, da slabša zmožnost napovedovanja odraža večjo nepredvidljivost. To smo poskušali eksperimentalno ugotoviti s pomočjo primerjave rezultatov različnih metod za napovedovanje časovnih vrst. Napovedovanje števila potnikov na postajališču smo vključili v napovedovanje časa vožnje avtobusa, saj smo predpostavili vpliv števila potnikov na čas vožnje avtobusa med dvema zaporednima postajališčema.

1.2 Pregled sorodnih del

Na področju napovedovanja časa vožnje vozil v javnem potniškem prometu in na avtocestnih odsekih je bilo napisanih več znanstvenih člankov, v katerih so uporabili različne metode za gradnjo napovednih modelov.

Osredotočili smo se na znanstvena dela, ki so napovedovala čas vožnje avtobusa ali drugih vozil na določenem predelu cestišča. Večinoma se uporabljata dve tehnologiji za pridobivanje podatkov: AVL (angl. *automatic vehicle location*) in APC (angl. *automatic passenger counters*). AVL je sistem, ki samodejno sporoča lokacijo vozila in čas, APC pa šteje potnike, ki vstopijo oziroma izstopijo iz vozila. Številna znanstvena dela so se za gradnjo natančnejših napovednih modelov osredotočala na uporabo podatkov, pridobljenih s tehnologijama AVL in APC.

Cong Bai in soavtorji [2] so primerjali različne metode za napovedovanje časa vožnje avtobusa. Kot težavo so vzeli 4 zaporedna avtobusna postajališča v mestu Shenzhen na Kitajskem. Del cestišča med avtobusnima postajališčema so poimenovali segment. Na segmentih vozi več različnih avtobusov, med katerimi lahko potniki izbirajo, vendar imajo različen čas prihoda na naslednje postajališče. Med seboj so primerjali modele napovedovanja, ki so jih pridobili z naslednjimi metodami napovedovanja: ANN (angl. *artificial neural network*), SVM (angl. *support vector machines*), Kal-

man filter in dinamična ANN-Kalman ter SVM-Kalman. Dinamični metodi ANN-Kalman in SVM-Kalman se od ANN in SVM razlikujeta v delu prilagajanja napovedi z metodo Kalman. Najprej so izračunali model ANN oziroma SVM. Dobljeni model so posodabljali z metodo Kalman in novimi podatki ter ga tako prilagajali trenutnim razmeram. Pri napovedovanju so si pomagali tudi s časi vožnje ostalih avtobusov na segmentu. Najbolj točne napovedi so dosegli z dinamičnima modeloma, in sicer na segmentu 1 in 3 s SVM-Kalmanovim modelom in segmentu 2 z ANN-Kalmanovim modelom. V skupnem seštevku je model SVM-Kalman za malenkost boljši kot model ANN-Kalman. Za izboljšanje napovedi so predlagali vključitev vremenskih podatkov in časov vožnje ostalih vozil.

Erik Jenelius in soavtorji [3] so s pomočjo statistične metode napovedovali čas vožnje avtobusa. Podatke so zbirali s pomočjo nizkofrekvenčne GPS-sonde (angl. *global positioning system*) na avtobusih in taksijih, kjer so se vozila po istem delu cestišča. Za oceno parametrov statističnega modela so kot pojasnjevalne spremenljivke uporabili povezane atribute, kot so omejitev hitrosti, število pasov na cesti, avtobusna postajališča, cestne znake in pogoje potovanja (vreme, čas v dnevu, delavnik oziroma konec tedna, letni čas). Tako so dodali vplive zunanjih dejavnikov na čas vožnje ter dobljene čase vožnje s tem poskušali obrazložiti. Model so ocenili z izračunom največje verjetnosti časa vožnje, statistično pomembnost rezultatov pa so ocenili s standardno napako. Model so testirali na podatkih z omrežja v Stockholmu na Švedskem in ugotovili, da imajo povezani atributi in pogoji potovanja velik vpliv na čas vožnje in so med seboj pozitivno korelirani. Še posebej velik vpliv so imeli vremenski pojavi, kot sta dež in sneg.

Avtorji člankov [4, 5, 6] so se namesto časov vožnje odločili za napovedovanje hitrosti vozil na izbranih odsekih cestišča, saj je od hitrosti odvisen čas vožnje. Z napovedano hitrostjo in znano dolžino cestišča so izračunali čas vožnje vozil. Napovedovanje hitrosti vozil je avtorjem omogočalo lažjo primerjavo različnih odsekov cestišča. Vsi trije članki so podrobneje predstavljeni v nadaljevanju.

V člankih [4] in [5] so avtorji napovedovali čas vožnje s pomočjo regresijskih metod. Za primer so uporabili avtocestni odsek v dolžini 2,5 km severnozahodne houstonske avtoceste v članku [4] in 50 km vzhodne avtoceste v Los Angelesu v članku [5], na katerem so pridobivali podatke o času in lokaciji vozil. Osredotočili so se na avtoceste in vsa vozila v cestnem prometu. Iz zbranih podatkov so odstranili proste dneve, konce tednov in določene dneve, kjer so manjkali podatki, tako jim je preostalo 32 dni podatkov. Med seboj so primerjali tri metode za napovedovanje: (1) lokalni linearni regresijski model, (2) k-NN (angl. *k-nearest neighbors*) in (3) linearni regresijski model s PCA (angl. *principal component regression*). Rezultati so pokazali, da najbolj točne napovedi dobijo z metodo lokalne linearne regresije. Linearna regresija s PCA je bila na drugem mestu in najslabše model k-NN.

Cilj avtorjev v članku [5] je bil narediti metodo, ki bo hitro procesirala, bo prilagodljiva za različne velikosti podatkov in bo delovala v realnem času. Primerjali so enake metode kot v članku [4]. Najboljše se je izkazala metoda linearne regresije, kateri se je najbolj približala metoda k-NN.

Čas vožnje na obvoznici mesta so napovedovali tudi v članku A Kind of Urban Road Travel Time Forecasting Model with Loop Detectors [6]. Podatke so zbirali 24 ur z 2-minutnim intervalom na odseku dolžine 600 m in tako dobili podatke o obsegu prometa, hitrosti vozil in zasedenosti cestišča. V izogib napakam so podatke pregledali, jih uredili ter odstranili napake in manjkajoče podatke. Osredotočili so se predvsem na iskanje točk spremembe v določenih delih dneva. Točke predstavljajo določeno uro v dnevu, pri kateri se očitno spremeni vrednost ena ali več opazovanih spremenljivk. S pomočjo teh točk so razdelili podatke na različne časovne intervale, ki imajo podobne lastnosti. Za napovedovanje so uporabili metodo ARIMA s parametri $p = 2$, $d = 0$ in $q = 0$, z različnimi uteževalnimi funkcijami (kvadrat korena, kvadrat, krivulja stopnje rasti in linearna funkcija). Med seboj so primerjali napovedovanje s celotnimi podatki in posameznimi intervali, ki so se izkazali za koristne, saj so bile napake manjše. Za funkcijo z najmanjšo napako se je izkazala linearna utežitvena funkcija.

1.3 Prispevki

V magistrskem delu smo primerjali različne metode za napovedovanje števila potnikov na postajališču in časov vožnje avtobusov med dvema zaporednima postajališčema. Pri raziskavi smo se omejili na 17 skrbno izbranih avtobusnih postajališč na območju mesta Ljubljane.

Analizirali smo vpliv zunanjih dejavnikov, konkretno vremenskih razmer, na uspešnost napovedovanja časovnih vrst. Poleg zunanjih dejavnikov smo analizirali še vpliv števila potnikov in števila avtobusov na postajališču ter njihov vpliv na napovedno točnost.

Po posameznih avtobusnih postajališčih smo ugotavljali nepredvidljivost oziroma nezmožnost uspešnega napovedovanja števila potnikov v različnih koledarskih mesecih, dnevih v tednu in urah v dnevu. Tudi za čas vožnje avtobusa med dvema zaporednima postajališčema smo ugotavljali nepredvidljivost oziroma nezmožnost uspešnega napovedovanja, poleg izbranih postajališč pa še po posameznih avtobusnih linijah na postajališčih v različnih koledarskih mesecih, dnevih v tednu in urah v dnevu.

1.4 Pregled magistrskega dela

Magistrsko delo sestavlja šest poglavij. Poglavje Priprava in analiza podatkov (2) je razdeljeno na tri podpoglavja. V podpoglavjih opišemo in analiziramo podatke o potnikih, avtobusih in vremenskih dejavnikih. V poglavju Metode za napovedovanje (3) smo opisali uporabljene metode za napovedovanje časovnih vrst ter njihove prednosti in slabosti. Ocenjevanje točnosti napovedi in postopek evalvacije, ki smo jo uporabili za evalvacijo metod za napovedovanje časovnih vrst, je opisan v poglavju Metode za evalvacijo (4).

Glavne prispevke magistrskega dela smo predstavili v poglavju Rezultati (5). Razdelili smo ga na podpoglavji o napovedovanju števila potnikov na postajališču in napovedovanju časa vožnje avtobusa med dvema zaporednima postajališčema. Pri obeh smo opisali rezultate napovedovanja in jih analizirali.

V poglavju Zaključki (6) smo zapisali končne ugotovitve pri napovedovanju števila potnikov na postajališču in časa potovanja avtobusa med dvema zaporednima postajališčema ter možnosti za izboljšave. Delu smo dodali prilogo Izbrana postajališča in linije (A), v kateri so zapisana vsa analizirana postajališča in pripadajoče linije.

Poglavje 2

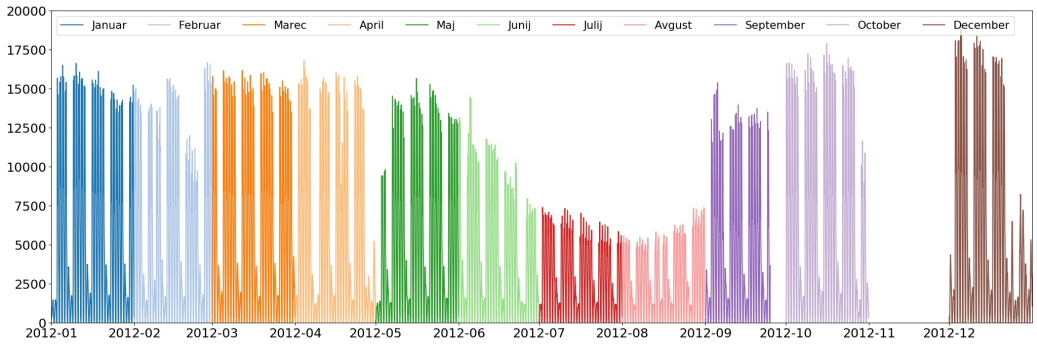
Priprava in analiza podatkov

Podatke, ki smo jih v magistrskem delu uporabili za analizo in napovedovanje časovnih vrst, smo pridobili s strani podjetja LPP. Podjetje LPP zbira več različnih podatkov o storitvah, ki jih ponuja. Med storitvami je tudi javni mestni potniški promet. Pridobili smo podatke o potnikih javnega mestnega potniškega prometa in podatke o zamudah avtobusov za leto 2012. Iz prejetih podatkov smo ugotovili, da je bilo leta 2012 v mestu Ljubljana 471 postajališč in 43 različnih linij.

2.1 Podatki o potnikih

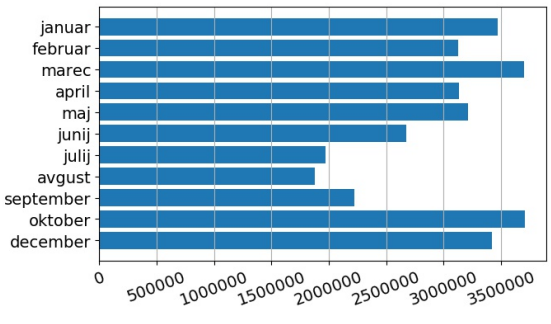
Podatke je podjetje LPP zbiralo z elektronskim sistemom enotnih kartic Urbana, ki jih potniki uporabljajo za plačevanje voženj z avtobusom in ostalih storitev. Podatki so za leto 2012 zbrani mesečno in so zapisani v formatu CSV (angl. *comma-separated values*). Pri pregledu podatkov smo ugotovili, da manjkajo zapisi za zadnjih 5 dni meseca septembra in celoten mesec november. Pridobljeni podatki predstavljajo 82 % vseh zabeleženih podatkov o potnikih[1]. Slika 2.1 prikazuje število zabeleženih potnikov na vseh linijah v urnih intervalih. Število skupnih zapisov po mesecih, ki smo jih uporabili v magistrskem delu, je prikazano na Sliki 2.2.

[1] V letnem poročilu 2012 [7] piše, da so prepeljali 39.437.496 potnikov.



Slika 2.1: Število potnikov na uro.

	Število potnikov
januar	3.473.618
februar	3.128.137
marec	3.699.523
april	3.132.984
maj	3.211.393
junij	2.678.369
julij	1.970.772
avgust	1.881.327
september	2.224.597
oktober	3.708.400
december	3.422.185



Slika 2.2: Število potnikov po mesecih.

Tabela 2.1: Neobdelani podatki o potnikih.

Datum	2012-01-01 00:00:22	2012-01-01 00:01:41	2012-01-01 00:01:42
Številka kartice Urbana	13793105545483123	12929983407514739	13793105799044979
Vozilo	Vozilo 361	Vozilo 351	Vozilo 351
Številka vozila	176	62	62
Smer linije	B LETALIŠKA-BTC-BLEIWEISOVA	N GAMELJNE-BAVARSKI DVOR	N GAMELJNE-BAVARSKI DVOR
Številka linije	914	1021	1021
Ime postajališča	KODROVA	PODGORA	PODGORA
Številka postajališča	2003	1922	1922
Tip kartice Urbana	MESEČNA	MESEČNA	MESEČNA
Številka tipa kartice Urbana	12	12	12
Tip vozovnice	ŠOLSKA A	ŠTUDENT A	ŠTUDENT A
Številka tipa vozovnice	14	35	35

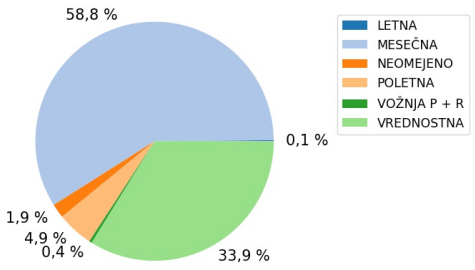
Neobdelane podatke o potnikih, primer lahko vidimo v Tabeli 2.1, smo uredili tako, da smo izločili zapise, ki niso vsebovali vseh podatkov ali podatki niso bili veljavni. Tako smo izločili 1.689.789 zapisov, kar predstavlja 5 %

Tabela 2.2: Obdelani podatki o potnikih.

Datum	Število potnikov	Število avtobusov	Ime postajališča	Številka postajališča	Smer linije
2012-01-01 00:00:00	10	5	BAVARSKI DVOR (V)	600011	ČRNUČE-DOLGI MOST
2012-01-01 01:00:00	7	2	BAVARSKI DVOR (V)	600011	ČRNUČE-DOLGI MOST
2012-01-01 05:00:00	6	1	BAVARSKI DVOR (V)	600011	ČRNUČE-DOLGI MOST
2012-01-01 06:00:00	5	3	BAVARSKI DVOR (V)	600011	ČRNUČE-DOLGI MOST
2012-01-01 07:00:00	7	4	BAVARSKI DVOR (V)	600011	ČRNUČE-DOLGI MOST

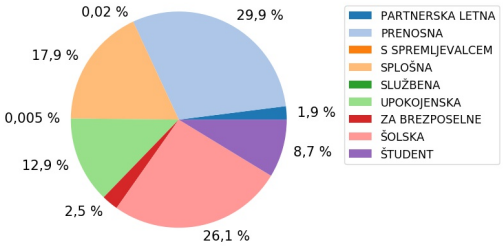
vseh prejetih podatkov. Podatke smo zaradi lažje obdelave omejili na stolpce: številka kartice Urbana, številka vozila, smer linije in ime postajališča. Izbrane podatke smo združili po dnevu na enourne, glede na postajališče in linijo avtobusa. Iz podatkov o potnikih smo izračunali, koliko različnih avtobusov je v določeni uri prišlo na posamezno postajališče. Primer obdelanih podatkov, ki smo jih uporabili za napovedovanje števila potnikov na postajališču Bavarski dvor, vidimo v Tabeli 2.2.

Tip Urbane	Število potnikov
LETNA	45.106
MESEČNA	19.002.875
NEOMEJENO	623.151
POLETNA	1.584.048
VOŽNJA P + R	121.540
VREDNOSTNA	10.938.917



Slika 2.3: Število potnikov za različne tipe Urbane.

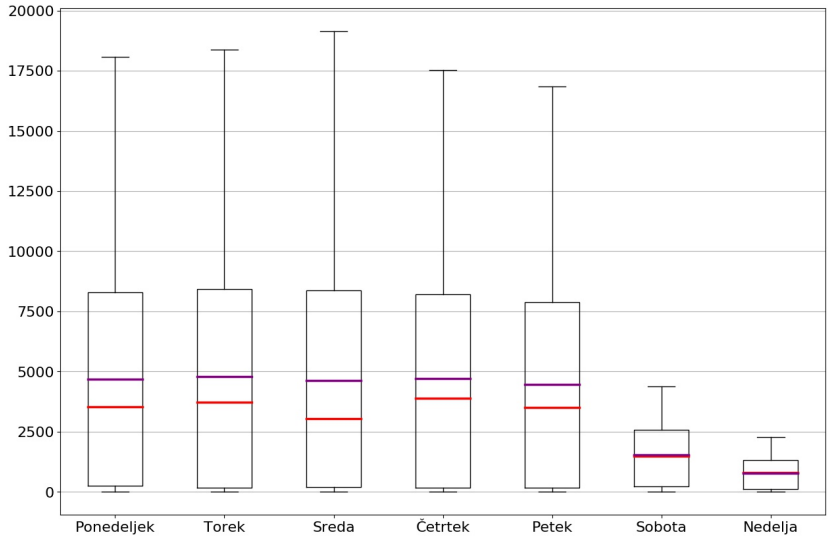
Tip vozovnice	Število potnikov
PARTNERSKA LETNA	648.606
PRENOSNA	9.742.902
S SPREMLJEVALCEM	6.485
SPLOŠNA	5.815.010
SLUŽBENA	1.525
UPOKOJENSKA	4.200.879
ZA BREZPOSELNE	801.115
ŠOLSKA	8.488.589
ŠTUDENT	2.826.194



Slika 2.4: Število potnikov za različne tipe vozovnic.

Podjetje LPP ponuja več različnih vrst elektronskih kartic Urbana, s katerimi lahko plačujemo vožnje z avtobusom. Slika 2.3 prikazuje različne tipe

kartice Urbana ter kolikokrat je bila uporabljena v letu 2012. Največkrat, 19 milijonkrat, je bila uporabljena mesečna kartica Urbana, sledi ji vrednostna, ki je bila uporabljena malo manj kot 11 milijonkrat. Najmanjkrat je bila uporabljena letna kartica Urbana, in sicer malo več kot 45 tisočkrat. Na elektronsko kartico Urbana si lahko potnik, glede na njegov družbeni/socialni status, naloži oziroma kupi različne tipe vozovnic. Tipe vozovnic, in kolikokrat je bil posamezni tip uporabljen, prikazuje Slika 2.4. Potniki so najpogosteje uporabljali prenosno vozovnico, sledila ji je šolska vozovnica. S skoraj 6 milijoni uporabe je na tretjem mestu vozovnica tipa Splošna. Najmanj potnikov je uporabljalo vozovnico tipa Partnerska letna.

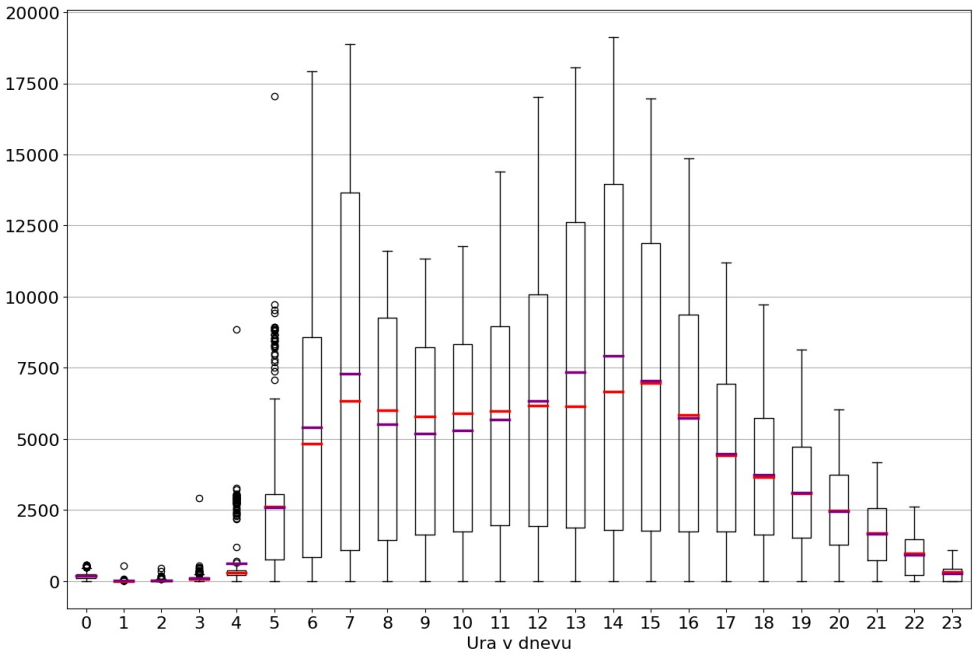


Slika 2.5: Število potnikov na uro glede na dan v tednu.

Iz podatkov smo razbrali, da so pri podjetju LPP v povprečju zabeležili 4.164 potnikov storitve prevoza z avtobusi na uro. Največje število potnikov v eni uri so zabeležili 5. decembra med 14. in 15. uro, in sicer 19.134 potnikov. Podatke o številu potnikov na uro smo združili po dnevih v tednu ter tako dobili, koliko potnikov se je na določen dan v tednu peljalo z avtobusom. Razpon števila potnikov na uro glede na dan v tednu je predstavljen na Sliki 2.5. S slike lahko vidimo, da se je največ potnikov vozilo ob sredah,

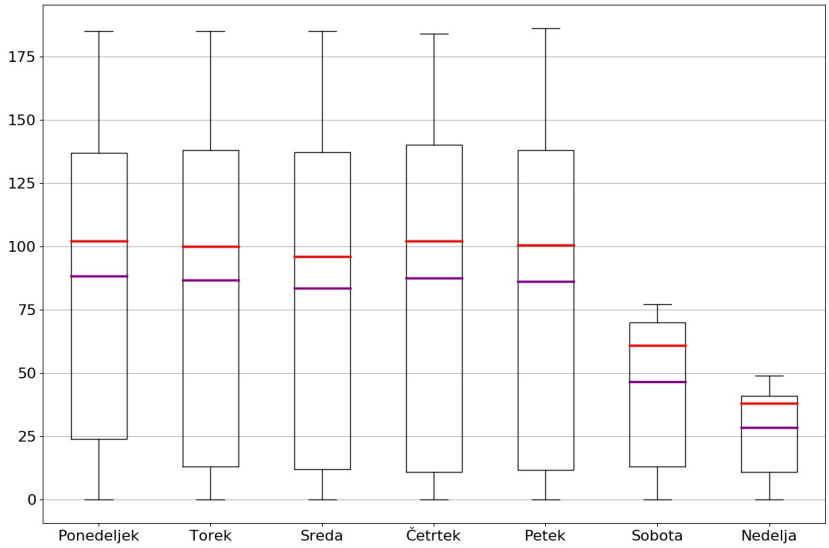
najmanj pa ob nedeljah. V povprečju je bilo največje število potnikov ob torkih, in sicer 4.785 potnikov, v mediani pa ob četrtkih, 3.886 potnikov.

Na Sliki 2.5 in vseh nadaljnjih slikah, ki prikazujejo dan v tednu oziroma uro v dnevu, je vsak vertikalni interval sestavljen iz dveh vodoravnih črtic, ki predstavljata minimalno in maksimalno vrednost, ter osrednjega pravokotnika. Spodnja stranica pravokotnika predstavlja prvi kvartil, zgornja pa tretji kvartil. Znotraj vsakega pravokotnika je označena povprečna vrednost z vijolično črto in mediana z rdečo črto. Krogi nad maksimalno vrednostjo in pod minimalno vrednostjo predstavljajo osamelce, ki se od prvega oziroma tretjega kvartila razlikujejo vsaj za faktor 1,5.



Slika 2.6: Število potnikov na uro glede na uro v dnevu.

Slika 2.6 prikazuje razpon števila potnikov na uro v letu 2012 v določeni uri v dnevu. V dnevu vidimo, da sta dve skrajnosti, med 7. in 8. uro 18.882 potnikov ter med 14. in 15. uro, ko je bilo 19.134 potnikov. To so ure v dnevu, ko se največ ljudi odpravi v službo in šolo ali iz nje. V povprečju je bilo največ potnikov med 14. in 15. uro, in sicer 7.922 potnikov. V mediani

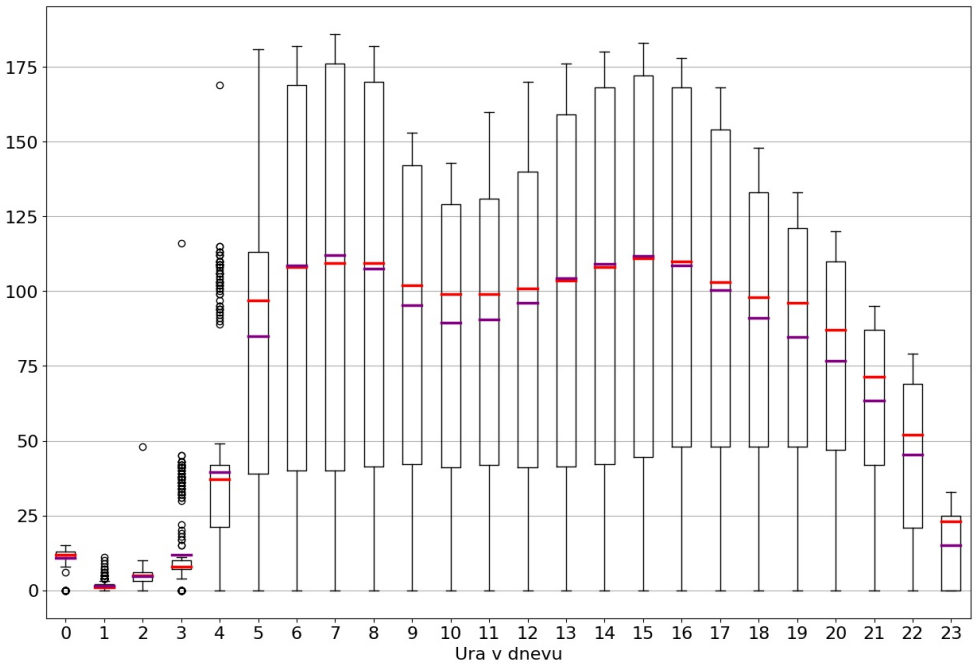


Slika 2.7: Število avtobusov na uro glede na dan v tednu.

pa je bilo največ prepeljanih 6.963 potnikov med 15. in 16. uro. Opazimo, da je med polnočjo in 5. uro zjutraj zelo malo potnikov, obstajajo le določena odstopanja.

Povprečno število avtobusov na uro, ki se ustavijo na različnih postajališčih je 81. Razpon števila avtobusov na uro po dnevih v tednu in urah v dnevu je prikazan na Sliki 2.7 in Sliki 2.8. Od ponedeljka do petka vidimo, da vsak dan vozi približno enako število avtobusov in imajo enak razpon, manjše število avtobusov pa vozi ob sobotah in nedeljah, ko tudi manj ljudi uporablja avtobusne storitve. Število avtobusov glede na uro v dnevu je predvsem povezano s številom potnikov v določeni uri, kar se opazi, če primerjamo Sliko 2.6 in Sliko 2.8. V nočnih urah med 23. in 5. uro zjutraj vidimo, da je število avtobusov bistveno manjše kot v preostalih delih dneva. Enako kot pri potnikih ima število avtobusov dve skrajnosti, in sicer med 7. in 8. uro ter med 15. in 16. uro, ko smo zabeležili malo več kot 180 avtobusov na različnih postajališčih.

Na osnovi števila potnikov, ki so vstopili na avtobus na določenem postajališču, smo izbrali postajališča. Izbrana postajališča smo uporabili za



Slika 2.8: Število avtobusov na uro glede na uro v dnevu.

Tabela 2.3: Izbrana postajališča.

Število potnikov	Ime postajališča	Številka postajališča
1693000	POŠTA	601011
1658349	BAVARSKI DVOR (V)	600011
1564239	BAVARSKI DVOR (Z)	600012
1216077	KONZORCIJ	601012
695347	AMZS	103031
613169	SLOVENIJA AVTO	803011
598169	DRAMA (Z)	602022
591425	AJDOVŠČINA	600022
584998	DRAMA (V)	602021
524836	HOTEL LEV	700012
459834	KINO ŠIŠKA	802021
403517	RAZSTAVIŠČE	100021
330079	KLINIČNI CENTER	402031
6010	GALJEVICA	603094
5989	KOSEZE	803142
4125	POD ROŽNIKOM	702012
4073	OB SOTOČJU	303032

Slika 2.9: Shema linij in izbrana postajališča.

analizo in napovedovanje števila potnikov na postajališču in časa vožnje avtobusa med dvema zaporednima postajališčema. Tabela 2.3 prikazuje imena izbranih postajališč, številko postajališča in število vseh potnikov na postajališču v letu 2012. Izbrali smo prvih 13 postajališč z največ potniki ter 4 postajališča približno na sredini seznama, da bi dosegli reprezentativen vzorec. Izbrana postajališča smo kategorizirali glede na oddaljenost postajališča od Prešernovega trga, ki se nahaja v strogem centru Ljubljane:

1. skupina (razdalja manjša od 1 km): Pošta, Konzorcij, Ajdovščina, Drama (V), Drama (Z), Bavarski dvor (V), Bavarski dvor (Z), Hotel Lev,

2. skupina (razdalja med 1 in 2 km): Klinični center, Razstavišče, Pod Rožnikom,

3. skupina (razdalja med 2 in 3 km): Galjevica, Kino Šiška, Ob sotočju,

4. skupina (razdalja več kot 3 km): AMZS, Slovenija avto, Koseze.

Na Sliki 2.9 so prikazane glavne linije in z rdečo označena postajališča, ki smo jih izbrali za napovedovanje števila potnikov na postajališču in časa vožnje avtobusov med dvema zaporednima postajališčema.

2.2 Podatki o avtobusih

Podjetje LPP podatke o avtobusih zbira s pomočjo sistema TELARGO[2], ki so ga uvedli leta 2005. Sistem omogoča nadzorovanje, sledenje in komunikacijo z vozilom in voznikom. V določenih intervalih sistem v nadzorni center sporoča podatke o položaju avtobusa. V nadzornem centru spremljajo položaj avtobusa, razliko med dvema avtobusoma na enaki liniji, čas vožnje avtobusa med postajališčema in čas avtobusa na postajališču.

Tabela 2.4: Neobdelani podatki o avtobusih.

	0	1	2
Registrska številka vozila	LJ LPP-061	LJ LPP-061	LJ LPP-061
Čas prihoda na postajališče	2012-01-03 04:11:54	2012-01-03 04:12:18	2012-01-03 04:13:12
Čas odhoda s postajališča	2012-01-03 04:12:04	2012-01-03 04:12:27	2012-01-03 04:13:27
Zamuda prihoda na postajališče	13,0	−78,0	−70,0
Številka linije	21,0	21,0	21,0
Opis linije	JEŽICA-BERIČEVO	JEŽICA-BERIČEVO	JEŽICA-BERIČEVO
Smer linije	BERIČEVO	BERIČEVO	BERIČEVO
Številka postajališča	104052	104102	104113
Ime postajališča	Sava	Kolodvor Črnuče	Primožičeva
Vožnja	332	332	332
Neznano			

Prejeti podatki so za vsak mesec leta 2012 zapisani v formatu CSV. Skupaj je 16.471.509 zapisov neobdelanih podatkov, primer lahko vidimo v Tabeli 2.4. Iz podatkov smo zaradi zagotavljanja zasebnosti izločili podatke o vozniku avtobusa. Izločili smo tudi vse zapise, ki niso vsebovali vseh podatkov ali so bili ti neveljavni. Izločenih je bilo 1.666.515 zapisov, kar predstavlja

[2]http://bus.lpp.si/Default.aspx?culture=sl-SI

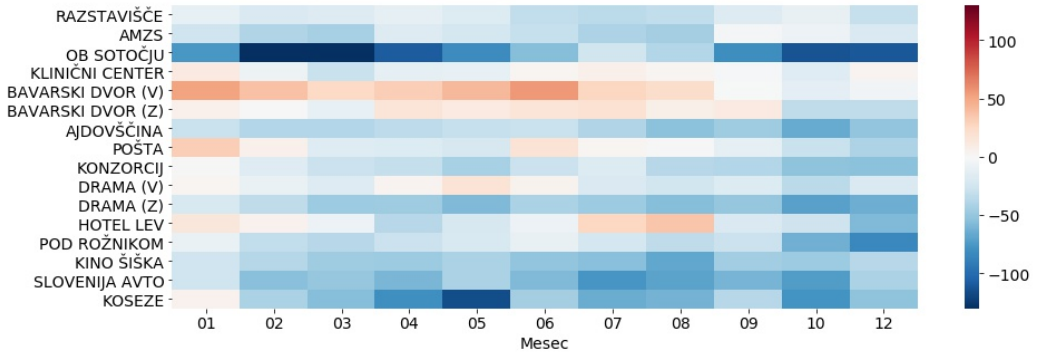
10 % vseh zapisov. Podatki in imena zadnjih petih atributov so bila neurejena in vsebovala nerelavantne podatke, zato smo jih uredili. Pri imenih postajališč, opisih linij in smereh linij smo odstranili presledke pred besedilom in jih zapisali z velikimi črkami. Za izračun časa vožnje smo datum in čas prihoda oziroma odhoda avtobusa s postajališča razdelili na ločena stolpca za datum in čas, kar nam je omogočalo povezovanje avtobusov med dvema zaporednima postajališčema. Iz podatkov o zamudah smo pridobili podatke o vseh linijah skupaj z imeni in številkami postajališč, da smo jih povezali s podatki o potnikih.

Tabela 2.5: Obdelani podatki o avtobusih.

Čas prihoda na postajališče	Čas vožnje	Ime linije	Številka postajališča	Ime postajališča
2012-01-18 03:00:00	31,0	ČRNUČE-DOLGI MOST	600011	BAVARSKI DVOR (V)
2012-01-18 04:00:00	32,0	ČRNUČE-DOLGI MOST	600011	BAVARSKI DVOR (V)
2012-01-18 05:00:00	34,0	ČRNUČE-DOLGI MOST	600011	BAVARSKI DVOR (V)

Čas vožnje avtobusa od enega do drugega postajališča, ki smo ga potrebovali za napovedovanje časa vožnje avtobusa med dvema zaporednima postajališčema, nam ni bil dan v podatkih, zato smo ga morali izračunati. Osredotočili smo se na izbrana postajališča, ki so vidna v Tabeli 2.3, ter za ta postajališča izračunali čase vožnje avtobusov med prejšnjim in izbranim postajališčem. Čas vožnje smo izračunali tako, da smo iz podatkov izluščili izbrano postajališče in postajališče pred njim ter iskali najmanjšo razliko v času za enake linije. To smo ponovili za vseh 17 postajališč in vsako linijo na postajališču. Izračunan čas vožnje smo zaradi napak omejili na 2.000 sekund. Enako kot podatke o potnikih smo tudi podatke o času vožnje združili po urah. Tako smo dobili podatke o času vožnje vseh linij avtobusov za 17 izbranih postajališč. Primer podatkov za avtobus na liniji Črnuče-Dolgi most za postajališče Bavarski dvor (V) lahko vidimo v Tabeli 2.5.

Vsako postajališče ima različno število linij, nekatere linije se ponavljajo pri različnih postajališčih. Seznam izbranih postajališč in linij smo navedli v Dodatek A. Nismo napovedovali za vse linije na postajališčih, saj nismo pri vseh linijah uspeli najti ustreznega avtobusa na prejšnjem postajališču, da

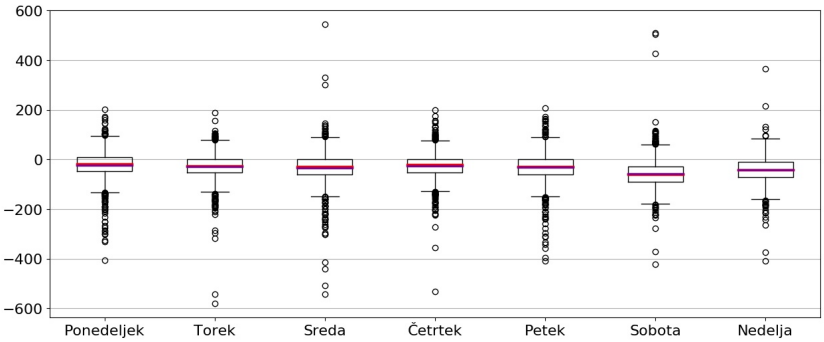


Slika 2.10: Mediana zamud za izbrana postajališča.

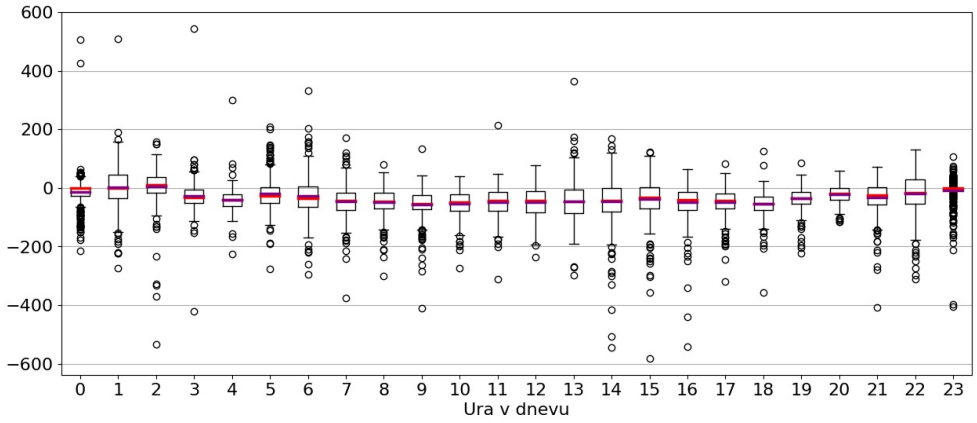
bi izračunali čas vožnje. Podatki o avtobusih za določen mesec so manjkali, nekatere linije pa so vozile le določen del leta 2012. Težave smo imeli tudi z usklajevanjem podatkov o potnikih in času vožnje avtobusa, saj za določene ure nismo imeli podatkov o številu potnikov ali podatkov o času vožnje avtobusa, kar nam je zmanšalo število linij pri določenih postajališčih. Tako lahko opazimo, da za postajališče 603094 – Galjevica nimamo nobene linije, saj nismo imeli dovolj informacij. Iz začetnih 2.188.719 zapisov smo izločili 825.964 neustreznih oziroma neujemajočih zapisov, kar predstavlja 38 % zapisov za izbranih 17 postajališč.

Mediana zamude avtobusov na vseh postajališčih in linijah v letu 2012 je bila -37 sekund, kar pomeni, da so avtobusi v povprečju prišli 37 sekund prekmalu na postajališče. Kot je vidno na Sliki 2.10, nekatere vrednosti izbranih postajališč izstopajo. Avtobus je na postajališče Ob sotočju meseca februarja in marca v mediani pripeljal 140 in 192 sekund prezgodaj. Tudi na postajališču Koseze je avtobus v mesecu maju pripeljal v mediani 115 sekund prezgodaj. Preostalim postajališčem se v celotnem letu vrednosti zamud zmanjšujejo. Na začetku leta je bila mediana zamud izbranih postajališč skupaj enaka -3 sekunde, konec leta pa je znašala -40 sekund.

Mediana zamude je na postajališču Bavarski dvor največja med izbranimi postajališči in znaša 31 sekund, kar pomeni, da je avtobus prišel prepozno na postajališče. Najbolj točne prihode avtobusov je imelo postajališče Pošta



Slika 2.11: Zamude avtobusov na uro glede na dan v tednu.



Slika 2.12: Zamude avtobusov na uro glede na uro v dnevu.

z mediano pol sekunde in povprečno zamudo 9 sekund.

Na Sliki 2.11 in Sliki 2.12 je prikazan razpon zamud glede na dan v tednu oziroma uro v dnevu za izbrana postajališča v letu 2012. Pri Sliki 2.11, kjer so predstavljene zamude glede na dan v tednu, vidimo, da je v celotnem tednu približno enaka zamuda. V soboto in nedeljo je malo odstopanja, saj je avtobus v povprečju pripeljal malo manj kot minuto prezgodaj na postajališče.

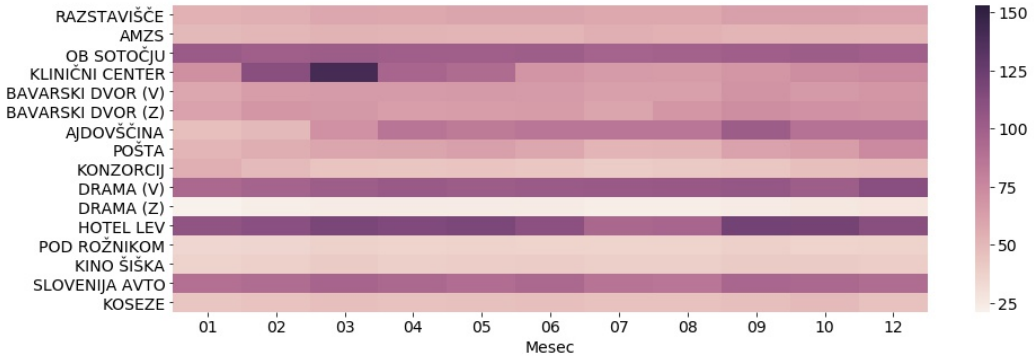
Razpon zamud glede na uro v dnevu za leto 2012 je prikazan na Sliki 2.12. Tekom dneva se pojavljajo določena nihanja v času zamude prihoda avtobusa na postajališče. Najbolj se to opazi v urah med 5. in 8. uro ter med 13. in

16. uro, ko je veliko prometa, saj se večina ljudi vozi v službo oziroma šolo in nazaj. To je povezano tudi s številom potnikov in številom avtobusov, saj je ob teh urah tudi največje število potnikov in avtobusov. Obstaja tudi veliko podatkov o zamudah, ki izstopajo od povprečja za več kot 400 sekund. Te vrednosti predvidevamo so ob izrednih dogodkih ali pa so napake v sistemu.

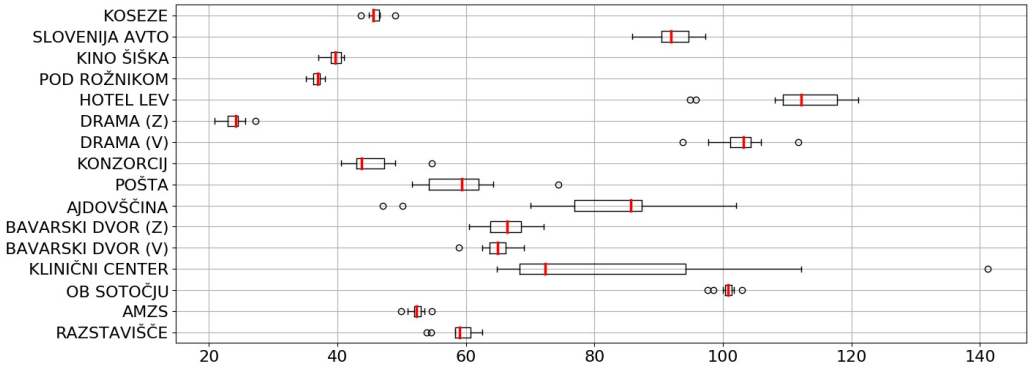
Tabela 2.6: Povprečen čas vožnje avtobusa med dvema zaporednima postajališčema.

Postajališče	Povprečen čas vožnje
RAZSTAVIŠČE	59
AMZS	52
OB SOTOČJU	101
KLINIČNI CENTER	87
BAVARSKI DVOR (V)	65
BAVARSKI DVOR (Z)	67
AJDOVŠČINA	78
POŠTA	60
KONZORCIJ	46
DRAMA (V)	103
DRAMA (Z)	25
HOTEL LEV	113
POD ROŽNIKOM	37
KINO ŠIŠKA	40
SLOVENIJA AVTO	92
KOSEZE	46

Za potrebe naše raziskave smo izračunali čase vožnje avtobusov med dvema zaporednima postajališčema. Povprečne vrednosti časov vožnje avtobusov med postajališči za celo leto 2012 smo prikazali v Tabeli 2.6, po posameznih mesecih pa na Sliki 2.13. Večina postajališč ima približno konstanten povprečen čas vožnje avtobusa med dvema zaporednima postajališčema skozi celotno leto 2012. Med izbranimi postajališči izstopa postajališče Klinični center, ki ima meseca februarja, marca, aprila in maja občutno daljše čase vožnje kot preostali del leta. Februarja je čas vožnje znašal 122 sekund, marca 153 sekund, aprila in maja pa 101 sekundo, skupno povprečje postajališča pa je bilo 87 sekund. Brez štirih mesecev, navedenih v prejšnjem stavku, bi znašal povprečen čas vožnje 70 sekund. Ugotovili smo, da je razlog



Slika 2.13: Povprečni čas vožnje avtobusa med dvema zaporednima postajališčema za izbrana postajališča.



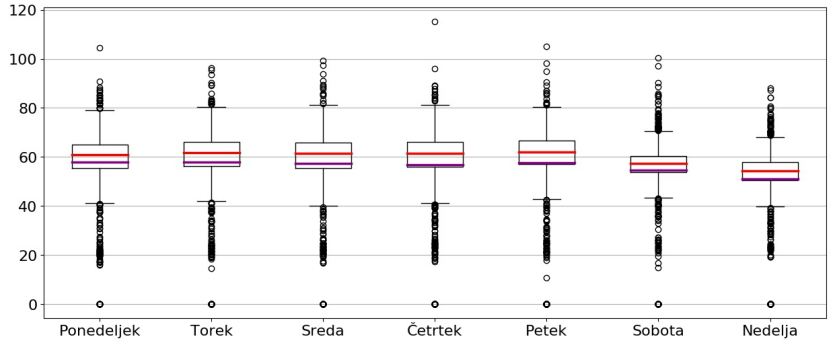
Slika 2.14: Razpon časov vožnje avtobusov med dvema zaporednima postajališčema za izbrana postajališča v letu 2012.

za povečanje časa vožnje sprememba prometne ureditve v mesecu marcu in aprilu v križišču Zaloška-Njegošova-Trubarjeva cesta[3]. Na postajališču Ajdovščina se je čas vožnje avtobusa povečal po mesecu februarju s 50 sekund na 83 sekund meseca aprila in imel približno tako vrednost do konca leta 2012.

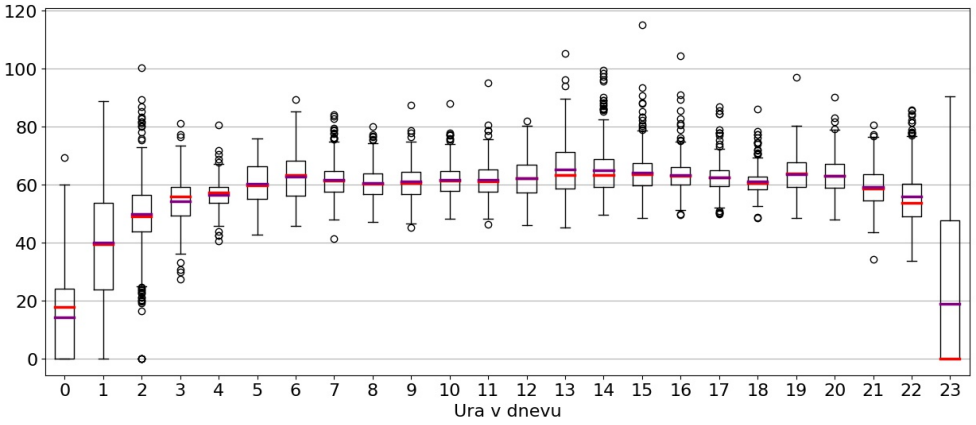
Najmanjši povprečen čas vožnje imajo avtobusi na postajališču Drama (Z), in sicer 25 sekund, saj se prejšnje postajališče nahaja na razdalji 330 m.

[3]https://goo.gl/QQXebk

Največji povprečni čas vožnje pa ima postajališče Hotel Lev, 113 sekund.



Slika 2.15: Čas vožnje avtobusov na uro glede na dan v tednu.



Slika 2.16: Čas vožnje avtobusov na uro glede na uro v dnevu.

Čas vožnje avtobusov na izbranih postajališčih se skozi leto, teden in dan spreminja. Avtobusi na izbranih postajališčih potrebujejo med 20 in 100 sekund, da pridejo do postajališča. Po posameznih dnevih v tednu ni velikega nihanja v času vožnje med dnevi, le ob sobotah in nedeljah je povprečen čas za približno 5 sekund krajši. Čase vožnje in njihov razpon glede na dan v tednu vidimo na Sliki 2.15.

Glede na uro v dnevu so časi voženj bolj raznoliki in se spreminjajo. Na Sliki 2.16 vidimo, da so časi vožnje v jutranjih urah med 6. in 8. uro ter

popoldanski med 13. in 15. uro daljši kot v preostalem delu dneva, saj je takrat več avtomobilov na cesti in potnikov na avtobusih. V nočnih urah se časi vožnje skrajšajo, saj je manj prometa in avtobusov.

2.3 Podatki o vremenskih dejavnikih

Pri napovedovanju smo kot zunanje dejavnike uporabili podatke o vremenu. Pridobili smo jih na spletni strani Agencije Republike Slovenije za okolje[4]. Izbrali smo informacije o:

- povprečni temperaturi zraka na višini 2 m [°C],
- povprečnem zračnem tlaku [hPa],
- povprečni relativni vlagi [%],
- povprečni hitrosti vetra [m/s],
- povprečna smeri vetra [°],
- količini padavin [mm],

iz arhiva samodejne postaje, postavljene v Ljubljani za Bežigradom. Primer neobdelanih podatkov, pridobljenih s spletnega arhiva, vidimo v Tabeli 2.7. Samodejna postaja je zbirala podatke vsake pol ure, mi pa smo jih agregirali na enourne. Vzorec obdelanih podatkov, ki smo jih uporabili kot zunanje dejavnike za napovedovanje števila potnikov na postajališču in časa vožnje avtobusa med dvema zaporednima postajališčema, je predstavljen v Tabeli 2.8.

Za lažje razumevanje smo zunanje dejavnike predstavili z grafi. Na Sliki 2.17 vidimo povprečno temperaturo za vsako uro. Povprečna letna temperatura je bila 11,8 °C, najvišja 36,7 °C je bila izmerjena 22. avgusta med 14. in 15. uro, najnižja -12,05 °C pa 9. februarja med 6. in 7. uro zjutraj. Slika 2.18 prikazuje spreminjanje relativne vlažnosti v letu 2012. Povprečna relativna

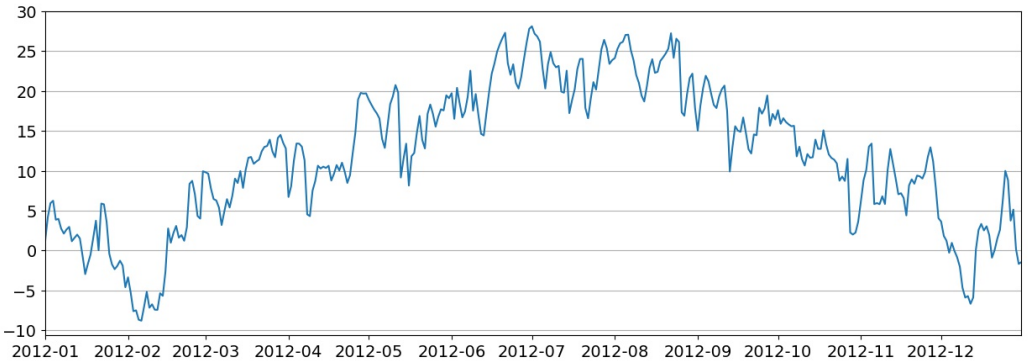
[4]http://meteo.arso.gov.si/met/sl/archive/

Tabela 2.7: Neobdelani podatki o vremenu.

	2012-01-01 00:00:00	2012-01-01 00:30:00	2012-01-01 01:00:00
station id	_1828	_1828	_1828
station name	LJUBLJANA-BEŽIGRAD	LJUBLJANA-BEŽIGRAD	LJUBLJANA-BEŽIGRAD
povp. tlak [hPa]	983,0	984,0	984,0
T [°C]	−0,4	−0,6	−0,5
rel. vla. [%]	92,0	92,0	93,0
količina padavin [mm]	0,0	0,0	0,0
hitrost vetra [m/s]	1,1	0,9	0,7
smer vetra [°]	30,0	348,0	13,0

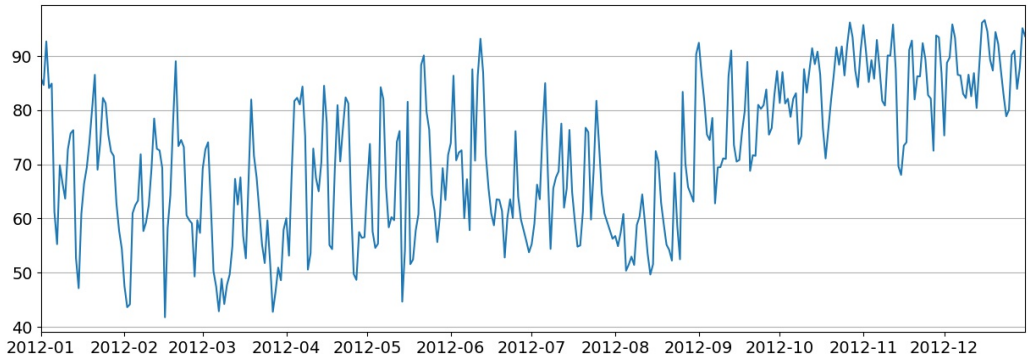
Tabela 2.8: Obdelani podatki o vremenu.

	Tlak	Temperatura	Vlažnost	Hitrost vetra	Smer vetra	Količina padavin
2012-01-01 00:00:00	983,5	−0,5	92,0	1,0	189,0	0,0
2012-01-01 01:00:00	984,0	−0,5	92,5	0,75	24,5	0,0
2012-01-01 02:00:00	984,0	−0,6	92,0	0,55	175,5	0,0

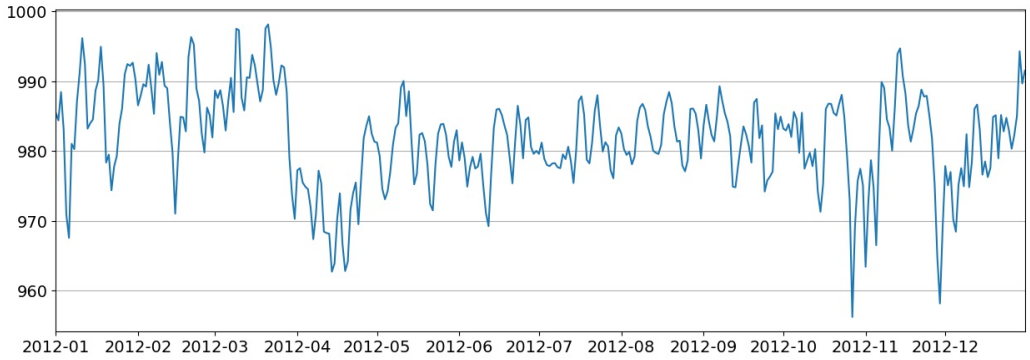


Slika 2.17: Povprečna temperatura zraka [°C].

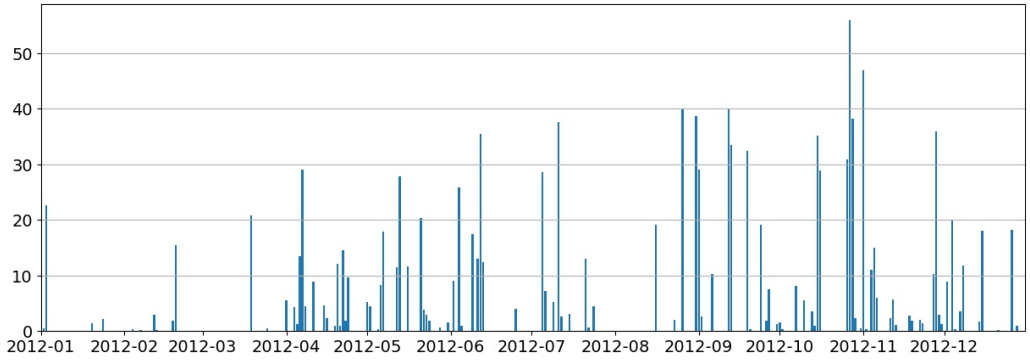
vlažnost je znašala 71,4 %. Opazimo lahko, da je bila relativna vlažnost od meseca septembra višja od povprečne relativne vlažnosti v preostalem delu leta. Zračni tlak je prikazan na Sliki 2.19, kjer opazimo, da je bil zračni tlak od začetka leta 2012 do meseca aprila višji, kot v preostanku leta. V povprečju je zračni tlak znašal 981,8 hPa. Skupna količina padavin v letu 2012 je znašala 1.243,9 mm. Povprečno je v Ljubljani padlo 3,39 mm padavin na dan. Največ padavin je padlo 27. oktobra, in sicer 56 mm. Slika 2.21



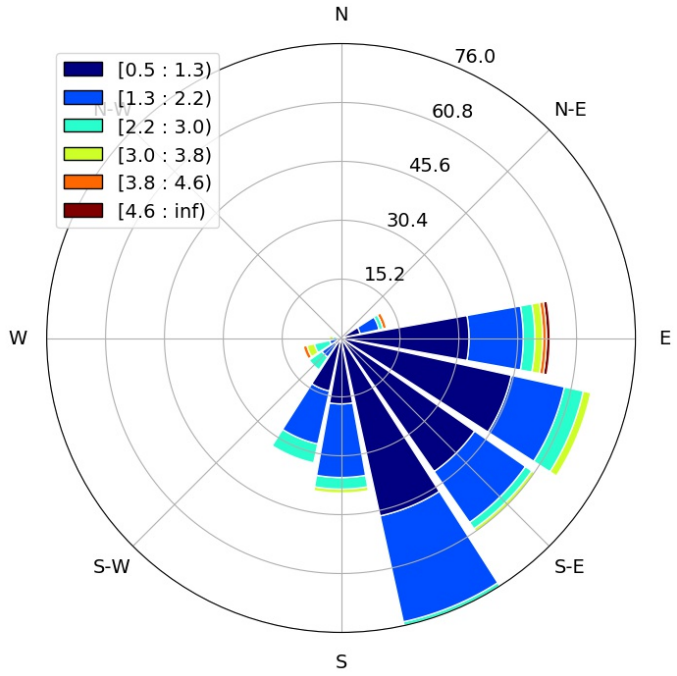
Slika 2.18: Povprečna relativna vlaga [%].



Slika 2.19: Povprečen zračni tlak [hPa].



Slika 2.20: Količina padavin [mm].



Slika 2.21: Povprečna smer in hitrost vetra [m/s].

predstavlja pogostost, smer in moč vetra, ki je pihal v Ljubljani. Moč vetra je predstavljena z barvo, in sicer temno modra predstavlja nizke hitrosti, rjava pa visoke hitrosti vetra, izražene v metrih na sekundo. Smer stolpcev s središča proti robu kroga kaže na to, iz katere smeri je pihal veter, in kot lahko opazimo, prevladuje veter jugozahodne smeri. Tretji podatek, ki ga predstavlja Slika 2.21, je, kolikokrat se je pojavil veter določene smeri in jakosti, kar prikazuje oddaljenost stolpca od središča kroga. Vse navedene vrednosti smo izračunali oziroma jih pridobili iz podatkov, ki smo jih imeli na voljo.

Poglavje 3

Metode za napovedovanje

Za različne raziskovalne probleme obstajajo različne metode napovedovanja. Metode delimo na kvalitativne in kvantitativne. Kvalitativne metode so subjektivne in temeljijo na mnenju. Kvantitativne metode za napovedovanje uporabljajo podatke, pridobljene v preteklem času, in na različne načine poskušajo napovedati prihodnost. Nekatere med njimi so preproste oziroma naivne metode, metode časovnih vrst, vzorčne metode ter metode umetne inteligence. V našem delu smo se osredotočili na metode za napovedovanje časovnih vrst. Te metode so relativno preproste, a pogosto bolj uspešne pri napovedovanju v primerjavi z metodami strojnega učenja, v primerjavi na primer z nevronskimi mrežami, ki zahtevajo veliko časa in računanja [8]. Med seboj smo primerjali naslednje metode za napovedovanje časovnih vrst:

- **RM** –referenčni model,
- **AR** – avtoregresijski model,
- **ARIMA** – avtoregresijski integriran model drsečih sredin,
- **ARIMAX** – avtoregresijski integriran model drsečih sredin s pojasnjevalnimi spremenljivkami,
- **VAR** – vektorski avtoregresijski model.

3.1 Referenčni model

Za referenčni model napovedovanja, označevali ga bomo s kratico "RM", v magistrskem delu smo izbrali naivni model. Naivni model deluje na osnovi preprostega pravila, vrednost y ob času $t-1$ je napoved za naslednji časovni korak t, kar lahko pokažemo s formulo

$$y_t = y_{(t-1)}.$$

Naivni model nam je služil kot referenčna točka za vse ostale modele. Model je preprost za implementacijo in potrebuje malo procesiranja, da dobimo napovedi. Ne potrebuje predhodnega učenja oziroma inteligence, kar poenostavi razumevanje modela. Pomembna je tudi ponovljivost, pri enakih vhodnih podatkih dobimo vedno enak rezultat napovedi, kar nam zagotovi deterministič nost modela.

Metoda deluje dobro za podatke, ki se veliko ne spreminjajo oziroma počasi naraščajo ali padajo. Včasih se uporablja pri ekonomskih in finančnih časovnih vrstah, ter za napovedovanje vremena [9, 10].

3.2 Avtoregresijski model

Avtoregresijski model – AR [11] se uporablja na različnih področjih v gospodarstvu za napovedovanje porabe električne energije, ekonomiji za napovedovanje cen in naravi za napovedovanje vremena. Uspešna je pri napovedovanju, kjer so vrednosti med seboj korelirane v času. Model je v osnovi linearna regresija podatkov z enim ali več preteklimi podatki enakega primera. AR je primer stohastičnega modela, ki vsebuje določeno stopnjo negotovosti in naključja. Naključnost je mišljena, da bomo lahko napovedali prihodnja gibanja s preteklimi podatki, ne bomo pa dobili stoodstotne natančnosti.

Z AR-jem napovedujemo spremenljivko, ki nas zanima, z linearno kombinacijo predhodnih vrednosti izbrane spremenljivke. Pojem avtoregresije pomeni regresijo spremenljivke same s sabo. Model uporablja določeno število

predhodnih vrednosti y_t kot predikcije vrednosti. Število predhodnih vrednosti se označuje s p in se imenuje odlog, model pa označimo z $AR(p)$. $AR(p)$ lahko zapišemo kot

$$y_t = c + \phi_1 y_{(t-1)} + \phi_2 y_{(t-2)} + ... + \phi_p y_{(t-p)} + \epsilon_t, \tag{3.1}$$

kjer je y_t napoved vrednosti ob času t, c je konstanta, $\phi_1, \phi_2, ..., \phi_p$ so posamezni koeficienti, $y_{(t-1)}, y_{(t-2)}, ..., y_{(t-p)}$ so pretekle vrednosti sprmenljivke y_t, ϵ_t je slučajni odklon ali napaka ($N(0, \sigma^2)$) in p število odlogov [12]. Konstanta c je definirana kot

$$c = (1 - \sum_{i=1}^{p} \phi_i)\mu,$$

kjer μ predstavlja povprečje preteklih vrednosti spremenljivke y_t, za katerega velja $Ey_t \neq 0$. Z uporabo odlogov, ki jih označimo z B, lahko izrazimo vrednost odloga kot

$$By_t = y_{(t-1)}, B^2 y_t = y_{(t-2)}, ..., B^p y_t = y_{(t-p)}. \tag{3.2}$$

Enačbo (3.1) lahko zapišemo v naslednji obliki

$$y_t - \phi_{(t-1)} y_{(t-1)} - \phi_{(t-2)} y_{(t-2)} - ... - \phi_{(t-p)} y_{(t-p)} = c + \epsilon_t,$$

in z apliciranjem (3.2) dobimo

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) y_t = c + \epsilon_t,$$

oziroma v krajši obliki

$$\phi(B) y_t = c + \epsilon_t, \tag{3.3}$$

kjer $\phi(B)$ predstavlja avtoregresijski operator odloka p

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p.$$

Tako lahko $AR(p)$ z enačbo (3.1) vidimo kot rešitev enačbe (3.3), ki je

$$y_t = \frac{1}{\phi(B)} \epsilon_t.$$

$\phi(B)$ je znan tudi kot karakteristični polinom procesa in koreni členov določajo ali je proces stacionaren ali ne [11]. Najpreprostejši model je $AR(1)$, v katerem gledamo za 1 časovni korak nazaj in zapišemo

$$y_t = c + \phi_1 y_{(t-1)} + \epsilon_t.$$

Lastnosti, ki veljajo za regresijski model $AR(1)$, so naslednje:

- če je $\phi_1 = 0$, je y_t enak slučajnemu odklonu oziroma napaki,
- če je $\phi_1 = 1$ in $c = 0$, je y_t enak modelu naključnega sprehoda (angl. *random walk*),
- če je $\phi_1 = 1$ in $c \neq 0$, je y_t enak modelu naključnega sprehoda z zamikom,
- če je $\phi_1 < 0$, bo y_t nihalo med pozitivnimi in negativnimi vrednostmi.

Avtoregresijski model torej deluje tako, da povezuje časovno vrsto z njeno preteklostjo. V magistrskem delu smo uporabili knjižnico StatsModels[1] [13], s privzetimi parametri. S pomočjo modela AR, ki smo mu podali podatke, smo zgradili model za napovedovanje časovnih vrst. Parameter p in koeficienti preteklih vrednosti so bili izbrani samodejno z modelom AR.

3.3 Avtoregresijski integriran model drsečih sredin

Avtoregresijski integriran model drsečih sredin – ARIMA [11, 14, 15] je kombinacija treh modelov: avtoregresije (AR), diferenciacije (I) in drsečih sredin (MA). ARIMA je posplošen oziroma izpeljan model iz modela ARMA, pri katerem mora biti časovna vrsta stacionarna. Model ARMA je prvi definiral in uporabil Herman Wold [16] leta 1938. Pomembno pri ARMA je stacionarnost časovne vrste, kar pomeni, da morata biti srednja vrednost in

[1] http://www.statsmodels.org/dev/generated/statsmodels.tsa.ar_model.AR.html

varianca časovne vrste konstantni, ker tako dobimo napovedi z majhno napako. Časovne vrste torej ne smejo vsebovati trendov ali sezonskega gibanja, saj to privede v nestacionarnost in neuporabnost modela. Vendar je to na določenih področjih, kot je promet, gospodarstvo, ekonomija težko, saj pogosto podatki vsebujejo trend ali sezonsko gibanje. Za časovne vrste s trendom oziroma sezonskim gibanjem sta George Box in Gwilyn Jenkisns [17] leta 1968 raziskala in ugotovila, da s pomočjo diferenciacije podatkov nestacionarno časovno vrsto spremenimo v stacionarno. Modelu ARMA sta dodala diferenciacijo podatkov (I) in tako dobila model ARIMA.

Model AR smo opisali v podpoglavju 3.2. Ideja pri modelu drsečih sredin (MA) je podobno kot pri modelu AR. Pri modelu MA novo vrednost napovemo z linearno kombinacijo napak, ki smo jih izračunali iz preteklih napovedi. Model uporabi določeno število q preteklih napak napovedi y_t in z njihovo pomočjo napove novo vrednost. Številu q pravimo tudi odklon, model pa označimo z MA(q). Formula modela ima obliko:

$$y_t = c + \epsilon_t + \theta_1\epsilon_{(t-1)} + \theta_2\epsilon_{(t-2)} + ... + \theta_q\epsilon_{(t-q)}, \tag{3.4}$$

kjer je y_t napovedana vrednost ob času t, c je konstanta, $\theta_1, \theta_2, ..., \theta_q$ so posamezni koeficienti in $\epsilon_t, \epsilon_{(t-1)}, \epsilon_{(t-2)}, ..., \epsilon_{(t-q)}$ so napake preteklih napovedi y_t [18]. Operator odloga, ki ga označimo z B, lahko zapišemo kot

$$B\epsilon_t = \epsilon_{(t-1)}, B^2\epsilon_t = \epsilon_{(t-2)}, ..., B^q\epsilon_t = \epsilon_{(t-q)}.$$

S preoblikovanjem enačbe (3.4) in uporabo operatorja odloga B dobimo novo enačbo

$$y_t = c + \epsilon_t + \theta_1 B\epsilon_t + \theta_2 B^2\epsilon_t + ... + \theta_q B^q\epsilon_t,$$

ki jo lahko zapišemo v obliki

$$y_t = c + (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q)\epsilon_t,$$

oziroma krajše

$$y_t = c + \theta(B)\epsilon_t,$$

kjer je $\theta(B) = (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q)$.

Glavna razlika med ARMA in ARIMA je diferenciacija podatkov, ki odstranjuje trende in sezonska gibanja. To je preprost postopek, ki trenutno opazovano vrednost odšteje od predhodne vrednosti. Postopek imenujemo tudi odvod vrednosti in ga zapišemo kot

$$y'_t = y_t - y_{t-1}. \tag{3.5}$$

Enačba (3.5) predstavlja diferenciacijo prvega reda in jo zapišemo kot $I(1)$. V primeru, da vrsta ne postane stacionarna po prvem redu, lahko nadaljujemo in izvedemo diferenciacijo drugega reda $I(2)$

$$\begin{aligned} y''_t &= y'_t - y'_{t-1} \\ &= y_t - 2y_{t-1} + y_{t-2}. \end{aligned}$$

Diferenciacije ponavljamo, dokler podatki ne postanejo stacionarni, kar se v večini primerov zgodi po prvi oziroma drugi stopnji. Diferenciacije podatkov lahko zapišemo tudi s pomočjo operatorja odklona B. Diferenciacijo prvega reda (3.5) tako zapišemo kot

$$y'_t = y_t - By_t = (1 - B)y_t,$$

kjer $(1 - B)$ predstavlja prvo diferenciacijo. Na splošno lahko d-to diferenciacijo zapišemo kot

$$I(d) = (1 - B)^d y_t.$$

Zdaj imam vse tri dele modela ARIMA opisane in jih lahko združimo skupaj v ARIMA(p,d,q), kjer je:

- p število odlogov modela AR,
- d število, kolikokrat smo izvedli diferenciacijo, da smo prišli do stacionarnih podatkov,
- q število odlogov modela MA.

Celoten model ARIMA lahko zapišemo kot

$$y'_t = c + \phi_1 y'_{(t-1)} + \phi_2 y'_{(t-2)} + ... + \phi_p y'_{(t-p)} + \theta_1 \epsilon_{(t-1)} + \theta_2 \epsilon_{(t-2)} + ... + \theta_q \epsilon_{(t-q)} + \epsilon_t, \tag{3.6}$$

kjer je y'_t diferenciacija podatkov, ki jo lahko izvedemo več kot enkrat. Pri zduževanju modelov se začnejo zapisi enačb zapletati. Lažje je če zapis enačbe (3.6) zapišemo z operatorjem odklona in dobimo naslednjo enačbo

$$(1 - \phi_1 B - ... - \phi_p B^p)(1 - B)^d y_t = c + (1 + \theta_1 B + ... + \theta_q B^q) e_t,$$

oziroma krajše zapisano

$$\phi(B) \nabla^d y_t = c + \theta(B) e_t,$$

kjer je $\nabla^d = (1 - B)^d$. Obstajajo določene kombinacije parametrov p, d in q, ki privedejo do posebnih primerov:

- $ARIMA(0, 0, 0)$ – beli šum,
- $ARIMA(0, 1, 0)$ – naključni sprehod,
- $ARIMA(p, 0, 0)$ – avtoregresijski model,
- $ARIMA(0, 0, q)$ – model drsečih sredin.

Model ARIMA smo uporabili iz knjižnice StatsModels[2] [13]. Funkciji ARIMA smo določili parametre p, d in q, ostalih parametrov nismo nastavljali.

Optimalne parametre p, d in q smo v našem delu poiskali tako, da smo za vsakega določili območje, ki ga določen parameter lahko izbere, in se sprehodili skozi vse kombinacije vrednosti parametrov. Te vrednosti so bile za napovedovanje števila potnikov na postajališču in časa vožnje avtobusa med dvema zaporednima postajališčema naslednje:

- p vrednosti med 0 in 10,

[2] http://www.statsmodels.org/dev/generated/statsmodels.tsa.arima_model.ARIMA.html

- d vrednost 0 in 1,
- q vrednost 0 in 1.

Med seboj smo primerjali rezultate napovedi s kriterijsko funkcijo RMSE, ter na osnovi najmanjše vrednosti RMSE izbrali optimalno kombinacijo parametrov p, d in q.

3.4 Avtoregresijski integriran model drsečih sredin s pojasnjevalnimi spremenljivkami

Avtoregresijski integriran model drsečih sredin s pojasnjevalnimi spremenljivkami – ARIMAX [11, 14, 19] je izpeljan iz modela ARIMA, opisanega v podpoglavju 3.3. ARIMA omogoča napovedovanje le na osnovi preteklih vrednosti napovedane spremenljivke, ARIMAX pa predpostavlja, da imajo na napoved vpliv tudi eksogene neodvisne spremenljivke, ki jih tako vključimo v napovedovanje. Črka X predstavlja eksogene podatke v imenu ARIMAX. Model lahko imenujemo tudi model prenosne funkcije [20] ali dinamični regresijski model, kot ga je poimenoval Pankratz [21].

Model ARIMAX pri napovedovanju uporablja pretekle vrednosti časovne vrste, ki jo napovedujemo, ter vrednosti eksogenih časovnih vrst, če med njima obstaja povezanost. Povezovanje eksogenih in napovedanih časovnih vrst lahko vodi do zmanjšanja napake napovedi. Enačbo modela $ARIMAX(p, d, q)$ zapišemo kot

$$\nabla^d y_t = c + \sum_{i=1}^{p} \phi_i \nabla^d y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{m=1}^{M} \beta_m x_{m,t} + \epsilon_t, \tag{3.7}$$

kjer je y_t napovedana vrednost ob času t, β_m je koeficient eksogenega podatka m in $x_{m,t}$ vrednost eksogenega podatka m ob času t. Z uporabo operatorja odloga enačba (3.7) dobi obliko

$$\nabla^d \phi(B) \nabla^d y_t = c + \beta x_t + \theta(B) \epsilon_t,$$

oziroma

$$\nabla^d y_t = c + \frac{\beta}{\nabla^d \phi(B)} x_t + \frac{\theta(B)}{\nabla^d \phi(B)} \epsilon_t,$$

kjer je $\phi(B) = 1 + \phi_1 B + ... + \phi_p B^p$ in $\theta(B) = 1 + \theta_1 B + ... + \theta_q B^q$.

Enako kot pri modelu ARIMA smo tudi za model ARIMAX iskali optimalne parametre p, d in q. Pri iskanju optimalnih parametrov smo se omejili na določen razpon vrednosti, ki je za napovedovanje števila potnikov na postajališču naslednji:

- p vrednosti med 0 in 25,
- d vrednost 0 in 1,
- q vrednost 0 in 1.

Eksogene spremenljivke, ki smo jih uporabili za napovedovanje, so bile tlak, temperatura, hitrost vetra in vlažnost. Za napovedovanje časa vožnje avtobusa med dvema zaporednima postajališčema smo izbrali drugačen razpon vrednosti spremenljivk p, d in q:

- p vrednosti med 0 in 10,
- d vrednost 0 in 1,
- q vrednost 0 in 1.

Prej navedenim eksogenim spremenljivkam pri napovedovanju števila potnikov na postajališču, smo dodali spremenljivki število potnikov na predhodnem postajališču in število avtobusov na postajališču, za katerega smo napovedovali čas vožnje avtobusov. Za eksogene spremenljivke smo predpostavili, da jih poznamo oziroma jih je preprosto točno napovedati za 1 uro vnaprej. Najboljše parametre smo v obeh primerih izbrali na osnovi napake, izračunane s kriterijsko funkcijo RMSE. Kombinacijo parametrov p, d in q z najmanjšo napako smo izbrali za napovedovanje.

Metodo ARIMAX smo uporabili iz knjižnice PyFlux[3] [22]. Parametri, razen p, d in q, so ostali privzeti in jih nismo spreminjali.

[3] http://pyflux.readthedocs.io/en/latest/arimax.html

3.5 Vektorski avtoregresijski model

Vektorski avtoregresijski model – VAR [11, 23, 24] je razširitev univariatnega modela AR, opisanega v podpoglavju 3.2, na multivariatnega z endogenimi spremenljivkami. Model VAR je prvi opisal Christopher Sims [25] leta 1980. Od modela ARIMAX se razlikuje predvsem v tem, da ne uporablja prihodnjih vrednosti dodatnih spremenljivk ter nima diferenciacije podatkov in modela drsečih sredin.

Vsaka spremenljivka v modelu ima svojo enačbo, s katero jo opisujejo njene lastne pretekle vrednosti, pretekle vrednosti ostalih spremenljivk v modelu in napaka. Enačbo modela VAR(p) z odklonom p in matriko časovne vrste Y, velikosti $T \times K$, kjer je T število opazovanj in K število endogenih spremenljivk, zapišemo kot

$$Y_t = C + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + ... + \Phi_p Y_{t-p} + E_t, \tag{3.8}$$

kjer je Y_t vektor napovedanih vrednosti ob času t dimenzije $K \times 1$, C je vektor konstant za posamezno enačbo dimenzije $K \times 1$, $\Phi_1, \Phi_2, ..., \Phi_p$ so posamezne matrike koeficientov dimenzije $K \times K$, $Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ so vektorji vrednosti endogenih spremenljivk dimenzije $K \times 1$ in E_t je vektor napak dimenzije $K \times 1$. Vektor E_t je enakomerno porazdeljena in neodvisna spremenljivka z ničelnim povprečjem in kovariančno matriko Σ ter jo zapišemo kot $E_t \sim N(0, \Sigma)$. Enačbo (3.8) zapišemo v matrični obliki s p-vrednostjo 1 in dvema endogenima spremenljivkama

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,1} \\ \phi_{2,1} & \phi_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}.$$

Z uporabo operatorja odloga B enačbo (3.8) zapišemo kot

$$Y_t = C + \Phi_1 B Y_t + \Phi_2 B^2 Y_t + ... + \Phi_p B^p Y_t + E_t,$$

oziroma krajše

$$\Phi(B) Y_t = C + E_t,$$

kjer je

$$\Phi(B) = I_k - \Phi_1 B - \Phi_2 B^2 - ... - \Phi_p B^p.$$

Model VAR je preprost za ocenjevanje in ima dobre sposobnosti za napovedovanje. Za razliko od modela ARIMAX ni potrebno posebej določati, katere spremenljivke so endogene in katere eksogene, vendar vse podamo skupaj. Spremenljivke, ki smo jih uporabili kot eksogene, so bile enake kot pri modelu ARIMAX. Algoritem modela VAR mora oceniti veliko število koeficientov, in sicer K enačb za vsako K spremenljivk in p odklonov za vsako spremenljivko v vsaki enačbi, torej je treba oceniti $K + pK^2$ koeficientov. Ta proces je dolgotrajen in zahteven.

Določanje parametra p smo izvedli z metodo izbiranja odloga, ki nam jo je omogočala knjižnica StatsModels[4] [13]. Med kriteriji Akaike, Hannan-Quinnovo (HQIC) in Bayesovo (BIC) smo se odločili za indormacijski kriterij Akaike [26] (AIC). AIC primerja med seboj modele z različnimi parametri ter jih oceni na osnovi enačbe

$$AIC = 2k - 2ln(\hat{L}), \tag{3.9}$$

kjer je k število opazovanih parametrov in $\hat{L}$ največja verjetnost vrednosti modela. Kriterij AIC izbere boljše parametre pri vzorcu z manj primeri kot HQIC in BIC. AIC izbere parametre, ki minimizirajo varianco napake pri napovednem modelu. AIC ima slabost prevelikega prilagajanja pri povečevanju števila parametrov v modelu. V delu smo omejili iskanje parametra p s kriterijsko funkcijo AIC na vrednosti med 0 in 15.

[4] http://www.statsmodels.org/dev/generated/statsmodels.tsa.vector_ar.var_model.VAR.html

Poglavje 4

Metode za evalvacijo

4.1 Ocenjevanje napovedne točnosti

Za ocenjevanje točnosti napovednih modelov smo uporabili dve kriterijski funkciji, povprečno absolutno napako (angl. *mean absolute error*, MAE) in koren povprečne kvadratne napake (angl. *root mean squared error*, RMSE). MAE in RMSE sta dve izmed najpogosteje uporabljenih mer, ki se uporabljajo za merjenje točnosti. Omenimo naj, da nas pri napovedovanju v praksi zanimajo predvsem absolutne napake pri napovedih in ne relativne napake naprimer glede na dejansko število potnikov na postajališču oziroma čas potovanja avtobusa med dvema zaporednima postajališčema. S tega vidika izbor kriterijskih funkcij MAE in RMSE smatramo kot najbolj ustrezen.

MAE meri povprečno velikost napak napovedanih vrednosti v primerjavi z dejanskimi vrednostmi. Enačbo MAE zapišemo kot

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j|,$$

kjer je n število vrednosti, y_j j-ta dejanska vrednost in $\hat{y}_j$ j-ta napovedana vrednost.

RMSE je kvadratno pravilo točkovanja, ki prav tako meri povprečno velikost napake. To je kvadratni koren povprečja kvadratnih razlik med napovedano in dejansko vrednostjo opazovane spremenljivke. Enačbo RMSE

zapišemo kot

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2},$$

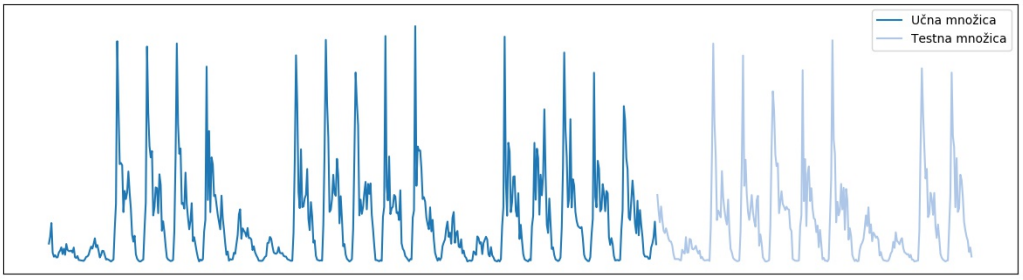
kjer je n število vrednosti, y_j j-ta dejanska vrednost in $\hat{y}_j$ j-ta napovedana vrednost.

Obe meri, RMSE in MAE, izražata povprečno napako napovedi v enakih enotah, kot smo jim podali vrednosti. V našem primeru sta enoti število potnikov in čas vožnje avtobusa v sekundah. Njuno območje rezultatov je med 0 in ∞ ter sta negativno orientirani, kar pomeni, da so nižje vrednosti boljše. Pri RMSE napake kvadriramo, preden jih povprečimo, zato imajo velike napake pri RMSE relativno večjo težo. RMSE je zato bolj uporaben v primerih, ko velike napake še posebej niso zaželjene. RMSE ima v primerjavi z MAE večje povečanje vrednosti, če se poveča vzorec, kar je lahko težavno pri primerjavi rezultatov RMSE, izračunanih na vzorcih različne velikosti.

4.2 Postopek evalvacije

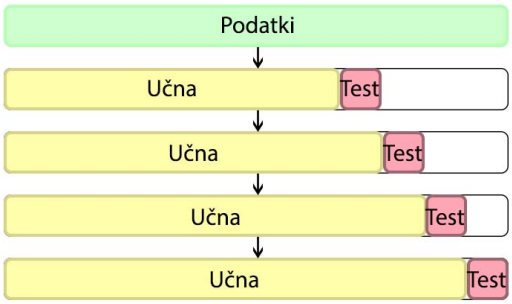
Metode za napovedovanje časovnih vrst, opisane v poglavju 3, smo evalvirali na enak način. Evalvirali smo po posameznih mesecih, posebej za vsako postajališče oziroma linijo. Podatke za posamezen mesec smo razdelili na učno in testno množico. Učna množica je vsebovala začetnih 66 % mesečnih podatkov, testna pa zadnjih 34 % podatkov. Primer razdelitve podatkov na učno in testno množico vidimo na Sliki 4.1.

Vsak model smo najprej naučili na učni množici in napovedali vrednost v naslednji uri. Napovedano vrednost smo shranili za kasnejšo primerjavo z dejanskimi vrednostmi ter računanjem točnosti modela. Po napovedi vrednosti v naslednji uri smo v vsaki iteraciji učni množici dodali dejansko vrednost napovedane vrednosti. Tako smo v vsaki iteraciji povečali učno množico za eno vrednost, testna množica pa je ostala enako velika. Model smo v vsaki iteraciji na novo naučili z razširjeno učno množico ter napovedali vrednost v naslednji uri. Tak postopek preverjanja v angleščini imenujemo *walk-forward*



Slika 4.1: Primer razdelitve podatkov.

validation, skico lahko vidimo na Sliki 4.2. Učno množico smo uporabili za iskanje optimalnih parametrov in koeficientov, testno pa za napovedovanje in preverjanje točnosti napovedi.



Slika 4.2: Skica postopka *walk-forward validation*.

Metoda AR je sama določila parameter p, ki določa število preteklih vrednosti za izračun napovedne vrednosti, na osnovi učne množice podatkov. Pri modelu VAR smo s kriterijsko funkcijo AIC na osnovi učne množice podatkov izbrali optimalen parameter p.

Za metodi ARIMA in ARIMAX smo optimalne parametre p, d in q iskali na vnaprej določenih omejenih intervalih za vsak parameter posebej. Vse kombinacije parametrov smo primerjali med seboj tako, da smo model naučili z določeno kombinacijo parametrov na učni množici podatkov ter primerjali napovedane vrednosti s testno množico. Parametre p, d in q, ki so dosegli najbolj točne napovedi, smo izbrali za optimalne. Tak postopek iskanja

optimalnih parametrov v angleščini imenujemo *grid search*.

Testno množico podatkov smo pri vseh metodah uporabili za ocenjevanje točnosti modela s kriterijskima funkcijama, opisanima v podpoglavju 4.1.

Poglavje 5

Rezultati

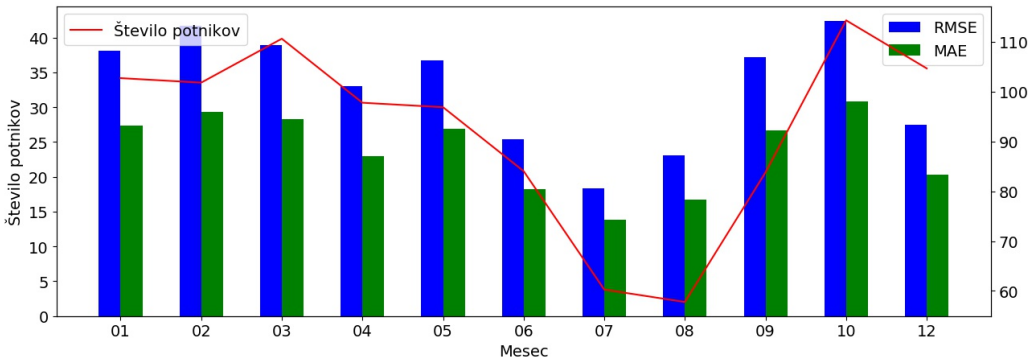
5.1 Napovedovanje števila potnikov

Število potnikov se čez dan, teden in mesec spreminja, temu je približno prilagojeno število in pogostost avtobusov na linijah. Na izbranih postajališčih (seznam postajališč vidimo v Tabeli 2.3) smo s pomočjo metod za napovedovanje časovnih vrst, opisanih v poglavju 3, napovedovali število potnikov na postajališču v naslednji uri. Pri tem smo uporabili podatke, opisane v poglavju 2. Primer uporabljenih podatkov, ki vsebujejo podatke o potnikih in vremenu, lahko vidimo v Tabeli 5.1. Pri metodi ARIMAX smo predpostavili, da lahko napovemo vrednosti vremenskih podatkov za naslednjo uro, saj teh s sodobnimi sistemi ni zahtevno dobro napovedati [27, 28, 29].

Za izbrana postajališča velja, da so napake napovedi v veliki meri odvisne od števila potnikov. Slika 5.1 prikazuje povprečne mesečne vrednosti RMSE in MAE vseh metod ter povprečno mesečno število potnikov na postajališču na uro. Vrednost RMSE in MAE pada oziroma raste skupaj s spremembo

Tabela 5.1: Primer uporabljenih podatkov.

	Število potnikov	Število avtobusov	Tlak	Temperatura	Vlažnost	Hitrost vetra	Količina padavin
2012-01-01 00:00:00	2,0	1	983,5	−0,5	92,0	1,0	0,0
2012-01-01 01:00:00	1,0	1	984,0	−0,5	91,5	0,4	0,0
2012-01-01 02:00:00	6,0	4	984,0	−0,65	90,0	0,3	0,0



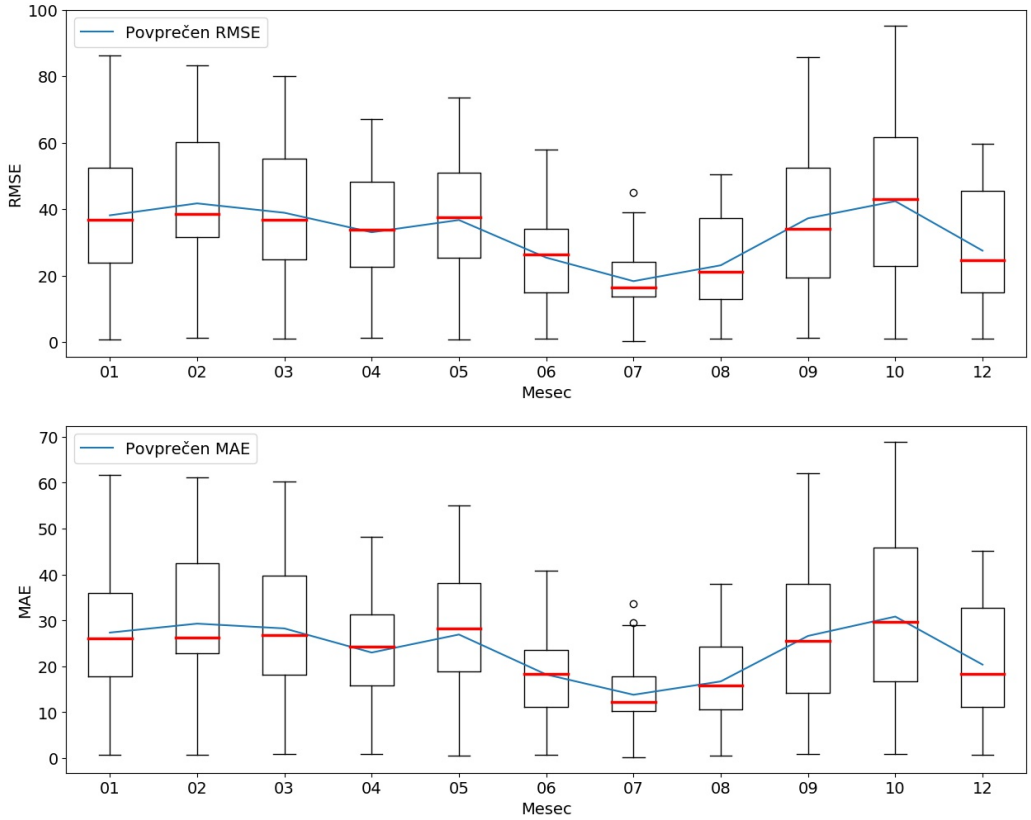
Slika 5.1: Povprečno število potnikov na uro in povprečen RMSE in MAE.

Tabela 5.2: Povprečne vrednosti RMSE, MAE in števila potnikov po mesecih.

	01	02	03	04	05	06	07	08	09	10	12
RMSE	38,1	41,7	38,9	33,0	36,7	25,4	18,3	23,1	37,2	42,4	27,5
MAE	27,3	29,3	28,3	23,0	26,9	18,2	13,8	16,7	26,6	30,8	20,4
Število potnikov	102	101	110	97	96	84	60	57	83	114	104

števila potnikov, kar pomeni, da je velikost napake napovedi odvisna tudi od števila potnikov. V poletnih mesecih, ko se je zmanjšalo število potnikov na postajališčih, je bila napaka napovedanih vrednosti manjša v primerjavi z ostalimi meseci v letu. V Tabeli 5.2 so zapisane povprečne vrednosti prikazane na Sliki 5.1. Povprečno je bilo največ 114 potnikov na uro v mesecu oktobru, ko je bila tudi napaka napovedanih vrednosti največja. Povprečno najmanj 57 potnikov na uro je bilo avgusta, vendar napaka napovedanih vrednosti števila potnikov ni bila najnižja. Najbolj točne napovedi smo dobili meseca julija, ko je bilo povprečno 60 potnikov na uro.

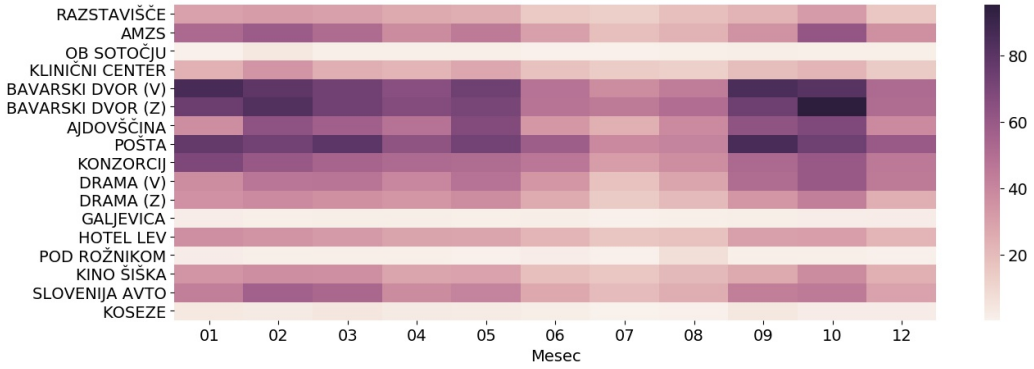
Slika 5.2 prikazuje mesečno porazdelitev RMSE in MAE vrednosti za izbrana postajališča ter povprečne mesečne vrednosti za vsako mero točnosti (modra črta). Posamezen mesec predstavlja razlike med povprečnimi vrednostmi, RMSE oziroma MAE, postajališč, kjer lahko vidimo najmanjšo in največjo izračunano napako napovedi števila potnikov v posameznem me-



Slika 5.2: Razlike v napakah napovedanih vrednosti med postajališči.

secu. Pravokotnik ponazarja območje med prvim in tretjim kvartalom, črta znotraj pravokotnika pa mediano vrednosti v mesecu. V poletnih mesecih, julija in avgusta, je napaka na vseh postajališčih najmanjša, iz tega sklepamo, da je za ta dva meseca, še posebej julij, lažje napovedovati število potnikov na postajališču kot preostali del leta.

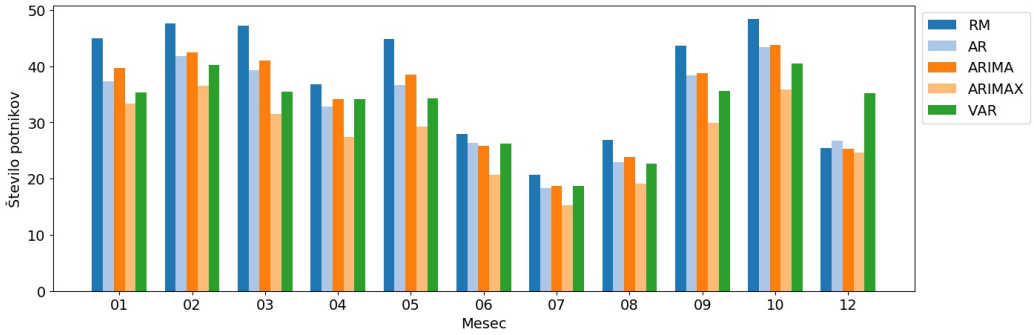
Izbrana postajališča so različno obremenjena s številom potnikov, kar privede do različnih napak pri napovedovanju. Povprečne vrednosti RMSE po posameznih postajališčih in mesecih smo prikazali s toplotno mapo na Sliki 5.3. Obe postajališči Bavarski dvor imata največje povprečne vrednosti RMSE, in sicer 66 potnikov, ki se od povprečja vseh postajališč (33) razlikuje za 33 potnikov. S povprečno 223 in 214 potniki na uro sta postajališči



Slika 5.3: Toplotna mapa povprečnih vrednosti RMSE po postajališčih in mesecih.

Bavarski dvor malo manj obremenjeni kot postajališče Pošta, ki ima največ potnikov na uro, in sicer 243. Povprečna napaka napovedanih potnikov postajališča Pošta ne odstopa veliko od postajališča Bavarski dvor in znaša 65 potnikov. Štiri postajališča, ki so imela najmanjšo povprečno napako, povprečne vrednosti RMSE pod 2,6 potnika, so Ob sotočju, Galjevica, Pod Rožnikom in Koseze. Na Sliki 5.3 imajo pravokotnike obarvane z belo oziroma zelo svetlo vijoličasto barvo. Omenjena štiri postajališča imajo tudi najmanjše število potnikov na uro med izbranimi postajališči, in sicer med 1 in 3 potniki na uro. Iz tega sklepamo, da postajališča z manjšim številom potnikov lažje napovedujemo kot tista z večjim številom potnikov, ki imajo večje napake.

Metode za napovedovanje časovnih vrst, ki smo jih uporabili pri napovedovanju števila potnikov na postajališču, so dosegle različne točnosti pri napovedovanju. Povprečne mesečne vrednosti RMSE, združene po metodah in izbranih postajališčih, so prikazane na Sliki 5.4. Vse metode so dosegle boljše povprečne napovedi od referenčne metode, razen v mesecu decembru sta bili metodi AR in VAR s slabšim rezultatom, kar lahko vidimo v Tabeli 5.3 in je razloženo v nadaljevanju. Rezultate z najmanjšo povprečno napako v celotnem letu smo dobili z metodo ARIMAX. Povprečna napaka



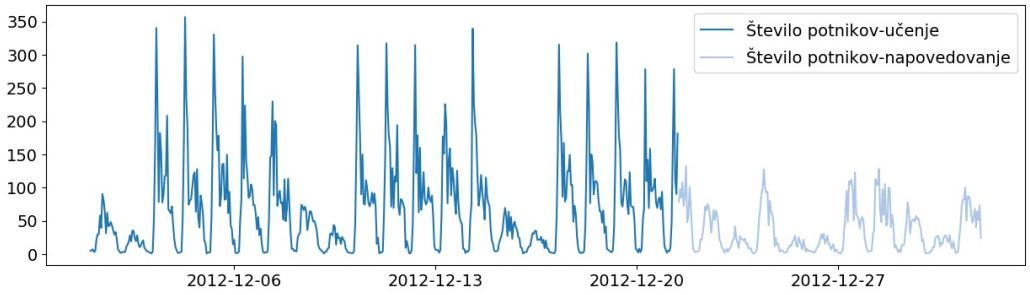
Slika 5.4: Povprečne vrednosti RMSE metod za napovedovanje časovnih vrst.

Tabela 5.3: Povprečne vrednosti RMSE metod za napovedovanje časovnih vrst po mesecih.

	01	02	03	04	05	06	07	08	09	10	12
RM	44,9	47,6	47,2	36,8	44,8	27,9	20,7	26,9	43,6	48,4	25,5
AR	37,4	41,8	39,3	32,8	36,7	26,4	18,2	22,0	38,4	43,4	26,8
ARIMA	39,7	42,5	41,0	34,1	38,6	25,9	18,8	23,9	38,8	43,8	25,3
ARIMAX	**33,4**	**36,6**	**31,6**	**27,4**	**29,3**	**20,7**	**15,3**	**19,1**	**29,9**	**35,8**	**24,6**
VAR	35,3	40,2	35,5	34,1	34,3	26,3	18,6	22,7	35,5	40,5	35,3

metode ARIMAX je bila 27,6 potnikov, kar je za 5,4 potnikov boljše od povprečja vseh metod skupaj. Naslednja metoda z najboljšo povprečno napako je metoda VAR, ki ima povprečno napako 32,6 potnikov, izstopa pa mesec december. Meseca decembra je metoda VAR dosegla najslabši rezultat, 35,3 potnikov, referenčni model pa 25,5 potnikov, razlog za to je razložen v nadaljevanju. Iz rezultatov je razvidno, da dodatni zunanji dejavniki vplivajo na izboljšanje napovedi števila potnikov na izbranih postajališčih pri metodi ARIMAX.

V mesecu decembru so metode za napovedovanje časovnih vrst dosegle podpovprečno napako napovedi, čeprav število potnikov ni upadlo v veliki meri. Metodi AR in VAR sta imeli večjo napako napovedanih vrednosti kot referenčni model. To lahko obrazložimo, če pogledamo Sliko 5.5, kjer vidimo začetni temno moder del, ki predstavlja podatke za učenje in iskanje

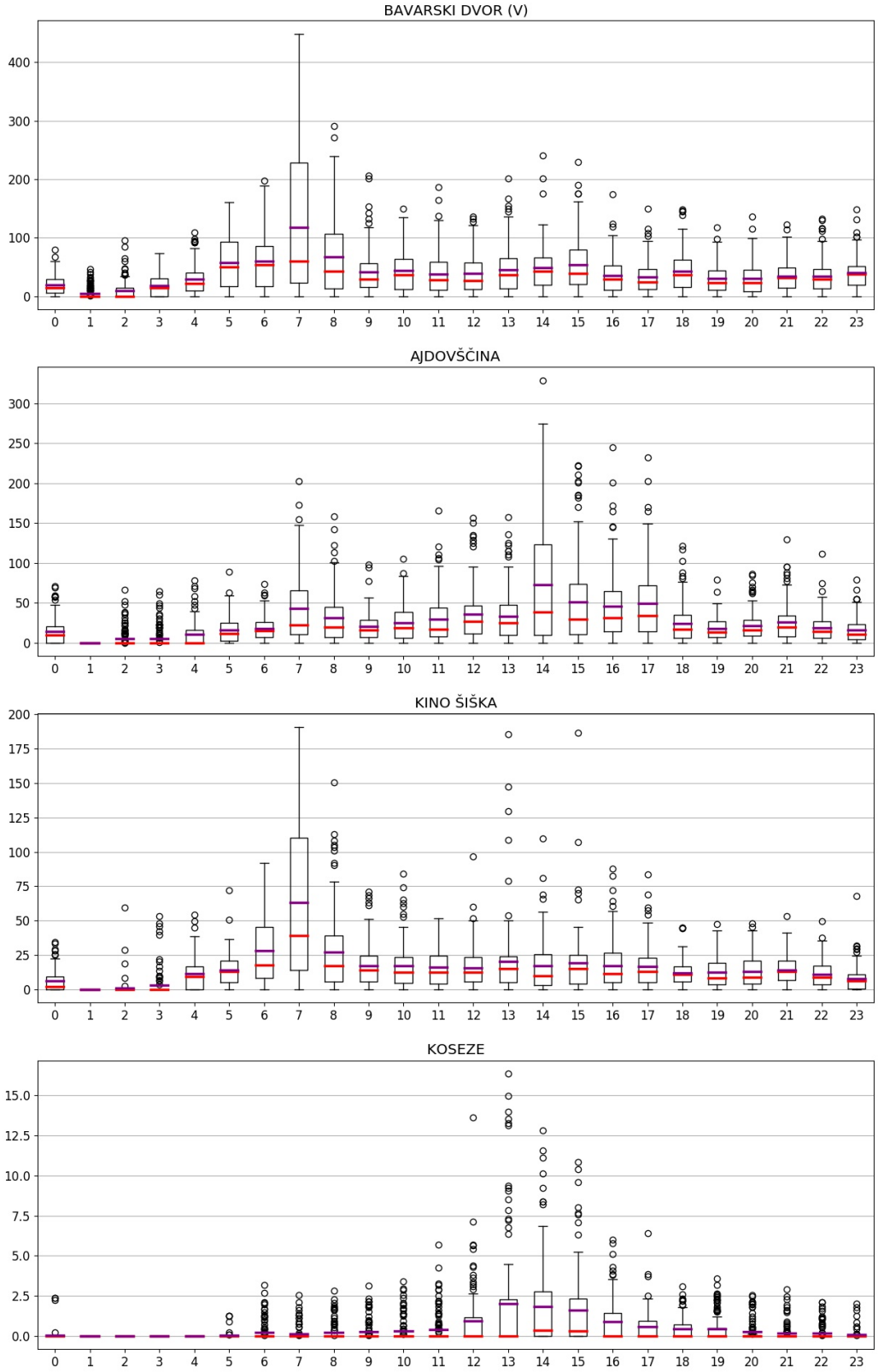


Slika 5.5: Število potnikov meseca decembra na izbranih postajališčih.

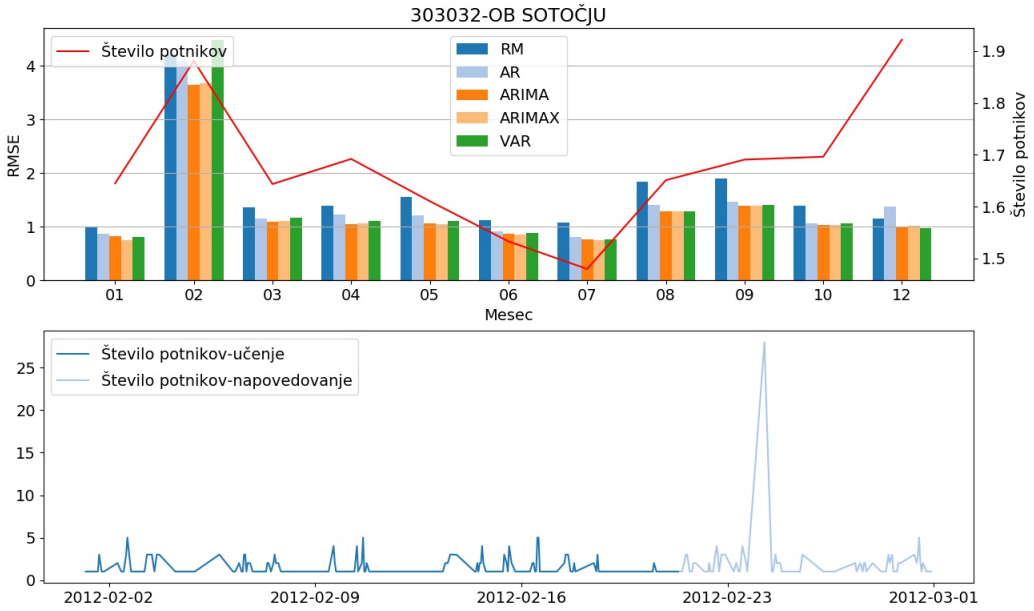
koeficientov metod, ter svetlo moder del, ki smo ga napovedali. Opazimo, da se v delu napovedovanja oblika grafa spremeni, saj je bilo obdobje počitnic, ki jih v našem modelu ne upoštevamo. Zmanjša se število potnikov, kar v danem primeru koristi referenčni metodi, da doseže napovedi z manjšo napako.

Pri pregledovanju rezultatov smo ugotovili, da na postajališčih z manjšim številom potnikov te lažje napovemo. Vendar prihaja do neenakomerne porazdeljenosti potnikov v mesecu in obstajajo določene ure, ob katerih je bilo v kratkem času veliko potnikov. To se odraža pri napaki napovedi, saj metode ne morejo napovedati nepredvidljivih dogodkov. Na Sliki 5.6 vidimo porazdelitev napake glede na uro v dnevu za postajališča Bavarski dvor (V), Ajdovščina, Kino Šiška in Koseze. Določene ure na postajališčih izstopajo, vendar velika odstopanja pri napakah se pojavijo le na postajališčih, ki imajo sicer manjše število potnikov. Na postajališčih z več potniki ta izstopanja ne vplivajo toliko, saj se porazdelijo čez celoten dan. Postajališča, kjer smo opazili odstopanja, so Ob sotočju, Pod Rožnikom in Koseze.

Vrednosti RMSE za uporabljene metode in število potnikov na uro za postajališče Ob sotočju je prikazano na zgornjem delu Slike 5.7, na spodnjem delu pa vidimo število potnikov na uro za mesec februar. Drugi del Slike 5.7 je razdeljen tudi na učne podatke, ki smo jih uporabili za ocenjevanje parametrov metod in podatke, ki smo jih napovedovali. Vidimo, da je med napovedanimi podatki dogodek, ko je število potnikov 24. februarja poskočilo



Slika 5.6: Razpon napake števila potnikov glede na uro v dnevu.

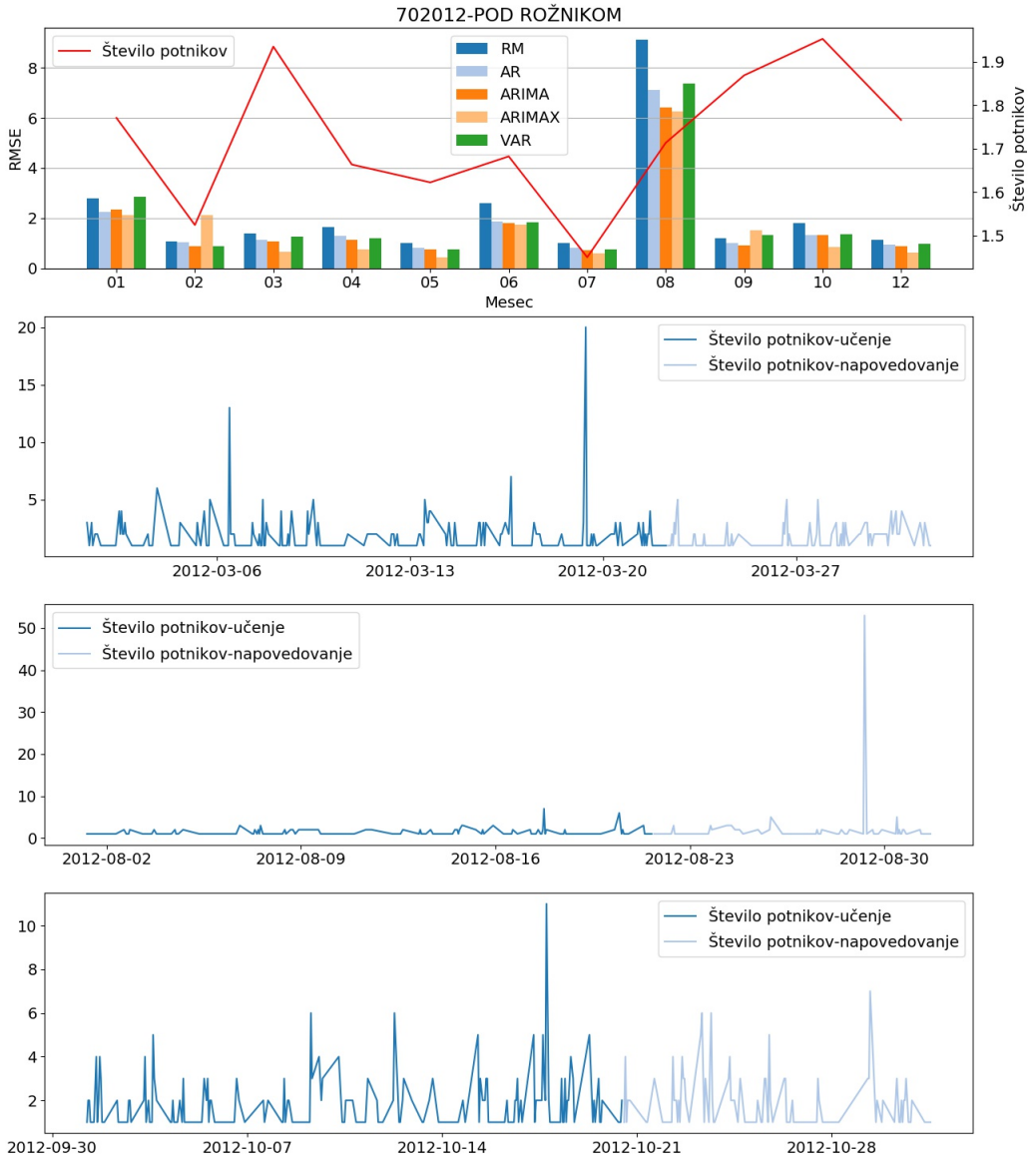


Slika 5.7: Število potnikov in rezultati napovedi za postajališče Ob sotočju.

na 28 potnikov med 6. ter 7. uro in na 23 potnikov med 7. ter 8. uro. Tega dela metode za napovedovanje časovnih vrst niso znale napovedati, in to je razlog, da je bila napaka napovedi enaka 4 potnikom, čeprav je povprečna napaka celotnega leta bila 1,4 potnika.

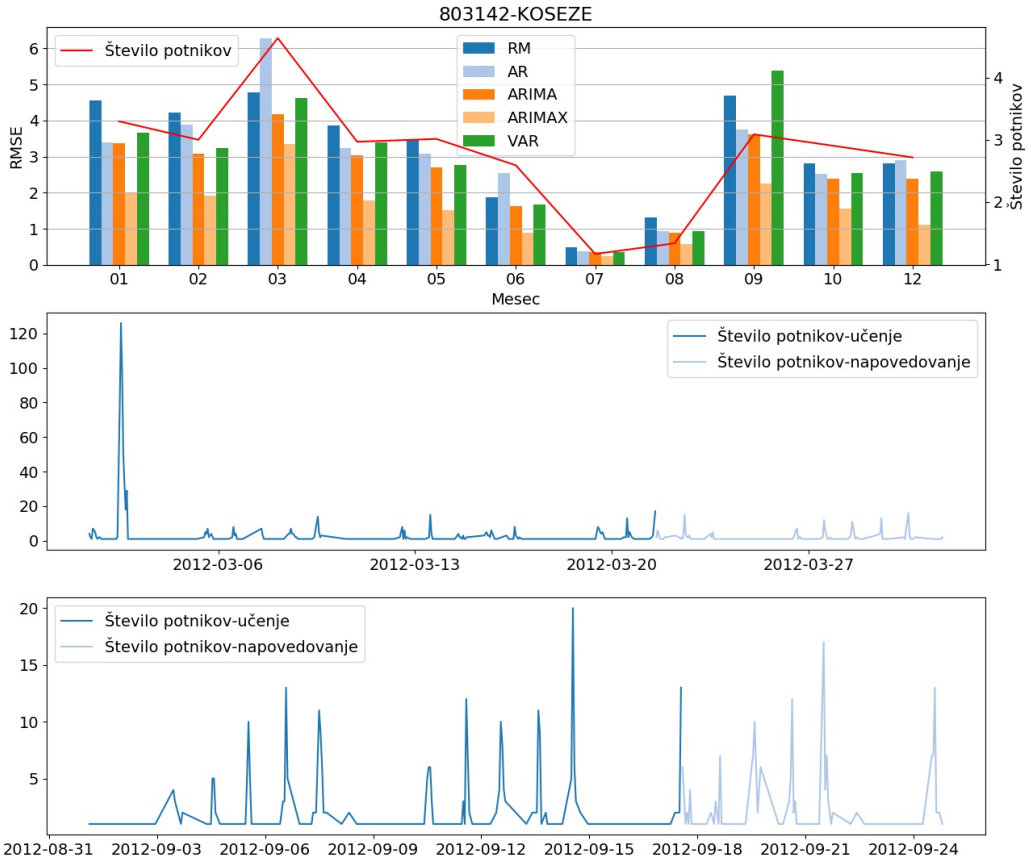
Podobno kot pri postajališču Ob sotočju (Slika 5.7) je tudi pri postajališču Pod Rožnikom v mesecu avgustu prišlo do nenavadnega povečanja števila potnikov, kar je prikazano na Sliki 5.8 v tretjem grafu. To povečanje je bilo 28. avgusta med 7. in 8. uro na 53 potnikov, povprečno v mesecu avgustu pa je bilo 1,7 potnika na uro. Na postajališču Pod Rožnikom je bila, v letu 2012, povprečna napaka 2 potnika, meseca avgusta pa je narasla na 7 potnikov. Podobno povečanje števila potnikov vidimo v mesecu marcu, septembru in oktobru, vendar v teh primerih ni prišlo do povečanja napake, saj je bilo povečano število potnikov v podatkih, ki smo jih uporabili za učenje.

Tretje postajališče, kjer smo opazili nihanja v povprečnemu številu potnikov na uro, je postajališče Koseze. Slika 5.9 prikazuje vrednosti RMSE za postajališče Koseze, na spodnjih dveh grafih pa vidimo število potnikov



Slika 5.8: Število potnikov in rezultati napovedi za postajališče Pod Rožnikom.

v mesecu marcu in septembru. Meseca marca je med podatki za učenje povečano število potnikov 2. marca med 12. in 15. uro, ko je bilo med 50 in 126 potnikov v eni uri, mesečno povprečje pa je 5 potnikov na uro. Vendar to ni poglavitno vplivalo na napako napovedi, saj so bili podatki v učni



Slika 5.9: Število potnikov in rezultati napovedi za postajališče Koseze.

množici. Na napako napovedi je vplivalo povečanje števila potnikov meseca septembra, kjer lahko vidimo izrazito povečanje števila potnikov med delovnimi dnevi. To lahko utemeljimo z začetkom šolskega leta, saj se v bližini postajališč nahaja vrtec in osnovna šola.

Postajališč z izjemami, kakršne so predhodno opisane, je bilo pet. Postajališča imajo različno razporejeno število potnikov v mesecu, tednu in delu dneva, kar moramo upoštevati pri napovedovanju in kasnejši analizi rezultatov. Večina od petih postajališč je imelo manjše število potnikov, s česar lahko sklepamo, da so bolj nepredvidljiva za napovedovanje, saj bi tovrstne dogodke težko vključili v napovedovanje.

Celotno gledano ne velja zmeraj, da je pri enakem številu potnikov po-

dobna napaka, saj včasih ni glavni vpliv na napoved število potnikov. Za primer vzemimo mesec junij in september, ki se po številu potnikov razlikujeta za enega potnika, njuni točnosti napovedi pa za skoraj 12 potnikov, kar predstavlja 32 % večjo napako meseca septembra. Podobno velja za posamezna postajališča, na primer postajališče Ajdovščina in Konzorcij se v točnosti napovedi razlikujeta le za 2 potnika, pri številu skupaj prepeljanih potnikov pa jih je bilo na postajališču Konzorcij za 51 % več, in sicer 1.216.077, kot na postajališču Ajdovščina. Obe postajališči sta v središču mesta. Postajališči Ajdovščina in Drama (Z) se v številu potnikov razlikujeta le za 1 %, v napaki napovedanih vrednosti pa za skoraj 18 potnikov, kar predstavlja 36 % manjšo točnost na postajališču Ajdovščina.

5.2 Napovedovanje časa vožnje avtobusov

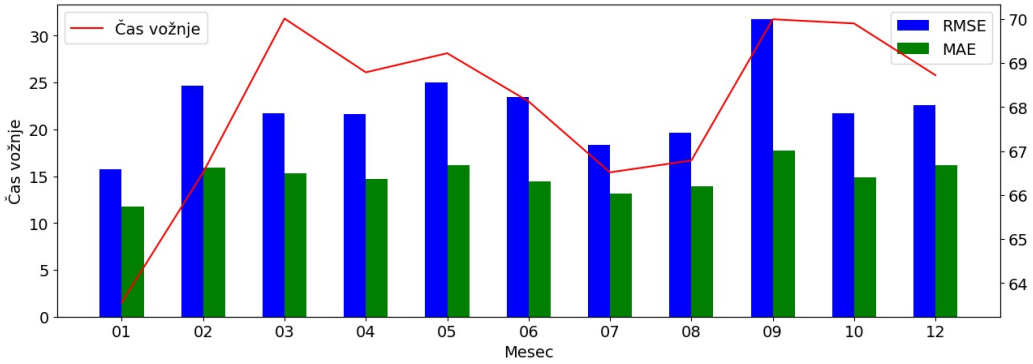
Avtobusi imajo na različnih linijah in postajališčih različen čas vožnje med dvema zaporednima postajališčema. Z metodami za napovedovanje časovnih vrst smo napovedovali čas vožnje avtobusa med dvema zaporednima postajališčema ter primerjali metode med seboj. Za napovedovanje smo uporabili podatke, opisane v poglavju 2, primer vidimo v Tabeli 5.4. Ker smo napovedovali čas vožnje avtobusa med dvema zaporednima postajališčema, smo uporabili podatek o številu potnikov na predhodnem postajališču, za katerega smo napovedovali in število avtobusov na postajališču, za katerega smo napovedovali. Pri napovedovanju smo predpostavili, da podatke o zunanjih dejavnikih lahko preprosto in natančno napovemo za naslednjo uro [27, 28, 29].

Slika 5.10: Skica za napovedovanje časa vožnje avtobusov.

Tabela 5.4: Primer uporabljenih podatkov.

	Čas vožnje	Število potnikov	Število avtobusov	Tlak	Temperatura	Vlažnost	Hitrost vetra
2012-01-18 04:00:00	33,0	1,0	1,0	994,0	−3,5	82,5	0,45
2012-01-18 05:00:00	38,5	22,0	8,0	994,0	−3,7	83,5	1,0
2012-01-18 06:00:00	41,5	64,0	31,0	995,0	−3,45	82,0	0,35

Napake napovedanih vrednosti so predvsem odvisne od časa vožnje avtobusa med postajališčema. Povprečen čas vožnje, RMSE in MAE je prikazan na Sliki 5.11. Vidimo, da se povprečen čas vožnje giba med 63 in 70 sekundami. Najdaljši čas vožnje je bil meseca marca, najkrajši pa januarja. Napake napovedanih vrednosti so sledile času vožnje. Največja povprečna napaka je bila meseca septembra, najmanjša pa januarja.

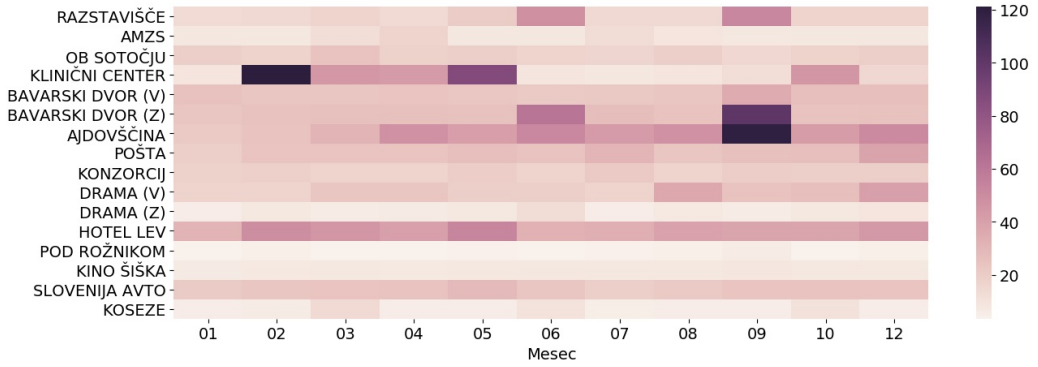


Slika 5.11: Povprečen čas vožnje na uro in povprečen RMSE in MAE.

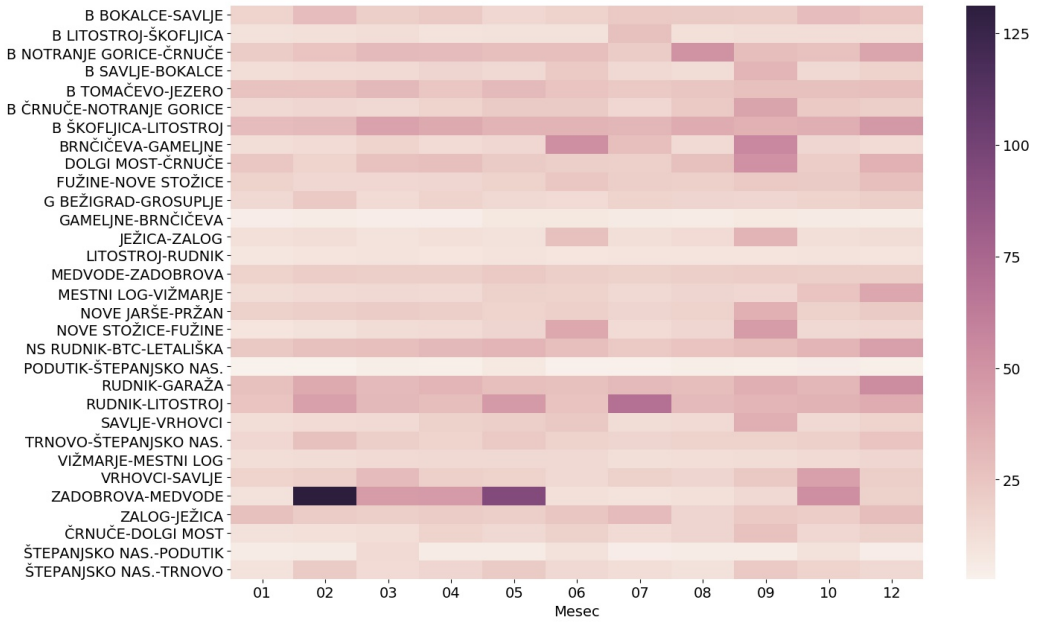
Tabela 5.5: Povprečen čas vožnje na uro in povprečen RMSE in MAE po mesecih.

	01	02	03	04	05	06	07	08	09	10	12
RMSE	15,7	24,7	21,7	21,6	24,9	23,4	18,4	19,6	31,8	21,7	22,6
MAE	11,8	15,9	15,4	14,7	16,1	14,5	13,2	13,9	17,7	14,8	16,2
Čas vožnje	63,6	66,7	70,0	68,8	69,2	68,1	66,5	66,8	69,9	69,9	68,7

Porazdelitev napake napovedanih vrednosti po postajališčih in posameznih mesecih vidimo na Sliki 5.12. Opazimo, da je za postajališče Klinični center v mesecu februarju in Ajdovščina v mesecu septembru bila napaka



Slika 5.12: Porazdelitev napake napovedanih vrednosti po postajališčih in mesecih.

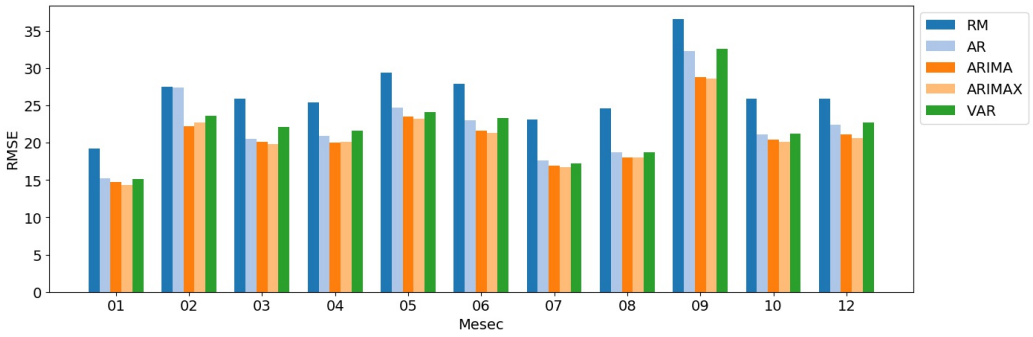


Slika 5.13: Porazdelitev napake napovedanih vrednosti po linijah in mesecih.

zelo odstopajoča od ostalih mesecev in postajališč. Če pogledamo povprečno vrednost RMSE na postajališče, je najtežje napovedovati postajališče Ajdovščina, ki ima povprečno vrednost RMSE 47,6 sekunde, sledi mu posta-

jališče Hotel Lev s povprečno vrednostjo RMSE 40,7 sekunde. Obe postajališči sta v središču mesta, kjer je velika prometna obremenjenost.

Slika 5.13 prikazuje temperaturno mapo vseh linij po mesecih za povprečne vrednosti RMSE. Opazimo, da je mesec september za vse linije težje napovedovati, saj povprečen RMSE znaša 25,8 sekunde, pri ostalih mesecih pa je RMSE približno 20 sekund. Linija Zadobrova–Medvode je imela meseca februarja veliko napako, in sicer 131,2 sekunde ter tudi povprečno vrednost RMSE ima ta proga največjega, to je 40,7 sekunde. Ta napaka je vzrok, da je v zgornjem grafu izstopajoča napaka pri postajališču Klinični center v mesecu februarju.



Slika 5.14: Povprečne vrednosti RMSE glede na metodo.

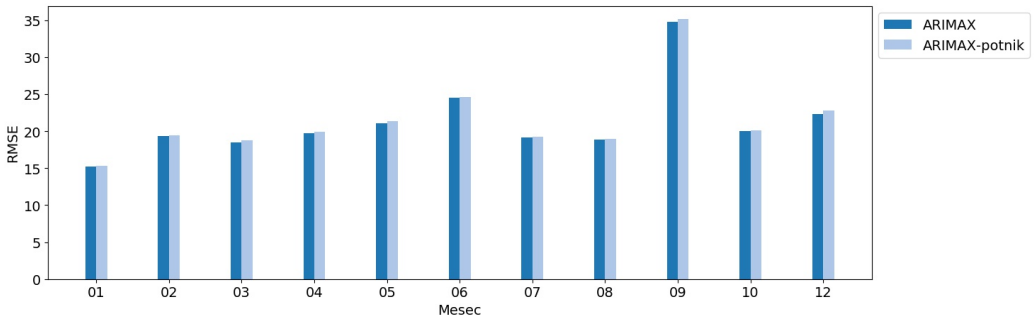
Tabela 5.6: Povprečne vrednosti RMSE glede na metodo po mesecih.

	01	02	03	04	05	06	07	08	09	10	12
RM	19,2	27,5	25,9	25,4	29,4	27,9	23,1	24,6	36,6	25,9	25,9
AR	15,2	27,4	20,5	20,9	24,7	22,9	17,6	18,7	32,3	21,1	22,4
ARIMA	14,7	**22,2**	20,1	**20,0**	23,5	21,7	16,9	18,0	28,8	20,4	21,1
ARIMAX	**14,3**	22,7	**19,8**	20,1	**23,2**	**21,3**	**16,8**	**17,9**	**28,5**	**20,1**	**20,6**
VAR	15,1	23,6	22,1	21,6	24,13	23,3	17,3	18,8	32,6	21,3	22,7

Vsi modeli so imeli bolj točno napoved od referenčnega modela. Najbolj točne napovedi sta dosegli metodi ARIMA in ARIMAX, saj imata zelo podobne rezultate. S Slike 5.14 vidimo, da je najtežje napovedovati mesec

september, najlažje pa mesec januar. Povprečna vrednost RMSE pri modelu ARIMA je 20,7 sekunde, pri modelu ARIMAX pa 20,5 sekunde. Iz te relativno majhne razlike lahko sklepamo, da pri napovedovanju časov vožnje avtobusa poznavanje vremenskih podatkov nima bistvenega vpliva na napovedne točnosti.

Na izbranih postajališčih smo evalvirali model ARIMAX tudi brez dodatnih podatkov o številu potnikov na predhodnem postajališču, saj smo želeli preveriti njihov vpliv na napovedane vrednosti. Rezultati se niso dosti razlikovali od rezultatov s podatki o številu potnikov, kar lahko vidimo na Sliki 5.15. Povprečen RMSE pri metodi ARIMAX je znašal 20,5 sekunde, brez dodatnega podatka o številu potnikov na predhodnem postajališču pa 21,4 sekunde. Iz tega lahko sklepamo, da ta podatek ne vpliva bistveno na napoved časa vožnje, ampak so ostali zunanji dejavniki vplivali na izboljšanje napovedi časa potovanja avtobusa med dvema zaporednima postajališčema (število avtobusov, vremenski podatki).



Slika 5.15: Primerjava modela ARIMAX za napovedovanje časa vožnje avtobusov z in brez eksogenega dejavnika števila potnikov na predhodnem postajališču.

Glede na povprečne čase vožnje avtobusov vidimo, da ne velja za vse mesece enako razmerje med časom vožnje in napako napovedanih vrednosti. Mesec marec in oktober imata enako število potnikov in enako napako, vendar ima tudi mesec september enak povprečni čas vožnje, v točnosti napovedi pa

se razlikuje za 32 %. Postajališči Hotel Lev in Drama (V) sta v središču mesta in imata čas vožnje različen le za 3 sekunde, njuna napaka napovedanih časov vožnje pa se razlikuje za 40 %, kar predstavlja 16 sekund.

Tako pri napovedovanju časa vožnje avtobusov med dvema zaporednima postajališčema, še zlasti pa pri napovedovanju števila potnikov na postajališču, poznavanje vremenskih razmer v bližnji prihodnosti, na primer v naslednji uri, močno vpliva na točnost napovedi. Najbolj točne napovedi smo dobili z metodo ARIMAX, ki v naših eksperimentih predvideva poznavanje vremenskih razmer v naslednji uri. Vendar pa doprinos poznavanja vremenskih podatkov na napovedno točnost pri napovedovanju časa vožnje avtobusov ni tako velik kot pri napovedovanju števila potnikov na postajališčih.

Ugotovili smo tudi, da število potnikov na predhodnem postajališču ne vpliva na čas vožnje avtobusa med dvema zaporednima postajališčema. Nekatera postajališča in proge je v določenih mesecih težje napovedovati, tako za ta postajališča in mesece predvidevamo, da so bolj nepredvidljiva.

Poglavje 6

Zaključki

Kakovostna priprava podatkov je predstavljala pomemben del pridobivanja ustreznih rezultatov. Velik del smo namenili analizi in pripravi podatkov o potnikih in časih vožnje avtobusa med dvema zaporednima postajališčema, saj je bilo podatkov zelo veliko. Velika količina podatkov nam je omogočala bolj reprezentativne rezultate na izbranih postajališčih. Iz podatkov smo ugotovili, da je število potnikov med tednom večje kot ob koncih tedna in da ima dan v povprečju dve skrajnosti, prva je zjutraj, ko se ljudje odpravijo v šole in službe, druga pa popoldne ob odhodu domov. Podobno velja za čas vožnje avtobusa, saj so ti ob jutranjih in popoldanskih konicah daljši.

Pri napovedovanju časa vožnje avtobusa med dvema zaporednima postajališčema, še zlasti pa pri napovedovanju števila potnikov na postajališčih, smo pokazali, da lahko dobro poznavanje vremenskih razmer v bližnji prihodnosti, denimo v naslednjih urah, pomembno vpliva na izboljšanje točnosti napovedi. To smo ugotovili s primerjavo metode ARIMAX s preostalimi metodami, ki smo jih uporabili. Pri uporabi vremenskih podatkov kot pojasnjevalnih spremenljivk je metoda ARIMAX dosegla občutno in konsistentno boljše napovedi.

Napovedne točnosti so v določenih mesecih boljše, še posebej velja za poletne mesece in mesece z manjšim številom potnikov in krajšim časom vožnje. Ugotovili smo, da med izbranimi postajališči nekatera težje napovemo ozi-

roma imajo manjšo točnost čez celotno leto. Enako velja za določene linije na izbranih postajališčih. Iz tega sklepamo, da so ta postajališča oziroma linije bolj nepredvidljive in jih je težje napovedovati.

Za postajališča v središču mesta je čas vožnje avtobusa praviloma težje napovedovati kot za postajališča izven središča mesta. Vendar obstajajo izjeme. Na primer za mesec september se je pokazalo, da je čas vožnje težje napovedovati tudi na postajališčih izven središča mesta. Za mesec marec in oktober so eksperimenti pokazali, da kljub tipično povečanemu prometu in daljšemu času vožnje avtobusov, teh ni težje napovedovati in sta ta dva meseca v tem smislu bolj predvidljiva. Pokazali smo tudi več eksperimentalnih rezultatov pri posameznih avtobusnih postajah, ki dajejo slutiti, da povečan promet sam po sebi ne pomeni nujno tudi večje nepredvidljivosti trajanja vožnje.

Za dodatno izboljšanje točnosti napovedi predlagamo vključitev dodatnih zunanjih dejavnikov, kot so razdalja med postajališči, delo na cesti, javni dogodki, kjer se zbere večje število ljudi, in podatki o ostalih vozilih v cestnem prometu. Napovedovanje bi poskušali izboljšati tudi z upoštevanjem praznikov in počitnic ter ločenim napovedovanjem vrednosti med tednom in koncem tedna. Dodatni zunanji dejavniki bi predvidoma še dodatno izboljšali napovedi modela ARIMAX.

Dodatek A

Izbrana postajališča in linije

- 100021-Razstavišče:
  - B ČRNUČE-NOTRANJE GORICE,
  - B TOMAČEVO-JEZERO,
  - BRNČIČEVA-GAMELJNE,
  - ČRNUČE-DOLGI MOST,
  - JEŽICA-ZALOG,
- 103031-AMZS:
  - B ČRNUČE-NOTRANJE GORICE,
  - BRNČIČEVA-GAMELJNE,
  - ČRNUČE-DOLGI MOST,
  - JEŽICA-ZALOG,
- 303032-Ob sotočju:
  - TRNOVO-ŠTEPANJSKO NAS.,
- 402031-Klinični center:
  - FUŽINE-NOVE STOŽICE,

  - NOVE JARŠE-ZELENA JAMA,
  - ŠTEPANJSKO NAS.-TRNOVO,
  - ZADOBROVA-MEDVODE,
  - ZALOG-BEŽIGRAD,
  - ZALOG-JEŽICA,

- 600011-Bavarski dvor:
  - B BOKALCE-SAVLJE,
  - B NOTRANJE GORICE-ČRNUČE,
  - DOLGI MOST-ČRNUČE,
  - FUŽINE-NOVE STOŽICE,
  - NOVE JARŠE-ZELENA JAMA,
  - NS RUDNIK-BTC-LETALIŠKA,
  - TRNOVO-ŠTEPANJSKO NAS.,
  - VRHOVCI-SAVLJE,
  - ZALOG-BEŽIGRAD,
  - ZALOG-JEŽICA,

- 600012-Bavarski dvor:
  - B ČRNUČE-NOTRANJE GORICE,
  - B SAVLJE-BOKALCE,
  - B TOMAČEVO-JEZERO,
  - BEŽIGRAD-ZALOG,
  - BRNČIČEVA-GAMELJNE,
  - ČRNUČE-DOLGI MOST,
  - JEŽICA-ZALOG,

  - NOVE JARŠE-PRŽAN,
  - NOVE STOŽICE-FUŽINE,
  - SAVLJE-VRHOVCI,

- 600022-Ajdovščina:
  - B ČRNUČE-NOTRANJE GORICE,
  - B SAVLJE-BOKALCE,
  - B TOMAČEVO-JEZERO,
  - BEŽIGRAD-ZALOG,
  - ČRNUČE-DOLGI MOST,
  - JEŽICA-ZALOG,
  - NOVE STOŽICE-FUŽINE,
  - SAVLJE-VRHOVCI,
  - ŠTEPANJSKO NAS.-TRNOVO,
  - ZELENA JAMA-NOVE JARŠE,

- 601011-Pošta:
  - B NOTRANJE GORICE-ČRNUČE,
  - B ŠKOFLJICA-LITOSTROJ,
  - DOLGI MOST-ČRNUČE,
  - FUŽINE-NOVE STOŽICE,
  - MESTNI LOG-VIŽMARJE,
  - NS RUDNIK-BTC-LETALIŠKA,
  - RUDNIK-GARAŽA,
  - RUDNIK-LITOSTROJ,
  - TRNOVO-ŠTEPANJSKO NAS.,

  - ZALOG-BEŽIGRAD,
  - ZALOG-JEŽICA,

- 601012-Konzorcij:
  - B ČRNUČE-NOTRANJE GORICE,
  - B LITOSTROJ-ŠKOFLJICA,
  - B SAVLJE-BOKALCE,
  - B TOMAČEVO-JEZERO,
  - BEŽIGRAD-ZALOG,
  - ČRNUČE-DOLGI MOST,
  - G BEŽIGRAD-GROSUPLJE,
  - JEŽICA-ZALOG,
  - LITOSTROJ-RUDNIK,
  - N BAVARSKI DVOR-RUDNIK,
  - NOVE STOŽICE-FUŽINE,
  - SAVLJE-VRHOVCI,
  - ŠTEPANJSKO NAS.-TRNOVO,
  - VIŽMARJE-MESTNI LOG,
  - ZELENA JAMA-NOVE JARŠE,

- 602021-Drama:
  - B NOTRANJE GORICE-ČRNUČE,
  - DOLGI MOST-ČRNUČE,
  - MESTNI LOG-VIŽMARJE,

- 602022-Drama:
  - B ČRNUČE-NOTRANJE GORICE,

  - B LITOSTROJ-ŠKOFLJICA,
  - B TOMAČEVO-JEZERO,
  - BEŽIGRAD-ZALOG,
  - ČRNUČE-DOLGI MOST,
  - G BEŽIGRAD-GROSUPLJE,
  - JEŽICA-ZALOG,
  - LITOSTROJ-RUDNIK,
  - N BAVARSKI DVOR-RUDNIK,
  - NOVE STOŽICE-FUŽINE,
  - ŠTEPANJSKO NAS.-TRNOVO,
  - VIŽMARJE-MESTNI LOG,
  - ZELENA JAMA-NOVE JARŠE,

- 700012-Hotel Lev:
  - B ŠKOFLJICA-LITOSTROJ,
  - MESTNI LOG-VIŽMARJE,
  - RUDNIK-LITOSTROJ,

- 702012-Pod Rožnikom:
  - B SAVLJE-BOKALCE,
  - SAVLJE-VRHOVCI,

- 802021-Kino Šiška:
  - B LITOSTROJ-ŠKOFLJICA,
  - GAMELJNE-BRNČIČEVA,
  - LITOSTROJ-RUDNIK,
  - MEDVODE-ZADOBROVA,

  - PODUTIK-ŠTEPANJSKO NAS.,
  - VIŽMARJE-MESTNI LOG,

- 803011-Slovenija avto:
  - MEDVODE-ZADOBROVA,
  - VIŽMARJE-MESTNI LOG,

- 803142-Koseze:
  - ŠTEPANJSKO NAS.-PODUTIK.

Literatura

[1] LPP, Shema avtobusnih linij 2012, [dostop: 6. 2. 2018] (2012).
URL http://www.lpp.si/sites/default/files/lpp_si/stran/datoteke/shema_linij.pdf

[2] C. Bai, Z.-R. Peng, Q.-C. Lu, J. Sun, Dynamic bus travel time prediction models on road with multiple bus routes, Vol. 2015, Hindawi Publishing Corp., 2015, pp. 63–65.

[3] E. Jenelius, H. N. Koutsopoulos, Travel time estimation for urban road networks using low frequency probe vehicle data, Vol. 53, Elsevier, 2013, pp. 64–81.

[4] H. Sun, H. X. Liu, H. Xiao, R. R. He, B. Ran, Short term traffic forecasting using the local linear regression model, in: 82nd Annual Meeting of the Transportation Research Board, Washington, DC, 2003.

[5] J. Rice, E. Van Zwet, A simple and effective method for predicting travel times on freeways, Vol. 5, IEEE, 2004, pp. 200–207.

[6] G. Zhu, L. Wang, P. Zhang, K. Song, A kind of urban road travel time forecasting model with loop detectors, International Journal of Distributed Sensor Networks 12 (2).

[7] LPP, Letno poročilo 2012, [dostop: 6. 2. 2018] (2013).
URL http://www.lpp.si/sites/default/files/lpp_si/stran/datoteke/lpp_letno_porocilo_2012.pdf

[8] K. Bandara, C. Bergmeir, S. Smyl, Forecasting across time series databases using long short-term memory networks on groups of similar series, arXiv preprint arXiv: 1710.03222.

[9] S. D. Campbell, F. X. Diebold, Weather forecasting for weather derivatives, Vol. 100, Taylor & Francis, 2005, pp. 6–16.

[10] H. Margusity, Persistence forecasting: Today equals tomorrow, [dostop: 6. 2. 2018].
URL http://ww2010.atmos.uiuc.edu/(Gh)/guides/mtr/fcst/mth/prst.rxml

[11] R. J. Hyndman, G. Athanasopoulos, Forecasting: principles and practice, OTexts, 2014.

[12] H. Lütkepohl, M. Krätzig, Applied time series econometrics, Cambridge university press, 2004.

[13] S. Seabold, J. Perktold, Statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference, 2010.

[14] G. E. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, Time series analysis: forecasting and control, John Wiley & Sons, 2015.

[15] M. J. De Smith, Statistical Analysis Handbook, 2015.

[16] H. Wold, A study in the analysis of stationary time series, Ph.D. thesis, Almqvist & Wiksell (1938).

[17] G. E. Box, G. M. Jenkins, Some recent advances in forecasting and control, 1968.

[18] J. Brownlee, Machine learning mastery, , [dostop: 6. 2. 2018] (2014).
URL http://machinelearningmastery.com/

[19] J. D. Hamilton, Time series analysis, Vol. 2, Princeton university press Princeton, 1994.

[20] G. E. Box, G. C. Tiao, Intervention analysis with applications to economic and environmental problems, Vol. 70, Taylor & Francis, 1975, pp. 70–79.

[21] A. Pankratz, Forecasting with dynamic regression models, Vol. 935, John Wiley & Sons, 2012.

[22] R. Taylor, Pyflux: An open source time series library for python, [dostop: 21. 3. 2018] (2016).
URL http://pyflux.readthedocs.io/en/latest/index.html

[23] H. Lütkepohl, New introduction to multiple time series analysis, Springer Science & Business Media, 2005.

[24] E. Zivot, J. Wang, Vector autoregressive models for multivariate time series, Springer, 2006, pp. 385–429.

[25] C. A. Sims, Macroeconomics and reality, Econometrica: Journal of the Econometric Society (1980) 1–48.

[26] H. Akaike, Factor analysis and aic, Psychometrika 52 (3) (1987) 317–332.

[27] D. Orrell, L. Smith, J. Barkmeijer, T. Palmer, Model error in weather forecasting, Nonlinear processes in geophysics 8 (6) (2001) 357–371.

[28] R. Ronda, G. Steeneveld, B. Heusinkveld, J. Attema, A. Holtslag, Urban finescale forecasting reveals weather conditions with unprecedented detail, Bulletin of the American Meteorological Society 98 (12) (2017) 2675–2688.

[29] S. Nurunnahar, D. B. Talukdar, R. I. Rasel, N. Sultana, A short term wind speed forcasting using svr and bp-ann: A comparative analysis, in: Computer and Information Technology (ICCIT), 2017 20th International Conference of, IEEE, 2017, pp. 1–6.