UNIVERZA V NOVI GORICI
POSLOVNO-TEHNIŠKA FAKULTETA

MODELIRANJE IN NAPOVEDOVANJE GOSTOTE PROMETA V KRIŽIŠČU

MAGISTRSKO DELO

Lea Manfreda

Mentor: prof. dr. Juš Kocijan

Nova Gorica, 2014

ZAHVALA

So stvari, ki jih na življenjski poti lahko dosežemo sami, in so stvari, pri katerih potrebujemo pomoč in podporo.

Posebej bi se zahvalila mentorju na Poslovno-tehniški fakulteti Univerze v Novi Gorici, prof. dr. Jušu Kocijanu, ki mi je nudil strokovno pomoč, nasvete, razlago, spodbudo in potrpežljivost.

Zahvalila bi se tudi družini, teti, fantu in ožjim prijateljem za pomoč, podporo in spodbude pri celotnem študiju in nastajanju magistrskega dela.

Iskrena hvala!

NASLOV

Modeliranje in napovedovanje gostote prometa v križišču

IZVLEČEK

Magistrsko delo opisuje reševanje problema s področja matematičnega modeliranja dinamičnih sistemov, statistike in prometa. V delu smo prikazali možno rešitev problema v primeru odpovedi enega izmed senzorjev na križišču v velikem mestu. V velikih mestih so na križiščih postavljeni senzorji, ki merijo intenzivnost prometa v določenih časovnih intervalih. Na podlagi pridobljenih podatkov se izvaja nadzor urbanega omrežja. V primeru odpovedi senzorja je napovedovanje intenzivnosti prometa na podlagi preteklih merjenih vrednosti ena izmed možnih rešitev do odprave problema na senzorju.

Namen dela je bil izdelati matematične modele za enokoračno napovedovanje gostote prometa v izbranem križišču v Pragi na Češkem. Izdelani matematični modeli morajo biti dovolj točni, da bi se jih v primeru napake na senzorju lahko zares uporabilo za napovedovanje intenzivnosti prometa do popravila senzorja.

Kot osnovo za izdelavo modela za enokoračno napovedovanje smo uporabili metodo regresije. Na konkretnem primeru smo uporabili dve metodi eksperimentalnega modeliranja. Izbrani metodi sta avtoregresijski model (model AR) in poseben model drsečega povprečja (model MA). Dobljene matematične modele smo nato ovrednotili s cenilko srednje vrednosti kvadratov napake (SMSE), histogrami napak in grafičnimi prikazi napovedi.

Cilj dela je bil dosežen, saj so napovedi dovolj točne, vsi histogrami napak imajo približno normalno oziroma Gaussovo porazdelitev, vizualni pregled grafov pa dosega pričakovanja, saj napoved sledi srednjim vrednostim izmerjenih podatkov.

KLJUČNE BESEDE

matematični model, enokoračna napoved, regresija, prometno križišče

TITLE

Modeling and forecasting traffic intensity at road intersections

ABSTRACT

The master's thesis describes the problem solution from the field of mathematical modeling of dynamic systems, statistics and traffic. In this thesis we presented a possible solution to a problem which may occur if one of the sensors at an intersection in a large city fails. Sensors which measure traffic intensity at specific time intervals are placed at intersections in large cities. The urban network is monitored on the basis of the acquired data. In the event of a sensor failure traffic-intensity prediction based on previously measured values is one of the possible solutions until the sensor issue is fixed.

The purpose of this thesis was to create mathematical models for one-step-ahead prediction of traffic density at a chosen intersection in the city of Prague in the Czech Republic. The constructed mathematical models must be accurate enough so that they can actually be used to predict traffic intensity in the event of a sensor error until the sensor is fixed.

We used the regression method as the basis for creating the model for one-step-ahead prediction. In our case we used two methods of experimental modelling. The chosen methods are autoregressive model (AR model) and a special moving average model (MA model). We then validated the derived mathematical models with a standardised mean square error estimator (SMSE), error histograms and graphical presentations of forecasts.

The objective of the thesis was achieved because the forecasts are accurate enough, all error histograms have an approximately normal or Gaussian distribution and the visual review of the model responses is up to expectations because the forecast follows the median values of the measured data.

KEYWORDS

mathematical model, one-step-ahead prediction, regression, traffic intersection

KAZALO

KAZALO SLIK

KAZALO TABEL

1 UVOD

Delo prikazuje modeliranje intenzivnosti prometa za enokoračno napovedovanje v izbranem križišču v Pragi na Češkem. V delu se prepletajo statistika, matematično modeliranje dinamičnih sistemov in tehnologija prometa. Postopek, ko sistem iz dejanskega sveta analiziramo na inženirski način tako, da oblikujemo matematični model, imenujemo modeliranje. Model (Mathematical model, 2014) pomaga razumeti sistem in omogoča analizo vpliva različnih komponent na sistem. Matematični model mora imeti vse glavne značilnosti dejanskega sistema.

Upravljanje prometa v velikih mestih zmanjšuje zastoje vozil in s tem povezane nevšečnosti. Zanesljivo vodenje urbanega omrežja je odvisno od točne in pravočasne informacije o prometu. Število vozil v križiščih beležijo senzorji. Včasih se na senzorju zgodi napaka ali senzor odpove. V tem primeru se nadzor in vodenje prometa izvajata napačno. V času, ko vzdrževalno osebje popravlja senzor, je ena izmed možnih rešitev problema napovedovanje intenzivnosti prometa na podlagi preteklih vrednosti.

Namen dela je izdelati delujoče in točne modele za enokoračno napovedovanje gostote prometa v izbranem križišču v Pragi. Cilj dela je izdelati dovolj točne napovedi intenzivnosti prometa v izbranem križišču z matematičnim modeliranjem različnih dni s podobnimi lastnostmi. Enokoračne napovedi morajo dosegati dovolj točne napovedi, da se matematični model lahko uporabi v primeru, ko eden izmed senzorjev v izbranem križišču odpove. Pridobljene enokoračne napovedi bomo zato ovrednotili in jim določili točnost.

Na začetku dela bomo predstavili problem, na katerega se navezuje celotno delo. Sledi definicija in opis regresije kot osnovnega orodja za razvoj modelov za enokoračno napovedovanje. Opisali bomo tudi časovne vrste in metode eksperimentalnega modeliranja. Nato sledi poglavje o modeliranju prometa v križišču, kjer se bomo reševanja problema lotili z izbranima metodama iz eksperimentalnega modeliranja časovnih vrst na konkretnem problemu. Najprej bomo meritve modelirali in vrednotili s posebnim modelom drsečega povprečja. Sledi podpoglavje o modeliranju z avtoregresijskim modelom, ki je razdeljeno na tri podpoglavja: analizo delovnih dni, vikendov in praznikov. Zadnje poglavje,

modeliranje prometa v križišču, bomo zaključili z enokoračno napovedjo celega tedna z uporabo matematičnih modelov, opisanih v prejšnjih podpoglavjih. Sledi zaključek in povzetek glavnih ugotovitev.

2 OPIS PROBLEMA

Opisan problem napovedovanja gostote prometa v primeru napake senzorja je povzet po članku (Kocijan in Prikryl, 2010).

Prometni zastoji so resen problem, ki se ga trudijo rešiti prometni inženirji po vsem svetu. Zastoji povečujejo negotovosti v času potovanja, ki vodijo do stresa med udeleženci v prometu in do nevarnih prometnih situacij. Vodenje prometa zmanjšuje zastoje in s tem povezane nevšečnosti. Zanesljivost sistema za vodenje je odvisna od točne in pravočasne informacije o prometu. Sistem vodenja prilagodi stanje semaforjev in razbremeni križišča na podlagi informacije o prometu. Druga možnost pa je, da se voznike o prometnem zamašku obvesti in se jih preusmeri na drugo pot.

Na križiščih v večjih mestih so postavljeni senzorji, s katerimi se meri stanje prometa. Senzorji običajno merijo število vozil v določenem časovnem intervalu, kar imenujemo intenzivnost prometa. Iz pridobljenih podatkov o prometu se izvaja nadzor urbanega omrežja.

Včasih se na senzorju pojavi napaka (na primer ustavljeno vozilo na senzorju) ali pa induktivna zanka, ki je glavni element senzorja, preprosto odpove. V tem primeru senzor sporoči napako vzdrževalnemu osebju. Reakcijski čas osebja lahko traja od nekaj ur do nekaj dni.

Tipičen scenarij avtomatskega vodenja mestnega križišča poteka tako, da senzorji merijo promet na omejenem nadzorovanem območju. Tako sistem za vodenje in nadzor ugotovi trenutne potrebe prometa na določenem križišču. V primeru, da kakšen senzor zataji ali postane nezanesljiv, avtomatski sistem vodenja odpove ali ukrepa napačno. Zato se za primer napake pretekli podatki o prometu shranjujejo. V primeru napake senzorja se lahko na podlagi shranjenih podatkov promet napove do odprave problema. Namen napovedi je torej nemoteno upravljanje prometa. Ta začasna rešitev lahko bistveno izboljša stanje prometa v obdobju, preden se nedelujoči senzor popravi. V službi za nadzor prometa poskušajo z matematičnimi modeli odkrivati senzorske napake sproti, v dejanskem času.

V magistrskem delu bomo iz podatkov, dobljenih s senzorji, izdelali matematični model. Glavni namen tega matematičnega modela bo napovedovanje intenzivnosti prometa v prihodnosti na podlagi preteklih podatkov.

Kot primer modeliranja intenzivnosti prometa, ki se meri s številom vozil na časovno enoto, smo uporabili meritve prometa na izbranem križišču v Pragi na Češkem. Za to lokacijo v Pragi smo se odločili zato, ker smo imeli direkten dostop do podatkovne baze o gostoti prometa na izbranem križišču. Podatkovna baza je bila zaradi velikega števila prometa tekom dneva primerna in dovolj obsežna za analizo. Izbrano križišče je glavna povezava do nakupovalnega centra (slika 1) in je večji del dneva zelo obremenjeno. Križišče ima vgrajene senzorje, ki beležijo število vozil s časom vzorčenja 90 sekund. Z analizo intenzivnosti prometa bomo preučili njen običajni potek in poskušali z matematičnim modelom napovedati potek prometa v naslednjem časovnem intervalu.

Slika 1: Obravnavano križišče v mestu Praga, Češka (Google maps, 2014)

2.1 Sorodne analize prometa

Podobno študijo, kot bo izvedena v tem delu, vendar z drugo metodo modeliranja, zasledimo v delu (Kocijan in Prikryl, 2010). Prispevek povzema možne uporabe modelov časovnih vrst v primeru odpovedi senzorjev. V njem je predlagana metoda za računski model, ki služi kot mehki senzor (angl. soft sensor). Model je izveden na podlagi Gaussovih procesov in je verjetnostni regresijski model.

Modeliranje iz podatkov so uporabili tudi avtorji dela (Sun in drugi, 2002) za napoved povprečne hitrosti vozil. Avtorji so raziskali, kakšna je povprečna hitrost vozil na odseku ceste. Povprečna hitrost vozil jih je zanimala zato, da bi dobili natančnejši izračun časa prihoda do želenega cilja v računalniških programih za navigacijo. Tudi tu so avtorji uporabili linearni avtoregresijski (AR) model. Linearna regresija se uporablja za oceno odnosa med napovedanim prometom in zgodovino prometa ter sedanjim prometom. Tak model je bil v delu (Sun in drugi, 2002) izbran zato, ker je dal v primerjavi z drugimi metodami najboljše napovedi.

3 REGRESIJA

Za boljše razumevanje nadaljevanja magistrskega dela je v tem poglavju predstavljena metoda regresije.

Znano vprašanje eksperimentalne znanosti je, kako nek sklop spremenljivk vpliva na druge spremenljivke (Turkman, 2008). Nekateri odnosi so deterministični in jih je enostavno razložiti. Drugi pa so preveč zapleteni, da bi jih razumeli na preprost način, saj imajo še naključno komponento. V tem primeru primerjamo te odnose s preprostimi funkcijami ali naključnimi procesi z uporabo empiričnih modelov. Med vsemi metodami, ki so na voljo za razumevanje teh kompleksnih odnosov, je linearna regresija najpogosteje uporabljana metoda.

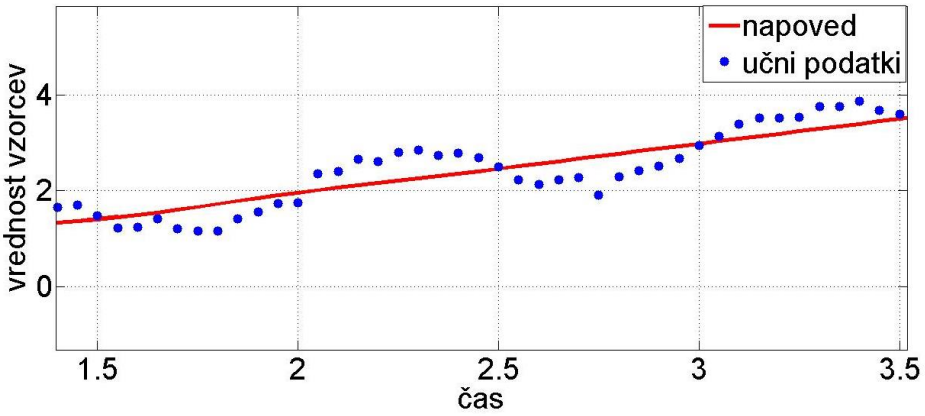
Skupna značilnost te metode je, da povzame funkcionalnost in parametrično razmerje med spremenljivkami. To je pogosto linearna regresija z neznanimi parametri, ki jih je treba oceniti iz razpoložljivih podatkov. Pri metodi ločimo dve vrsti spremenljivk: neodvisne spremenljivke in odvisne spremenljivke. V literaturi imajo te spremenljivke različna imena. Neodvisne (nadzorovane) spremenljivke so tiste, ki jim lahko določimo želeno vrednost ali imajo vrednosti, ki jih je mogoče nadzorovati brez napake. Naš cilj je ugotoviti, kako spremembe neodvisnih spremenljivk vplivajo na vrednosti odvisnih (odzivnih) spremenljivk.

Regresija torej opisuje povezanost dveh slučajnih spremenljivk in vpliv ene na drugo. Na voljo imamo naslednji dve možnosti:

- slučajni vzorec: to je zaporedje neodvisnih in enako porazdeljenih vrednosti naključnih spremenljivk. Pri tej možnosti izmerimo dve spremenljivki, nobene od njiju pa ne nadziramo. Primer je izpust emisij v ozračje in njihov vpliv na okolje;
- eksperimentalnim spremenljivkam določamo vrednosti. Eni spremenljivki določimo vrednost (neodvisna spremenljivka) in merimo izid (odvisna spremenljivka). Primer je vpliv novega zdravila na prostovoljca ali bolnika.

Prvi vtis o povezanosti spremenljivk daje njun grafični prikaz. Dvodimenzionalni diagram je prikaz parov spremenljivk v koordinatnem sistemu. Primer takšnega

diagrama prikazuje slika 2. S slike razberemo, da se vrednosti vzorcev spreminjajo periodično s počasnim linearnim naraščanjem.



Slika 2: Dvodimenzionalni diagram povezanosti spremenljivk

3.1 Linearna regresija

Pri linearni regresiji napovedujemo vrednosti ene spremenljivke glede na vrednosti druge spremenljivke. Linearna enorazsežna regresija (Kejžar, 2011) poskuša najti najustreznejšo premico skozi dane točke. Najbolj prilegajoča linija se imenuje regresijska premica. Model povezanosti med odvisno in neodvisno spremenljivko opisuje enačba (1). Z enačbo (2) pa je opisana matrična oblika linearne regresije, ki jo uporabimo, da zapišemo zvezo med več vzorci vrednosti spremenljivk.

$$E(y|x) = \alpha + \beta x \tag{1}$$

$$var(y|x) = \sigma^2 \tag{2}$$

Pri tem je:

$E(y|x)$ – pričakovana vrednost spremenljivke y v odvisnosti od spremenljivke x,

$var(y|x)$ – varianca spremenljivke y v odvisnosti od x,

σ^2 – varianca kvadrata standardnega odklona,

y – odvisna spremenljivka,

x – neodvisna spremenljivka,

α – presečišče (točka, kjer regresijska premica seka ordinato),

β – regresijski nagib (naklonski kot regresijske premice).

Model predpostavlja, da je pri neodvisni spremenljivki x odvisna spremenljivka y normalno porazdeljena s pričakovano vrednostjo (enačba (1)) in varianco (enačba (2)). Splošna formula linearne regresije je opisana z enačbo (3).

$$y = \alpha + \beta x + \varepsilon, \tag{3}$$

kjer je:

ε – napaka modela, naključna spremenljivka.

Odstopanja okrog premice so naključna. Napaka, ki se slučajno spreminja okoli premice (enačba (4)), ima normalno porazdelitev s povprečjem 0.

$$\varepsilon \sim \aleph(0, \sigma^2) \tag{4}$$

Opravka imamo torej z dvema spremenljivkama, ki imata lahko N vrednosti ($y \in \{y_1, y_2, \dots y_N\}, x \in \{x_1, x_2, \dots x_N\}$). To daje N izračunov, kar prikazuje enačba (5).

$$\left.\begin{matrix} y_1 = \alpha + \beta x_1 \\ y_2 = \alpha + \beta x_2 \\ \vdots \\ y_N = \alpha + \beta x_N \end{matrix}\right\} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \alpha + \beta \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \tag{5}$$

V matričnem zapisu to zapišemo z enačbo (6). Za matrični zapis uporabimo vektorje in matrike **y, X** in **θ**.

$$\mathbf{y} = \mathbf{X\theta}, \tag{6}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}, \qquad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \tag{7}$$

kjer je:

y_i – vrednost odvisne spremenljivke,

x_i – vrednost neodvisne spremenljivke,

$\mathbf{y}$ – vektor odvisne spremenljivke,

$\mathbf{X}$ – matrika vhodnih vrednosti,

$\boldsymbol{\theta}$ – vektor koeficientov regresijske premice.

Predpostavljamo, da je spremenljivost okrog premice povsod enaka, torej isto povprečje in ista varianca napake. To se imenuje homoscendastičnost. Če ta pogoj ni izpolnjen, je model heteroscendastičen.

3.1.1 Lastnosti regresije

V tem podpoglavju bomo opisali, na kaj vse je treba biti pozoren pri regresijski analizi, katere grafične predstavitve so priporočljive in kako merimo točnost napovedi.

- Ocenjevanje vzorčenja: sistematično vzorčenje

Iz vzorca naključno izberemo vzorce z nespremenljivim periodičnim intervalom (Systematic sampling, 2014). Ta interval se imenuje interval vzorčenja. Interval vzorčenja se izračuna tako, da se podatke deli do želene velikosti vzorca. Sistematično vzorčenje je vzorčenje z omejitvami. Iz serije podatkov vzamemo za vzorec vsak k-ti element.

- Natančnost in točnost z vidika vzorčenja

V postopkih vzorčenja sta točnost in natančnost dva različna statistična kazalca. Natančnost vzorčenja (Stamatopoulos, 2002) je navadno izražena z indeksom procentov (med 0 in 100%). Procent označuje bližino, ki temelji na vzorcu parametrov cenilke do prave vrednosti podatkov. Če število vzorcev raste in so vzorci reprezentativni, se natančnost vzorčenja povečuje. Po neki količini števila vzorcev se natančnost bistveno ne izboljšuje več. Točnost vzorčenja se nanaša na raznolikost vzorcev, ki se uporabljajo. Meri se s koeficientom variacije (relativni indeks spremenljivosti), ki računa varianco in srednjo vrednost vzorca.

- Vpliv belega šuma na model

Pri obdelavi signalov je beli šum (White noise, 2014) naključni signal s spektralno gostoto konstante moči. Beli šum se nanaša na statistične modele signalov in ne na točno določen signal. Izraz se uporablja pri diskretnih signalih, kjer se vzorci obravnavajo kot zaporedje serijsko nepovezanih naključnih spremenljivk z ničelno srednjo vrednostjo in končno varianco. Odvisno od želenega rezultata se pogosto predpostavlja, da so vzorci neodvisni in imajo enako porazdelitev verjetnosti. Če ima vsak vzorec normalno porazdelitev s srednjo vrednostjo nič, potem je signal Gaussov beli šum.

- Ocenjevanje vpliva stopnje polinoma na točnost napovedi

Aproksimacija (Curve fitting, 2014) je proces oblikovanja krivulje ali matematične funkcije, da se ta čim bolj približa nizu podatkovnih točk. Prilagajanje krivulje dosežemo z interpolacijo (natančno prilagajanje podatkov) ali glajenjem (funkcija, ki se približa podatkom). Oboje je povezano z regresijsko analizo, ki se osredotoča na statistično sklepanje. Metoda najmanjših kvadratov (Curve fitting, 2014) je ena izmed najbolj uporabljenih za določanje matematičnega modela. Obstaja več metod, da dobimo približno ali točno ujemanje funkcije modela z večanjem stopnje matematičnega modela. Na primer, prvi red modela (linearna funkcija) je omejen le na enem mestu namesto na dveh, kar da neskončno število rešitev. Tako pridemo do problema, kako primerjati in izbrati le eno rešitev. Iz tega razloga je navadno najbolje, da izberemo čim enostavnejši model. Izbrani model naj daje čim natančnejše ujemanje vseh podatkov. Navadno se točnost napovedi do določene stopnje kompleksnosti modela izboljšuje, kasneje pa vedno manj.

- Ocenjevanje vpliva redukcije časa vzorčenja na model

Zmanjševanje velikosti vzorca, ne da bi izgubili točnost, je mogoče doseči na dva načina (Chin in Lee, 1996). Prvi je, da izboljšamo razmerje signal – šum (spremenljivka ε). To dosežemo tako, da zmanjšamo šum (ε), okrepimo signal ali zmanjšamo spremenljivost (zmanjša hrup in okrepi signal). Druga možnost je uporaba drugih statističnih tehnik, ki izluščijo več informacij iz podatkov. Primer take metode je večkratna uporaba istih podatkov.

- Grafična predstavitev obnašanja modela

Grafične predstavitve združujejo teorije grafov in teorije verjetnosti. Te predstavitve zarišejo splošni okvir, ki predstavlja modele in interakcijo med spremenljivkami. Poznamo več različnih vrst grafov, ki uporabniku posredujejo informacije o ustreznosti različnih vidikov modela. Za pregled funkcionalnosti modela uporabimo na primer graf raztresenih ostankov v primerjavi s predvidenimi. Histogram daje vpogled v porazdelitev napak napovedi modela. Iz histograma vidimo, ali je model točen in kje se nahaja največ napak. Želena oblika histograma je normalna porazdelitev oziroma tako imenovana Gaussova porazdelitev. Če porazdelitev ni normalna, osnovne predpostavke regresijskega modela niso izpolnjene, rezultati pa posledično pristranski.

- Pregled razlik med izmerjenimi in napovedanimi vrednostmi

Točnost napovedi (Mean squared error, 2014) pogosto vrednotimo s funkcijo SMSE (Standardised Mean Square Error). V statistiki je srednja vrednost kvadrata napake s cenilko eden od mnogih načinov, ki meri razliko med vrednostmi cenilke in dejansko vrednostjo vzorca, ki se ga ocenjuje. SMSE je funkcija, ki ustreza pričakovani vrednosti kvadratov napake. Funkcija prikazuje povprečje kvadratov napak. Funkcija SMSE (enačba (8)) torej izračuna povprečno vrednost razlik meritve in napovedi in jih deli z varianco meritev. Bližje kot je vrednost funkcije SMSE nič, boljša je napoved. Bližje kot je vrednost SMSE ena, slabša je napoved. Za dobro napoved naj bo vsota kvadratov razlike čim manjša.

$$SMSE = \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} y_i^2} \tag{8}$$

Pri tem je:

y_i – meritev v i-tem koraku,

$\hat{y}_i$ – napovedana vrednost v i-tem koraku,

N – red modela.

3.2 Polinomska regresija

Polinomska regresija je oblika večrazsežne regresije. Pri tej vrsti regresije imamo eno izhodno spremenljivko in več vhodnih spremenljivk, ki jih imenujemo regresorji. Splošno enačbo polinomske regresije, ki je ena izmed oblik večrazsežne regresije, prikazuje enačba (9).

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \varepsilon, \tag{9}$$

kjer je:

y – napovedana vrednost,

a_i – koeficient regresije,

x – merjene vrednosti,

n – potenca modela (red modela),

ε – napaka modela, naključna spremenljivka.

Spremenljivki y in x naj imata N različnih vrednosti (enačba (10)).

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = a_0 + a_1 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} + \cdots + a_N \begin{bmatrix} x_1^n \\ x_2^n \\ \vdots \\ x_N^n \end{bmatrix} \tag{10}$$

To v vektorski obliki zapišemo z enačbo (11).

$$\mathbf{y} = \mathbf{X\theta}, \tag{11}$$

kjer so:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^n \end{bmatrix}, \qquad \boldsymbol{\theta} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{bmatrix}, \tag{12}$$

kjer je:

y_i – vrednosti spremenljivke y,

x_i – vrednosti spremenljivke x,

$\mathbf{x} = [1, x, x^2, x^3, \ldots x^n]$ – regresorski vektor,

a_0 – regresorski koeficienti,

$\mathbf{y}$ – vektor odvisne spremenljivke,

$\mathbf{X}$ – matrika regresorskih vektorjev,

$\boldsymbol{\theta}$ – vektor koeficientov regresije.

Enačba (13) je prirejena za matrično računanje polinomske regresije. Vrednosti regresijskih koeficientov $\boldsymbol{\theta}$ se po metodi najmanjših kvadratov izračunajo z enačbo (14).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}, \tag{13}$$

$$\widehat{\boldsymbol{\theta}} = [\mathbf{X}^{\mathbf{T}} \quad \mathbf{X}]^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}} \quad \mathbf{y}, \tag{14}$$

kjer je:

$\mathbf{X}^{\mathbf{T}}$ – transponirana (zasukana) matrika regresorskih vektorjev,

$\widehat{\boldsymbol{\theta}}$ – izračunana matrika koeficientov regresije.

Podrobnosti in izpeljave metode najmanjših kvadratov najdemo v (Korenjak, 2010).

Ko pridobimo podatke za modeliranje, jih razdelimo na učne podatke in testne podatke. Učni podatki so tisti podatki, na katerih se matematični model uči. Testni podatki pa so podatki, na katerih se matematični model ne uči in služijo zgolj za preverjanje točnosti napovedi. Testni podatki so torej neodvisni podatki, ločeni od matematičnega modela, namenjeni preverjanju napovedi. Pogosto je množica učnih podatkov večja od množice testnih podatkov. Z delitvijo podatkov lahko dosežemo, da je model čim bolj neodvisen od uporabljenih podatkov.

Ko podatke analiziramo z regresijsko analizo, najprej pregledamo njihov odnos. Včasih ugotovimo, da je odnos med odvisno in neodvisno spremenljivko opisan s krivuljo in ne linearno. V tem primeru enorazsežna linearna regresija ne bo najboljša

izbira za napovedovanje in opisovanje odnosa med spremenljivkami, prav tako ne bi bile pravilne hipoteze. V tem primeru se lahko uporabi polinomsko regresijo. Pri tej regresijski metodi sta izbira reda modela in ocena ujemanja odvisni od presoje uporabnika. Pogosto se uporablja za vrednotenje točnosti funkcijo SMSE. Na modelu preizkusimo različne stopnje polinoma in primerjamo njihove vrednosti SMSE. Izberemo tisto stopnjo, pri kateri je še viden velik skok med razliko v vrednosti SMSE (McDoland, 2009).

Primer:

Imamo 50 učnih podatkov na intervalu od 0 do 1. Odvisna spremenljivka je definirana z enačbo (15).

$$y = \sin(2\pi x) + \varepsilon, \tag{15}$$

kjer je: $\varepsilon \sim \aleph(0, \sigma_\varepsilon^2)$.

Iz vhodnih podatkov sinusne funkcije s šumom (napako) smo predpostavili matematični model, ki ima obliko polinoma (enačba (16)).

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \varepsilon \tag{16}$$

Za vsako spremenljivko smo vzeli 50 vzorcev, kar lahko zapišemo z enačbo (17).

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{50} \end{bmatrix} = a_1 \begin{bmatrix} x_1 \\ \vdots \\ x_{50} \end{bmatrix} + a_2 \begin{bmatrix} x_1^2 \\ \vdots \\ x_{50}^2 \end{bmatrix} + \cdots + a_n \begin{bmatrix} x_1^n \\ \vdots \\ x_{50}^n \end{bmatrix} \tag{17}$$

Izračunati želimo koeficiente regresije (spremenljivke a_i) oziroma vektor regresijskih koeficientov $\boldsymbol{\theta}$, ki zajema vse koeficiente regresije. Za lažje razumevanje si še enkrat oglejmo enačbo (18) za 50 vzorcev spremenljivk. Vsaka spremenljivka ima svoj vektor (enačba (19)).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} \tag{18}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{50} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^n \\ 1 & x_2 & x_2^2 & x_2^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{50} & x_{50}^2 & x_{50}^n \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \tag{19}$$

Koeficiente regresije izračunamo po enačbi (20).

$$\widehat{\boldsymbol{\theta}} = [\mathbf{X}^\mathbf{T} \quad \mathbf{X}]^{-\mathbf{1}}\mathbf{X}^\mathbf{T} \quad \mathbf{y} \tag{20}$$

Za matematični model smo v našem primeru predpostavili model s polinomom 6. stopnje (končna enačba ima 7 koeficientov regresije). Zanimalo nas je, kako dobro funkcija napove odvisno spremenljivko (y) znotraj meje učnih podatkov na intervalu od 0 do 1 ter če interval razširimo še na območje od –0,2 do 0 in od 1 do 1,2. Tvorili smo vrednosti testnih podatkov za vrednosti neodvisne spremenljivke x na intervalu med –0,2 in 1,2 z vzorčenjem 0,01, učni podatki pa so v intervalu od 0 do 1 z vzorčenjem 0,02. Izračunali smo tudi pas zaupanja.

V verjetnostni statistiki kovarianca (Covariance, 2014) opisuje, kako se dve naključni spremenljivki spreminjata. Če se vrednosti obeh spremenljivk odzivajo podobno (višajo oziroma nižajo), je kovarianca pozitivna. V nasprotnem primeru (ko se spremenljivke odzivajo nesorazmerno) je kovarianca negativna. Kovarianco smo izračunali tako, da smo najprej izračunali kovarianco koeficientov regresije z enačbama (21) in (22).

$$cov\{a_i\} = \widehat{\sigma_\varepsilon}[\mathbf{X^T X}]^{-1} \tag{21}$$

$$\widehat{\sigma_\varepsilon} = \frac{1}{N-(n+1)}(y - y_0)^T(y - y_0) \tag{22}$$

Vrednosti varianc parametrov so diagonalne vrednosti matrike $cov\{a_i\}$.

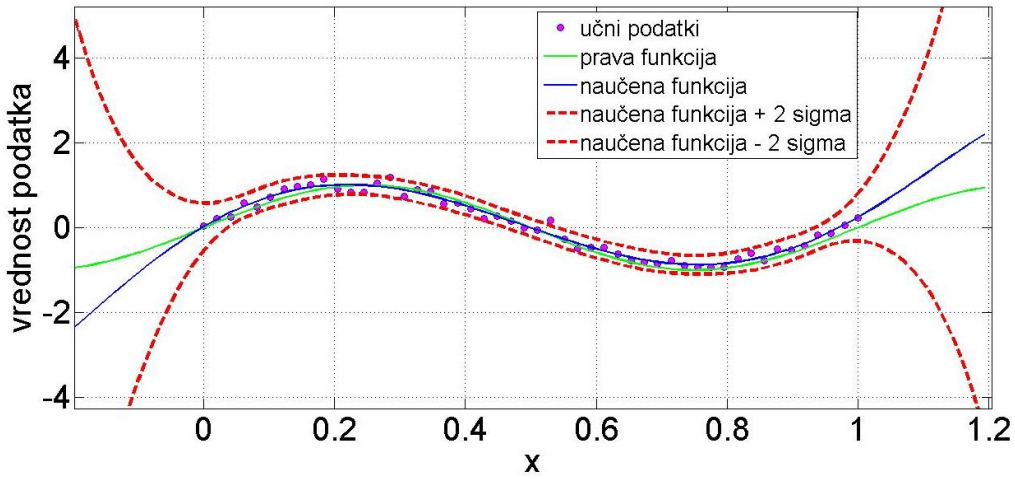
Rezultat iz enačbe (21) smo vstavili v enačbo za izračun kovariance odvisne spremenljivke za testno vrednost opisane z regresijskim vektorjem $\mathbf{x_0}$ z enačbo (23).

$$var\{y_0\} = \mathbf{x_0} cov\{a_i\} \mathbf{x_0^T} \tag{23}$$

Vrednost σ smo izračunali z enačbo (24).

$$\sigma = \sqrt{var\{y_0\}} \tag{24}$$

Nato smo napovedali odvisne spremenljivke (torej tiste, ki smo jih dobili, ko smo v model vstavljali vrednost $\mathbf{X_0}$) in jim prišteli in odšteli vrednost 2σ. Z grafa razberemo, da je pas zaupanja tam, kjer so učni podatki, zelo ozek, drugje pa se izrazito razširi. Pri vrednostih $\pm 2\sigma$ je verjetnost, da bodo podatki v pasu, približno 95%. S slike razberemo, da sta le 2 točki, pri katerih sta vrednosti izven pasa zaupanja. Zelena linija na sliki 3 predstavlja originalno funkcijo brez šuma (ε), modra linija pa predstavlja naučeno funkcijo, ki je motena s šumom.



Slika 3: Preizkus učenja modela na sinusnem signalu

3.3 Časovne vrste

Časovne vrste so zaporedje numeričnih podatkovnih točk v časovno zaporednem vrstnem redu (Time series, 2014), ki se navadno pojavljajo v enakih intervalih. Časovne vrste so torej zaporedje števil, zbranih v rednih časovnih presledkih v določenih časovnih obdobjih.

Časovne vrste so pogosto ponazorjene z dvodimenzionalnimi grafi. Analiza časovnih vrst obsega metode za analizo časovnih vrst za pridobivanje pomembnih statističnih in drugih značilnosti podatkov. Napovedovanje prihodnosti na podlagi preteklih vrednosti je primer uporabe modela časovnih vrst. Model časovne vrste lahko dobimo z regresijsko analizo.

Podatki v časovni vrsti imajo podatke v časovni razvrstitvi. Časovna razvrstitev pomeni, da so podatki razvrščeni po točno določenem vrstnem redu, ki ga predpisuje čas. Vrstni red podatkov ima velik vpliv na kasnejšo uporabo podatkov in njihov pomen. Pri strojnem učenju običajnih regresijskih modelov je vsak podatek le eden izmed podatkov, ki se ga mora računalnik naučiti. Zaporedje teh podatkov v običajni podatkovni seriji ni pomembno. To razlikuje analizo časovnih vrst od drugih skupin analize podatkov, kjer ni naravne časovne porazdelitve pri opazovanju modela. Analiza časovnih vrst se razlikuje tudi od analize prostorskih podatkov, kjer se ugotovitve navadno nanašajo na prostorske lokacije. Stohastičen model časovne vrste

na splošno odraža dejstvo, da so opazovanja časovno blizu gledana tesneje povezana kot opazovanja daleč narazen.

Na spreminjanje vrednosti podatkov v časovni vrsti vplivajo številni dejavniki (Pfajfer in Arg, 1998). V nadaljevanju naštevamo tipične elemente časovnih vrst, s katerimi opisujemo omenjene dejavnike:

- trend: dejavniki, ki na model vplivajo na daljše časovno obdobje. Primer: stalno naraščanje ali upadanje vrednosti podatkov;
- ciklične spremembe: dolgoročni dejavniki, ki nimajo stalne periode in se po določenem času ponovijo;
- periodične spremembe: dejavniki, ki se pojavljajo v enakih časovnih razmikih. Primer: dnevne, tedenske, mesečne, letne periode;
- iregularne spremembe: vsebujejo posamične vplive, ki imajo ponavljajočo se naravo, in slučajne vplive, ki so različno izraziti tekom časa.

3.3.1 Regresija v kontekstu časovnih vrst

Iz podatkov časovnih vrst so podatki meritev razporejeni glede na čas, zato moramo spoštovati njihov vrstni red (Baum, 2012). To je treba še posebej upoštevati, če ima model dinamično obnašanje.

Večrazsežna regresija časovnih vrst se uporablja takrat, ko želimo točnost napovedi povečati. Točnost povečamo z dodatnimi informacijami. Večrazsežno regresijo uporabimo, ko nastopa v medsebojni odvisnosti več spremenljivk (enačba (25)). V modelu navadno ne uporabimo neposredno izražene časovne komponente. Čas, navadno označen s spremenljivko t, namreč nadomestimo z vzorčnimi vrednostmi kT, kjer sta k zaporedno število vzorca in T čas vzorčenja. Ker je T ves čas enak, ga lahko iz zapisa izpustimo. V enačbi tako uporabimo zakasnjene vrednosti podatkov. Vrednost pri $(k-1)$ pomeni, da vzamemo pri času k en zakasnjen vzorec iz serije podatkov. Regresijski problem nastavimo tako, da je odvisna spremenljivka $y(k)$ za čas $k = 1, \dots n$ pod vplivom neodvisnih spremenljivk $y(k-1), y(k-2), \dots, y(k-n)$, ki so znane (Kejžar, 2011). Regresorji so torej zakasnjene vrednosti spremenljivke y oziroma vrednosti spremenljivke v prejšnjih časovnih intervalih. Konstante a_i so neznani regresijski koeficienti. Regresijski koeficienti so realna

števila, katerih ocene določimo iz vzorčnih podatkov. Vrednost ε je napaka oziroma šum, za katerega predpostavimo normalno porazdelitev s srednjo vrednostjo 0 in nespremenljivo varianco. Z ε ponazorimo odstopanja med modelom in merjenimi vrednostmi.

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + \cdots + a_n y(k-n) + \varepsilon(k), \quad (25)$$

kjer je:

$y(k)$ – napovedana vrednost,

a_i – koeficient regresije,

$y(k-1), \ldots, y(k-n)$ – merjene vrednosti,

n – red modela,

k – zaporedno število vzorca.

4 METODE EKSPERIMENTALNEGA MODELIRANJA ČASOVNIH VRST

Matematični modeli so temelj analize in načrtovanja večine naravoslovnih in tehničnih metod. Modele pridobimo s teoretičnim pristopom, ki temelji na fizikalnih zakonih, ali z eksperimentalnim pristopom, ki temelji na meritvah, dobljenih iz sistema. Identifikacija sistemov se ukvarja s problematiko gradnje matematičnih modelov za dinamične sisteme, ki jih dobimo iz izmerjenih podatkov sistema. Tako dobljene modele se uporablja za simulacijo, napovedovanje ali za načrtovanje digitalnih sistemov vodenja.

Postopek eksperimentalnega modeliranja (Knudsen, 2004) se prične z določanjem strukture modela. Struktura modela se lahko določi iz osnovnih fizikalnih zakonov ali empiričnih ugotovitev. Sledi izbira eksperimentalnega vzorca. Na tej stopnji je pomembno, da imamo ustrezne vhodne signale. Izberemo jih na podlagi predznanja o sistemu, ki ga modeliramo, ali glede na to, kako je signal oblikovan (optimizirane karakteristike podatkov). Zadnji korak je preizkus modela.

4.1 Modeli s pogojeno srednjo vrednostjo

V nadaljevanju dela so opisani eksperimentalni modeli, ki imajo pogojeno srednjo vrednost (Guangyi, 2011). Opisali bomo štiri najpogosteje uporabljene modele časovnih vrst, ki so pogojene s srednjo vrednostjo: avtoregresijski model, model drsečega povprečja, avtoregresijski model drsečih povprečij in avtoregresijski zbirni model drsečih povprečij.

4.1.1 Avtoregresijski model

V statistiki in obdelavi signalov avtoregresijski model (angl. Autoregressive Model – AR) predstavlja tako naključne kot tudi nenaključne procese. Opisuje lahko različne časovne procese v naravi, gospodarstvu ali na drugih področjih. Sedanja vrednost modela je odvisna od preteklih vrednosti iste časovne vrste. Model AR določa, da je izhodna spremenljivka linearno odvisna od svoje prejšnje vrednosti (Autoregressive model, 2014).

Model AR reda n ali okrajšano AR(n) predstavlja enačba (26).

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + \cdots + a_n y(k-n) + \varepsilon, \qquad (26)$$

kjer je:

$y(k)$ – napovedana vrednost,

a_i – koeficient avtoregresije,

$y(k-1), \ldots, y(k-n)$ – merjene vrednosti,

n – red modela,

ε – napaka modela, naključna spremenljivka,

k – zaporedno število vzorca.

4.1.2 Model drsečega povprečja

Model drsečega povprečja (angl. Moving Average Model – MA) (Shumway in Stoffer, 2010) reda q ali okrajšano MA(q) je definiran z enačbo (27). Predpostavljamo, da je ε Gaussov beli šum s povprečjem 0 in varianco σ_ε^2.

$$y(k) = \varepsilon(k) + b_1\varepsilon(k-1) + b_2\varepsilon(k-2) + \cdots + b_q\varepsilon(k-q), \qquad (27)$$

kjer je:

$y(k)$ – napovedana vrednost,

b_q – koeficient drsečega povprečja, ($b_q \neq 0$),

q – največja zakasnitev,

ε – napaka modela, naključna spremenljivka,

k – zaporedno število vzorca.

Model je enak kot neskončno drseče povprečje, definirano kot linearni proces.

4.1.3 Avtoregresijski model drsečih povprečij

Časovno zaporedje ($y(k);\ k = 0, 1, 2, \ldots$) ponazorimo kot model ARMA (n, q) (angl. Autoregressive moving average model – ARMA) z enačbo (28), kjer je $a_i \neq$

$0, b_q \neq 0$ in $\sigma_\varepsilon^2 > 0$. Parameter n je red modela AR, parameter q pa red modela MA.

$$y(k) = a_1 y(k-1) + \cdots + a_n y(k-n) + b(k) + b_1 \varepsilon(k-1) + \cdots + b_q \varepsilon(k-q) \quad (28)$$

Če ima $y(k)$ konstantno srednjo vrednost, jo zapišemo z enačbo (29)

$$\alpha = \mu(1 - a_1 - \cdots - a_n);\ \mu \neq 0 \quad (29)$$

in enačbo (28) razširimo v:

$$y(k) = \alpha + a_1 y(k-1) + \cdots + a_n y(k-n) + \varepsilon(k) + b_1 \varepsilon(k-1) + \cdots + b_q \varepsilon(k-q), \quad (30)$$

kjer je:

α – srednja vrednost.

Predpostavljamo, da je ε Gaussov beli šum s srednjo vrednostjo 0 in varianco σ_ε^2. Ko je $q = 0$, se model imenuje AR(n) z redom n, ko je $n = 0$, se model imenuje MA(q) z redom q (Shumway in Stoffer, 2010).

4.1.4 Avtoregresijski zbirni model drsečih povprečij

Modele ARIMA (angl. Autoregressive Integrated Moving Average – ARIMA) sta prva uporabila George Box in Gwilyn Jenkins (Zobovnik, 2011). Razviti so bili zato, ker so imeli statistiki probleme z napovedovanjem časovnih vrst, ki so vsebovale določen trend. Bistvo modela ARIMA je, da odpravlja pomanjkljivosti večrazsežnostne regresije, ki predvideva, da je časovna vrsta stacionarna. To pomeni, da mora časovna vrsta imeti konstantno aritmetično sredino in varianco. Nepravilnemu delovanju statističnega modela se izognemo na ta način, da nestacionarno vrsto spremenimo v stacionarno z različnimi stopnjami avtoregresije, drsečega povprečja in diferenciranja podatkov.

Pri analizi časovnih vzorcev je ARIMA posplošitev modela ARMA. Model ARIMA je torej sestavljen iz avtoregresijskega modela (AR), drsečega povprečja (MA) in

diferenciranja podatkov. Vsak izmed delov modela ARIMA se na svoj način odziva na napake oziroma šume.

Model ARIMA zapišemo kot ARIMA (n, d, q), kjer je:

n – red avtoregresije,

d – kolikokrat smo diferencirali časovno vrsto pred analizo (transformacija),

q – red drsečega povprečja.

Model ARIMA je torej sestavljen iz treh delov (Hyndman, 2001), ki so:

- avtoregresijski del ali model AR (podrobnejši opis v podpoglavju 4.1), opisan z enačbo (31):

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + \cdots + a_n y(k-n); \quad (31)$$

- model drsečega povprečja ali model MA (podrobnejši opis v podpoglavju 4.2), opisan z enačbo (32):

$$y(k) = \varepsilon(k) + b_1 \varepsilon(k-1) + b_2 \varepsilon(k-2) + \cdots + b_q \varepsilon(k-q); \quad (32)$$

- stopnja diferenciranja.

  Stopnja diferenciranja v modelu ARIMA določi, ali bodo vrednosti modelirane neposredno ali se bo modeliralo razlike zaporednih podatkov. Če je $d = 0$, se opazovanja modelira neposredno. Če je $d = 1$, se modelira razlike podatkov. V praksi je d redko večji od 2.

Te tri dele modela ARIMA (ARIMA model, 2014) lahko združimo in zapišemo z enačbo (33).

$$L^d y(k) = a_1 y(k-1) + \cdots + a_n y(k-n) + \varepsilon(k) + b_1 \varepsilon(k-1) + \cdots + b_q \varepsilon(k-q), \quad (33)$$

kjer je:

L^d – operator diferenciranja stopnje d, kjer je $Ly(k) = y(k) - y(k-1)$,

$y(k)$ – napovedana vrednost,

d – kolikokrat je diferencirana časovna vrsta,

a_i – koeficient avtoregresije,

b_i – koeficient drsečega povprečja.

Primer uporabe:

Cene stanovanj v Hong Kongu (Raymond, 2013) preučujejo že mnoga leta. Ugotovili so, da je tržna cena v vsakem trenutku vedno povezana z vzorcem preteklih cenovnih gibanj nepremičnin. Zato za napovedovanje cen uporabljajo empirični model ARIMA. Rezultati nakazujejo, da se cene ciklično spreminjajo. Napovedovanje lahko zagotovi kratkoročno usmeritev pri nakupu nepremičnin, saj napove, ali bodo nihanja v ceni velika ali majhna. To je za investitorje zelo pomemben podatek.

4.2 Modeli, pogojeni z varianco

V nadaljevanju dela so predstavljeni eksperimentalni modeli družine avtoregresijskih modelov s pogojeno heteroscendastičnostjo (angl. AutoRegressive Conditional Heteroskedasticity – ARCH), ki imajo pogojeno varianco (Conditional variance, 2014). V statistiki je pogojena varianca verjetnostne porazdelitve. Pogojna odstopanja so pomemben del modelov ARCH.

Modele ARCH (GARCH, 2014) se uporablja za opazovanje časovnih vrst v ekonometriji. Uporabljamo jih, kadar imamo domnevo ali potrditev, da bo katerakoli točka v časovnem vzorcu imela značilno varianco ali velikosti napake. Takšni modeli se imenujejo modeli ARCH, čeprav se za strukture modela, ki imajo podobno osnovo, uporablja tudi druge kratice. Modeli ARCH se pogosto uporabljajo pri modeliranju finančnih časovnih vrst (obdobja velikih nihanj in obdobja relativnega mirovanja).

Model ARCH se lahko oceni z uporabo metode najmanjših kvadratov. V nadaljevanju bo opisan postopek za modeliranje zakasnjenih dolžin ARCH residualov (razlik med napovedmi in meritvami).

Spremenljivka $\varepsilon(k)$ označuje napako (ostanke ali residuale). Ti ostanki so razdeljeni na stohastičen del, opisan z belim šumom $z(k)$, in časovno spremenljivo standardno deviacijo $\sigma(k)$, kar prikazuje enačba (34).

$$\varepsilon(k) = \sigma(k)z(k) \tag{34}$$

Spremenljivka ε^2 je beli šum. Spremenljivko $\sigma^2(k)$ dobimo po enačbi (35), ob upoštevanju, da je $b_0 > 0$ in $b_i \geq 0$ ter $i > 0$.

$$\sigma^2(k) = b_0 + b_1\varepsilon^2(k-1) + b_2\varepsilon^2(k-2) + \cdots + b_q\varepsilon^2(k-q), \tag{35}$$

kjer je:

$\sigma^2(k)$ – časovno odvisen kvadrat standardnega odklona,

$\hat{\varepsilon}^2(k-i)$ – kvadrat napake.

V postopku modeliranja iz podatkov naprej določimo model AR z enačbo (36).

$$y(k) = a_0 + a_1y(k-1) + a_2y(k-2) + \cdots + a_ny(k-n) + \varepsilon(k) \tag{36}$$

Naslednji korak je izračun kvadratov napake $\hat{\varepsilon}^2$, kjer je q dolžina ARCH zakasnitev (enačba (37)).

$$\hat{\varepsilon}^2(k-i) = \hat{b}_0 + \hat{b}_1\varepsilon^2(k-1) + \hat{b}_2\varepsilon^2(k-2) + \cdots + \hat{b}_q\varepsilon^2(k-q), \tag{37}$$

kjer je:

$\hat{\varepsilon}^2(k-i)$ – ocenjeni kvadrat napake,

$\hat{b}_i$ – ocenjeni koeficient drsečega povprečja.

Podobno osnovo, kot jo ima model ARCH, imajo tudi nekateri drugi modeli, ki so opisani v nadaljevanju. Vsak izmed teh modelov je posebna oblika modela ARCH in se ga uporabi za specifične namene modeliranja.

- Posplošena avtoregresivna pogojna heteroscendastičnost (angl. Generalized AutoRegressive Conditional Heteroskedasticity – GARCH)

Model GARCH (posplošeni model ARCH) je analogen modelu ARMA, tako kot je ARCH analogen modelu AR (Tsay, 2002). Od modela ARCH se razlikuje po tem, da so v model vključene tudi zakasnjene vrednosti pogojene variance same.

- Eksponentna posplošena avtoregresivna pogojna heteroscendastičnost (angl. Exponental Generalized AutoRegressive Conditional Heteroskedasticity – E-GARCH)

E-GARCH (Sivec, 2009) je eksponentni model GARCH. Te vrste model modelira logaritem pogojene variance. Poleg modeliranja logaritma ima E-GARCH tudi možnost zajemanja asimetrij variance. Modeli ARCH ne ločijo pozitivnih in negativnih vrednosti slučajnih napak, model E-GARCH pa je sposoben razlikovati med negativnimi in pozitivnimi napakami.

- GJR (model Glosten, Jagannathan in Runkle)

Model GJR (GJR models, 2014) je inačica modela GARCH, ki dopušča odprte možnosti za modeliranje asimetrične nestanovitnosti podatkov. Povezava med modelom GJR in modelom GARCH je velika. Pri oblikovanju modela GJR bodo velike negativne spremembe bolj modelirane kot pozitivne spremembe.

4.3 Napovedovanje

Eden izmed namenov modeliranja je napovedovanje. Bodočnost napovedujemo zato, da se pravočasno in pravilno odločamo, kadar nismo prepričani, kakšna bo prihodnost. Pri sprejemanju strateških odločitev smo v negotovosti in na nek način z odločanjem delamo napovedi. Mogoče se tega ne zavedamo, vendar so naše odločitve pogojene glede na naše pričakovanje. Za dobro napovedovanje je potrebna dobra presoja prenesena na prepoznavo ustreznih informacij in njihovo uporabo za napovedovanje procesa.

"Napovedovanje je zahtevno, posebej če se nanaša na prihodnost." (Niels Bohr, Nobelov nagrajenec za fiziko) (Courvalin, 2005, str. 1503). Ta citat služi kot

opozorilo na pomen vrednotenja modela za napovedovanje iz vzorcev. Običajno je preprosto poiskati model, ki ustreza danim podatkom. Povsem druga stvar je poiskati model, ki pravilno napoveduje iz preteklih podatkov in bo še naprej pravilno napovedoval prihodnost. (Course Outline, 2012)

Najpogosteje se napovedi uporablja v statistiki, obdelavi signalov, avtomatskem vodenju, razpoznavanju vzorcev, ekonometriji, matematičnih financah, pri vremenskih napovedih, napovedovanju potresov, elektroencefalografiji, astronomiji in telekomunikacijah. (Time series, 2014)

Proces napovedovanja se prične z zbiranjem ustreznih podatkov za analizo želenega modela, sledi preučevanje vzorcev podatkov (trend, ciklične in periodične spremembe). Ko vemo, kako se model obnaša, sledi izbira metode za napovedovanje oziroma izbira strukture modela. Delovanje modela nato preverimo na podlagi preteklih vrednosti. V modelu vrednotimo napake modela s preučevanjem modela. Nato v model vključimo naključne oziroma majhne napake. Točnost in natančnost napovedi redno preverjamo s preteklimi (testnimi) podatki, ki jih nismo uporabili za učenje modela.

Poznamo dve vrsti napovedovanja (Johnson, 2013).

- Kvantitativno napovedovanje: Uporablja se ga predvsem takrat, ko imamo množico števil. Primeri: napovedi vremena, ekonometrija, proračunske napovedi in druge. Podatki običajno izvirajo iz zgodovine statističnih podatkov agencij, podjetij in drugih obsežnih zbiralcev informacij. Pri kvantitativnem napovedovanju sta najbolj uporabljeni dve metodi. Prva je tehnika časovnih vrst, s katero predvidevamo prihodnost glede na pretekle podatke o spremenljivki, ki jo napovedujemo. Druga je relacijska metoda, ki deluje po principu, da je prihodnost odvisna od več različnih dejavnikov. Ti dejavniki lahko vplivajo tudi na trenutno napoved.
- Kvalitativno napovedovanje: Napovedovanje temelji na mnenjih in presojah in se navadno ne zanaša na zgodovino. Uporablja se za določitev prednosti, slabosti, priložnosti in nevarnosti, kar imenujemo tudi SWOT (angl. Strenght, Wikness, Opportunity, Threat) analiza (Lawrence, 2009) in napovedovanje uspešnosti. Ena izmed priljubljenih metod je metoda Delphi (Hsu in

Sandford, 2007). Pri metodi Delphi napovedujemo trende in posledice odločitev. Ta metoda ne zahteva strokovnega znanja. Rezultati so večinoma pridobljeni preko anket, za katere anketirancem plačujejo. Rezultati so kasneje analizirani glede na povprečno mnenje anketirancev.

Razliko med kvantitativnim in kvalitativnim napovedovanjem lahko povzamemo v naslednjih stavkih. Kvantitativno napovedovanje temelji na matematičnih modelih. Za napovedovanje uporablja pretekle trende. Kvalitativna analiza temelji na miselnih modelih. Velikokrat se jo uporablja tudi takrat, ko nimamo dovolj podatkov, da bi lahko naredili kvantitativno analizo. (Johnson, 2013)

5 MODELIRANJE PROMETA V KRIŽIŠČU

Matematične modele bomo izdelali zato, da bi rešili problem v primeru odpovedi senzorjev za meritve prometa v izbranem križišču v Pragi. Spremenljivka, ki jo napovedujemo, je podatek o intenzivnosti prometa (število vozil/90 s) v križišču ob določenem času. Neodvisna spremenljivka je čas. Meritve so beležene s časom vzorčenja 90 s. Kot je že opisano v drugem poglavju, se v primeru odpovedi enega izmed senzorjev za napovedovanje uporabi matematični model.

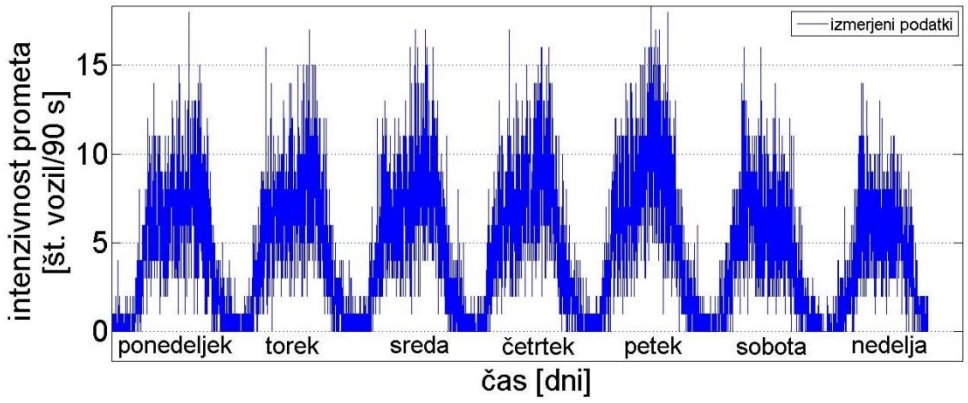
Kateri model izberemo, je odvisno od predpostavk o tem, ali imajo podatki stalen trend. Analizirali bomo ponedeljke, torke, srede, četrtke, petke, praznike in vikende (sobota, nedelja). Modelirali bomo dni, pri katerih smo opazili podobne vrednosti signalov in ponovitve obnašanja modela. V tem primeru se osredotočamo na modele, pri katerih je napoved $\hat{y}$ odvisna od preteklih vrednosti y. V večini primerov se za najboljši model izkaže model, ki uporablja bodisi samo AR obliko bodisi samo MA obliko. Zaradi omejenega obsega dela bomo obdelali le ta dva modela. Model MA, ki ga bomo uporabljali, je posebne oblike in ga bomo uporabili v kombinaciji z modelom AR.

5.1 Kombinacija modela MA in modela AR

Za model MA smo se odločili zato, ker v našem primeru omogoča hiter in relativno enostaven vpogled v obnašanje modela. Uporaba drsečega povprečja pomeni, da v vsakem določenem trenutku ugotovimo povprečje opazovanih vrednosti (Smooting, 2014).

Glajenje signala pomaga pri tem, da bolje vidimo vzorce obnašanja in trende v časovnih vrstah. Glajenje omogoča, da bo model bolje napovedoval trend.

Najprej smo pregledali obnašanje intenzivnosti prometa po posameznih dneh (slika 4). Pri nekaterih dneh smo tekom tednov opazili večja nihanja, pri nekaterih pa manjša. Za prikaz metode smo izbrali četrtke, saj se ob teh dneh promet spreminja najbolj enakomerno.



Slika 4: Intenzivnost prometa v enem tednu

Uporabili smo poseben model MA, opisan z enačbo (38):

$$y(k) = b_1 y(k-T) + b_2 y(k-2T) + b_3 y(k-3T), \tag{38}$$

kjer je:

T – perioda, ki je enaka 1 tednu,
b_i – koeficient drsečega povprečja,
$y(k)$ – vrednost intenzivnosti prometa,
k – zaporedno število vzorca.

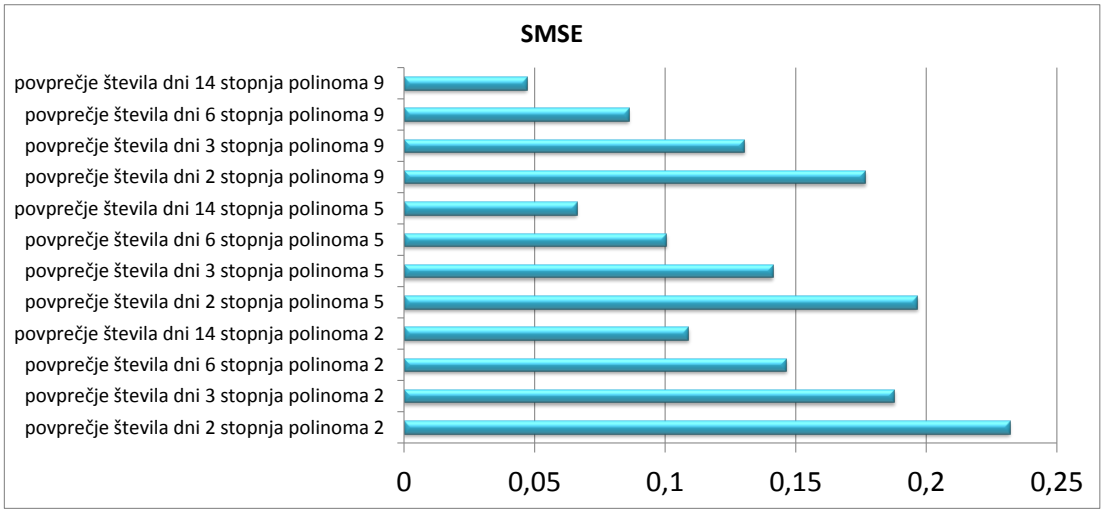
Od navadnega modela MA (enačba (27)) se razlikuje po tem, da v enačbah ne uporabljamo napake modela oziroma šuma (ε), ampak merjene vrednosti intenzivnosti prometa ob določenem času in njihovo povprečje (enačba (39)). Izbrali smo povprečje treh istovrstnih dni iz treh zaporednih tednov, ker na ta način signala ne pogladimo preveč in so vrednosti še zmeraj točne in ne preveč popačene.

$$y(k) = \frac{1}{3} y(k-T) + \frac{1}{3} y(k-2T) + \frac{1}{3} y(k-3T) \tag{39}$$

Zajemanje vzorcev poteka na vsakih 90 s, torej je skupno število vzorcev v 24 urah 960. Za izbrane tri dni smo odčitane vzorce ob istih časovnih odčitkih sešteli in delili s tri (enačba (40)). Tako smo dobili povprečje treh dni v treh zaporednih tednih ob določenem času.

$$b_1 = b_2 = b_3 = \frac{1}{3} \tag{40}$$

Na ta način smo pridobili podatkovno bazo, ki vsebuje glajene vrednosti intenzivnosti prometa za različne dni (isti dan v tednu iz več zaporednih tednov). Nato smo te glajene vrednosti zmodelirali s polinomsko regresijo, kjer je bil čas neodvisna spremenljivka. Preizkusili smo različne stopnje polinomov in s tem rede modelov, kar prikazujeta slika 5 in tabela 1.



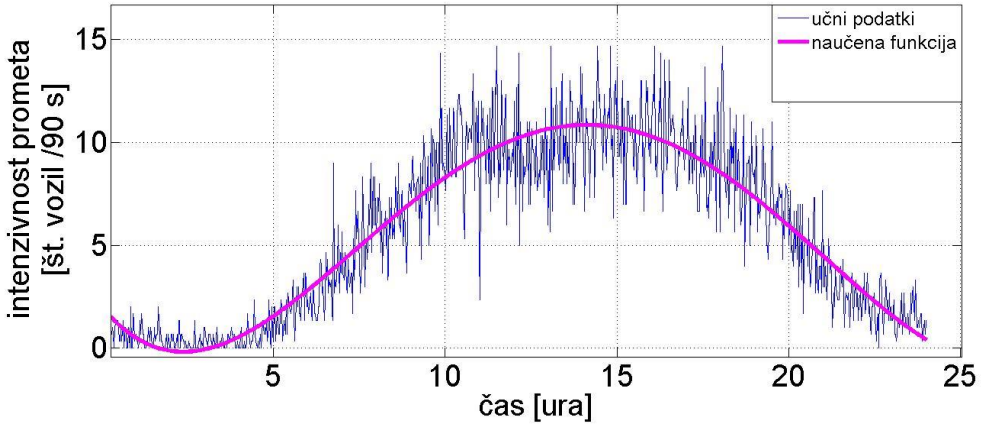
Slika 5: Primerjava različnih stopenj polinoma in glajenih vrednosti prometa z vrednostjo funkcije SMSE

Tabela 1: Primerjava vpliva stopnje polinoma na glajene vrednosti prometa za istovrstne dni tekom zaporednih tednov

2. stopnja polinoma		5. stopnja polinoma		9. stopnja polinoma	
Povprečje števila dni	Vrednost SMSE	Povprečje števila dni	Vrednost SMSE	Povprečje števila dni	Vrednost SMSE
2 dneva	0,232	2 dneva	0,197	2 dneva	0,177
3 dni	0,188	3 dni	0,141	3 dni	0,130
6 dni	0,146	6 dni	0,100	6 dni	0,086
14 dni	0,109	14 dni	0,066	14 dni	0,047

Za povprečje treh dni v treh zaporednih tednih smo se odločili zato, ker to povprečje signale dovolj pogladi. Povprečje manj dni je preveč grobo in vrednosti signalov ne pogladi dovolj. Dobro je, da se pri modeliranju odločimo za neko srednjo vrednost, saj te navadno dajejo najboljše rezultate. Izbrali smo povprečje treh dni z uporabo 5. stopnje polinoma. Za 5. stopnjo polinoma smo se odločili zato, ker je to zadnja stopnja, pri kateri je še opazen skok vrednosti SMSE. Z višanjem stopnje polinoma

se vrednosti SMSE ne izboljšujejo več bistveno. Na sliki 6 so prikazani podatki intenzivnosti prometa za povprečen četrtek, ki so zglajeni s posebno metodo drsečega povprečja (MA) treh dni. Z modro črto so označeni učni podatki, rdeča črta pa prikazuje naučeno funkcijo.



Slika 6: Odziv modela s polinomom 5. stopnje za četrtek na podlagi podatkov, dobljenih kot povprečje 3 dni iz treh zaporednih tednov

Problem pri predstavljeni metodi je, da za napoved potrebuje bazo podatkov, ki sega precej nazaj. V našem primeru je to 3 tedne. Ker pa tako dolgo hranjenje podatkov na strojni opremi v križišču ni želeno, smo morali poiskati še alternativni model, ki bo temeljil na manjši količini podatkov za napoved. Tak model je model AR.

5.2 Model AR

Pomemben del modela AR je izbira reda modela. Red modela smo izbrali tako, da smo točnost napovedi ovrednotili s cenilko standardne srednje vrednosti kvadratov napake SMSE, grafičnimi prikazi in histogrami.

Izmerjene podatke smo najprej analizirali za vsak dan posebej. Ugotovili smo, da se promet ob določenih dneh in urah približno ponavlja, opazili smo, da se pojavljajo tudi napačne meritve. Odločili smo se analizirati dneve, ki so si med seboj najbolj podobni in pri njih opazimo ponovitve signalov. Analizirali smo tudi dni, ki imajo različne vrednosti intenzivnosti prometa (na primer prazniki) od povprečnih delovnih dni. Zanima nas, kako dobra je napoved, če so si dnevi med seboj različni oziroma če so vrednosti signalov zelo različne.

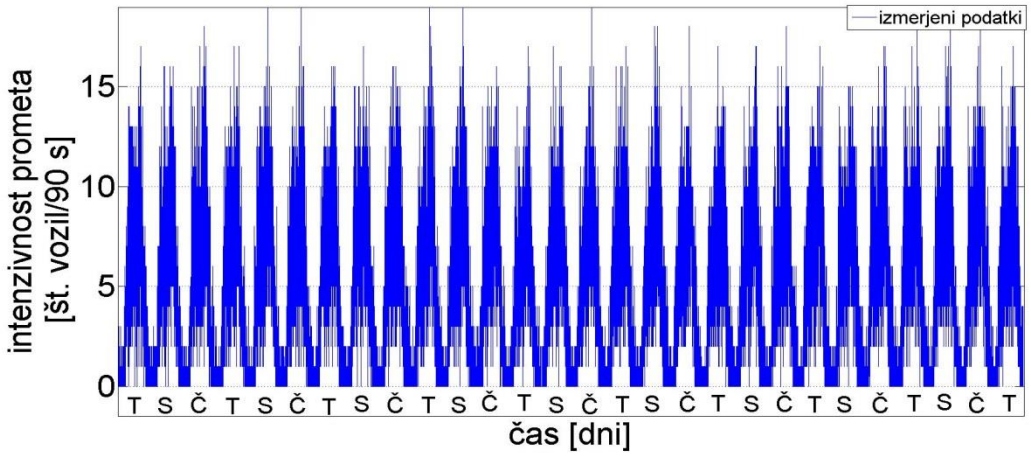
Sledilo je izločanje napačnih meritev, saj te pokvarijo samo napoved modela. Napake so razvidne iz prekomernega števila vozil ali izjemno nizkega števila vozil tekom dneva.

Podatke smo razdelili na učne (⅔ podatkov) in testne (⅓ podatkov). Iz učnih podatkov se je model učil in nato napovedal ⅓ podatkov, ki smo jih primerjali z ⅓ testnih podatkov. Na ta način smo ovrednotili točnost napovedi. Napoved smo vrednotili s funkcijo SMSE. Napovedane vrednosti smo vrednotili tudi s histogramom napak in grafom napake meritve. Histogram napak prikazuje izračun razlike med testnimi podatki in napovedanimi. Predpostavljamo, da ima šum (spremenljivka ε) normalno porazdelitev. Če to drži, je pričakovana porazdelitev histograma napak normalna ali Gaussova porazdelitev. Graf napake meritve se od histograma razlikuje po tem, da histogram omogoča vpogled v celoten model in njegove napake. Graf napake meritve v odvisnosti od časa pa omogoča vpogled v napake za posamezen dan ob določenem času.

5.2.1 Delovni dnevi

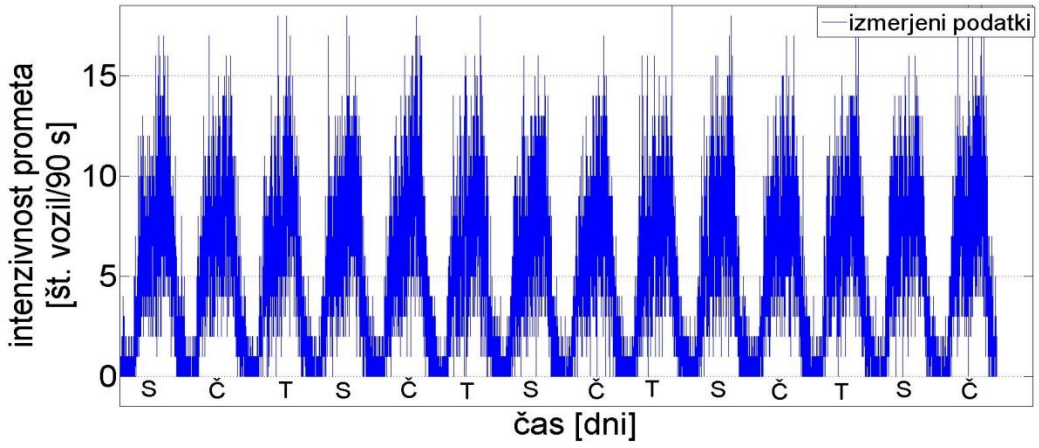
Pri delovnih dnevih smo opazili podobno obnašanje modela, kar prikazuje slika 7. Zato smo se odločili, da bomo analizirali te dni. Pri ponedeljkih in petkih smo opazili nekoliko višjo zasedenost križišča. Sklepamo, da imajo ponedeljki nekoliko drugačno obnašanje signalov zato, ker je začetek tedna in se to navadno nanaša na podaljšanje dopustov in večjih nakupovanj. Podobno sklepamo tudi za petke. Ocenili smo, da so odstopanja ponedeljkov in petkov od ostalih dni dovolj majhna, da lahko tudi za te dneve uporabimo isti model. Ta ocena se je pri vrednotenju modelov pokazala za ustrezno.

Podatke, očiščene vseh napak in odstopanj, smo razdelili na učne in testne podatke. Vsi podatki skupaj zajemajo 42 dni. Učni podatki (slika 7) zajemajo 28 dni, kar je ⅔ vseh podatkov.



Slika 7: Učni podatki za delovne dni

Testni podatki (slika 8), ki so ločeni od modela AR in služijo zgolj za pregled točnosti enokoračnega napovedovanja modela, zajemajo 14 dni, kar je ⅓ vseh podatkov. Model bo podal enokoračno napoved za naslednjih 14 dni, oziroma toliko, kolikor je dolžina testnih podatkov. Enokoračna napoved napove za toliko dni, kolikor je testnih podatkov, saj na ta način preverjamo točnost napovedi.



Slika 8: Testni podatki za delovne dni

Najprej smo s spreminjanjem reda modela iskali najustreznejšega. Spreminjanje reda modela pomeni spreminjanje števila zakasnjenih vzorcev, ki jih upošteva model. Za primer vzemimo 3. red modela (slika 9). Pri redu 3 model upošteva 3 zakasnjene vzorce. Enačba (41) prikazuje model 3. reda.

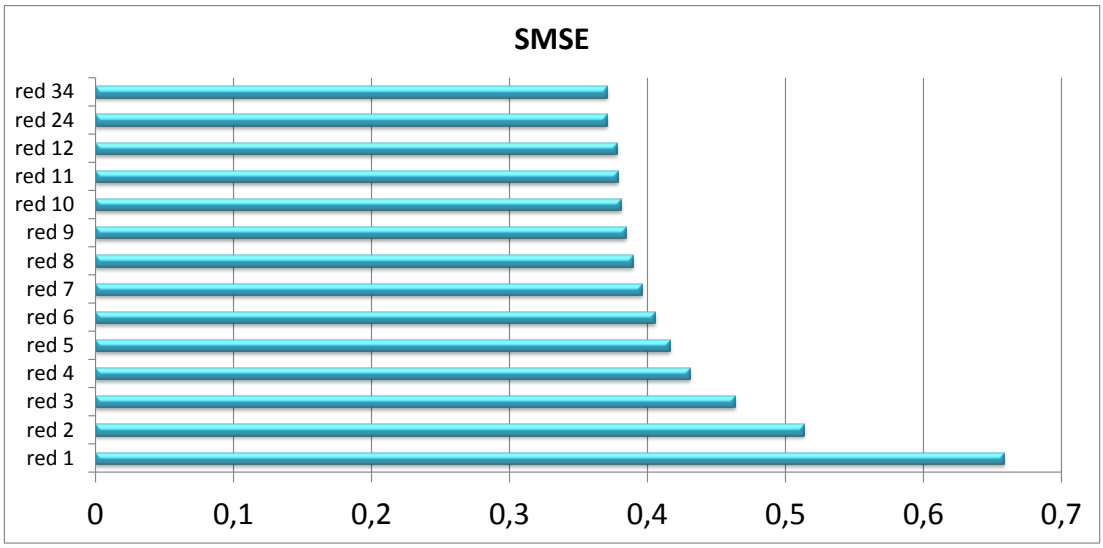
$$\hat{y}(k) = 0{,}3137y(k-1) + 0.3135y(k-2) + 0{,}3209y(k-3) + \varepsilon, \tag{41}$$

kjer je:

$y(k-1), \ldots, y(k-n)$ – merjene vrednosti,

n – red modela.

Ugotavljamo, da se z večanjem reda modela točnost napovedi na začetku izboljšuje, nato pa zmeraj manj. Cilj je uporabiti čim manjši oziroma primeren red in to tako, da bo enokoračna napoved še vedno dovolj točna. Pri redu 1 je vrednost SMSE = 0,659, pri redu 3 je vrednost SMSE = 0,463, pri redu 5 je vrednost SMSE = 0,412, pri redu 9 pa je vrednost SMSE = 0,384, kar prikazuje slika 9. Po primerjavi smo ugotovili, da je najprimernejši 5. red, saj se vrednost SMSE z višanjem reda ne zmanjšuje več bistveno.



Slika 9: Vrednotenje točnosti modela (SMSE) z redom modela

Naslednji korak je bila redukcija podatkov. Redukcija podatkov pomeni, da zmanjšamo število podatkov. Čas nadomestimo z vzorčnimi vrednostmi kT, kjer sta k zaporedno število vzorca in T čas vzorčenja (podrobnejši opis v podpoglavju 3.3.1). Enačba (42) prikazuje primer, ko upoštevamo redukcijo podatkov 2 oziroma

čas vzorčenja na 180 s. V primeru, da se odločimo za redukcijo podatkov 3, to pomeni, da podatke izvzamemo na vsakih 270 s, kar bi v enačbi zapisali kot $k3T$.

$$\hat{y}(k) = a_1 y(k2T - 1) + a_2 y(k2T - 2) + \cdots + a_n y(k2T - n) + \varepsilon(k), \tag{42}$$

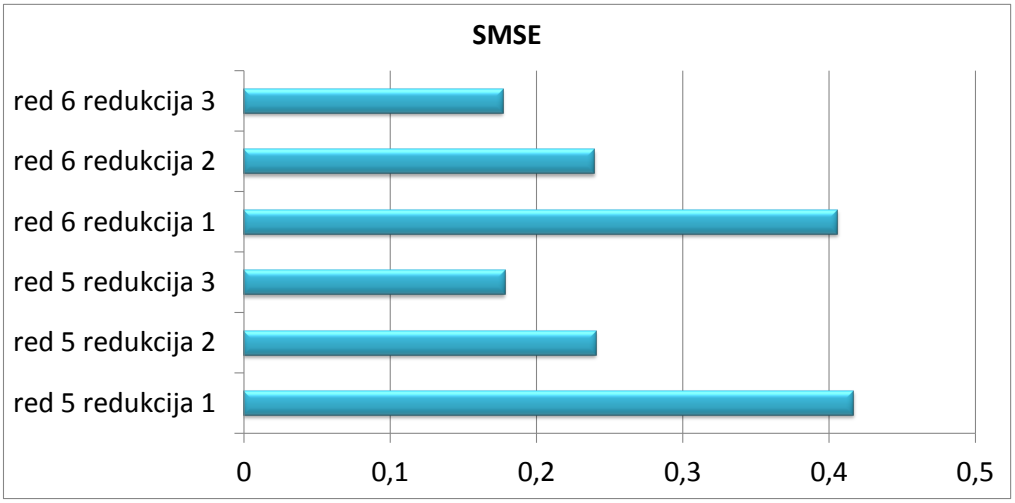
kjer je:

$T = 90$ s – perioda časovnega vzorca,

$y(k2T - 1), \ldots, y(k2T - n)$ – merjene vrednosti,

$k2T$ – zaporedno število vzorca, izvzetega na 180 s.

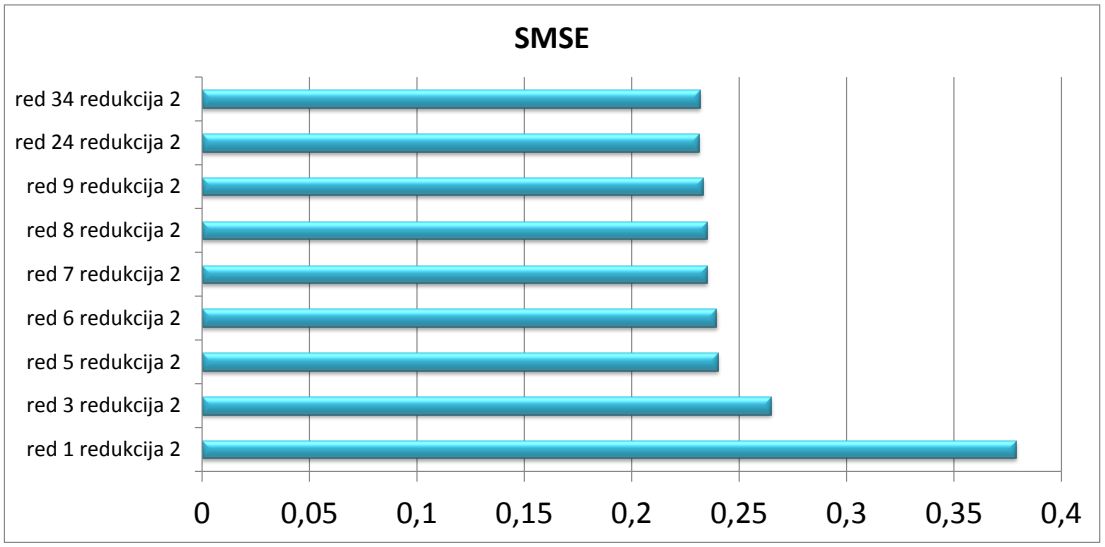
Redukcija podatkov zmanjša šum in s tem omogoča izboljšanje regresijskega modela. Na sliki 10 je prikazana analiza 5. in 6. reda z različnimi časi vzorčenja. Za primerjavo 5. in 6. reda smo se odločili zato, ker imata najprimernejše vrednosti SMSE (slika 9). Pri redu 5 z redukcijo časa vzorčenja 2 se točnost poveča skoraj za tretjino. Pri redukciji časa vzorčenja 3 pa se vrednost SMSE zniža za več kot polovico. Z večanjem redukcije časa vzorčenja se točnost napovedi izboljšuje. Treba je poudariti, da previsoka redukcija podatkov model poslabša, kar posledično zmanjša točnost napovedi. Zato je najbolje, da se odločimo za neko srednjo vrednost.



Slika 10: Vrednotenje točnosti modela (SMSE) s spreminjanjem časa vzorčenja

Ugotavljamo, da je matematični model z redukcijo časa vzorčenja 2 najustreznejši, ker dovolj zgladi signal in izboljša napoved. Točnost enokoračne napovedi smo še

enkrat preverili z višanjem reda modela (slika 11, tabela 2) in redukcijo časa vzorčenja 2 oziroma vzorčenjem na 180 s.



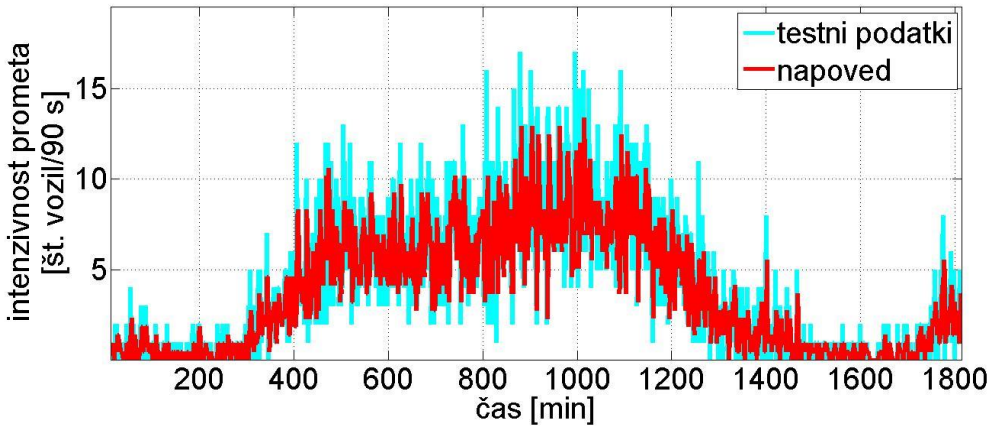
Slika 11: Vrednotenje točnosti modela (SMSE) s spreminjanjem reda in časa vzorčenja na 180 s

Tabela 2 prikazuje vrednosti funkcije SMSE pri izbranih redih in z redukcijo časa vzorčenja 2 oziroma vzorčenjem podatkov na 180 s. S slike 11 in spodnje tabele lahko razberemo, da je 5. red modela najprimernejši. Po 5. redu s časom vzorčenja na 180 s (redukcija podatkov 2) se točnost napovedi ne izboljšuje več bistveno.

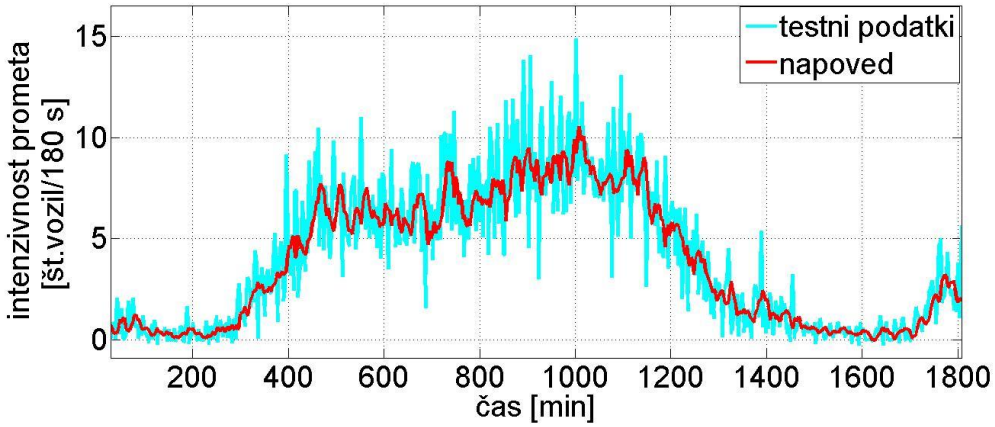
Tabela 2: Primerjava redov s časom vzorčenja 2, vpliv na točnost napovedi

Red	Redukcija	SMSE
1	2	0,379
2	2	0,303
3	2	0,265
4	2	0,251
5	2	0,240
6	2	0,239
7	2	0,235
8	2	0,235
9	2	0,233
10	2	0,232
11	2	0,231
12	2	0,231
24	2	0,231
34	2	0,232

Za boljšo predstavo o tem, kako redukcija časa vzorčenja in višanje reda pogladita signal, je na slikah 12 in 13 prikazan primer. Na sliki 12 je prikazan signal z uporabo 2. reda modela in brez spremembe časa vzorčenja. Na sliki 13 pa je prikazan signal za izbrani 5. red in redukcijo časa vzorčenja 2. S slik je razvidno, da višji red in redukcija časa vzorčenja signal zelo pogladita in tudi napoved je bolj gladka.



Slika 12: Izsek primerjave napovedi glede na različne rede modela in časa vzorčenja podatkov (2. red modela)

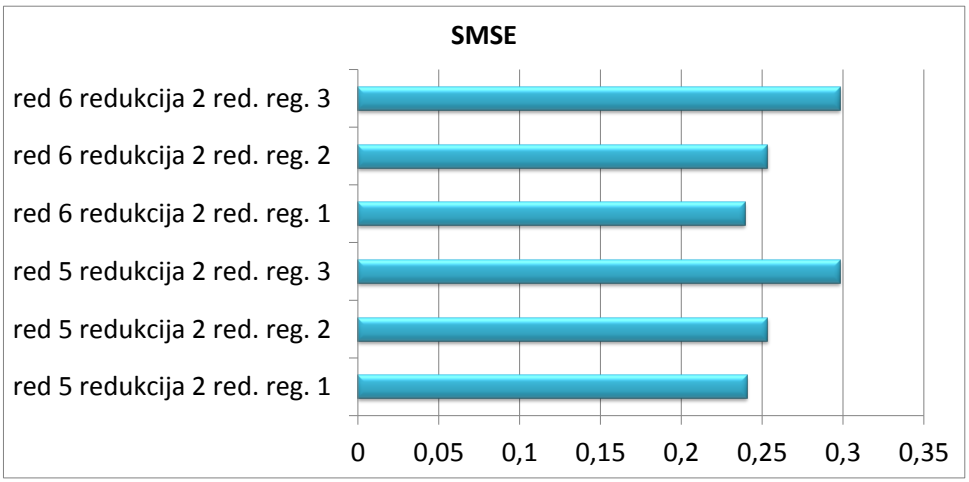


Slika 13: Izsek primerjave napovedi glede na različne rede modela in časa vzorčenja podatkov (5. red modela, redukcija časa vzorčenja 2)

Preizkusili smo tudi redukcijo regresorjev. Redukcija regresorjev pomeni izločitev vsakega n-tega regresorja iz začetne množice regresorjev. Enačba (43) prikazuje redukcijo regresorjev 2. Pri redukciji regresorjev 2 se iz vektorja koeficientov regresije $\boldsymbol{\theta}$ (podrobnejši opis v podpoglavju 3.2) izvzame vsak drugi koeficient (a_i).

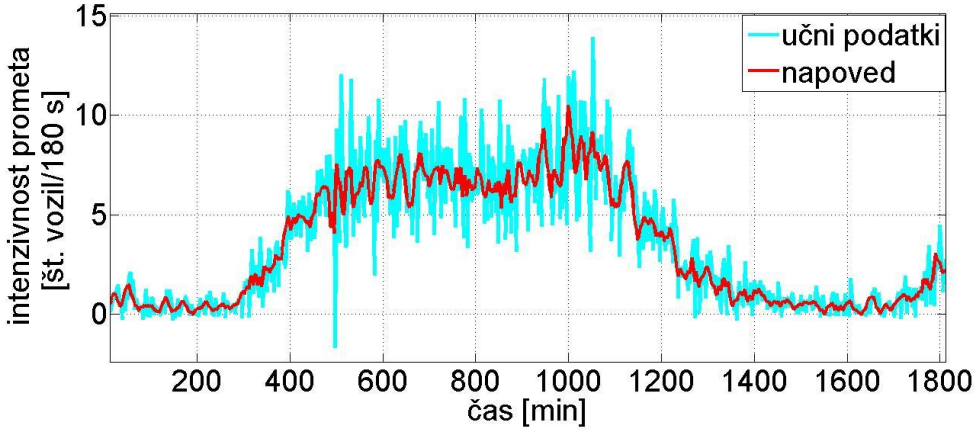
$$\hat{y}(k) = a_2 y(k-2) + a_4 y(k-4) + \cdots + a_{2n} y(k-2n) + \varepsilon \qquad (43)$$

Za lažjo predstavo vpliva redukcije smo izbrali najprimernejša reda glede na vrednosti funkcije SMSE. To sta 5. in 6. red modela ter redukcija časa vzorčenja 2. Vpliv redukcije regresorjev prikazuje slika 14. Odločili smo se, da redukcije regresorjev ne bomo uporabili, saj se je enokoračna napoved brez redukcije regresorjev izkazala za točnejšo.



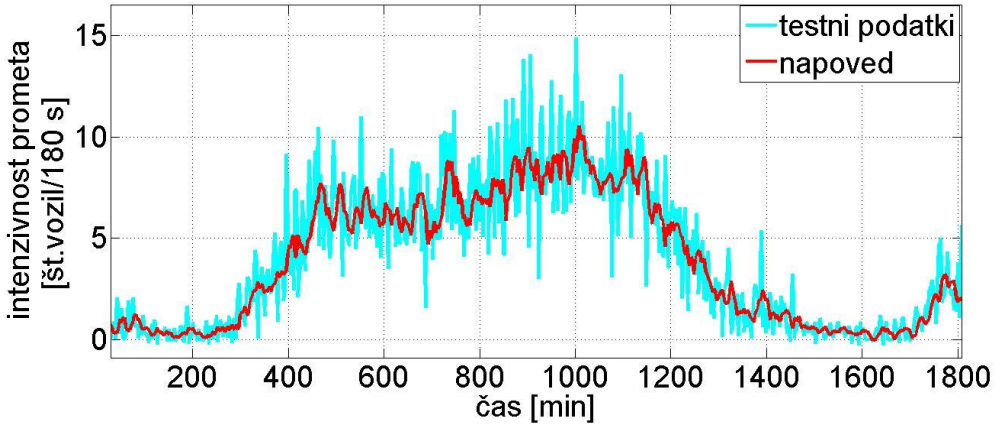
Slika 14: Vpliv redukcije regresorjev na točnost napovedi (SMSE)

Na sliki 15 so z modro črto prikazani surovi podatki gostote prometa na izbranem križišču. Izbrali smo 5. red modela, zato se napoved prične s petimi začetnimi vrednostmi. Napoved na tem grafu je odvisna od učnih podatkov, kar pomeni da prikazuje le, kako se model uči in omogoča vpogled, ali je napoved primerna. Učni podatki zajemajo ⅔ vseh podatkov.



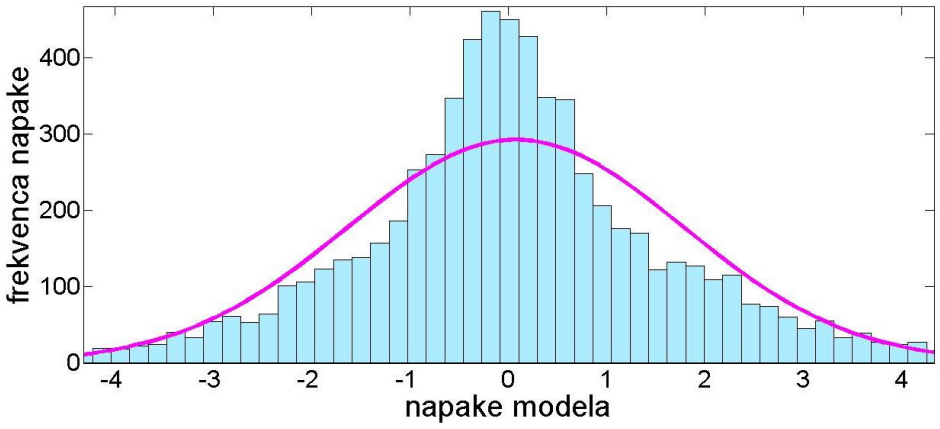
Slika 15: Izsek napovedi modela na učnih podatkih

Napoved smo vrednotili na ⅓ podatkov, ki so neodvisni od napovedi. Na grafu so prikazani zato, da lahko vizualno vrednotimo in ocenimo, ali je napoved primerna oziroma ali dosega želena pričakovanja. S slike 16 je razvidno, da napoved poteka v želeni smeri, saj napoved sledi srednji vrednosti testnih podatkov.



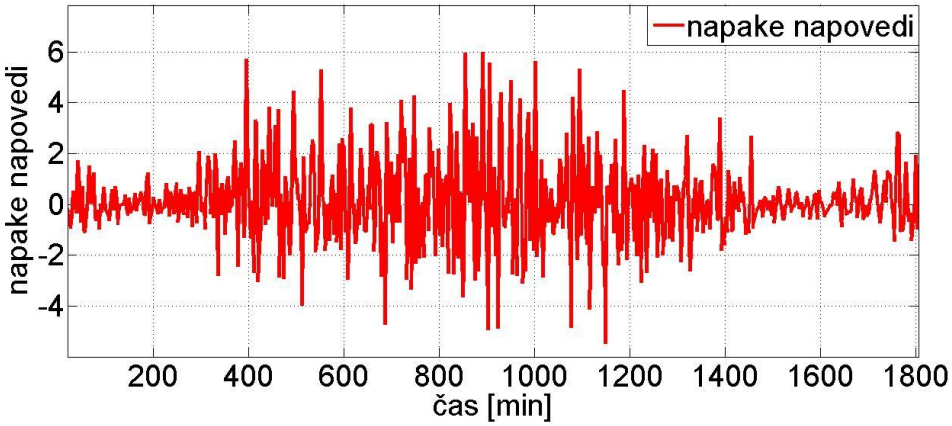
Slika 16: Izsek napovedi na testnih podatkih (neodvisni podatki)

Histogram napak je naslednja metoda, s katero lahko vrednotimo točnost napovedi. Že sama oblika histograma napak narekuje, ali je model sprejemljiv. Histogram napak, ki ima normalno ali Gaussovo porazdelitev, je prvi znak, da je model sprejemljiv. Iz histograma za model 5. reda je razvidno (slika 17), da je 461 vozil napovedalo pravilno (brez napake). Z napako ± 2 je napovedanih od 109 do 119 vozil, z napako ± 4 pa je napovedanih 17 do 25 vozil. Rdeča črta na histogramu prikazuje funkcijo porazdelitve intenzivnosti prometa.



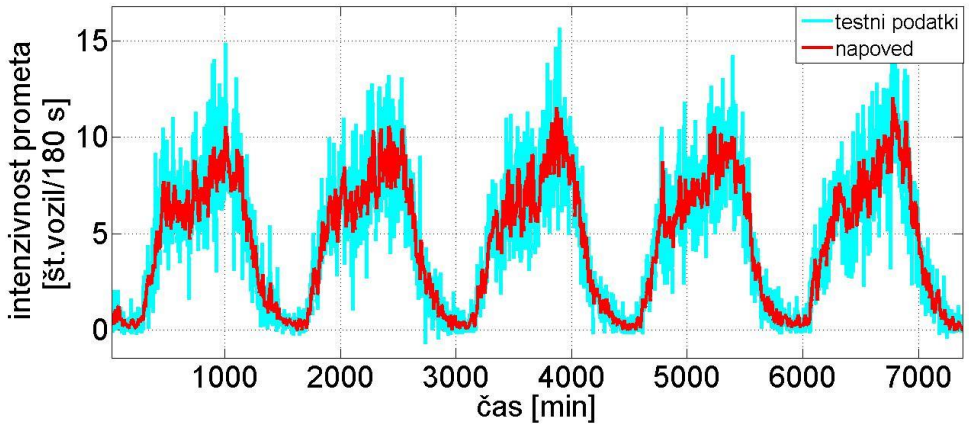
Slika 17: Vrednotenje točnosti napovedi s histogramom

Podobno kot histogram napak na sliki 17 napako napovedi prikazuje graf na sliki 18. Od histograma napak se razlikuje po tem, da napake napovedi prikazuje v odvisnosti od časa. Napaka napovedi je izračunana tako, da od napovedi odštejemo testne podatke. Omogoča torej vpogled, kdaj je največ napak in kako so velike. S slike 18 lahko razberemo, da matematičen model dnevno zgreši za največ ± 6 vozil.



Slika 18: Izsek napak napovedi ob določenem času

Na sliki 19 je prikazana napoved za en korak naprej naslednjih 5 dni z izbranim 5. redom modela in redukcijo časa vzorčenja 2 ($T = 180\ s$). Z grafa lahko razberemo, da je napoved dobra. Modra črta prikazuje dejanske (testne) podatke, rdeča črta pa ponazarja napovedi za en korak naprej naučenega matematičnega modela. Napoved dnevno zgreši največ do 6 vozil. Enokoračna napoved napoveduje in velja za vse delovne dni.



Slika 19: Izsek napovedi petih dni na testnih podatkih za delovne dni

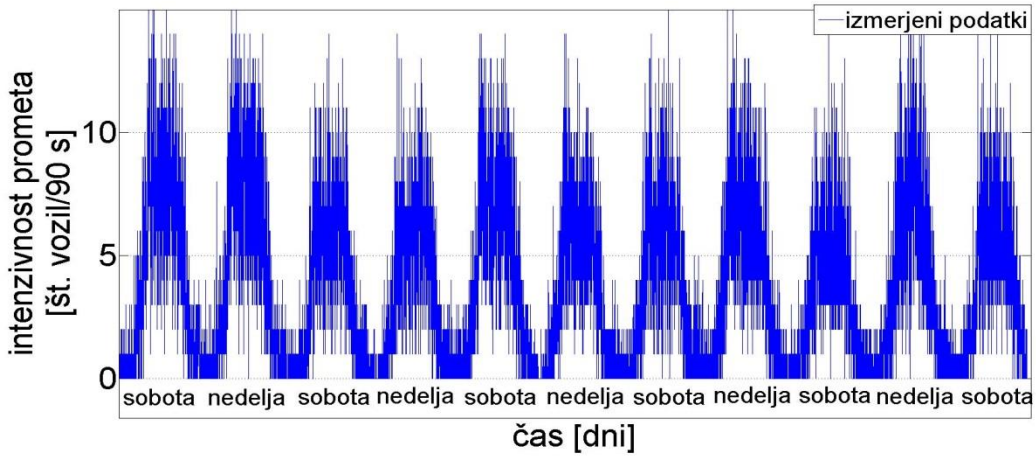
Na koncu postopka primerjave in izbire najprimernejšega reda, časa vzorčenja in redukcije regresorjev sledi še zapis izbranega modela (enačba (44)). Za izbrani 5. red in čas vzorčenja na 180 s je vrednost SMSE = 0,240.

$$\hat{y}(k) = 0{,}1651y(k-1) + 0{,}2347y(k-2) + 0{,}1828y(k-3) + 0{,}2346y(k-4) + 0{,}1666y(k-5) + \mathcal{E}\,;T = 180\text{ s} \quad (44)$$

5.2.2 Vikendi

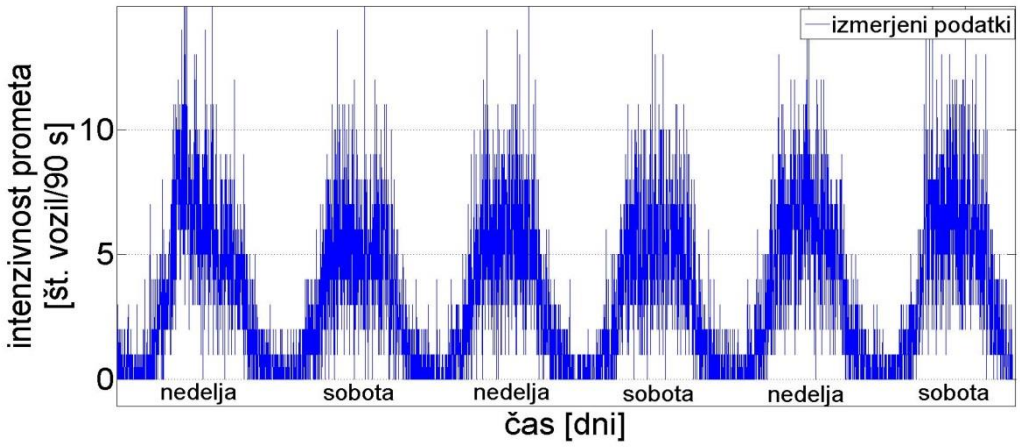
Z analizo podatkov smo tudi pri sobotah in nedeljah opazili podobnost obnašanja modela (sliki 20, 21). Iz serije podatkov smo izbrali vikende in iz izbranih dni izločili tiste, pri katerih smo opazili prevelika odstopanja tekom dneva. Tako je ostala le prečiščena serija sobot in nedelj, pripravljena za nadaljnjo analizo.

Ponovno smo podatke razdelili na učne in testne. Vsi podatki zajemajo 17 dni. Učni podatki (slika 20) zajemajo nekaj več kot ⅔ vseh podatkov, kar je 11 dni.



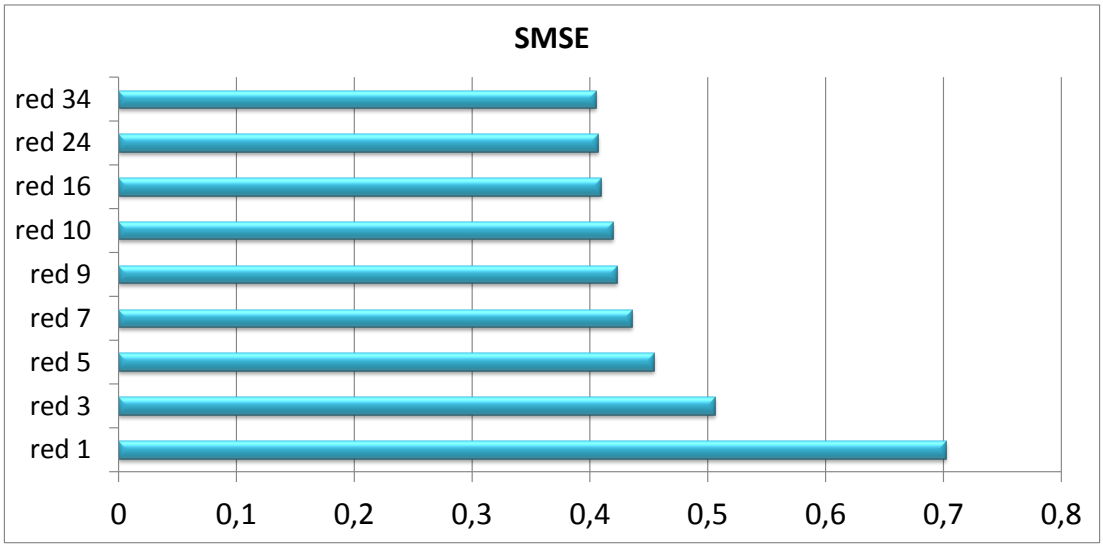
Slika 20: Učni podatki (vikendi)

Testni podatki (slika 21) zajemajo slabo ⅓ podatkov oziroma 6 dni. Napovedali bomo intenzivnost prometa za nadaljnjih 6 dni in točnost napovedi ovrednotili na testnih podatkih, ki so neodvisni od modela.



Slika 21: Testni podatki (vikendi)

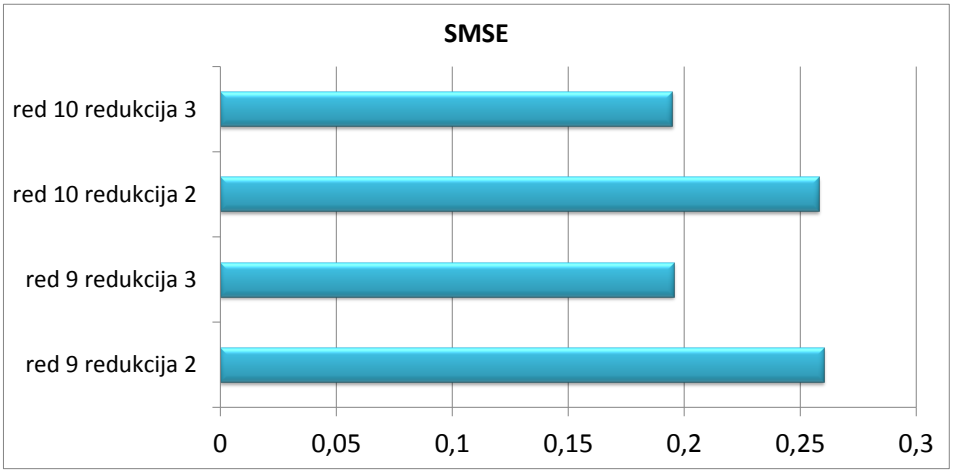
Strukturo matematičnega modela smo določali s spreminjanjem reda modela (opis v podpoglavju 5.2.1) oziroma s spreminjanjem zakasnjenih vzorcev, ki jih upošteva model s funkcijo SMSE. Na sliki 22 opazimo koleno pri 9. redu modela. Od tega reda naprej se vrednosti SMSE ne izboljšujejo več bistveno.



Slika 22: Vrednotenje točnosti modela (SMSE) z redom modela

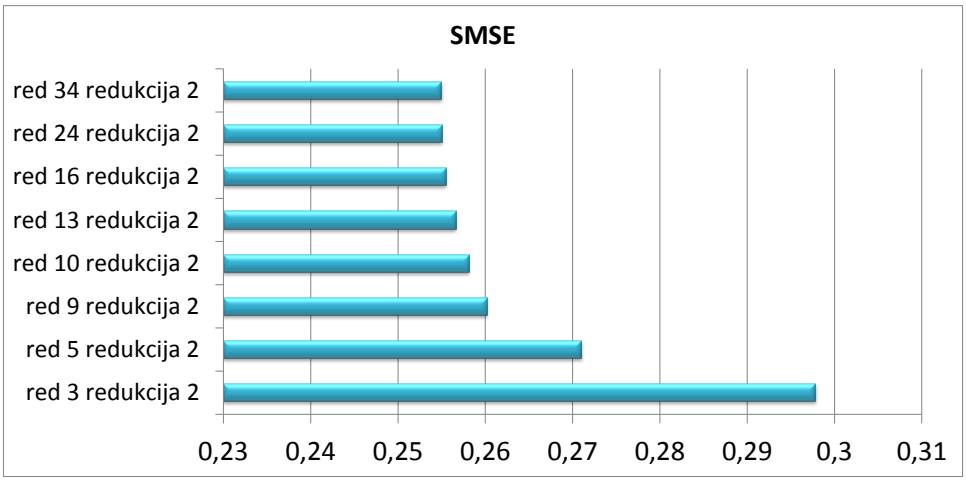
Naslednji korak pri analizi sobot in nedelj je spreminjanje časa vzorčenja. Kot smo ugotovili že v podpoglavju 5.2.1, se z večanjem časa vzorčenja model izrazito izboljšuje (slika 23). S tem, ko večamo čas vzorčenja, se enokoračna napoved slabša, zato smo se odločili, da analiziramo redukcijo časa vzorčenja 2 ($T = 180$ s) in 3 ($T = 270$ s) z 9. in 10. redom modela, saj pri teh redih vrednost SMSE še izrazito

pada (slika 22). Vrednosti SMSE sicer z redukcijo časa vzorčenja 3 padejo, vendar ne dovolj izrazito. Če bi se vrednost SMSE zmanjšala za polovico, bi izbrali čas vzorčenja na 270 s (redukcija 3), tako pa je čas vzorčenja 180 s (redukcija 2) primernejši.



Slika 23: Vrednotenje SMSE izbranih redov s spreminjanjem časa vzorčenja

Časa vzorčenja torej ne bomo povečevali na več kot 180 s oziroma redukcijo časa vzorčenja 2, saj bi s tem enokoračno napoved poslabšali, saj se tako zmanjša število podatkov, s tem pa točnost napovedi pada. Še enkrat smo preverili, kako na vrednosti SMSE vpliva red modela s časom vzorčenja na 180 s. S slike 24 in tabele 3 razberemo, da sta 9. in 10. red po vrednostih SMSE res najprimernejša. Podrobnejša analiza za izbiro najprimernejšega reda sledi v nadaljevanju.

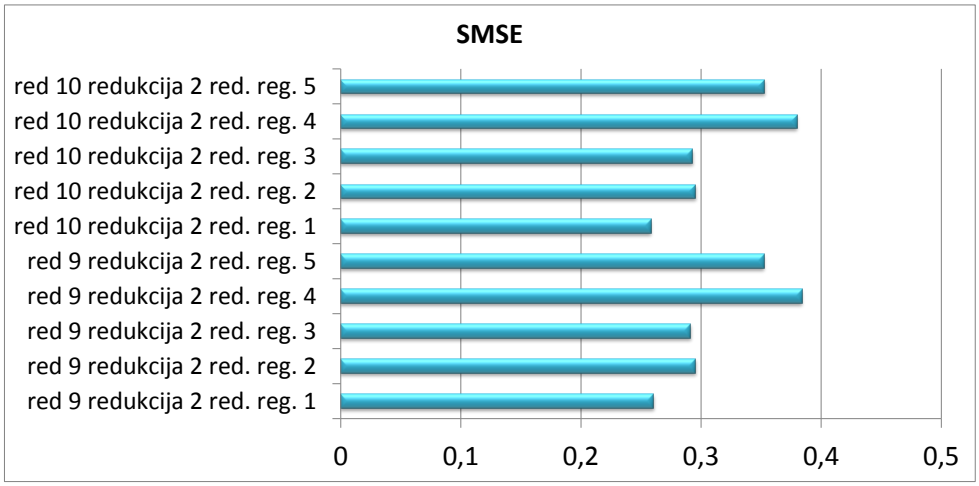


Slika 24: Vrednotenje točnosti modela (SMSE) z različnimi redi in časom vzorčenja 2

Tabela 3: Vrednosti SMSE glede na red in čas vzorčenja

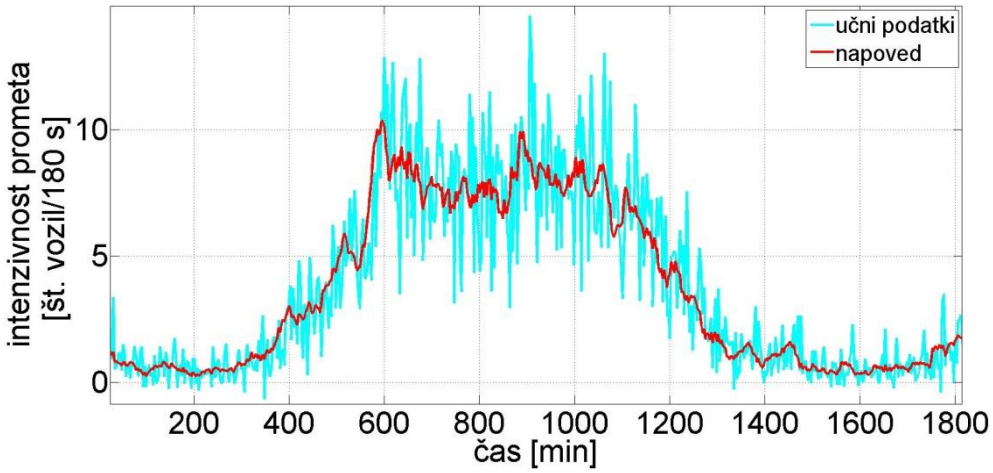
Red	Redukcija	SMSE
3	2	0,298
5	2	0,271
9	2	0,260
10	2	0,258
13	2	0,257
16	2	0,255
24	2	0,255
34	2	0,255

Preverili smo tudi, kako vpliva redukcija regresorjev na matematični model (podrobnejši opis v podpoglavju 5.2.1). Slika 25 prikazuje, da višanje redukcije regresorjev točnost napovedi le niža. Zato smo se ponovno odločili, da redukcije regresorjev ne bomo uporabili.



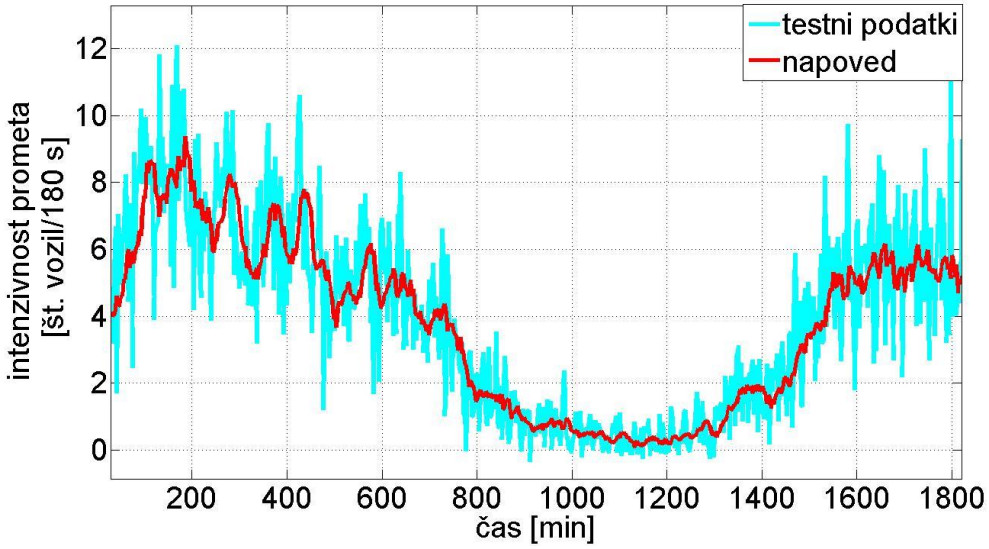
Slika 25: Vrednotenje točnosti napovedi s spreminjanjem redukcije regresorjev

Odločali smo se med 9. in 10. redom modela. Izbrali smo 9. red modela, čas vzorčenja na 180 s (redukcija podatkov 2) in brez redukcije regresorjev. Najprej smo grafično pregledali, ali se model uči pravilno (slika 26). S slike razberemo, da se model dobro uči, saj je rdeča črta, ki ponazarja napoved znotraj učnih podatkov (obarvanih modro) in sledi njihovim vrednostim.



Slika 26: Izsek učenja modela na učnih podatkih

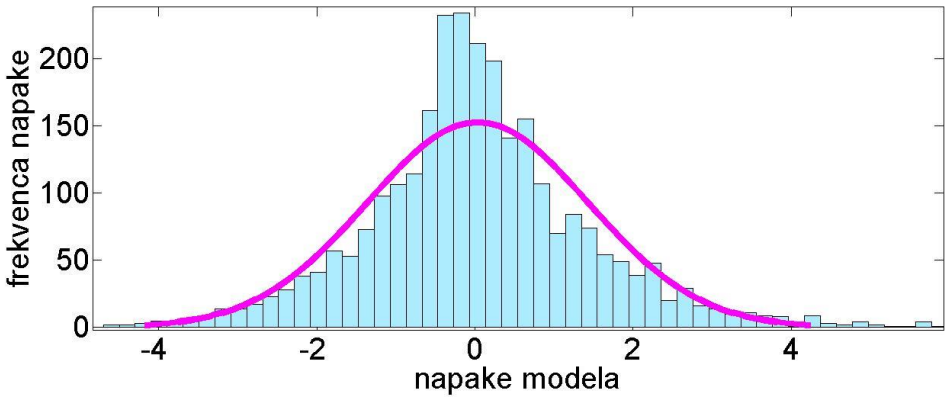
Sledil je še pregled napovedi na testnih podatkih. Tudi ta napoved dosega želena pričakovanja, kar prikazuje slika 27. Napoved, označena z rdečo barvo, sledi srednji vrednosti testnih podatkov, označenih z modro barvo.



Slika 27: Izsek obnašanja napovedi na testnih (neodvisnih) podatkih

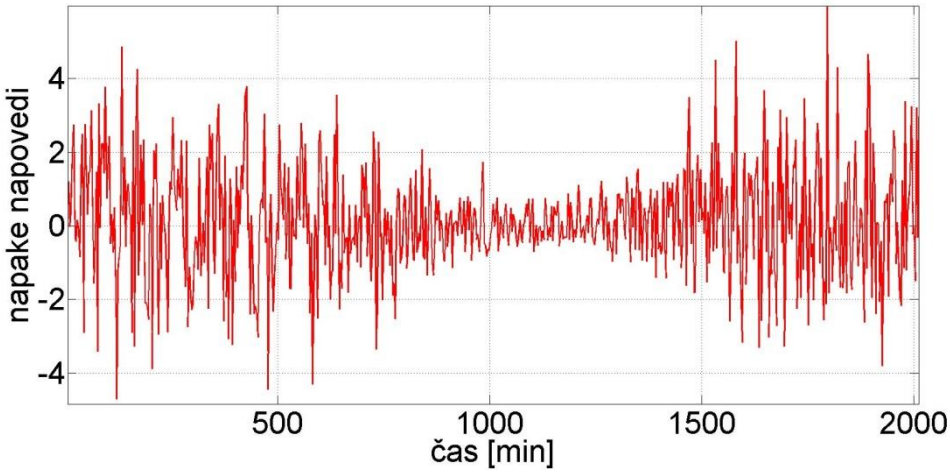
Poleg uporabe funkcije SMSE in grafičnih prikazov smo točnost napovedi ovrednotili tudi s histogramom napak. Na sliki 28 je prikazan histogram napak z normalno porazdelitvijo. Že oblika histograma sporoča, da je matematični model zadovoljiv, saj ima približno Gaussovo porazdelitev. Brez napake je model napovedal 235 vozil. Z napako ± 2 je napovedanih od 35 do 40 vozil, z napako ± 4 pa

od 1 do 4 vozila. Rdeča črta na histogramu prikazuje funkcijo porazdelitve intenzivnosti gostote prometa.



Slika 28: Vrednotenje modela s histogramom napak

Napake napovedi lahko pregledamo tudi za vsak dan posebej (slika 29) z grafičnim prikazom časovne odvisnosti napake napovedi. S slike razberemo, da dnevno model zgreši največ ± 5 vozil na dan.



Slika 29: Izsek napake napovedi za izbran model

Enačba (45) prikazuje zapis izbranega modela 9. reda z redukcijo podatkov 2 oziroma vzorčenjem na 180 s. Vrednost SMSE je za izbran matematični model SMSE = 0,260.

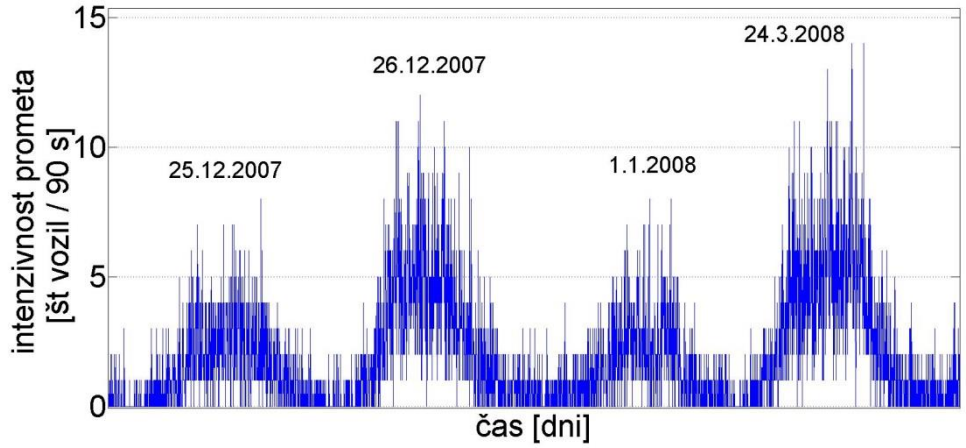
$$\hat{y}(k) = 0{,}1165(k-1) + 0{,}1378y(k-2) + 0{,}1112y(k-3) +$$
$$0{,}131y(k-4) + 0{,}0734y(k-5) + 0{,}1142y(k-6) +$$

$$0{,}1029y(k-7) + 0{,}0819y(k-8) + 0{,}1194y(k-9) + \mathcal{E};\ T = 180\ \text{s} \quad (45)$$

5.2.3 Prazniki

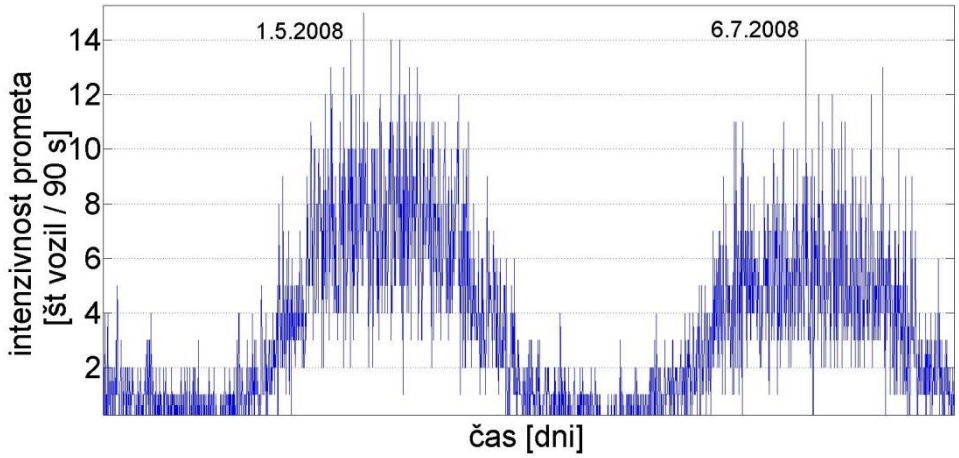
Prazniki nimajo podobnega obnašanja modela kot preostali dnevi v prejšnjih analizah. Zanimalo nas je, kako točna bo napoved glede na to, da si dnevi med seboj niso podobni in so različni od preostalih dni. Za razliko od prejšnjih analiz v podpoglavju 5.2 smo tokrat izbrali dneve (prazniki), ki so si različni. Zato iz serije podatkov ni bilo treba izločati dni, ki so si med seboj podobni, saj je bil namen analize modeliranje različnih prazničnih dni.

Ponovno smo izbrane praznike razdelili na učne in testne podatke. Skupno število praznikov zajema 6 dni. Od tega sta ⅔ učnih podatkov (4 dni), kar prikazuje slika 30.



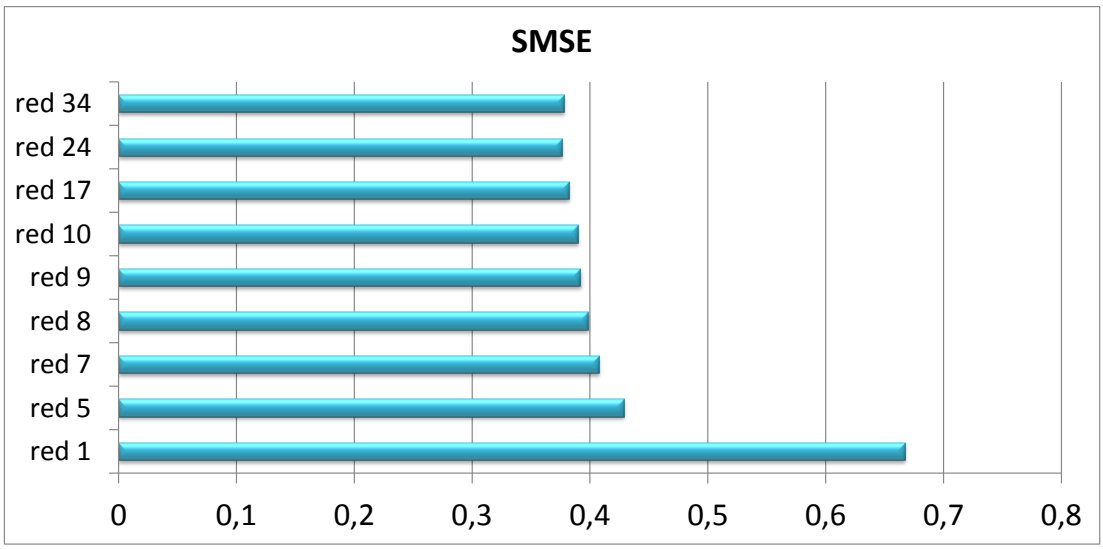
Slika 30: Prazniki, zajeti v učne podatke

Testni podatki pa zajemajo ⅓ vseh podatkov (2 dneva), kar je prikazano na sliki 31.



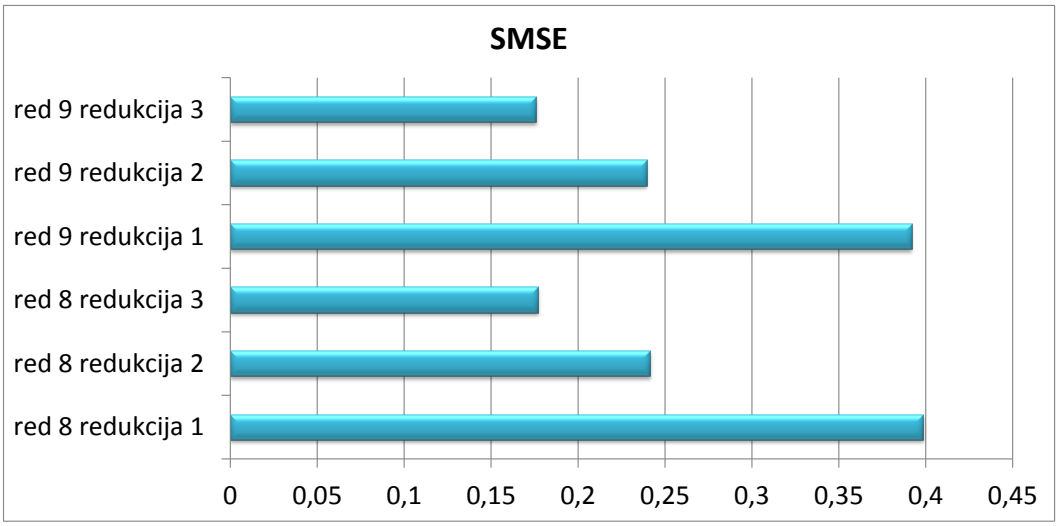
Slika 31: Dnevi za testne podatke

Vpliv reda smo, kot že rečeno, vrednotili s funkcijo SMSE. Ugotavljamo, da se z višanjem reda točnost modela do 8. reda hitro izboljšuje (slika 32), kasneje pa se vrednost SMSE z višanjem reda ne spreminja več bistveno.



Slika 32: Vpliv reda na točnost napovedi

Če primerjamo 8. red in redukcijo časa vzorčenja 2 ter 9. red in redukcijo časa vzorčenja 2, opazimo, da ni bistvenih razlik v točnosti napovedi (slika 33, tabela 4). Zato smo se odločili za 8. red in redukcijo časa vzorčenja 2. Vrednost SMSE je pri teh pogojih enaka 0,242.

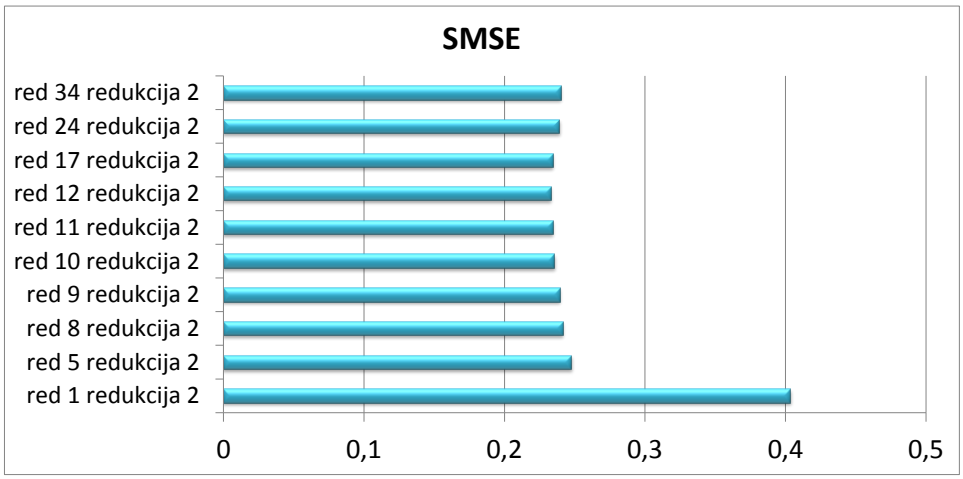


Slika 33: Vpliv reda in časa vzorčenja na točnost napovedi

Tabela 4: Vrednosti SMSE za izbrane rede in čas vzorčenja

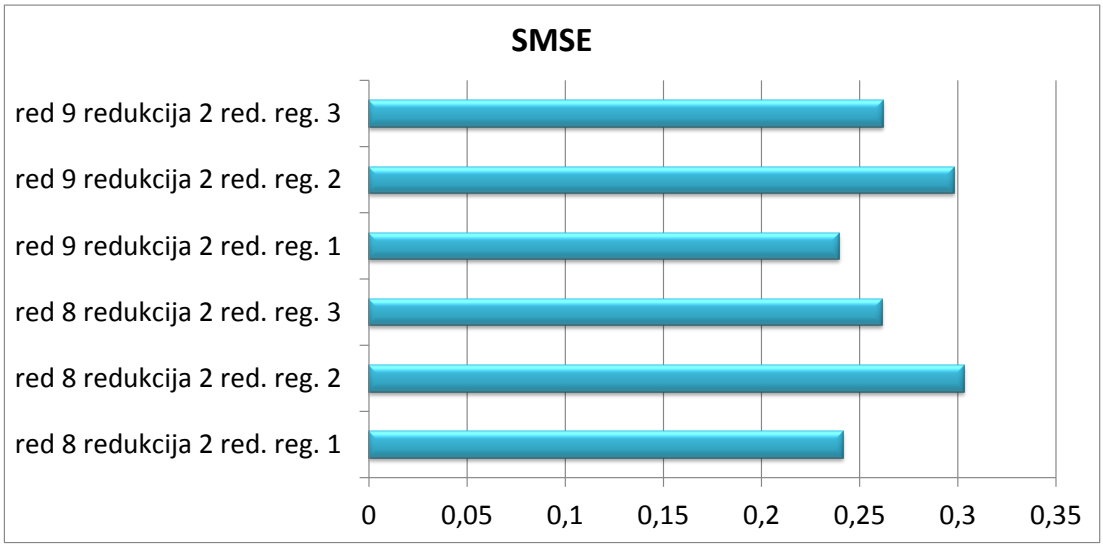
Red	Redukcija	SMSE
8	1	0,399
8	2	0,242
8	3	0,177
9	1	0,392
9	2	0,239
9	3	0,176

Odločili smo se za redukcijo podatkov 2 (čas vzorčenja na 180 s). Ponovno smo točnost napovedi primerjali z različnimi redi pri času vzorčenja na 180 s (slika 34). Tudi tukaj smo ocenili, da sta vrednosti SMSE najprimernejši pri 8. in 9. redu.



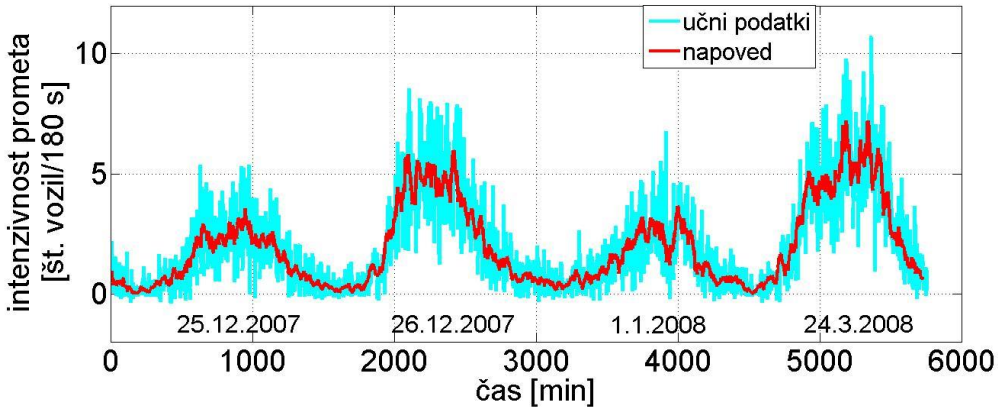
Slika 34: Vpliv reda na točnost napovedi pri redukciji podatkov 2

Višanje redukcije regresorjev (podrobnejši opis v podpoglavju 5.2.1), kot smo ugotovili že v podpoglavjih 5.2.1 in 5.2.2, le slabša točnost napovedi, kar prikazuje slika 35. Zato smo se odločili, da redukcije regresorjev ne bomo uporabili.



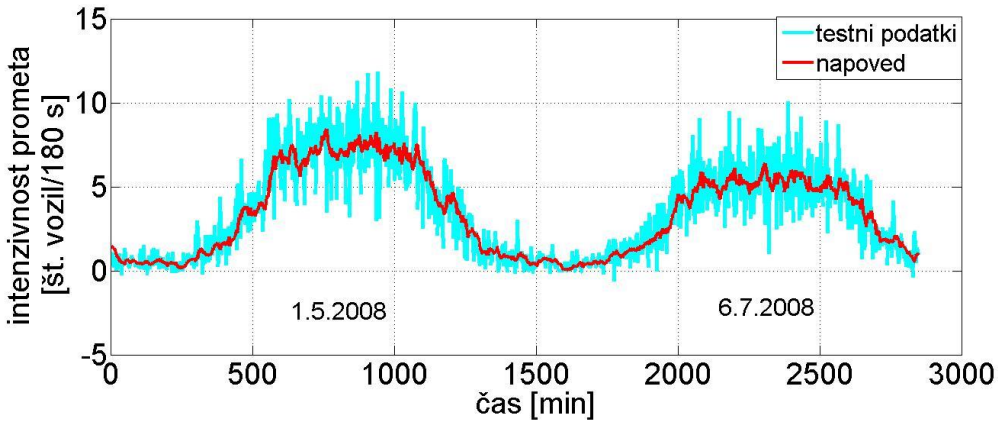
Slika 35: Vpliv redukcije regresorjev na točnost napovedi

Na sliki 36 je prikazana napoved na učnih podatkih. Z rdečo črto je označena napoved, z modro pa učni podatki. Napoved sledi gibanju učnih podatkov.



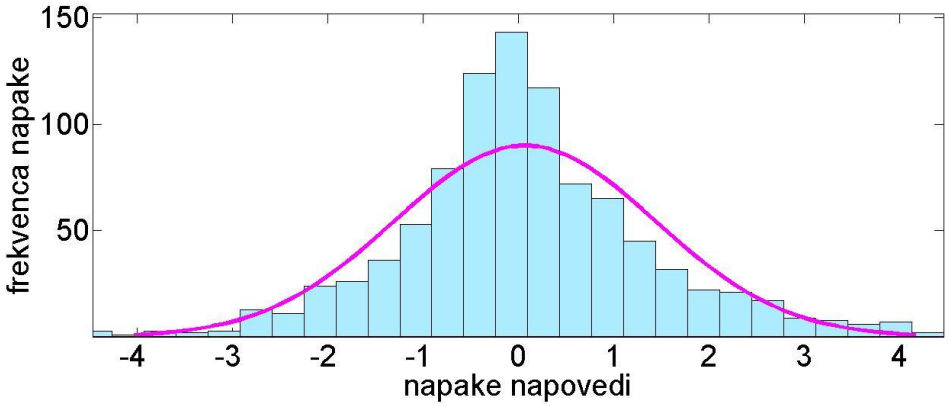
Slika 36: Učenje modela na učnih podatkih

Sledila je še vizualna ocena točnosti napovedi z grafičnim pregledom odstopanja med testnimi podatki in napovedjo (slika 37). Tudi tu opažamo, da se napoved giblje okoli srednje vrednosti testnih podatkov, kar potrjuje, da matematični model deluje v skladu s pričakovanji.



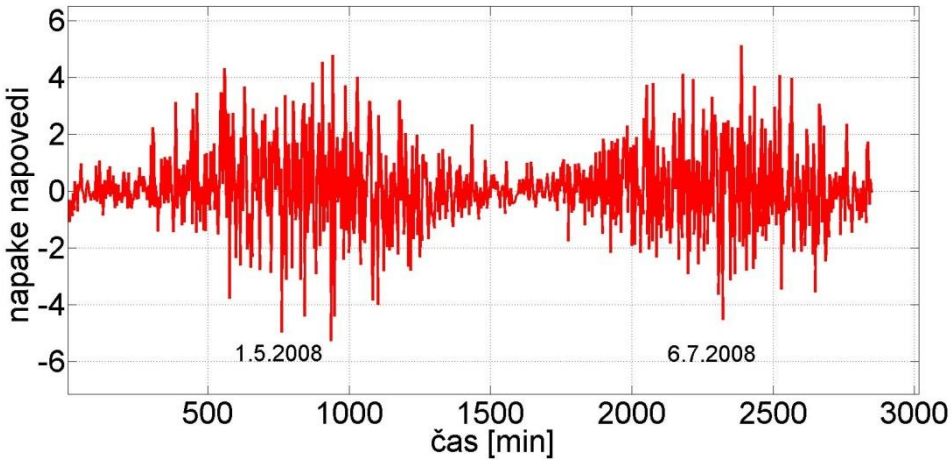
Slika 37: Napoved na testnih (neodvisnih) podatkih

Naslednji korak pri vrednotenju napovedi je histogram napak. Slika 38 prikazuje histogram napak za izbrani 8. red modela z redukcijo podatkov 2 oziroma vzorčenjem na 180 s. Histogram napak ima Gaussovo obliko. To je prvi znak, da je izbrani model primeren. Brez napake je napovedanih 145 vozil, z napako ± 2 je napovedanih od 16 do 22 vozil, z napako ± 4 pa od 3 do 4 vozila. Rdeča črta na histogramu prikazuje funkcijo intenzivnosti gostote prometa.



Slika 38: Vrednotenje modela s histogramom napak

Napaka napovedi je, kot že rečeno, razlika med testnimi podatki in napovedjo. Omogoča nam vpogled v napake napovedi ob določenem času za izbran dan. S slike 39 je razvidno, da napoved zgreši za največ ± 5 vozil na dan.



Slika 39: Napake napovedi za izbrani model

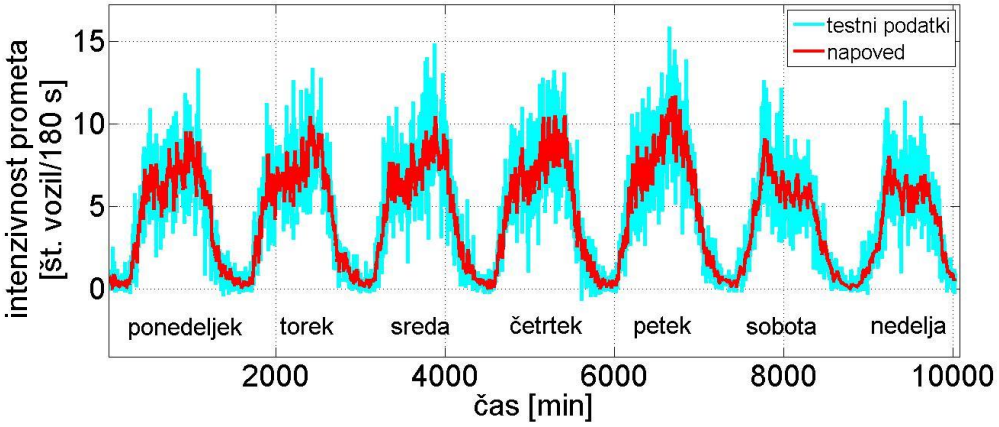
Izbrali in ovrednotili smo 8. red modela s časom vzorčenja 2 in brez redukcije regresorjev. Enačba (46) prikazuje zapis izbranega modela. Model je dovolj točen, saj vsa vrednotenja dosegajo želena pričakovanja. Vrednost SMSE je pri izbranih pogojih 0,242.

$$\hat{y}(k) = 0{,}1237y(k-1) + 0{,}1511y(k-2) + 0{,}1122y(k-3) + 0{,}1771y(k-4) + 0{,}1142y(k-5) + 0{,}1188y(k-6) + 0{,}674y(k-7) + 0{,}1116y(k-8) + \varepsilon\,;\; T = 180\ \text{s} \qquad (46)$$

5.2.4 Celoten teden

Izdelali smo matematične modele za delovne dni, vikende in praznike. Za vodenje prometa v primeru napake na senzorju je treba pri enokoračnem napovedovanju upoštevati za različne dni različne matematične modele.

Za vpogled v enokoračno napoved celotnega tedna smo na testnih podatkih iz dveh matematičnih modelov sestavili napoved celotnega tedna. Slika 40 prikazuje napoved delovnih dni z upoštevanjem 5. reda modela in redukcijo podatkov 2. Napoved sobote in nedelje pa upošteva izbrani 9. red modela in redukcijo podatkov 2 (čas vzorčenja na 180 s). V primeru, da bi bil v tednu tudi praznik, bi v napoved celotnega tedna vključili še tretji model z upoštevanjem izbranih kriterijev za praznike (8. red modela in redukcije podatkov 2).



Slika 40: Enokoračna napoved za ves teden z dvema modeloma na testnih podatkih

Kljub dobrim rezultatom ne moremo trditi, da je naša uporabljena metoda najboljša. Vedno je treba analizirati vsako metodo posebej in se na podlagi rezultatov v dani situaciji odločiti za najustreznejšo. Zaključimo lahko, da so pridobljeni modeli zadovoljivi, saj delajo dovolj točne enokoračne napovedi, vrednosti SMSE na testnih podatkih pa nikoli ne presegajo vrednosti 0,3. Iz modeliranja smo se veliko naučili. Samega križišča, na katerem analiziramo intenzivnost gostote prometa, sicer nismo nikoli obiskali in si ga zato nismo dobro predstavljali, zemljevidi pa ne dajejo dobrega vpogleda v dejansko križišče. Pri tem problemu nam je bila v veliko pomoč aplikacija Google Street View, saj smo si križišče lahko ogledali in pridobili občutek, kot da smo tam zares bili. Naloga modeliranja intenzivnosti prometa v križišču računsko ni zapletena, saj programski paket Matlab omogoča hitre izračune. Vseeno smo porabili relativno veliko časa, saj smo program velikokrat dopolnjevali in izpopolnjevali. Poleg tega smo naredili tudi mnogo analiz, saj smo želeli modele, ki dosegajo dovolj visoko točnost.

6 ZAKLJUČEK

V delu smo obravnavali problem napovedovanja intenzivnosti prometa, ki bi ga lahko uporabili v primeru odpovedi senzorjev na enem izmed križišč v Pragi. V primeru, da se na senzorju zazna napaka, se lahko za vodenje prometa do odprave problema uporabljajo enokoračne napovedi modela, dobljenega iz preteklih meritev intenzivnosti prometa.

V Sloveniji se nadzor prometa izvaja samo na hitrih cestah in avtocestah. V območju mestnega prometa pa do sedaj po našem vedenju še ni bilo daljinskega nadzora prometa. Smiselno pa bi bilo, da se v vseh večjih in dnevno obremenjenih križiščih v Sloveniji namesti senzorje, ter tako sprosti promet tudi v prometno največjih konicah dneva.

Prikazali smo osnove regresije in časovnih vrst, ki so osnova za enokoračno modeliranje. Predstavili smo tudi metode eksperimentalnega modeliranja časovnih vrst. Izbrani metodi eksperimentalnega modeliranja časovnih vrst smo uporabili pri enokoračnem napovedovanju intenzivnosti prometa v izbranem križišču. Sledila je izdelava enokoračnih napovedi z modeli AR in MA na dejanskem križišču v Pragi. Z modelom AR smo analizirali in izdelali enokoračne napovedi za delovne dni, vikende in praznike, s posebnim modelom MA pa smo izračunali povprečje enega dne tako, da smo na določen dan iz več zaporednih tednov za izbran dan izračunali njegovo aritmetično povprečje. Nato smo napovedi pogladili z uporabo modela AR.

Namen magistrskega dela, da izdelamo modele, ki bodo točni in bodo lahko uporabljeni za vodenje prometa v primeru odpovedi senzorja, je dosežen, saj smo uspešno izdelali modele za enokoračno napovedovanje intenzivnosti prometa v križišču. Napovedi smo ovrednotili in povzemamo, da so vse vrednosti SMSE nižje od 0,3, vsi histogrami napak imajo približno normalno oziroma Gaussovo porazdelitev. Vizualni pregledi napovedi in učenja modelov pa ponazarjajo, da so napovedi zadovoljive, saj enokoračne napovedi sledijo srednjim vrednostim podatkov. Povzamemo lahko, da je dosežen tudi cilj dela, saj so napovedi dovolj točne in v skladu s pričakovanji.

Naloga modeliranja nam je vzela precej časa, saj smo računalniški program večkrat popravljali in izpolnjevali. Naredili smo mnogo analiz, saj smo želeli dovolj točne napovedi. Izvedbeno pa naloga ni bila zapletena, saj smo vse izračune pridobili s programskim paketom Matlab, ki omogoča uporabniku prijazno delo.

Edini problem je bil, da si dejanskega križišča v Pragi nismo nikoli ogledali v živo in si ga zato nismo dobro predstavljali. Rešitev je omogočala spletna aplikacija Google Street View. Z njo smo pregledali križišče in dobili občutek, kot da smo tam tudi zares bili.

V delu smo uporabili metode, ki se lahko uporabijo na raznovrstnih področjih. Uporabljene eksperimentalne metode niso prilagojene izključno za napovedovanje intenzivnosti prometa v križišču. Metodo za modeliranje se lahko uporabi tudi za analizo podatkov in izdelavo enokoračnih napovedi v energetiki, ekonomiji, organizaciji in drugje.

Za nadaljnje delo bi lahko enokoračne napovedi ovrednotili s sedanjo intenzivnostjo prometa v Pragi in na ta način še enkrat preverili točnost enokoračnega napovedovanja in posledično ustreznost izdelanih matematičnih modelov.

7 LITERATURA

ARIMA model (2014). Pridobljeno 20. 5. 2014 s svetovnega spleta: http://www.mathworks.com/help/econ/arima-model.html

Autoregressive model (2014). Pridobljeno 14. 3. 2013 s svetovnega spleta: http://en.wikipedia.org/wiki/Autoregressive_model#Definition

Baum F. C. (2012). Econometrics methods. Course materials. Boston: Boston college.

Chin R., Lee Y. B. (1996). Principles and practice of clinical trial medicine. Elsevier Science: Academic Press.

Conditional variance models (2014). Pridobljeno 25. 5. 2014 s svetovnega spleta: http://www.mathworks.com/help/econ/what-is-a-conditional-variance-model.html

Covariance (2014). Pridobljeno 24. 5. 2014 s svetovnega spleta: http://en.wikipedia.org/wiki/Covariance

Curve fitting (2014). Pridobljeno 7. 4. 2014 s svetovnega spleta: http://en.wikipedia.org/wiki/Curve_fitting

GARCH (2014). Pridobljeno 13. 3. 2014 s svetovnega spleta: http://en.wikipedia.org/wiki/GARCH#GARCH

GJR models (2014). Pridobljeno 20. 3. 2014 s svetovnega spleta: http://www.mathworks.com/help/econ/gjr-model.html

Google map (2014). Řevnická, Praha. Pridobljeno 10. 3. 2014 s svetovnega spleta: https://www.google.com/maps/search/%C5%98evnick%C3%A1Praha+/@50.053893,14.2889586,18z/data=!3m1!4b1

Guangyi M. (2011). MA, AR and ARMA models. Economic forecasting. Lecture notes no. 5. Department of economics: Texas A&M University.

Hsu C., Sandford B. A. (2007). The Delphi technique. Practical asessment, research & evaluation, vol. 12, str. 10.

Hyndman R. J. (2001). ARIMA processes. Datajobs: Data science knowledge repository. A central knowledge resource for data scientists/analytic experts.

Johnson S. (2013). The differences between qualitative and quantitative forecasting techniques. Pridobljeno 10. 3. 2014 s svetovnega spleta: http://www.ehow.com/info_12042327_differences-between-qualitative-quantitative-forecasting-techniques.html

Kejžar N. (2011). Regresija. Zapiski in gradivo s predavanj. Inštitut za biostatistiko in medicinsko informatiko. Ljubljana: Medicinska fakulteta.

Knudsen M. (2004). Experimental modelling of dynamic systems. Department of control engineering, Aalborg University.

Kocijan J., Prikryl J. (2010). Soft sensor for faulty measurements detection and reconstruction in urban traffic, Melecon 2010, 15th IEEE Mediterranian Electromechanical Conference, Valleta, Malta, IEEE, str. 172–176.

Korenjak A. (2010). Regresijska analiza. Diplomsko delo. (Fakulteta za naravoslovje in matematiko, Univerza v Mariboru). Maribor. [A. Korenjak], 2010.

Lawrence G. F. (2009). The SWOT Analysis: Using your strength to overcome weaknesses, using opportunities to overcome threats. Create space.

Mathematical model (2014). Pridobljeno 21. 6. 2014 s svetovnega spleta: http://en.wikipedia.org/wiki/Mathematical_model

McDoland H. J. (2009). Handbook of biological statistics. Second edition. Maryland: Sparky House Publishing.

Mean squared error (2014). Pridobljeno 7. 4. 2014 s svetovnega spleta: http://en.wikipedia.org/wiki/Mean_squared_error

Pfajfer L., Arh F. (1998). Statistika 1. Univerza v Ljubljani, Ekonomska fakulteta. Ljubljana.

Raymond Y. C. (2013). An application of the ARIMA model to real-estate prices in Hong Kong, Hong Kong: The University of Hong Kong, str. 152–162.

Shumway R. H., Stoffer D. S. (2010). Time Series Analysis and its applications with R examples. EZ – third edition, Springer.

Sivec V. (2009). Empirična analiza volatilnosti borznih donosov. Diplomsko delo. (Univerza v Ljubljani, Ekonomska fakulteta), Ljubljana: [V. Sivec].

Smooting time series (2014). The Pennsylvania State Univeristy. Online statistic course. Pridobljeno 25. 5. 2014 s svetovnega spleta: https://onlinecourses.science.psu.edu/stat510/?q=node/70

Sun H., Liu Xx. H., Xiao H., He R. R., Ran B. (2002). Short-term traffic forecasting using the local linear regression model. 82nd TRB annual meeting, str. 1-18.

Stamatopoulos C. (2002). Sampled-based fishery surveys: a tehnical handbook. FAO fishers tehnical paper. No. 425. Rome, 2002.

Systematic sampling (2014). Pridobljeno 7. 4. 2014 s svetovnega spleta: http://www.investopedia.com/terms/s/systematic-sampling.asp

Time Series (2014). Pridobljeno 14. 4. 2014 s svetovnega spleta: http://en.wikipedia.org/wiki/Time_series

Tsay R. S. (2002). Analysis of financial time series. John Wiley & Sons.

Turkman K. F. (2008). Linear Regression. Department of Statistics and Operation Research. Faculty of Science. Portugal: University of Lisbon.

White noise (2014). Pridobljeno 6. 3. 2014 s svetovnega spleta: http://en.wikipedia.org/wiki/White_noise

Zobovnik K. (2011). Uporaba programa SPSS pri napovedovanju električne energije. Diplomsko delo. (Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko), Prebold: [K. Zobovnik].