Analiza multivarijatnih vremenskih serija putem objašnjivih metoda strojnog učenja

Kučinić, Ivana

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
SVEUČILIŠNI DIPLOMSKI STUDIJ

Ivana Kučinić

DIPLOMSKI RAD

Zagreb, rujan 2022.

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
SVEUČILIŠNI DIPLOMSKI STUDIJ

Ivana Kučinić

ANALIZA MULTIVARIJATNIH VREMENSKIH SERIJA PUTEM OBJAŠNJIVIH METODA STROJNOG UČENJA

DIPLOMSKI RAD

Voditelj diplomskog rada:	doc. dr. sc. Željka Ujević Andrijić
Članovi ispitnog povjerenstva:	doc. dr. sc. Željka Ujević Andrijić
	prof. dr. sc. Dragana Mutavdžić Pavlović
	doc. dr. sc. Mario Lovrić

Zagreb, rujan 2022.

SADRŽAJ

IZJAVA

Izjavljujem da sam ovaj diplomski rad pod nazivom „Analiza multivarijatnih vremenskih serija putem objašnjivih metoda strojnog učenja“ izradila samostalno uz stručnu pomoć i pod nadzorom mentorice doc. dr. sc. Željke Ujević Andrijić te uz podršku doc. dr. sc. Marija Lovrića u Chemical codes d.o.o.

Ovaj diplomski rad izrađen je na Zavodu za mjerenje i automatsko vođenje procesa Fakulteta kemijskog inženjerstva i tehnologije Sveučilišta u Zagrebu pod mentorstvom doc. dr. sc. Željke Ujević Andrijić.

Zahvaljujem se svojoj mentorici doc. dr. sc. Željki Ujević Andrijić na stručnoj pomoći, susretljivosti i savjetima tijekom pisanja ovog rada.

Posebnu zahvalnost dugujem doc. dr. sc. Mariju Lovriću na izdvojenom vremenu, iskazanom povjerenju, nesebičnoj pomoći i prenesenom znanju koji je uvelike pridonio izradi ovog rada.

I na kraju, najveću zaslugu za svoja postignuća pripisujem svojim roditeljima koji su mi osigurali školovanje i omogućili da mi ništa ne nedostaje. Hvala im što su uvijek bili uz mene i pružali podršku kada je bilo najpotrebnije. Zahvaljujem se i bratu koji me svojim savjetima usmjeravao na pravi put. Ovaj diplomski rad je za Vas i zbog Vas.

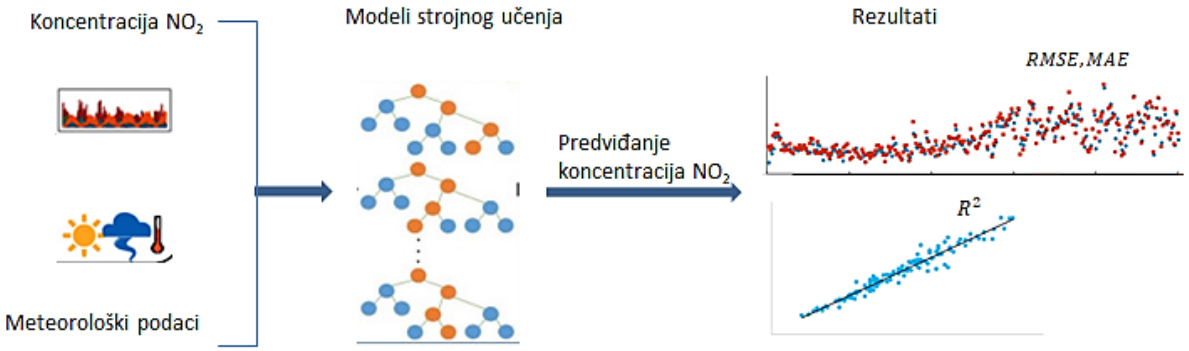
SAŽETAK

Visoke koncentracije dušikova dioksida (NO_2) u zraku, posebno u jako urbaniziranim područjima, negativno utječu na mnoge aspekte ljudskog zdravlja. U ovom radu primjenjuju se metode strojnog učenja za predviđanje koncentracija NO_2 u zraku. Meteorološki podaci i koncentracije NO_2 odabrane su i analizirane s četiri mjesta u gradu Grazu. Koncentracije NO_2 postavljene su kao ciljne varijable *Prophet* i *Random Forest* modela za predviđanje vrijednosti u periodu od 15. ožujka 2019. do 15. ožujka 2020. godine. Razvijeni modeli strojnog učenja pokazali su dobru razinu generalizacije za predviđanje koncentracija NO_2 u zraku. Kombinacija *Prophet* značajki i Random Forest modela pokazala se najboljom za razvoj modela predviđanja koncentracija NO_2. Najbolji rezultati ostvareni su na mjernoj postaji Zapad gdje je vrijednost koeficijenta determinacije modela R^2 = 0,65, dok je najlošiji rezultat ostvaren na postaji Don Bosco gdje je R^2 = 0,50. Značajke koje najviše utječu na razvoj modela su značajke sezonalnosti, a najveći utjecaj ima godišnja sezonalnost.

Ključne riječi:

Onečišćenje zraka, dušikov dioksid, strojno učenje, multivarijatne vremenske serije, *Prophet* algoritam, algoritam slučajnih šuma

GRAFIČKI SAŽETAK:



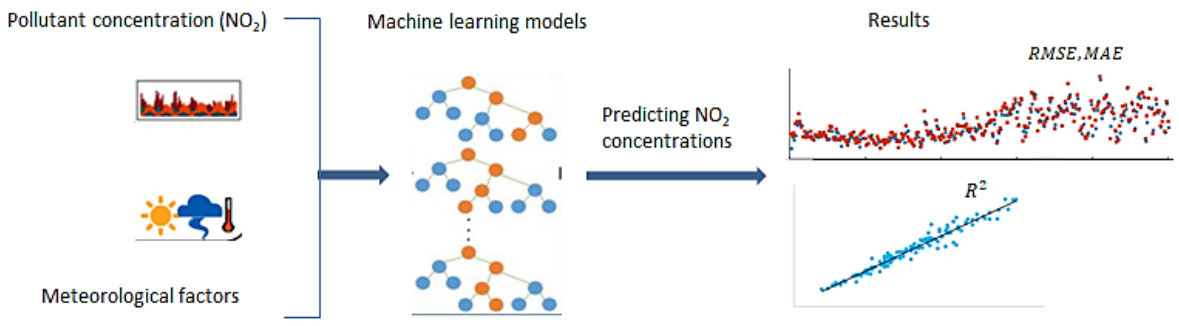
ABSTRACT

High concentrations of nitrogen dioxide (NO_2) in the air, especially in heavily urbanised areas, have a negative impact on many aspects of human health. In this work, machine learning methods were applied for the prediction of NO_2 concentrations in the air. Meteorological data and concentrations of NO_2 were selected and analyzed from four places in the city of Graz. Concentrations of NO_2 are set as target variables of *Prophet* and *Random Forest* models for predicting values from March 15th 2019 to March 15th 2020. The machine learning models showed a good level of generalization for predicting the NO_2 concentrations. The combination of *Prophet* features and *Random Forest* model proved to be the best for developing the NO_2 concentration prediction model. The best results were achieved at the West measuring station where the value determination coefficient R^2 is 0,65, and the worst result was achieved at Don Bosco station where R^2 is 0,50. The features that most affect the development of the model are the features of seasonality and the greatest impact has the yearly seasonality.

Key words:

Air pollution, nitrogen dioxide, machine learning, multivariate time series, *Prophet* model, *Random Forest* model

GRAPHICAL ABSTRACT:



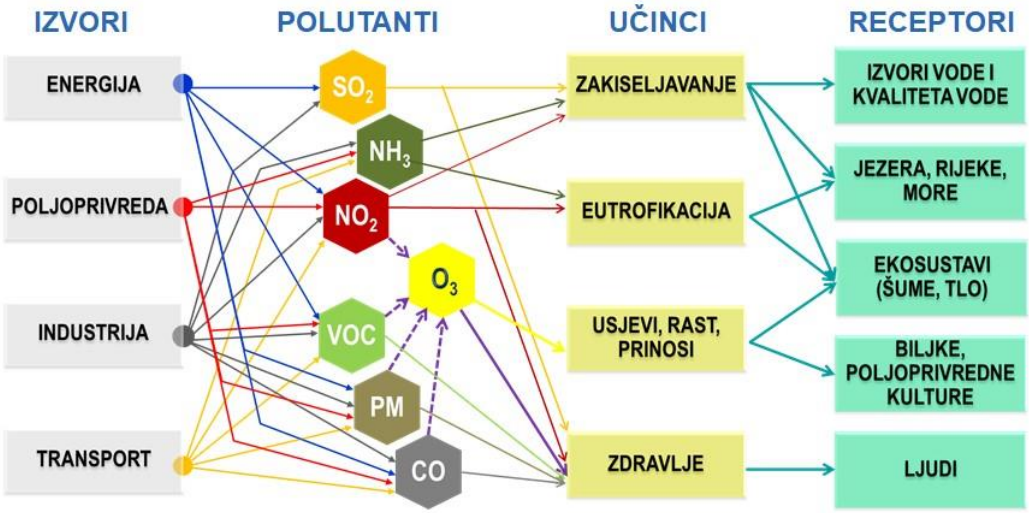
1. UVOD

Ubrzanim razvojem tehnologije i industrije uzrokovane su sve veće klimatske promjene, nastanak ozonskih rupa i povećava se onečišćenje zraka čime je dosta narušena kvaliteta ekosustava. Onečišćenje zraka jedan je od glavnih ekoloških problema koji štetno utječe na ljudsko zdravlje izazivajući kardiovaskularne bolesti i visoke stope smrtnosti.[1] Osim utjecaja na ljudsko zdravlje, onečišćenje zraka loše utječe i na okoliš, a također pridonosi trošenju ozonskog omotača koji štiti Zemlju od UV zračenja.[2] Velika gustoća naseljenosti znatno utječe na onečišćenje zraka u gradovima i urbaniziranim područjima. Prema podacima Svjetske zdravstvene organizacije preko 90% populacije udiše zrak koji odstupa od dopuštenog ograničenja i koji sadrži povišene koncentracije onečišćujućih tvari.[3] Na kvalitetu zraka utječu štetne tvari koje onečišćuju zrak poput dušikovih i sumporovih oksida, prizemnog ozona, ugljikovog monoksida, ugljikovog dioksida, amonijaka, hlapivih komponenti te lebdećih čestica.[2] Onečišćenje zraka uglavnom je posljedica potrošnje i proizvodnje energije, a najčešći izvori su motorna vozila, industrijska postrojenja i šumski požari.[2] Kako bi se onečišćenje zraka smanjilo potrebno ga je pravovremeno identificirati odnosno imati u realnom vremenu dostupne podatke o iznosu, lokaciji i vremenu nastanka onečišćenja. U nastojanju da se omogući točno predviđanje onečišćenja primjenjuju se različite metodologije na različitim skupovima podataka. Algoritmi strojnog učenja pokazali su se vrlo učinkovitim u rješavanju problema predviđanja onečišćenja zraka tijekom mnogih znanstvenih istraživanja i rada u industriji.[4] Predmet ovog rada, grad Graz, jedno je on najzagađenijih područja Austrije. Za pronalazak optimalno učinkovitog modela predviđanja onečišćenja zraka korištene su različite kombinacije podataka prikupljenih na četiri različite lokacije grada Graza. Na mjernim postajama mjerene su koncentracije dušikova dioksida NO_2 te meteorološki podaci – temperatura, tlak, relativna vlažnost zraka, oborine, brzina i smjer vjetra te radijacija. Slijed izrade modela započinje analizom i predobradom dostupnih podataka, a dalje se nastavlja odabirom algoritma strojnog učenja, razvojem (učenjem) modela, validacijom i testiranjem te konačnom procjenom modela. Svaka faza razvoja modela u ovom radu se provodi pomoću programskog jezika Python.

2. OPĆI DIO

2.1. Onečišćenje zraka

Onečišćenje zraka predstavlja kontaminaciju okoline uzrokovano bilo kojim kemijskim, fizikalnim ili biološkim agensom koji mijenja ili utječe na prirodna obilježja atmosfere. Postoje različiti izvori koji dovode do onečišćenja, ali uglavnom su najučestaliji i najopasniji antropogeni izvori, tj. izvori koji nastaju ljudskim djelovanjem. Do onečišćenja zraka dolazi kada su određeni plinovi te krute ili tekuće čestice suspendirani u zraku.[5] Čestice i plinovi mogu dospjeti u zrak putem ispušnih plinova automobila i ostalih prijevoznih sredstava, emisijom plinova iz proizvodnih pogona poput tvornica, pasivnim dimom cigarete, ali i iz prirodnih izvora.[2] U prirodne izvore onečišćenja zraka ubrajaju se šumski požari, aeroalergeni poput peludi, prašine, erupcije vulkana, mikroorganizmi, meteorska prašina i slično. Na slici 1. prikazani su izvori i emisije onečišćujućih tvari te njihovi štetni učinci kao što su eutrofikacija, zakiseljavanje tla zbog taloženja kiselina koje sadrže veće količine sumpora i dušika te smanjeni poljoprivredni prinosi zbog izloženosti usjeva visokim koncentracijama ozona. Prema podacima Svjetske zdravstvene organizacije neki od najzagađenijih svjetskih gradova su Karachi (Pakistan), New Delhi (Indija), Beijing (Kina), Lima (Peru) i Cairo (Egipat).[6] Osim siromašnijih područja i zemalja u razvoju, onečišćenje je prisutno i u svjetski razvijenim metropolama poput Los Angelesa (California).[6]



Slika 1. Izvori i emisije onečišćujućih tvari te njihovi učinci.[7]

2.1.1. Onečišćujuće tvari u zraku

Postoji mnogo različitih tvari koje onečišćuju zrak, a prema američkoj Agenciji za zaštitu okoliša utvrđeno je šest glavnih pokazatelja onečišćenja zraka koji se nadziru kako se ne bi prekoračile njihove dopuštene koncentracije obzirom na ljudsko zdravlje i/ili okoliš. Glavne onečišćujuće tvari u zraku su ugljikov monoksid (CO), olovo (Pb), prizemni ili fotokemijski ozon (O_3), lebdeće čestice (PM), sumporovi oksidi (SO_x) i dušikovi oksidi (NO_x).[8] Tablica 1. prikazuje dozvoljene granične koncentracije za šest glavnih onečišćujućih komponenti zraka.

Tablica 1. Granične vrijednosti onečišćujućih tvari utvrđene od strane Agencije za zaštitu okoliša.[8]

ONEČIŠĆUJUĆA KOMPONENTA		RAZDOBLJE	GRANIČNE VRIJEDNOSTI
Ugljikov monoksid (CO)		1 sat	35 ppm
Olovo (Pb)		3 mjeseca	0.15 μg/m^3
Prizemni ozon (O_3)		8 sati	0.070 ppm
Lebdeće čestice (PM)	PM $_{2.5}$	1 dan	35 μg/m^3
	PM $_{10}$	1 dan	150 μg/m^3
Sumporov dioksid (SO_2)		1 sat	75 ppb
Dušikov dioksid (NO_2)		1 sat	100 ppb

2.1.1.1. Ugljikov monoksid

Ugljikov monoksid je plin bez boje i mirisa koji u velikim količinama može biti jako štetan i smrtonosan. Većina ugljikova monoksida nastaje nepotpunim sagorijevanjem fosilnih goriva, a najčešći izvori su dimnjaci i plinske peći te automobili i ostala vozila.[9]

2.1.1.2. Olovo

Olovo je najjeftiniji tehnički metal koji se zbog svoje toksičnosti nastoji sve manje upotrebljavati. Glavni izvori olova u zraku su zrakoplovi s klipnim motorom koji se pokreću na olovno gorivo. Velik dio olova u zrak dospijeva obradom rude i metala, iz spalionica otpada te proizvodnjom olovnih baterija. Nakon što je Agencija za zaštitu okoliša regulirala olovo njegovim uklanjanjem iz benzina motornih vozila, razine olova u zraku smanjile su se za 98% u periodu od 1980.g. do 2014.g.[10]

2.1.1.3. Prizemni ozon

Prizemni ozon nastaje kemijskim reakcijama između dušikovih oksida (NO_x) i hlapljivih organskih spojeva u prisutnosti sunčeve svjetlosti, odnosno kemijskim reakcijama onečišćujućih tvari koje emitiraju automobili, elektrane, industrijski kotlovi, rafinerije, kemijska postrojenja i drugi izvori. Kada se čestice zraka kombiniraju s ozonom nastaje smog, vrsta zagađenja koja izgleda kao zadimljena magla.[2] Ponekad planine ili visoke građevine sprječavaju širenje onečišćenog zraka pa je zbog povišene prisutnosti smoga otežan vid.

2.1.1.4. Lebdeće čestice

Pojam lebdećih čestica označava generički izraz koji se koristi za vrstu onečišćujućih tvari u zraku, a sastoji se od mješavine krutih čestica i/ili kapljica tekućine raspršene u zraku. Obzirom na veličinu čestica dijele se na:

1. grube – čestice promjera 10 μm i manje,
2. fine – čestice promjera 2 μm i manje,
3. ultra fine – čestice promjera 100 nm i manje.

Lebdeće čestice nastaju iz širokih raspona prirodnih i antropogenih aktivnosti, a neki od izvora su tvornice, spalionice otpada, motorna vozila, požari te prirodna prašina.[11]

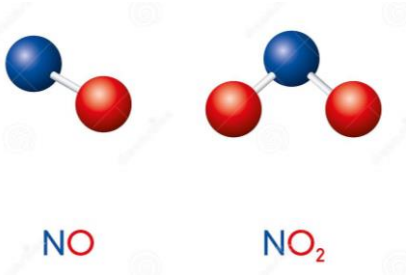
2.1.1.5. Sumporovi oksidi

Skupinu sumporovih oksida čine sumporov monoksid, sumporov dioksid i sumporov (VI) oksid. Uglavnom dolaze iz električnih komunalnih usluga, posebno onih koji sagorijevaju ugljen i iz industrijskih postrojenja koja svoje proizvode crpe iz sirovina kao što su metalne rude, ugljen i sirova nafta ili koje sagorijevaju ugljen i naftu za proizvodnju topline u procesu.[12]

2.1.1.6. Dušikovi oksidi

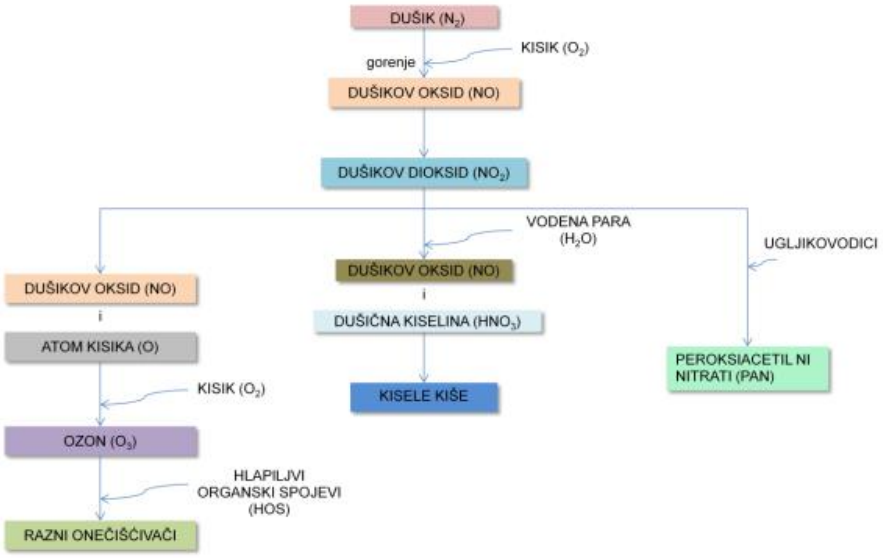
Dušikovi oksidi u atmosferi su prisutni u obliku spojeva dušikova(I) oksida, N_2O, dušikova(II) oksida, NO te dušikova(IV) oksida, NO_2. Većinski dio dušikovih oksida NO_x koji je emitiran u atmosferu je dušikov(II) oksid koji oksidacijom prelazi u dušikov(IV) oksid.[13]

Dušikov dioksid, NO_2, najštetniji je spoj iz skupine dušikovih oksida NO_x. Pri standardnim uvjetima dušikov dioksid je otrovan plin crveno-smeđe boje karakterističnog i nadražljivog mirisa.[14]



Slika 2. Trodimenzionalne strukture dušikova monoksida (lijevo) i dušikova dioksida (desno).[15]

Dušikov dioksid djeluje kao jako oksidacijsko sredstvo, ali isto tako može djelovati kao redukcijsko sredstvo pri čemu se oksidira u nitrate. U vodi se otapa i disproporcionira na dušikov(II) oksid i nitratnu kiselinu, a otapanjem u lužinama daje smjesu nitrita i nitrata.[14] Glavni izvori dušikova dioksida su kemijska industrija, elektrane i promet zbog čega je karakterističan za onečišćenje urbanih područja, a u zrak dospijeva uglavnom sagorijevanjem goriva.[13] Količina dušika u zraku je velika pa u motoru pri visokim temperaturama i tlakovima oksidacijom atmosferskog dušika nastaju dušikovi oksidi. Prvo nastaje dušikov monoksid, a izgaranjem uz suvišak kisika nastaje otrovan dušikov dioksid.[16] Slika 2. prikazuje trodimenzionalne strukture dušikova monoksida i dušikova dioksida. Prisutnost dušikovih oksida u atmosferi je mala, ali imaju značajan utjecaj na okoliš i zdravlje. Jedan su od uzroka nastanka kiselih kiša jer dušikov dioksid fotokemijski reagira s radikalima i stvara dušičnu kiselinu odnosno najvažniju sastavnicu kiselih kiša.[2] Također reagira s hlapljivim organskim spojevima uslijed čega nastaju ozon i smog. U vodenoj okolini uzrokuje eutrofikaciju, tj. smanjuje količinu kisika zbog čega izumiru životinjske vrste. Na slici 3. shematski je prikazano kako nastaje dušikov dioksid i kako dolazi do opasnosti stvaranja prizemnog ozona, kiselih kiša i nitrata.



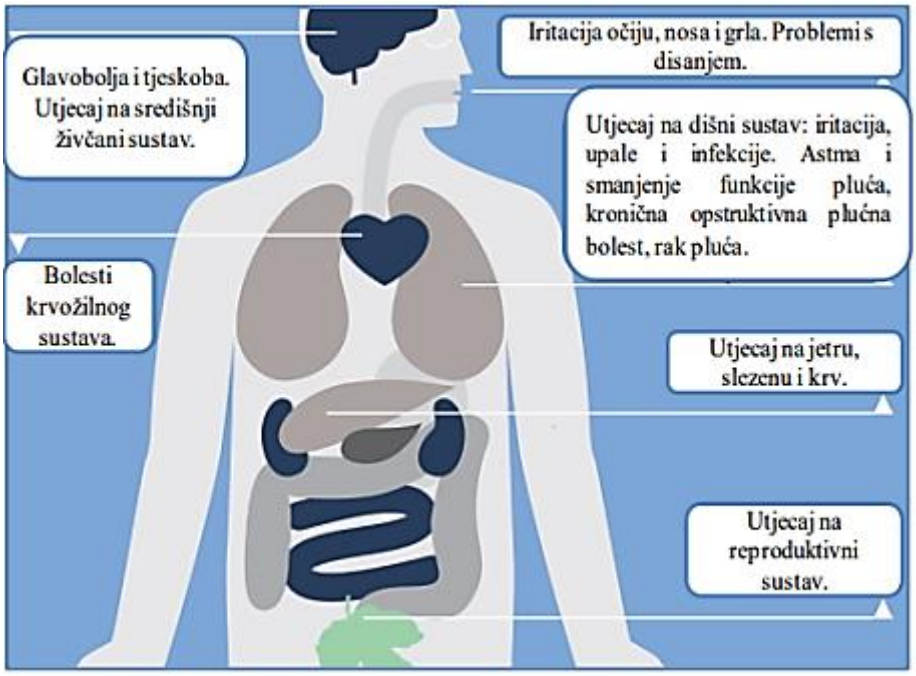
Slika 3. Shema nastanka dušikova dioksida i njegovih negativnih posljedica na ekosustav.[17]

2.1.2. Utjecaj onečišćenja zraka na ljudsko zdravlje

Čovjek bez zraka može izdržati svega nekoliko minuta i kao takav neophodan je za život. Onečišćenje zraka uzrokuje široki raspon zdravstvenih tegoba i bolesti utječući na niz različitih sustava organa. Privremene kratkoročne tegobe uključuju bolesti kao što su upala pluća ili bronhitis te brojne nelagode poput iritacija nosa, grla, očiju ili kože. Ozbiljnija je situacija s kroničnim i dugoročnim bolestima, primjerice s kroničnim srčanim bolestima, rakom pluća, respiratornim infekcijama u djece i kroničnim bronhitisom u odraslih kao i pogoršanjem već postojećih bolesti srca i pluća ili astme čiji ishodi mogu biti smrtonosni.[1] Ljudi dolaze u kontakt s onečišćujućim tvarima iz zraka prvenstveno udisanjem i gutanjem, a nešto rjeđi kontakt ostvaruje se putem kože. Onečišćenje zraka u velikoj mjeri doprinosi onečišćenju hrane i vode čijom konzumacijom štetne tvari dospijevaju u organizam. Brojne studije opisuju da sve vrste onečišćenja zraka, u visokoj koncentraciji, mogu utjecati na dišne putove. Ipak, slični učinci primijećeni su i kod dugotrajne izloženosti nižim koncentracijama onečišćujućih tvari.[1]

Onečišćujuće tvari u zraku razlikuju se po kemijskom sastavu, svojstvima, postojanosti i njihovim učincima na zdravlje ljudi, životinja i okoliša. Na slici 4. sažeto je prikazan utjecaj onečišćenog zraka na ljudsko zdravlje. Plinovite onečišćujuće tvari uglavnom utječu na respiratorni sustav, ali mogu i izazvati hematološke probleme te karcinom.[1] Dioksini, tvari koji nastaju tijekom nepotpunog izgaranja ili spaljivanjem

materijala koji sadrže klor, imaju tendenciju taloženja na tlu i u vodi te preko biljaka ulaze u prehrambeni lanac gdje se bio-akumuliraju. Teški metali, kao elementi u tragovima, neophodni su za održavanje normalnih metaboličkih reakcija, ali pri višim koncentracijama mogu postati toksični. Većina teških metala je opasna jer imaju tendenciju bio-akumulacije odnosno nakupljanja u ljudskom organizmu. Veličina lebdećih čestica određuje mjesto taloženja u respiratornom sustavu. Grube čestice (PM_{10}) talože se uglavnom u gornjim dišnim putovima, dok fine i ultra fine čestice mogu dospjeti u plućne alveole, krvotok i mozak.[1] Istraživanja pokazuju da su fine i ultra fine čestice opasnije od grubih, većih čestica u smislu smrtnosti, kardiovaskularnih i respiratornih učinaka. Vrlo je važno napomenuti kako onečišćujuće tvari u zraku također mogu utjecati na fetus u razvoju pri čemu može doći do spontanog pobačaja te uzrokovanja kognitalnih malformacija i lezija živčanog sustava čime se oštećuju motoričke i kognitivne sposobnosti novorođenčeta.[19]



Slika 4. Utjecaj onečišćenog zraka na ljudski organizam.[20]

2.1.3. Utjecaj onečišćenja zraka na okoliš

Osim na ljude, onečišćenje zraka utječe na cijeli ekosustav. Zagađenje zraka može izravno kontaminirati površinu vode i tla. Na taj način negativno utječe na usjeve čime može smanjiti njihov prinos. Čestice sumporovog dioksida i dušikovog oksida u zraku u kontaktu s vodom i kisikom stvaraju kiselu kišu.[2] Kisela kiša utječe

na promjenu pH tla, narušava kvalitetu vode u rijekama i jezerima, a može uzrokovati i propadanje nekih građevina. Zakiseljavanje tla loše utječe na biljke jer štetne tvari oštećuju korijen biljke i na taj način se unose u ostatak biljke. Kao posljedica se javlja zastoj u rastu biljke, žutilo i prerano otpadanje lišća.[21] U zimskim mjesecima se na površini snijega nakuplja mnogo štetnih tvari iz zraka koje prilikom otapanja snijega dospijevaju u rijeke i jezera te povećavaju kiselost. Narušena kvaliteta vode posljedično može dovesti do pomora riba. Onečišćenje zraka negativno utječe na različite materijale, primjerice na bakrene cijevi, čelične mostove, tekstil, spomenike, a glavna posljedica je njihovo ubrzano trošenje. Prirodno i antropogeno onečišćenje zraka uzrokuje još jedan ekološki fenomen – globalno zatopljenje.[21] Ono se odnosi na porast temperature kojeg djelomično uzrokuju povećanja količine stakleničkih plinova u atmosferi.

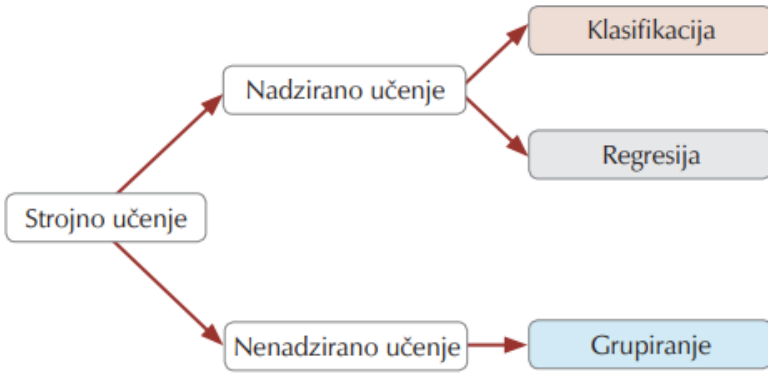
2.2. Strojno učenje

Strojno učenje omogućuje da opsežne i naizgled složene setove podataka jednostavno objasnimo pomoću modela koji uz izvorne podatke obuhvaća značajke koje nisu navedene u izvornim podacima, a bitno utječu na razvoj točnog modela. S obzirom da se za učenje modela koriste velike količine podataka algoritmi mogu biti i jednostavniji, a opet sposobni za uspješna predviđanja. Strojno učenje je proces koji je podijeljen na dvije glavne faze – fazu modeliranja i daljnju analizu modela kojom se definira struktura modela, odnosno bitni čimbenici u okviru istraživanog problema. Algoritmi strojnog učenja mogu se opisati nadziranim i nenadziranim učenjem.

2.2.1. Nadzirano i nenadzirano učenje

Nadzirano učenje izvodi se s ciljem pronalaska funkcije koja najbolje preslikava ulazne podatke ili deskriptivne varijable (neku matricu X) u izlazne podatke ili ciljne varijable (neka matrica Y ili vector y). Kod takvog učenja pod nadzorom modeli se treniraju, tj. razvijaju primjenom dostupnog skupa podataka (deskriptivnih varijabli) prikupljenih primjerice kroz eksperiment kako bi se mogli predvidjeti (ciljne varijable) budući rezultati na temelju poznatih ulaza. Nadzirano učenje primjenjuje se za predviđanje diskretnih odziva (metodama klasifikacije) i kontinuiranih odziva (regresijske metode). Kod diskretnih odziva ulazni podaci se razvrstavaju kategorijski, a kod kontinuiranih se metodama regresije predviđaju kontinuirane varijable.

Nenadzirano učenje ne poznaje ciljnu varijablu čime učenje postaje teže jer algoritam ima općeniti zadatak pronalaska zakonitosti, skrivenih uzoraka i intrinzičnih struktura u skupu podataka. Najčešća tehnika učenja bez nadzora je *klaster-analiza,* odnosno grupiranje gdje algoritam traži skrivene obrasce ili grupe podataka koje su međusobno sličnije. Na slici 5. prikazana je podjela metoda strojnog učenja.



Slika 5. Podjela metoda strojnog učenja.[22]

2.2.2. Odabir algoritma

Svaki algoritam ima individualan pristup učenju iz podataka, a njegov odabir ovisi o količini i vrsti podataka, o konačnom predviđanju te koja je primjena rezultata. Temelj predviđanja svakog modela je inferencija kojom se opisuje veza između deskriptivnih i ciljnih varijabli. Identificiranje svojstava koja su korelirana s ciljnim svojstvom je korisno jer takva svojstva mogu dati uvid u procese koji dovode do fenomena predstavljenog ciljnim svojstvom. Kako bi bilo moguće preslikati skup vrijednosti ulazne varijable u skup vrijednosti ciljne varijable mora postojati odnos između ulazne i ciljne varijable. Ako taj odnos ne postoji, najbolje što model može učiniti je zanemariti ulazne podatke i samo predvidjeti središnju tendenciju željene varijable. Suprotno tome, ako postoje izraženi odnosi, algoritmi strojnog učenja će vjerojatno generirati vrlo točne prediktivne modele.[23] Nakon obavljene korelacijske analize varijabli, algoritmi učenja pod nadzorom sudjeluju u stvaranju modela predviđanja. Neki od popularnih prediktivnih modela su modeli linearne regresije, neuronske mreže i modeli stabala odlučivanja. Regresijska analiza procjenjuje srednju vrijednost ciljnih numeričkih varijabli pod pretpostavkom da su sve ulazne varijable fiksne. Regresijskom analizom može se modelirati i pretpostaviti mnoštvo različitih vrsta odnosa između varijabli. Kada se pretpostavi linearni odnos između

varijabli, regresijska analiza se naziva linearna regresija, a njena najjednostavnija primjena je modeliranje odnosa između dvije varijable: ulazne X i ciljane Y varijable.

2.2.2.1. Neuronske mreže

Neuronske mreže su modeli strojnog učenja koji se koriste za rješavanje mnogih problema strojnog učenja. Neuronska mreža se sastoji od skupa neurona koji su međusobno povezani. Neuron uzima skup numeričkih vrijednosti kao ulaz i preslikava ih u jednu izlaznu vrijednost. Značajna razlika između neurona i linearne regresije je u tome što kod neurona izlaz funkcije višestruke linearne regresije prolazi kroz drugu funkciju, odnosno aktivacijsku funkciju. Postoje klasične unaprijedne mreže koje se dobro nose s velikom količinom podataka i vezama ulaz-izlaz, ali ne uspijevaju pronaći korelaciju među slijednim podacima. Osim njih, postoje i povratne neuronske mreže koje sadrže petlje, a rade na principu da se u određenom trenutku izlaz mreže računa na temelju ulaza mreže i izlaza mreže u prošlom koraku.

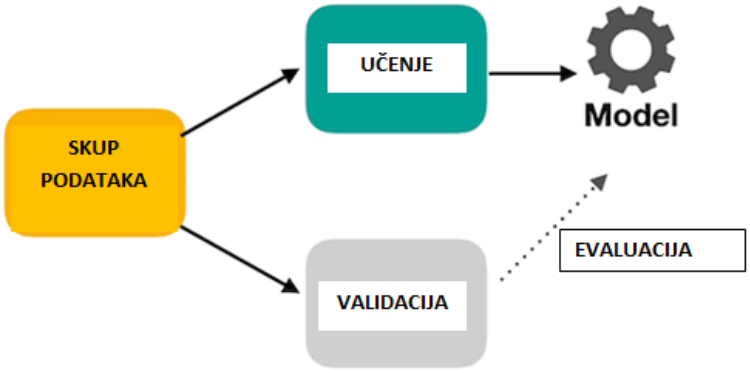
2.2.2.2. Stabla odlučivanja

Stablo odlučivanja kodira i ugrađuje skup *„ako-onda“* pravila u svoju strukturu. Prije donošenja svake odluke stablo odlučivanja provjerava ispunjavaju li varijable zadane uvjete. Svaki čvor u stablu provjerava varijablu, a proces provjere se spušta niz stablo do sljedećeg čvora grane koja pokazuje rezultat provjere vrijednosti varijable u početnom čvoru. Konačna odluka je predstavljena oznakom u konačnom čvoru gdje proces odlučivanja završava. Cilj algoritma za uvježbavanje stabla odlučivanja je pronaći skup pravila klasifikacije koja razvrstavaju ulazni skup podataka za uvježbavanje stabla u skupove s istom vrijednošću ciljne varijable. Jedna od prednosti stabala odlučivanja je to što ih je lako razumjeti i uz njihovu pomoć mogu se razviti napredniji modeli poput modela slučajnih šuma (engl. *random forest model*) koji se sastoje od skupa stabala odlučivanja gdje je svako stablo odluke trenirano na slučajnom poduzorku podataka. Previđanje koje algoritam daje za svaki pojedinačni upit je predviđanje koje daje većinski dio stabala slučajnih šuma.[23]

2.2.3. Vrednovanje modela

Nakon odabira skupa algoritama strojnog učenja, sljedeći korak je vrednovanje odnosno evaluacija razvijenih modela (slika 6). Cilj vrednovanja modela

je ocijeniti vladanje modela na podacima koji nisu primijenjeni za razvoj modela. Glavno pravilo za vrednovanje modela prema tome je da se modeli ne smiju vrednovati na istim podacima na kojima su naučeni. Procjenom vladanja probnih modela na određenom skupu podataka može se odrediti koji algoritam daje najbolji model. Nakon odabira najboljeg algoritma, skupovi podataka za učenje i validaciju mogu se kombinirati u veći skup za učenje, koji se zatim prosljeđuje najboljem algoritmu za izradu konačnog modela. Druga glavna komponenta procesa provjere valjanosti modela je odabir odgovarajućih statističkih pokazatelja. Ako je ciljni atribut numerički, jedna od statističkih mjera valjanosti modela može biti zbroj kvadrata pogrešaka na skupu podataka korištenom za procjenu modela.

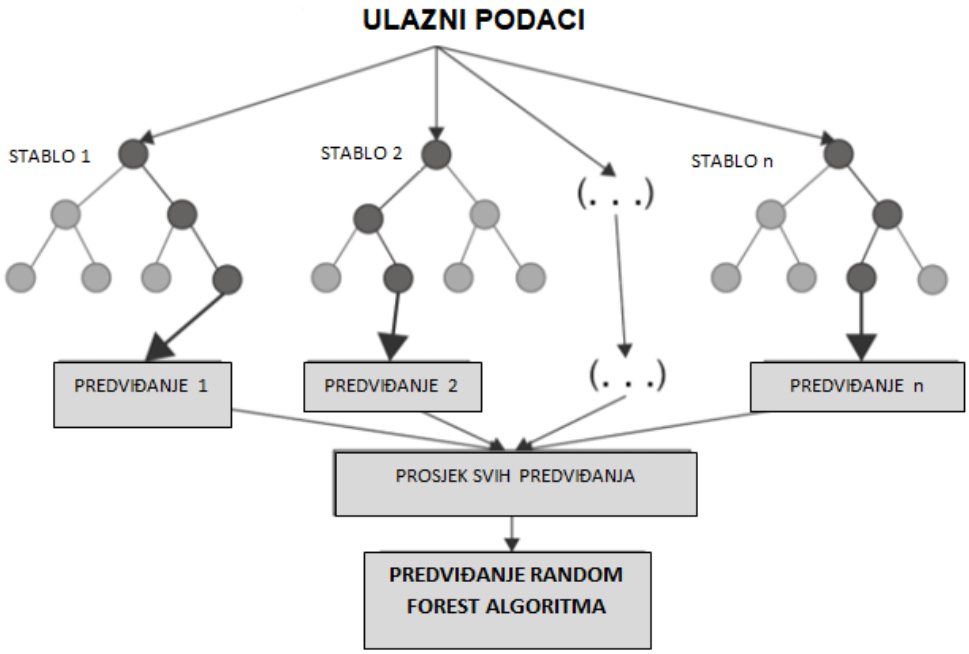


Slika 6. Podjela skupa podataka na podatke za učenje i validaciju.

2.2.4. Algoritam slučajnih šuma (engl. *Random Forest model*)

Algoritam slučajnih šuma je fleksibilan nadzirani algoritam strojnog učenja koji daje vrlo dobre rezultate odnosno predviđanja čak i bez podešavanja hiperparametara, tj. parametara koji određuju model i koji upravljaju složenošću modela. Zbog svoje raznolikosti, jednostavnosti te mogućnosti primjene za klasifikaciju i regresiju, jedan je od najčešće korištenih algoritama. Algoritam slučajnih šuma koristi mnogo stabala odluke, a temelji se na predviđanju prosječnih vrijednosti predikcija svake komponente stabla (slika 7). Općenito ima puno bolju točnost predviđanja od jednog stabla odluke.[23] Kod algoritma slučajnih šuma pri učenju odnosno treniranju podataka svako stablo uči iz nasumce odabranih uzoraka. Stabla odluke su veoma osjetljiva na skup podataka učenja tako da male promjene na tim podacima mogu rezultirati vrlo različitim strukturama stabala. Navedenu osjetljivost

algoritam nasumičnih šuma iskorištava dopuštajući svakom stablu da bude nasumično uzorkovan iz podataka zamjenom što u konačnici daje različita stabla. Pri radu algoritma može doći do problema pretjeranog i nedovoljnog učenja. Pretjerano učenje (engl. *overfitting*) događa se kada funkcija savršeno opisuje podatke s kojima je model razvijan, ali ne predviđa dobro nove skupove podataka. Stoga će na novoj funkciji vrijednost pristranosti modela biti vrlo niska, a novi podaci neće biti točno klasificirani/procijenjeni.[23] Problem pretjeranog učenja rješava se nasumičnim odabirom ciljnih varijjabli prilikom izgradnje stabla odnosno kada postoji čvor, razmatra sve moguće značajke i odabire onu koja daje najveće razlike između uzoraka u lijevom i desnom čvoru. Ovaj proces se izvodi kao algoritam klasteriranja. Ako sva stabla uvijek promatraju iste značajke, bit će vrlo slična. Uzimanje manje značajki način je dodavanja slučajnosti rezultatima i izbjegavanja pretjeranog učenja. Broj zadržanih značajki po stablu je podesivi parametar, a optimalna vrijednost može varirati ovisno o bazi podataka i problemu. Jedna od najvećih prednosti algoritma je njegova svestranost budući da se može koristiti za zadatke regresije i klasifikacije, a također je lako vidljiva relativna važnost koju dodjeljuje ulaznim značajkama. Osim toga, razumijevanje hiperparametara, kojih nema mnogo, prilično je jednostavno. Glavni nedostatak algoritma je spora provedba predviđanja. Veliki broj stabala može učiniti algoritam presporim i neučinkovitim za predviđanje u stvarnom vremenu. Generalno, ovaj algoritam se brzo trenira, ali se predviđanja nakon učenja odvijaju sporo, pogotovo ako se radi o velikom skupu podataka.[23]

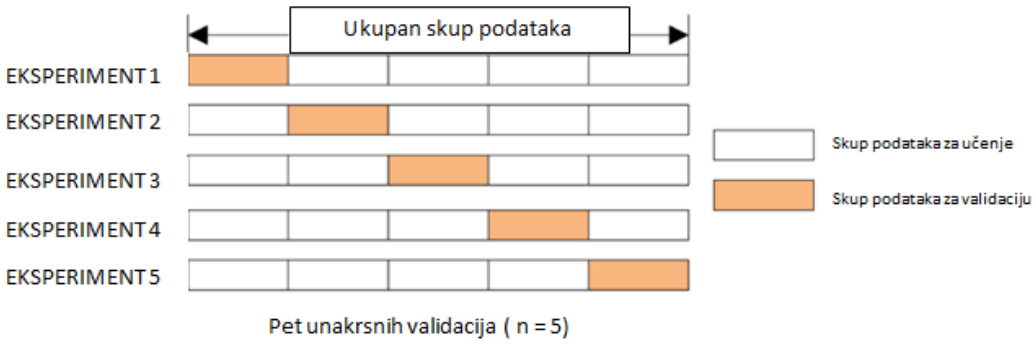


Slika 7. Opći prikaz algoritma slučajne šume.

2.2.4.1. Hiperparametri i unakrsno vrednovanje

Hiperparametri se često nazivaju parametrima jer su to dijelovi strojnog učenja koji se moraju ručno postaviti i podesiti. Razlika između hiperparametara i parametara modela je u tome što se parametri modela procjenjuju iz podataka, a hiperparametri modela se postavljaju ručno i koriste se nadalje za procjenu optimalnih parametara modela.[24] Najbolje hiperparametre obično nije moguće odrediti unaprijed već se njihov odabir bazira na metodi pokušaja i pogreške. Odabir hiperparametara više se oslanja na eksperimentiranje nego na teorijsko predznanje pa je najbolja metoda za određivanje optimalnih hiperparametara isprobavanje njihovih različitih kombinacija za procjenu vladanja svakog od modela. Ako se model optimizira samo prema skupu podataka za učenje tada će model postići vrlo dobre rezultate na podacima za učenje, ali neće moći generalizirati nove podatke.[23] Stoga standardni postupak optimizacije parametara uključuje unakrsnu validaciju, odnosno tehniku provjere valjanosti modela koja dijeli podatke u *n* podskupina, a zatim iterativno uklapa model *n* puta, svaki put trenirajući podatke na *n*-1 podskupina i testirajući na *n*-toj podskupini kao što je prikazano na slici 8. Primjerice, ako će se model validirati s *n* = 5, onda se u prvoj iteraciji model trenira na prve četiri podskupine podataka, a ocjenjuje na petoj. Drugi put se model trenira na prvoj,

drugoj, trećoj i petoj podskupini, a ocjenjuje na četvrtoj. Isti postupak ponavlja se još tri puta, pri tom svaki put procjenjujući na različitoj podskupini. Na kraju se računa prosječno vladanje na svakoj podskupini i dolazi se do konačnih statističkih pokazatelja validacije modela. Kao i kod većine problema u strojnom učenju, odabir parametara može se automatski optimizirati korištenjem metode pretraživanja rešetke (engl. *grid search*).



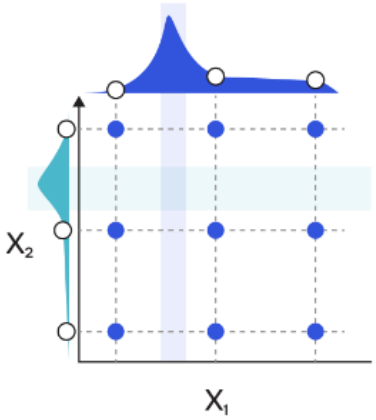
Slika 8. Načelo rada pet unakrsnih validacija.

Najvažnije značajke koje poboljšavaju predviđanje *Random Forest* modela i olakšavaju njegovo učenje su [24] :

a. *n_estimators* – broj stabala odluke u modelu
b. *max_depth* – maksimalna dubina stabla odluke
c. *min_samples_split* – najmanji broj uzoraka potrebnih za podjelu unutarnjeg čvora
d. *min_samples_leaf* – najmanji broj uzoraka potrebnih po čvoru kako bi podjela bila moguća
e. *max_features* – broj značajki koje treba uzeti u obzir pri traženju najbolje podjele
f. *n_jobs* – broj procesa koji se istovremeno obavljaju
g. *random_state* – vrijednost koja olakšava repliciranje rezultata na način da je podjela uvijek ista, odnosno izlaz je uvijek isti nakon svakog pokretanja koda modela

Metoda pretraživanja rešetke je metoda pronalaženja najbolje moguće kombinacije hiperparametara modela na kojoj model postiže najveću točnost

odnosno daje najbolje rezultate predviđanja. Metoda razmatra nekoliko kombinacija hiperparametra i odabire onu kombinaciju uz koju model daje najmanju pogrešku kao što je to prikazano na slici 9. Ova metoda posebno je korisna kada postoji samo nekoliko hiperparametara za optimizaciju, dok se u slučaju složenijih modela koriste ponderirane-slučajne metode.



Slika 9. Prikaz standardne metode pretraživanja rešetke za odabir optimalnih hiperparametara.[25]

2.2.5. Multivarijatne vremenske serije

Vremenska serija je niz podataka varijable od interesa koji su odijeljeni jednakim vremenskim periodima. Pretpostavka vremenske serije je da podaci u seriji dijele neku zajedničku unutarnju strukturu. Multivarijatni procesi su oni procesi kod kojih se tijekom vremena istovremeno promatra nekoliko povezanih procesa vremenskih serija umjesto promatranja samo jedne serije. Multivarijatni procesi vremenskih serija od velikog su interesa za različita područja, primjerice proučavanje istodobnog ponašanja struje i napona ili tlaka, temperature i volumena, ili u ekonomiji analiza varijacije kamatnih stopa, ponude novca, nezaposlenosti itd. Multivarijatne vremenske serije imaju više od jedne varijable koje ovise o vremenu pri čemu svaka varijabla ne ovisi samo o prošlim vrijednostima, već ima i određenu ovisnost o drugim varijablama. Ta se ovisnost koristi za predviđanje budućih vrijednosti. U proučavanju multivarijatnih procesa potreban je okvir za opisivanje ne samo svojstava pojedinih serija, već i korelacija između ispitivanih serija podataka. Svrha analiziranja i modeliranja serije podataka je razumijevanje dinamičkih odnosa tijekom vremena među serijama i dokazivanje točnosti predviđanja za svaku pojedinu seriju.

2.2.5.1. Trend i sezonalnost

Trend vremenske serije predstavlja trajnu, dugoročnu promjenu srednje vrijednosti serije. Trend postoji onda kada postoji dugoročno povećanje ili smanjenje unutar skupa podataka koje ne mora nužno biti linearno. Nakon identifikacije oblika trenda, moguće ga je pokušati modelirati pomoću značajke vremenskog koraka (engl. *time step feature*).

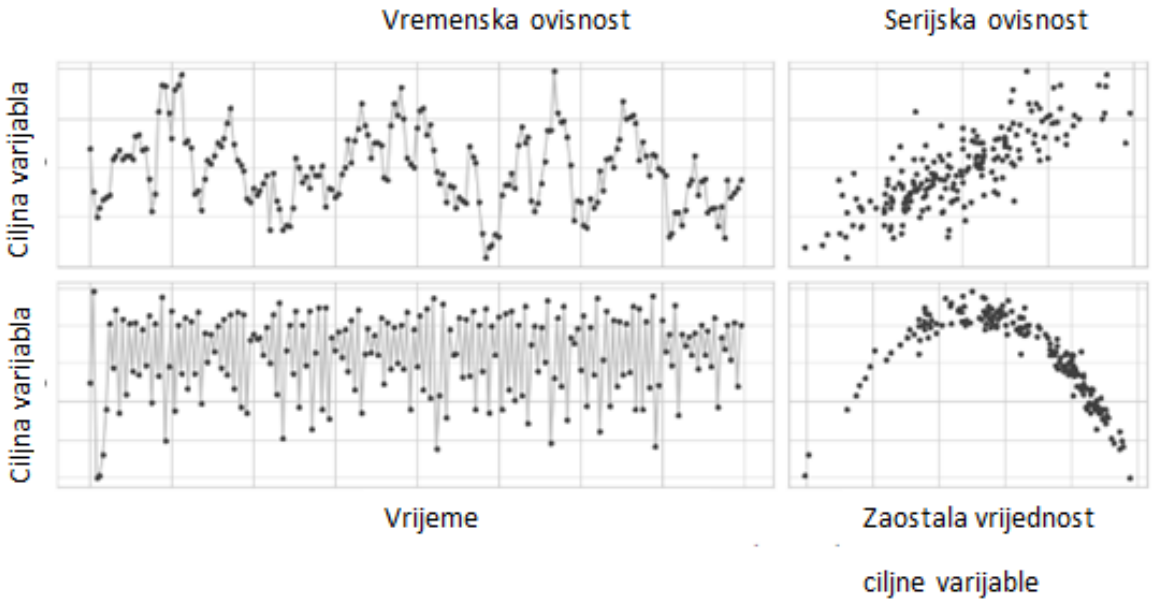
Kaže se da vremenska serija pokazuje sezonalnost kada dođe do redovite, povremene promjene srednje vrijednosti serije. Sezonske promjene uobičajeno slijede određena ponavljanja tijekom dana, tjedna ili godine. Sezonalnost je često potaknuta uobičajenim životnim ciklusima tijekom dana i godina ili konvencijama društvenog ponašanja koje odgovaraju određenom datumu i vremenu. Sezonalnost je uvijek fiksnog i poznatog razdoblja. Pojam ciklusa često se miješa sa pojmom sezonalnosti, ali oni su prilično različiti. Ciklus nastaje kada uzorak podataka pokazuje fluktuacije koje nisu fiksne, a ako je frekvencija nepromjenjiva ili povezana s nekim kalendarskim aspektom, tada je uzorak sezonski.[26] Dakle, ciklusi nisu nužno ovisni o vremenu kao što su primjerice godišnja doba. Ono što se događa u ciklusu manje se odnosi na određeni datum pojavljivanja, a više na ono što se dogodilo u nedavnoj prošlosti. Općenito, prosječna duljina ciklusa je dulja od duljine sezonskog uzorka, a veličine ciklusa obično su promjenjivije od veličina sezonskih uzoraka.

Postoje dvije vrste značajki koje se mogu koristiti za modeliranje sezonalnosti. Tjedne i dnevne pokazatelje najbolje je koristiti za prikaz sezone malog skupa podataka, a Fourierove značajke bolje se odnose na veće skupove podataka poput godišnjih sezona s dnevnim mjerenjima.[26] Sezonski pokazatelji su binarne značajke koje predstavljaju sezonske razlike u razini vremenske serije, a dobivaju se ako sezonsko razdoblje tretiramo kao kategoričku značajku. Fourierove značajke nastoje uhvatiti cjelokupni oblik sezonske krivulje sa samo nekoliko značajki. Ideja Fourierovih značajki je uključiti periodične krivulje u skup podataka za treniranje pri čemu krivulje imaju iste frekvencije kao sezona koju se pokušava modelirati. Krivulje su trigonometrijske funkcije sinusa i kosinusa te ako modeliraju godišnju sezonalnost imaju frekvencije jednom godišnje, dva puta godišnje, tri puta godišnje itd. Prednost Fourierovih značajki u odnosu na sezonske pokazatelje je u tome što je potrebno puno manje značajki kako bi se dobila dobra procjena godišnje sezonalnosti dok

sezonski pokazatelji zahtijevaju stotine značajki (po jedna za svaki dan u godini) čime je vrijeme računanja puno duže, pri čemu može doći do problema pretjeranog učenja.[27]

2.2.5.2. Vremenska serija kao značajka modela

Da bismo istražili moguću serijsku ovisnost, odnosno kontinuiranu ovisnost podataka u vremenskoj seriji, moramo stvoriti "zaostale" vrijednosti vremenskog niza (engl. *lag feature*) kao kopiju vremenske serije. Zaostajanje vremenskog niza znači pomicanje njegovih vrijednosti unaprijed za jedan ili više vremenskih koraka ili unatrag za jedan ili više koraka. Takva metoda naziva se metodom klizećeg prozora, a kao rezultat daje modelu više informacija što često može dovesti do boljih zaključaka.[27] Kada se grafički prikaže zaostala vremenska serija s pravom vremenskom serijom tada je često očita serijska ovisnost gledajući zaostajanje kao što je prikazano na slici 10. Najčešće korištena mjera koja ukazuje na kontinuiranu ovisnost podataka poznata je kao autokorelacija odnosno korelacija koju vremenska serija ima s jednim od svojih zaostalih vrijednosti. Budući da vremenske serije u stvarnom svijetu imaju često značajne nelinearne ovisnosti, moguća je transformacija u linearne odnose odgovarajućim algoritmom.



Slika 10. Usporedba općenitog prikaza vremenske i serijske ovisnosti.[27]

2.2.6. *Prophet* model

Prophet tehnika spada u jednostavan, a učinkovit način predviđanja vremenskih serija. *Prophet* je jednostavan aditivni regresijski model koji sadrži tri glavne komponente – cjeloviti općeniti trend, sezonalnost i praznike.[28] Može se opisati jednadžbom *(1):*

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t) \qquad (1)$$

gdje je *g(t)* funkcija trenda koja modelira linearne ili logističke promjene tijekom vremena, *s(t)* predstavlja periodične promjene u odnosu na povijesne podatke koje mogu biti dnevne, tjedne, mjesečne ili godišnje sezonalnosti, *h(t)* predstavlja predvidljive dane u godini koji se događaju nepravilno i *ε(t)* označava neovisne i identične šumove koji nisu prilagođeni modelu. Za procjenu uspješnosti predviđanja modela koriste se različiti statistički pokazatelji kao što su koeficijent determinacije (R^2), relativna pogreška (*RE*) i kvadratna pogreška srednje vrijednosti korijena (*RMSE*). Što je vrijednost R^2 bliža broju 1 to se predviđene vrijednosti bolje poklapaju s očekivanim točnim vrijednostima izlazne varijable modela. Isto tako, model je bolji što su manje vrijednosti pogreški *RE* i *RMSE*.[28] *Prophet* model daje najbolje rezultate kada se razvija s vremenskim serijama koje imaju snažne sezonske učinke i koje sadrže nekoliko sezona povijesnih podataka.[29] Prednosti *Prophet* modela su brojne, a neke od njih su točnost, brzina, potpuna automatizacija, dobra prilagodba sezonskim varijacijama te robusnost na podatke koji nedostaju.

Ulaz *Prophet* modelu uvijek je podatkovni okvir s dva stupca imenovanih *ds* (engl. *datestamp*) kao vrijeme kolone i *y* kao vrijednosti prediktivne varijable. Kao izlaz *Prophet* modela dobiva se podatkovni okvir s više stupaca, a od najvećeg interesa su sljedeći stupci:

- *ds* – vrijeme predviđene vrijednosti
- *yhat* – predviđena vrijednost mjerne varijable
- *yhat_lower* – donja granica predviđenih vrijednosti
- *yhat_upper* – gornja granica predviđenih vrijednosti

U dostupnim softverima *Prophet* pruža prikladnu funkciju za brzi prikaz rezultata predviđenih vrijednosti, a kod prikazanih rezultata očekuju se varijacije vrijednosti budući da se *Prophet* model oslanja na metodu MCMC (engl. *Markov Chain Monte*

Carlo) odnosno na stohastički proces što rezultira svaki put malo drugačijim vrijednostima predviđanja.[29]

2.2.7. Metoda permutacijske važnosti

U strojnom učenju važno je objasniti zašto se određeni model vlada na određeni način. Jedan od načina objašnjavanja vladanja modela je opisati koje su ulazne značajke modela odabrane i zašto su odabrane. Postoje mnoge metode za odabir ulaznih značajki, ali u ovom radu se usredotočuje samo na metodu permutacijske važnosti. Metodom odabira značajki testiraju se performanse modela nakon uklanjanja svake pojedine značajke koje se pri tom zamjene slučajnim šumom. Na taj se način može izravno usporediti važnost pojedinih značajki. Za određivanje kriterija uklanjanja značajki iz modela može se koristiti određeni kvantitativni prag. Metoda permutacijske važnosti koristi se kako bi se po potrebi smanjio broj ulaznih značajki, a uklanjanjem značajki trebale bi se poboljšati kvaliteta modela i njegova objašnjivost.

2.3. Python

Python je programski jezik pogodan za podučavanje programiranja, kojim se osim proceduralnih metoda mogu savladati i objektno orijentirane metode programiranja. Python jezik je osmišljen tako da ga se može početi koristiti bez svladavanja kompleksnih detalja, te prema daljnjoj potrebi u potpunosti ovladati njegovim sve složenijim značajkama. Pomoću ovog jezika mogu se savladati sva načela proceduralnog i objektno orijentiranog programiranja. Python omogućuje većini početnika da vrlo brzo svladaju jezik u mjeri potrebnoj za rješavanje problema. Mnogi od njih steknu dovoljno znanja i vještina za pisanje naprednih programa i rješavanje praktičnih problema za koje ne postoje gotovi programi ili steknu znanja koja će im pomoći u savladavanju drugih naprednih programskih jezika. Postoji nekoliko razloga zašto je Python postao sve popularniji u posljednjih nekoliko godina:

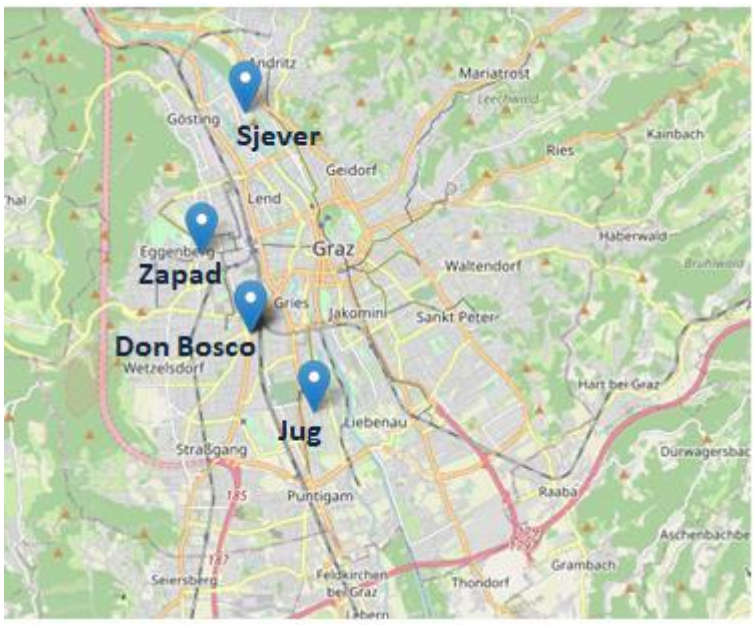
- potpuno je besplatan i lako se instalira na Windows, Linux i Mac OS operativne sustave,
- ima jednostavnu sintaksu,
- prikladan je za upotrebu zbog brze pripreme i ispitivanja programa.

Programski jezik Python je pogodan svima zbog svoje fleksibilnosti i moguće nadogradivosti pa se pomoću njega mogu razvijati jednostavne aplikacije kao i složenija znanstveno-istraživačka predviđanja, statistike i druge analize. Sve izglednija budućnost korištenja Pythona i njegovog razvoja potaknuta je i potrebama “Big Data" aplikacija.

3. MATERIJALI I METODE

3.1. Prikupljanje podataka u Grazu

Grad Graz drugi je grad po broju stanovnika u Austriji i glavni je grad savezne pokrajine Štajerske. Prostire se na površini od oko 127 četvornih metara i nalazi se u nekadašnjem alpskom području.[30] Graz se razvio uz važnu prometnicu između Italije i Panonije, na raskrižju rijeke Mure. Takav raskrižni položaj rano je utjecao na razvoj obrta i trgovine. Slijedeći tu tradiciju, Graz je danas poznat po sajmovima koji se održavaju svakog proljeća i jeseni. Grad Graz industrijsko je i gospodarsko središte južne Austrije. Poznat je po brojnim znamenitostima, zanimljivoj povijesti i bogatoj kulturi. Međutim, odnedavno je dobio pomalo zloglasan nadimak – Grad prašine (engl. „*City of Dust*“).[30] Razlog je taj što se nalazi u dolini između Štajerskih Alpa i Pohorja pa je prirodna cirkulacija zraka u tom području otežana, osobito tijekom hladnijih mjeseci. Kako bismo dobili realnu sliku kvalitete zraka, analizirani su dugoročni mjerni podaci od 1. siječnja 2014. do 15. ožujka 2020. s četiri mjerne postaje u gradu Grazu – Jug, Sjever, Zapad i Don Bosco. Glavni zadatak ove analize jest predviđanje vrijednosti koncentracija dušikovog dioksida uzimajući u obzir utjecaj ostalih mjerenih varijabli koje ukazuju na kvalitetu zraka u Grazu.



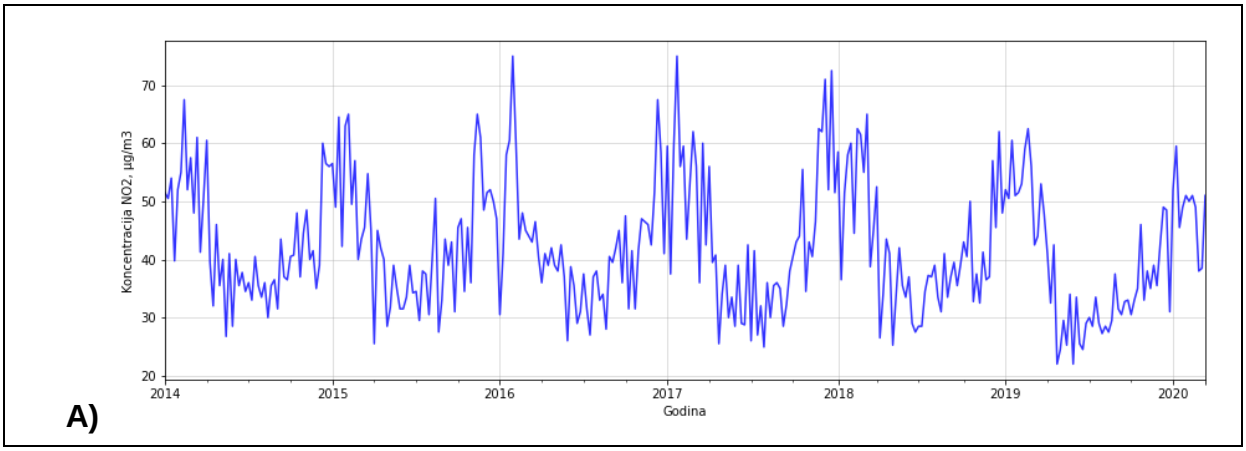
Slika 11. Karta grada Graza koja označava mjerne postaje: Jug – 47.041692º N, 15.433078º E; Sjever – 47.09437º N, 15.415122º E; Zapad – 47.069506º N, 15.403728º E; Don Bosco – 47.055617º N, 15.416539º E

3.1.1. Koncentracije dušikova dioksida

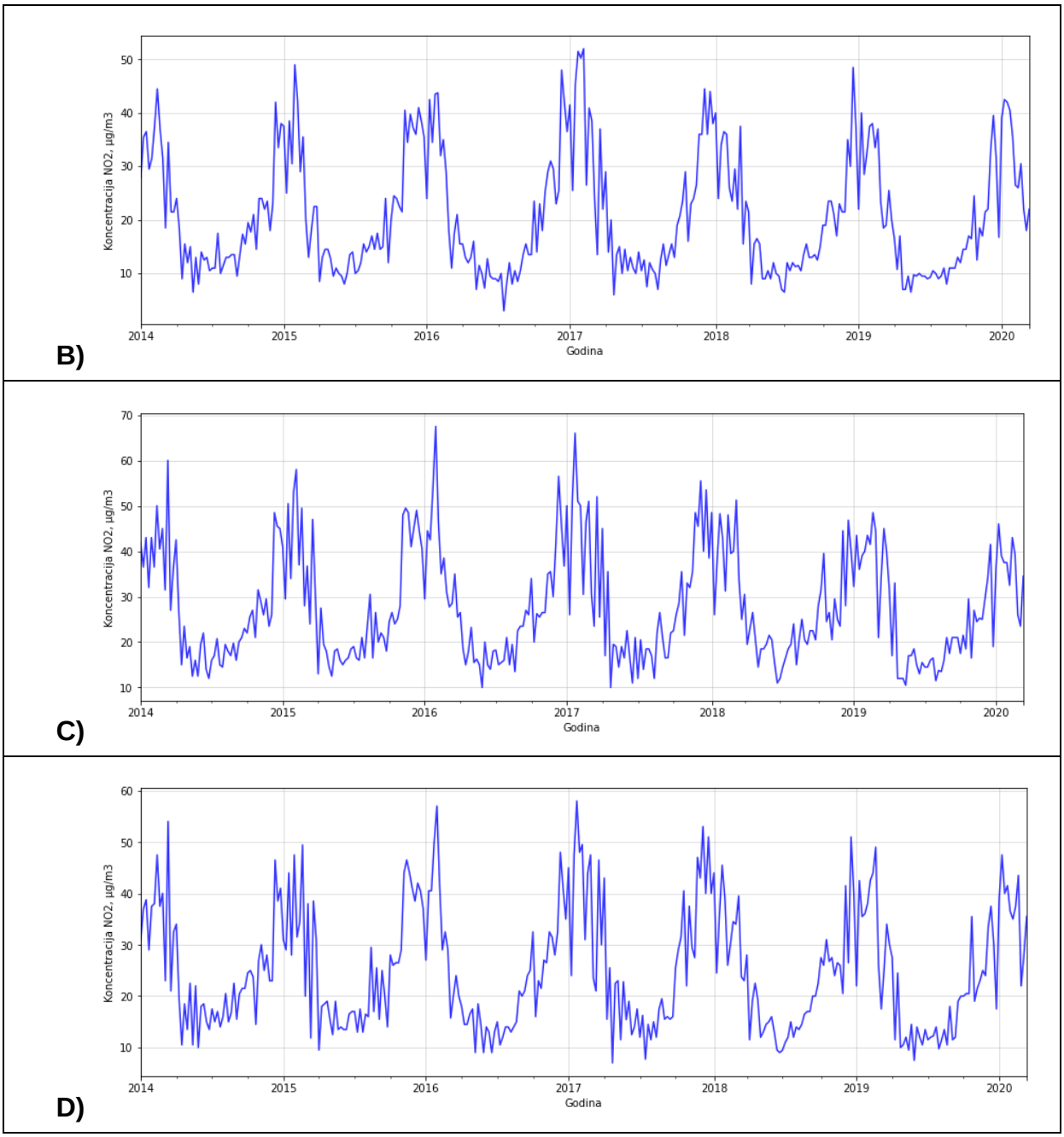
Koncentracije dušikovog dioksida (NO_2) mjerene su na četiri postaje u gradu Grazu. U ovom radu koncentracije NO_2 iskazane su u mikrogramima po kubičnom metru ($\mu g/m^3$). Slika 12. prikazuje tjedne koncentracije NO_2 za promatrano razdoblje od 01.01.2014. do 15.03.2020. mjerene na postajama Don Bosco, Sjever, Jug i Zapad. Prema grafovima, vidljivo je postojanje sezonalnosti odnosno vidi se da koncentracije NO_2 padaju i rastu ovisno o godišnjim dobima. Za vrijeme zimskih mjeseci koncentracije NO_2 na svim postajama imaju više vrijednosti, dok za vrijeme ljetnih mjeseci koncentracije NO_2 pokazuju najniže vrijednosti. Tijekom 2019. godine koncentracije NO_2 najniže su na mjernim postajama Don Bosco, Zapad i Jug, a tijekom 2016. godine na postaji Sjever. U tablici 2. navedene su najviše prosječne dnevne koncentracije NO_2 za svaku mjernu postaju. Zanimljivo je što su najviše vrijednosti koncentracije NO_2 zabilježene istog datuma na mjernim postajama Don Bosco i Jug kao i za mjerne postaje Sjever i Zapad. Prema slici 16. uočava se kako se koncentracije NO_2 povećavaju u vrijeme jutarnjih i večernjih prometnih gužvi, odnosno ujutro između 6 i 10 sati te navečer između 18 i 22 sata.

Tablica 2. Najviše prosječne dnevne koncentracije NO_2 prikazane za svaku postaju.

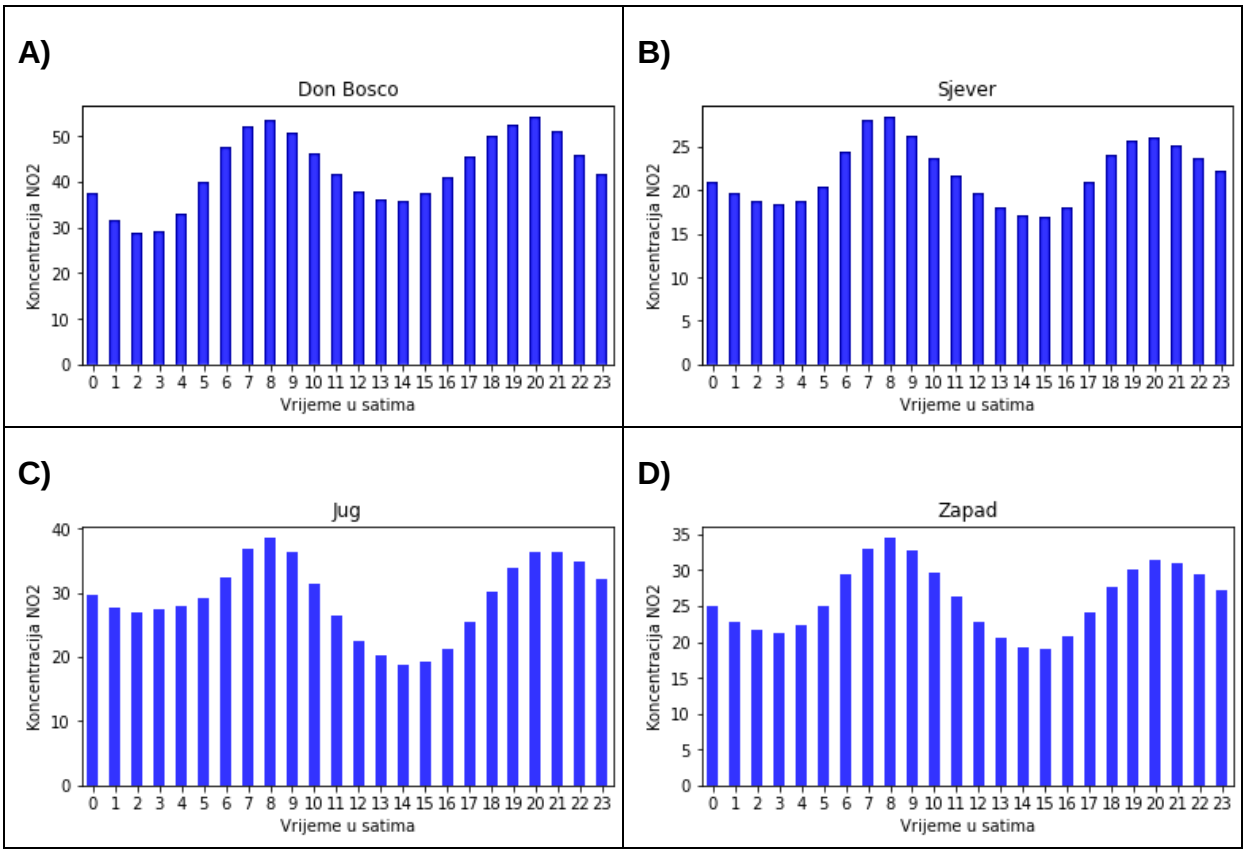
DATUM	MJERNA POSTAJA	NAJVIŠA PROSJEČNA DNEVNA KONCENTRACIJA NO_2 ($\mu g/m^3$)
16.01.2019.	Don Bosco	102,48
16.01.2019.	Jug	83,78
02.06.2014.	Sjever	67,48
02.06.2014.	Zapad	72,39



A)



Slika 12. Koncentracije NO_2 iskazane po tjednima za promatrano razdoblje mjerene na postajama: A) Don Bosco, B) Sjever, C) Jug, D) Zapad.



Slika 13. Koncentracije NO_2 iskazane kao prosjek po satima u promatranom razdoblju za mjerne postaje: A) Don Bosco, B) Sjever, C) Jug, D) Zapad.

3.1.2. Meteorološki podaci

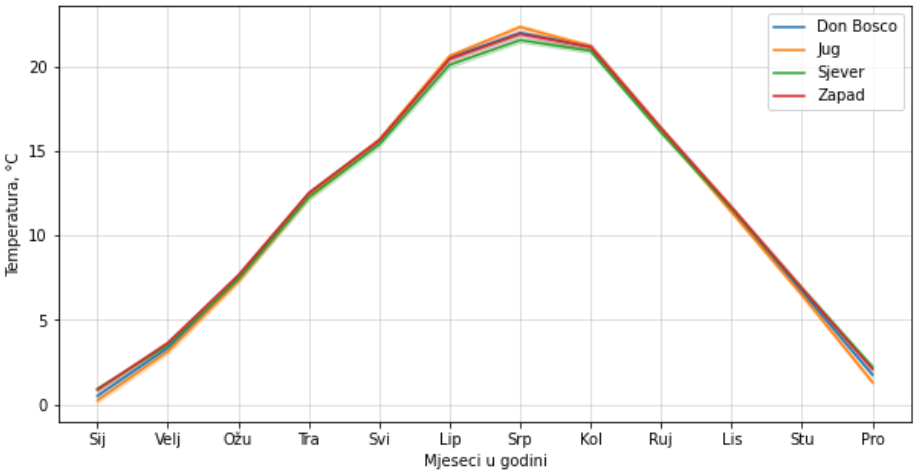
Tablica 3. Meteorološki podaci prikupljeni mjerenjem na četiri postaje u Grazu svakih sat vremena u periodu 01.01.2014. - 15.03.2020. Simbol x označava postojanje mjernog podatka za određenu mjernu postaju.

MJERNI PODATAK	DON BOSCO	JUG	SJEVER	ZAPAD
Temperatura, °C	x	x	x	x
Relativna vlažnost, %	x	x	x	x
Oborine, Lm^2			x	
Radijacija, Sv			x	
Tlak, mbar			x	
Smjer vjetra, deg		x	x	x
Brzina vjetra, ms^{-1}		x	x	x
Max. brzina vjetra, ms^{-1}		x	x	x

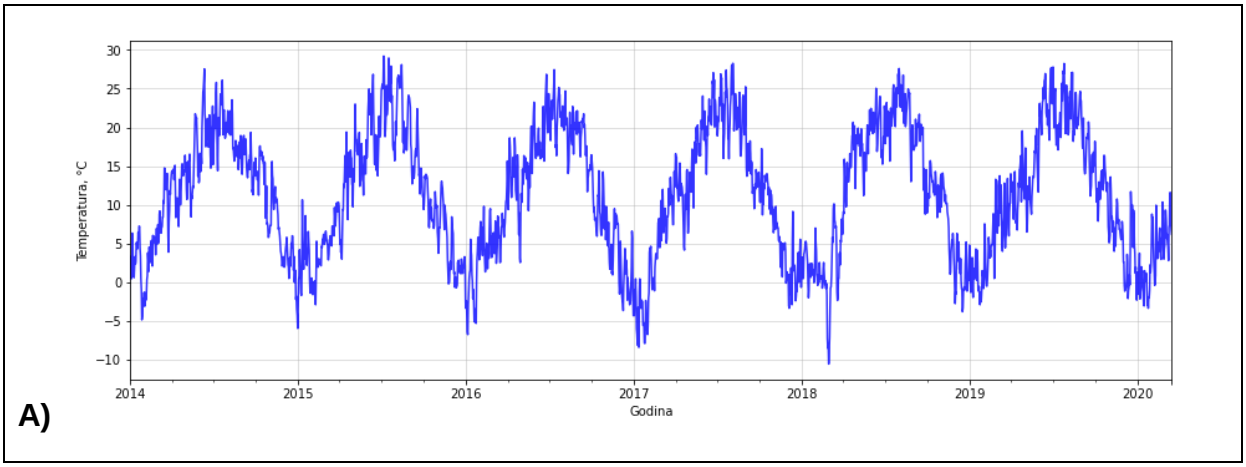
3.1.2.1. Temperatura zraka

Temperatura zraka je temperatura prizemnog sloja atmosfere mjerena na nadmorskoj visini od 2 m zbog utjecaja topline tla. Temperatura zraka varira tijekom dana i godine. Dnevne temperaturne promjene uvelike su uvjetovane dobom dana i

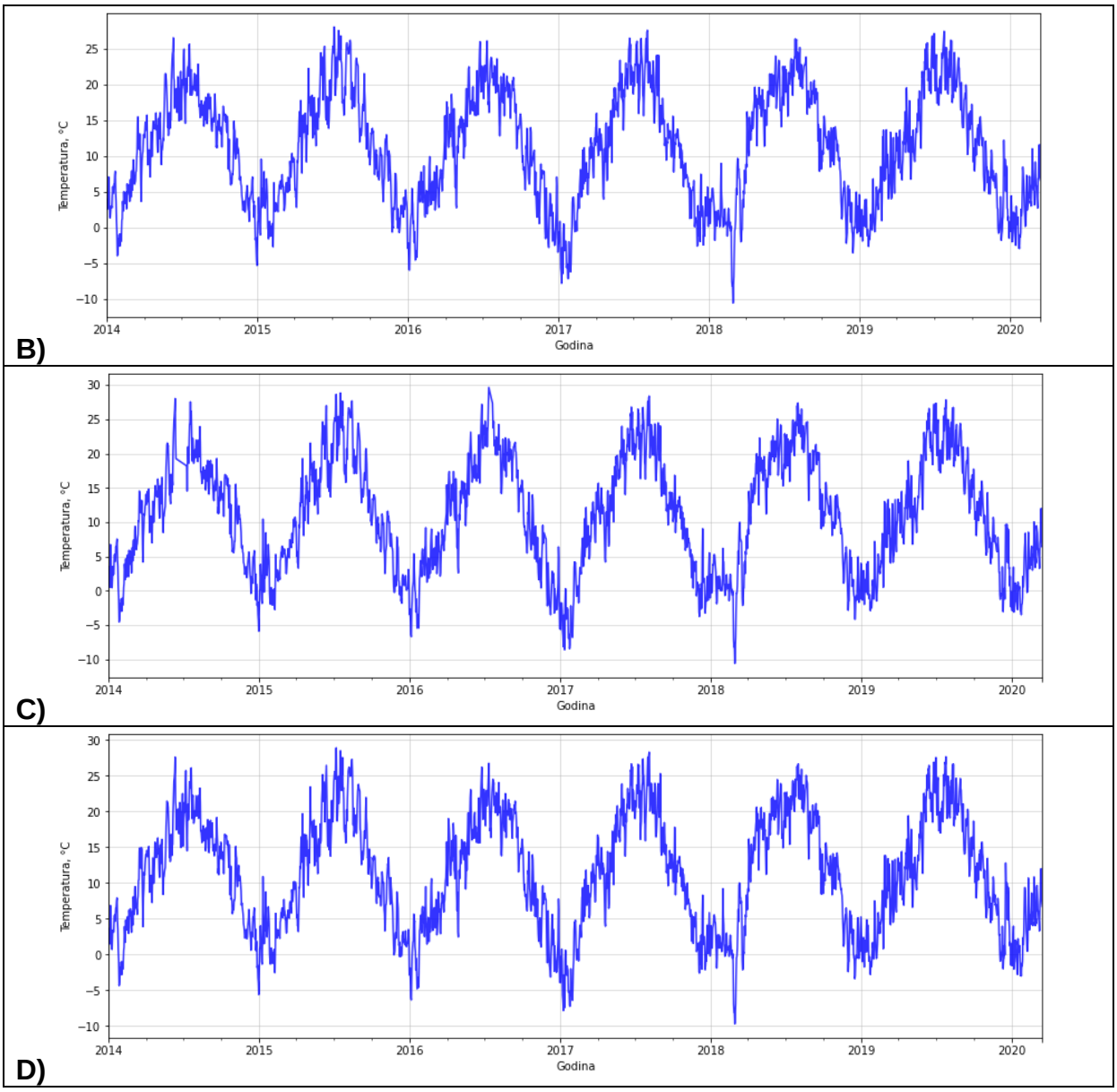
poremećajima koji se javljaju tijekom dana. Godišnji tijek ovisi o položaju Zemlje u odnosu na Sunce i klimatskim promjenama.[18] Prikupljeni podaci o temperaturi zraka u ovom radu izraženi su u Celzijevim stupnjevima i mjereni su na sve četiri postaje u gradu. Budući da mjerne postaje nisu međusobno jako udaljene nema ni velikih razlika u temperaturama između postaja. Prema slici 14. najhladniji mjesec u promatranom razdoblju je siječanj, a najtopliji je mjesec srpanj. Na slici 15. prikazane su temperature iskazane po tjednima za promatrano razdoblje za sve mjerne postaje. Tijekom 2016. godine temperature su nešto malo niže u odnosu na ostale godine, a početkom 2017.g. temperature su najniže obzirom na početak ostalih godina. Također, na svim postajama se tijekom početnih tjedana 2018. godine vidi nagli pad temperature, a zatim nagli skok temperature.



Slika 14. Temperature zraka prikazane po mjesecima za sve četiri mjerne postaje.



A)

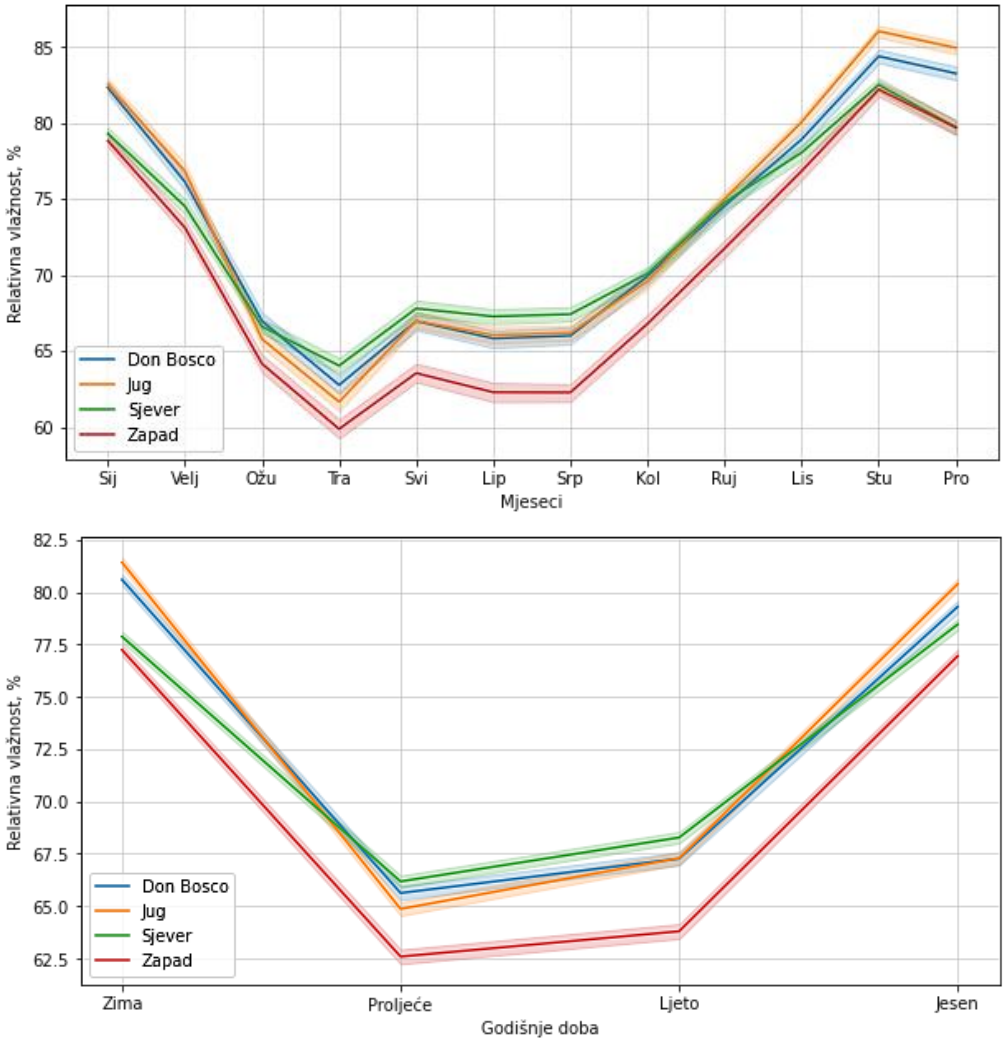


Slika 15. Temperature zraka iskazane po danima za promatrano razdoblje mjerene na postajama: A) Don Bosco, B) Sjever, C) Jug, D) Zapad.

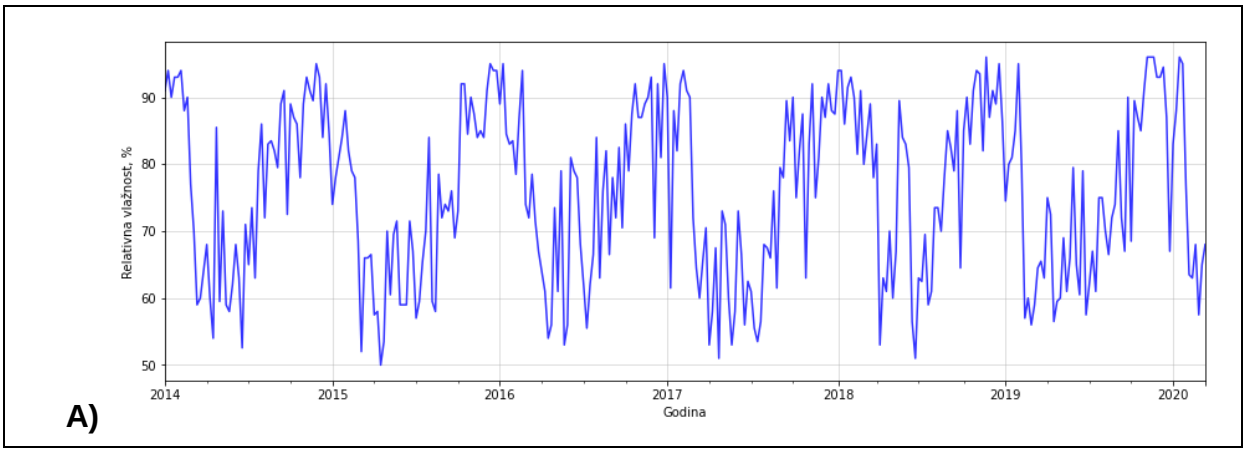
3.1.2.2. Relativna vlažnost zraka

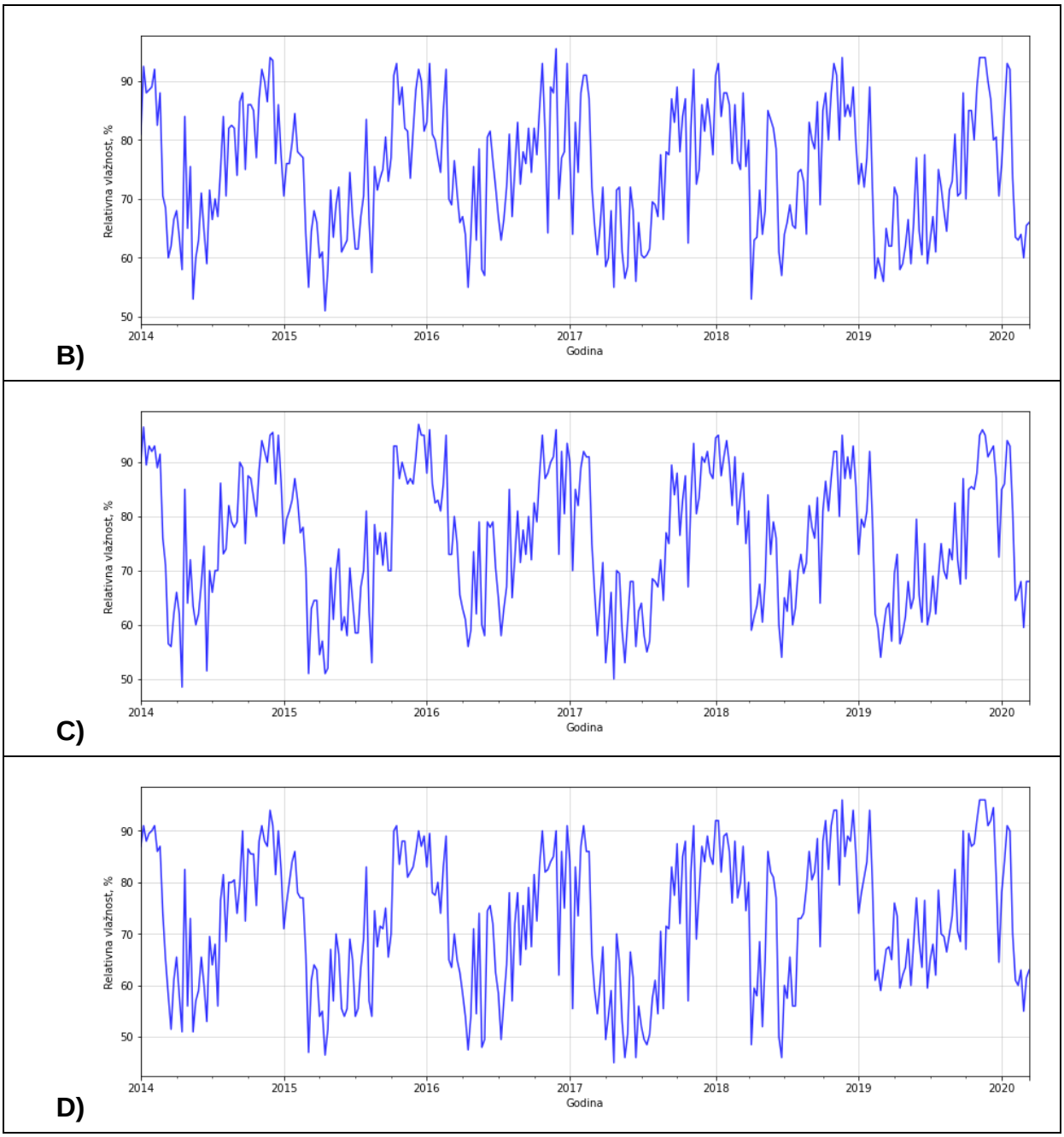
Fizikalna veličina koja se koristi za izražavanje udjela vodene pare u zraku naziva se relativna vlažnost zraka. Predstavlja omjer parcijalnog tlaka vodene pare i parcijalnog tlaka zasićene vodene pare pri određenoj temperaturi i tlaku.[18] Podaci o relativnoj vlažnosti zraka u ovom radu prikupljeni su na sve četiri postaje u promatranom razdoblju i izraženi su u postocima. Povećanu relativnu vlažnost zraka pokazuju sve četiri postaje tijekom zime odnosno najveća relativna vlažnost prisutna je u studenom, a najmanja relativna vlažnost zraka prisutna je tijekom proljeća odnosno u travnju (slika 16). Na slici 17. prikazane su relativne vlažnosti zraka

iskazane po tjednima za promatrano razdoblje na svim postajama. Relativne vlažnosti zraka na svim postajama uglavnom pokazuju jako slične promjene tijekom promatranih godina, a najviše vrijednosti se kreću oko 90%.



Slika 16. Relativna vlažnost zraka prikazana po mjesecima (gore) i godišnjim dobima (dolje) za sve četiri mjerne postaje u promatranom razdoblju.



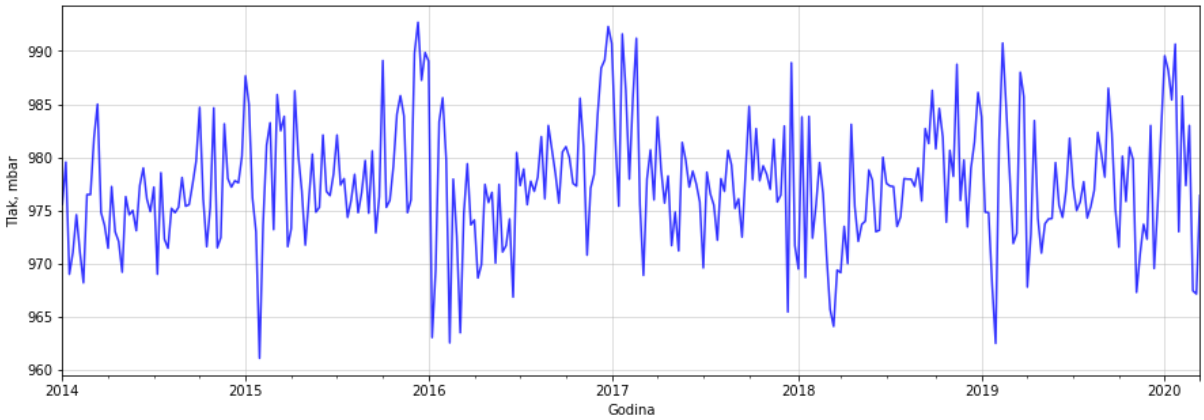


Slika 17. Relativna vlažnost zraka iskazana po tjednima za promatrano razdoblje na mjernim postajama: A) Don Bosco, B) Sjever, C) Jug, D) Zapad.

3.1.2.3. Tlak zraka

Fizikalna veličina koja opisuje djelovanje sile po površini odnosno djelovanje težine zraka na Zemljinu površinu naziva se tlak zraka. Da bismo bolje razumjeli pojam tlaka zraka, možemo zamisliti stupac zraka jediničnog presjeka koji se proteže od tla do vrha atmosfere. Tlak zraka jednak je težini stupca zraka. Zbog razlike u atmosferskom tlaku dolazi do strujanja zraka s jednog mjesta na drugo, a to uzrokuje pojavu vjetra.[18] Podaci o tlaku zraka prikupljeni su samo na mjernoj postaji Sjever i

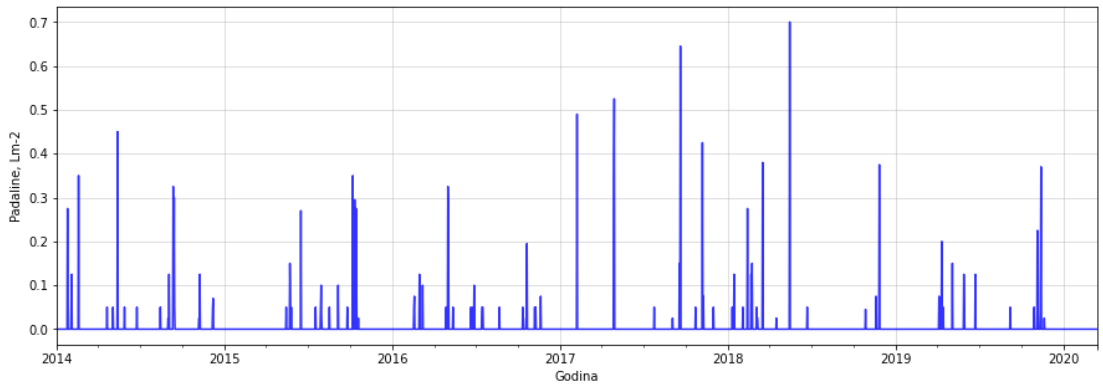
izraženi su u milibarima (slika 18). Krajem 2016. i 2017. godine vrijednosti tlaka dosežu najviše vrijednosti, a početkom 2015. tlak zraka doseže najniže vrijednosti.



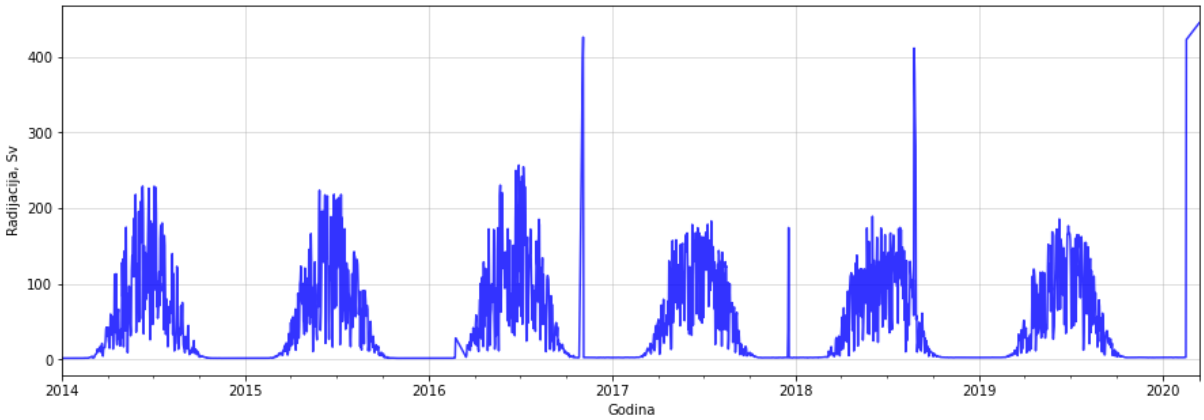
Slika 18. Tlak zraka prikazan po tjednima za promatrano razdoblje mjeren na postaji Sjever.

3.1.2.4. Padaline i radijacija

Padaline definiramo kao tekuću ili čvrstu vodu koja pada iz oblaka na tlo ili nastaje na tlu kondenzacijom vodene pare iz zraka. Padaline su vremenski i prostorno vrlo promjenljive, a mjere se pomoću kišomjera. Radijacija ili zračenje predstavlja prijenos energije putem čestica ili elektromagnetskih valova.[18] U ovom radu padaline su iskazane u litri po kvadratnom metru (Lm^{-2}), a zračenje pomoću mjerne jedinice Sivert (Sv). Slika 19. prikazuje padaline iskazane po danima tijekom promatranog razdoblja za mjernu postaju Sjever gdje se vidi kako je tijekom 2017. i 2018. godine bilo najviše padalina.



Slika 19. Padaline prikazane po danima za promatrano razdoblje mjerene na postaji Sjever.

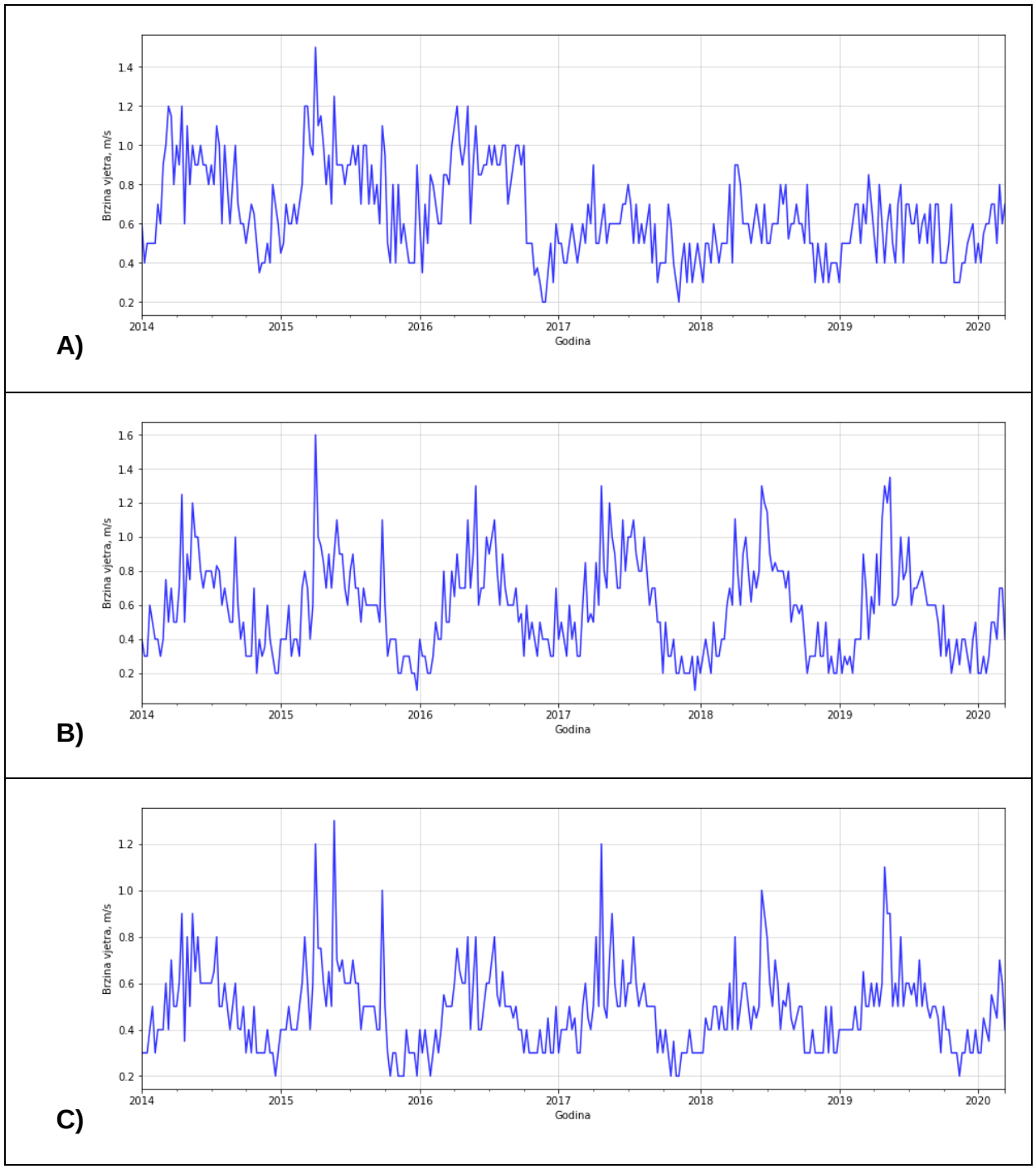


Slika 20. Radijacija prikazana po danima za promatrano razdoblje mjerena na postaji Sjever.

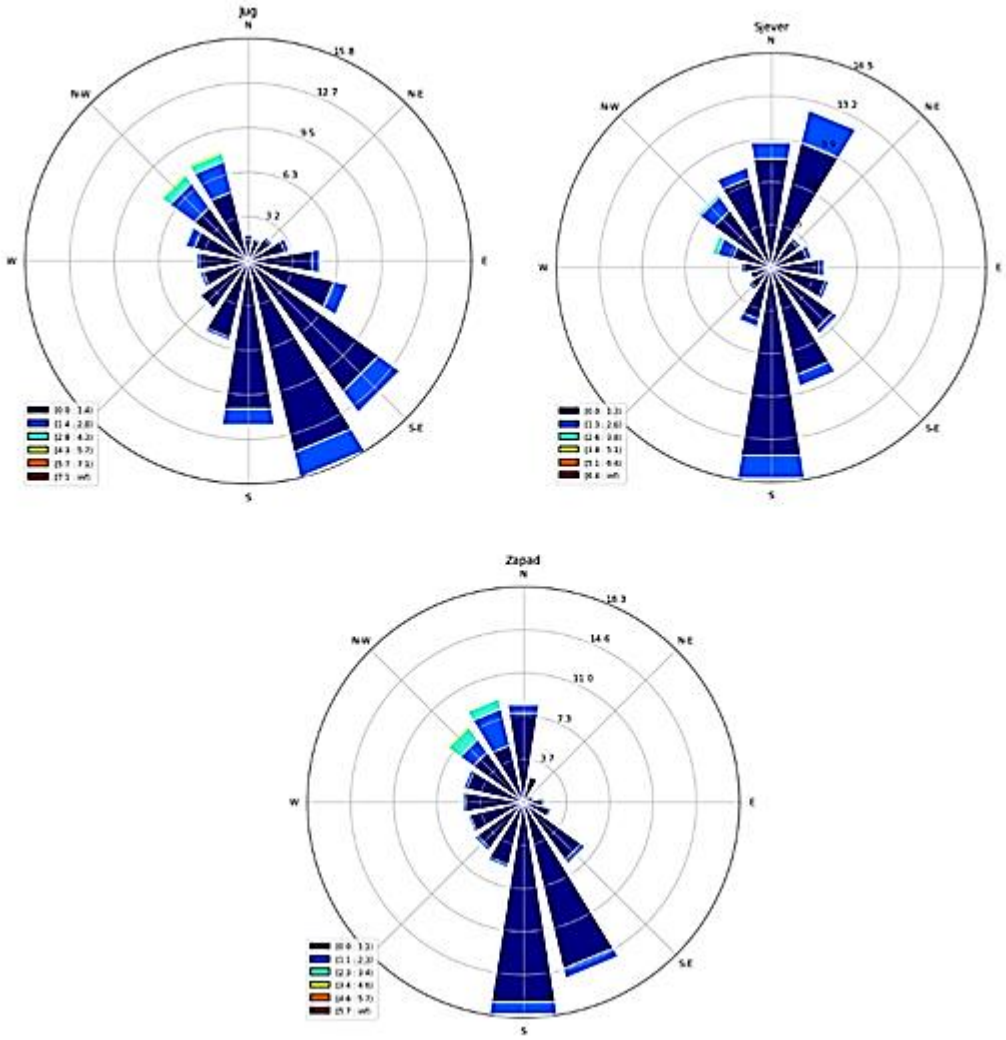
Na slici 20. prikazane su radijacije iskazane po danima na mjernoj postaji Sjever, a uočava se kako je nakon 2016. godine radijacija smanjena. Isto tako, jasno je uočljiva pojava naglog skoka radijacije krajem 2016. godine, krajem 2017. godine kao i sredinom 2018. godine.

3.1.2.5. Vjetar

Vjetar je strujanje zraka koje se kreće vodoravnim smjerom iz područja višeg tlaka u područje nižeg tlaka. Određen je smjerom i brzinom te može nastati iz niza razloga kao što su razlika tlaka između dva područja, Coriolisova sila, rotacija Zemlje ili sila trenja s podlogom. Na strujanje zraka utječu razlike u temperaturi kao i reljefno područje. Brzina vjetra mjeri se anemometrom i u ovom radu iskazana je u metrima po sekundi (ms^{-1}). Smjer vjetra u pravilu ovisi o Zemljinoj rotaciji i razlici atmosferskog tlaka, a ružom vjetrova grafički se prikazuje režim vjetra na određenom mjestu.[18] Prikupljeni podaci o smjeru vjetra iskazani su u stupnjevima (deg). Brzina i smjer vjetra mjereni su na tri mjerne postaje u Grazu – Jug, Sjever i Zapad. Graz nije izložen jakim vjetrovima: prosječne brzine vjetrova za mjerne postaje Jug, Sjever i Zapad iznose redom 0,78 ms^{-1}, 0,76 ms^{-1} i 0,62 ms^{-1}, a najveće izmjerene vrijednosti iznose redom 7,1 ms^{-1}, 6,4 ms^{-1} i 5,7 ms^{-1}. Brzina vjetra mjerena na svim postajama iskazana je po tjednom prosjeku što je prikazano na slici 21. gdje se vidi kako je od 2017. godine na postaji Sjever brzina vjetra smanjena u odnosu na prethodne godine. Na mjernim postajama Jug i Zapad promjena brzine vjetra se kroz godine uglavnom ne razlikuje već se ponegdje javljaju veći naleti brzine vjetra.

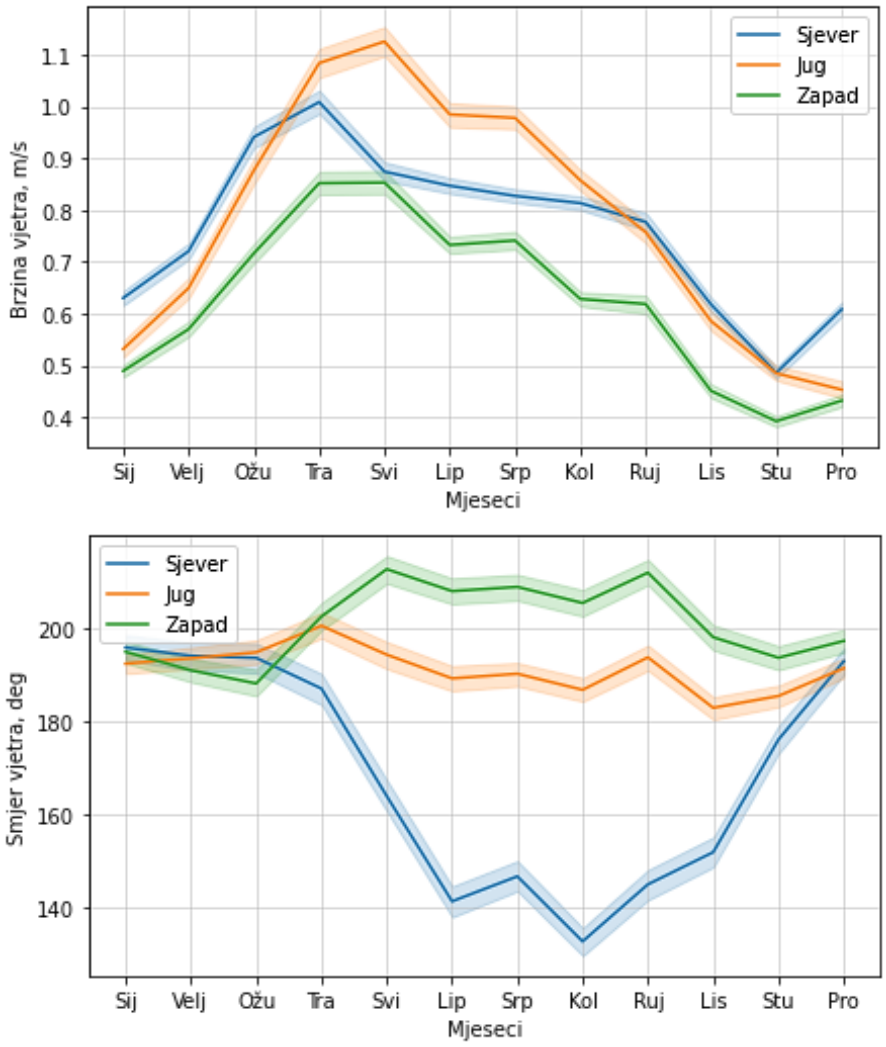


Slika 21. Brzina vjetra prikazana po tjednima za promatrano razdoblje mjerena na postajama: A) Sjever, B) Jug, C) Zapad.



Slika 22. Smjer i brzina vjetra prikazana po satima za promatrano razdoblje mjereni na postajama Jug (gore lijevo), Sjever (gore desno) i Zapad (dolje).

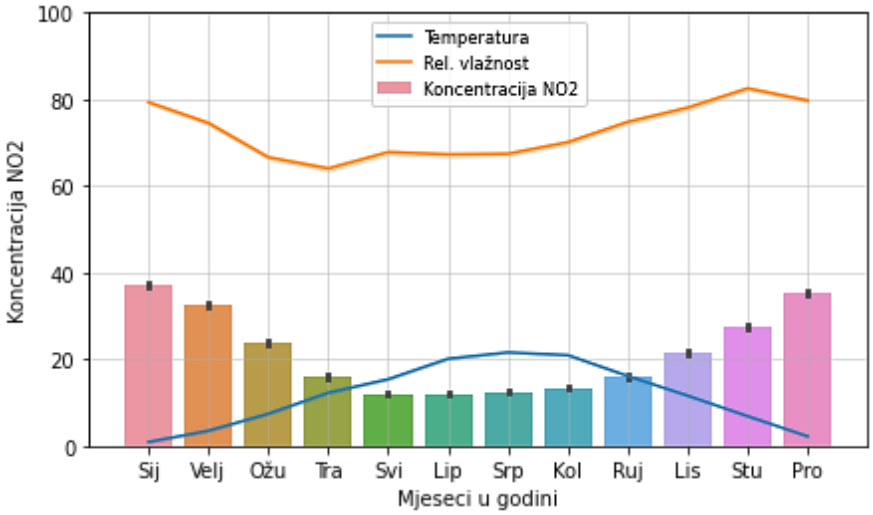
Analiza brzine vjetra u uzastopnim danima u godini pokazala je sezonsku varijabilnost (slika 23). U hladnijim mjesecima brzine vjetra su manje u odnosu na toplije mjesece. Što se tiče smjera puhanja vjetra, na postaji Sjever tijekom ljetnih mjeseci pušu vjetrovi u smjeru jugoistoka, na postaji Jug u južnom smjeru, a na postaji Zapad u smjeru jugozapada. Tijekom zimskih mjeseci na sve tri prethodno spomenute postaje pušu uglavnom slabi vjetrovi u smjeru juga.



Slika 23. Grafički prikazi tijeka brzine vjetra (gore) i smjera vjetra (dolje) po mjesecima u promatranom razdoblju za mjerne postaje Jug, Sjever i Zapad.

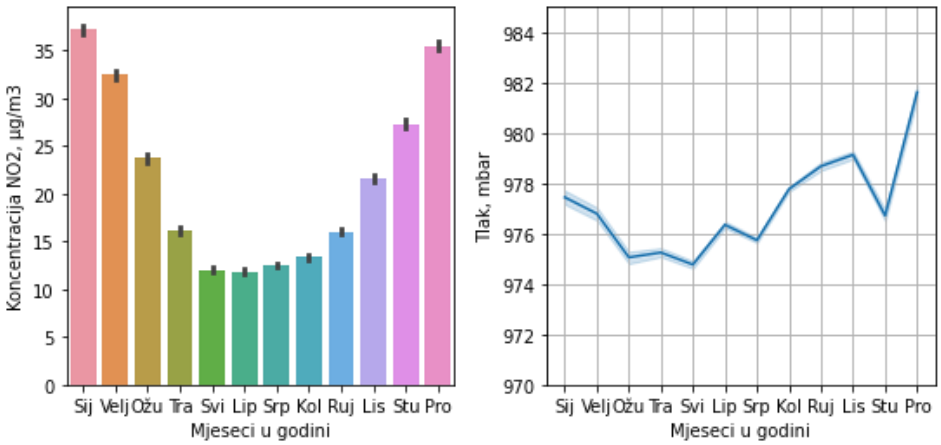
3.2. Utjecaj meteoroloških čimbenika na koncentraciju dušikova dioksida

Koncentracija dušikova dioksida mijenja se tijekom godine, a na to znatno utječu promjene meteoroloških čimbenika. Budući da su svi meteorološki čimbenici mjereni jedino na mjernoj postaji Sjever u nastavku slijede grafički prikazi s postaje Sjever koji predstavljaju općeniti utjecaj meteorologije na koncentraciju NO_2. Na slici 24. promatrajući utjecaj temperature na koncentraciju NO_2 uočeno je da postoji obrnuto proporcionalna ovisnost između temperature i koncentracije NO_2. Hladnije mjesece u godini prati povećanje koncentracije NO_2, dok je u toplijim mjesecima povoljnija situacija jer se koncentracija NO_2 smanjuje. Povećanjem relativne vlažnosti povećava se koncentracija NO_2 i obratno.

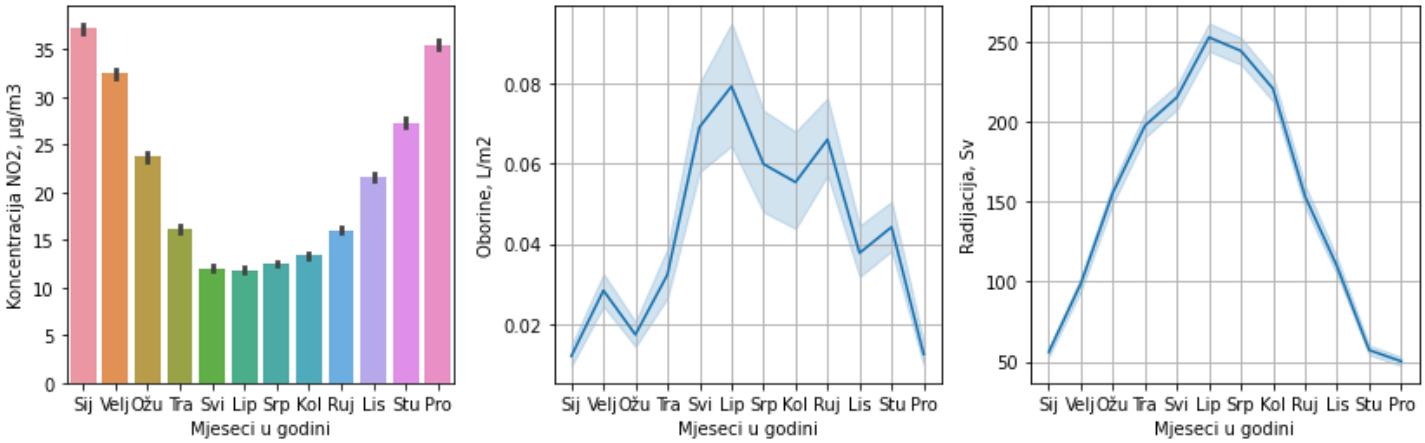


Slika 24. Grafički prikaz utjecaja temperature i relativne vlažnosti na koncentraciju NO_2.

Slika 25. usporedno prikazuje kako promjenu tlaka prati promjena koncentracije NO_2. Generalno, kod većih promjena tlaka koncentracije NO_2 viših su vrijednosti u odnosu na manje promjene tlaka kada su koncentracije NO_2 niže. Promatrajući utjecaj padalina i radijacije vidljivo je kako velike količine padalina i visoke vrijednosti radijacije prate male koncentracije NO_2 i obratno (slika 26).



Slika 25. Usporedba grafičkih prikaza tijeka koncentracije NO_2 (lijevo) i tlaka (desno) po mjesecima u promatranom razdoblju.



Slika 26. Usporedba grafičkih prikaza tijeka koncentracije NO_2 (lijevo), oborina (sredina) i radijacije (desno) po mjesecima u promatranom razdoblju.

4. EKSPERIMENTALNI DIO

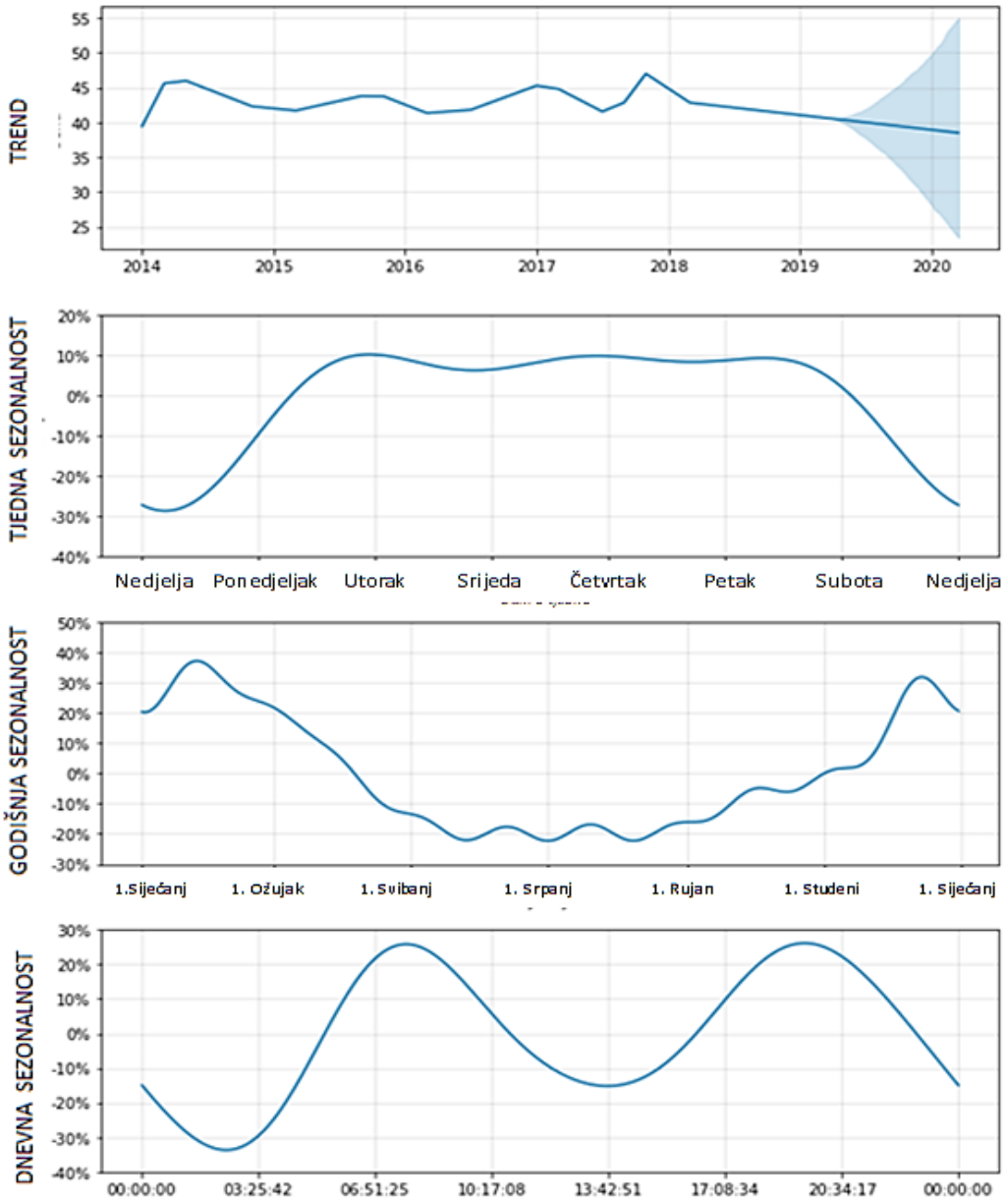
4.1. Modeliranje

Prophet modeli i *Random Forest* modeli, razvijani u ovom radu, pored svoje osnovne namjene obučavani su u svrhu njihove usporedbe kada podaci sadrže i kada ne sadže ekstremne vrijednosti. Kasnije su u radu (poglavlje *5. Rezultati i rasprava*) detaljnije opisane ekstremne vrijednosti za svaku mjernu postaju.

4.1.1. *Prophet* modeli

Izrada *Prophet* modela započinje učitavanjem potrebnih biblioteka i podataka koji sadrže koncentracije NO_2 mjerene svakih sat vremena na četiri mjerne postaje u promatranom razdoblju. Skup podataka se zatim dijeli na dva skupa podataka, gdje jedan služi za treniranje/obučavanje modela, a drugi za testiranje/validiranje modela. Skup podataka za treniranje modela započinje 1. siječnja 2014. u 00:00:00h, a završava 15. ožujka 2019. u 00:00:00h, dok skup podataka za testiranje modela traje od 15. ožujka 2019. u 00:00:00h do 15. ožujka 2020. u 00:00:00h budući da nakon 15. ožujka 2020. započinje „*lockdown*“ u Austriji kao posljedica pandemije SARS-CoV-2 virusa čime se mijenja svakodnevna rutina, a to se bitno odražava na koncentracije NO_2 u zraku. Kao ulaz *Prophet* modelu kreira se podatkovni okvir koji sadrži stupce *'ds'* i *'y'*. Stupac *'ds'* označava vrijeme, a stupac *'y'* odnosi se na vrijednosti ciljne, tj. prediktivne varijable (koncentracija NO_2). Prophet modeli kreirani su za svaku mjernu postaju pojedinačno, ali na isti način i na istim skupovima podataka. Kako su u promatranom skupu podataka uočene godišnje, tjedne i dnevne sezonalnosti iste su uzete u obzir pri izradi *Prophet* modela. Sljedeći korak odnosi se na uklapanje modela (engl. *model fitting*) instanciranjem novog Prophet objekta, a odvija se na skupu za treniranje modela. Pomoću metode *Prophet.make_future_dataframe* dobiven je odgovarajući podatkovni okvir koji se proteže u budućnost 365 dana kako bi se dobila predviđanja koncentracija NO_2 po satima za narednih godinu dana. Metoda predviđanja dodjeljuje svakom retku *'ds'* kolone u budućnosti predviđenu vrijednost koncentracije NO_2 koju imenuje kao *'yhat'.* Stvarne i predviđene vrijednosti kao i minimalne i maksimalne granice predviđenih vrijednosti (intervale nesigurnosti) moguće je grafički prikazati pozivanjem funkcije *Prophet.plot.* Kako bi se predočile komponente predviđanja korištena je metoda

Prophet.plot_components koja omogućuje uočavanje promjena u trendu, dnevnoj, tjednoj i godišnjoj sezonalnosti. Komponente predviđanja za postaju Don Bosco prikazane su slikom 27. gdje se vidi kako u skupu podataka postoji dnevna, tjedna i godišnja sezonalnost na svim mjernim postajama, a uočen je i trend pada koncentracije NO_2 nakon 2018.g. Komponente predviđanja za preostale mjerne postaje dane su u dodatku 1.



Slika 27. Komponente predviđanja *Prophet* modela s uključenim ekstremnim vrijednostima za Don Bosco.

4.1.2. Temporalni podaci

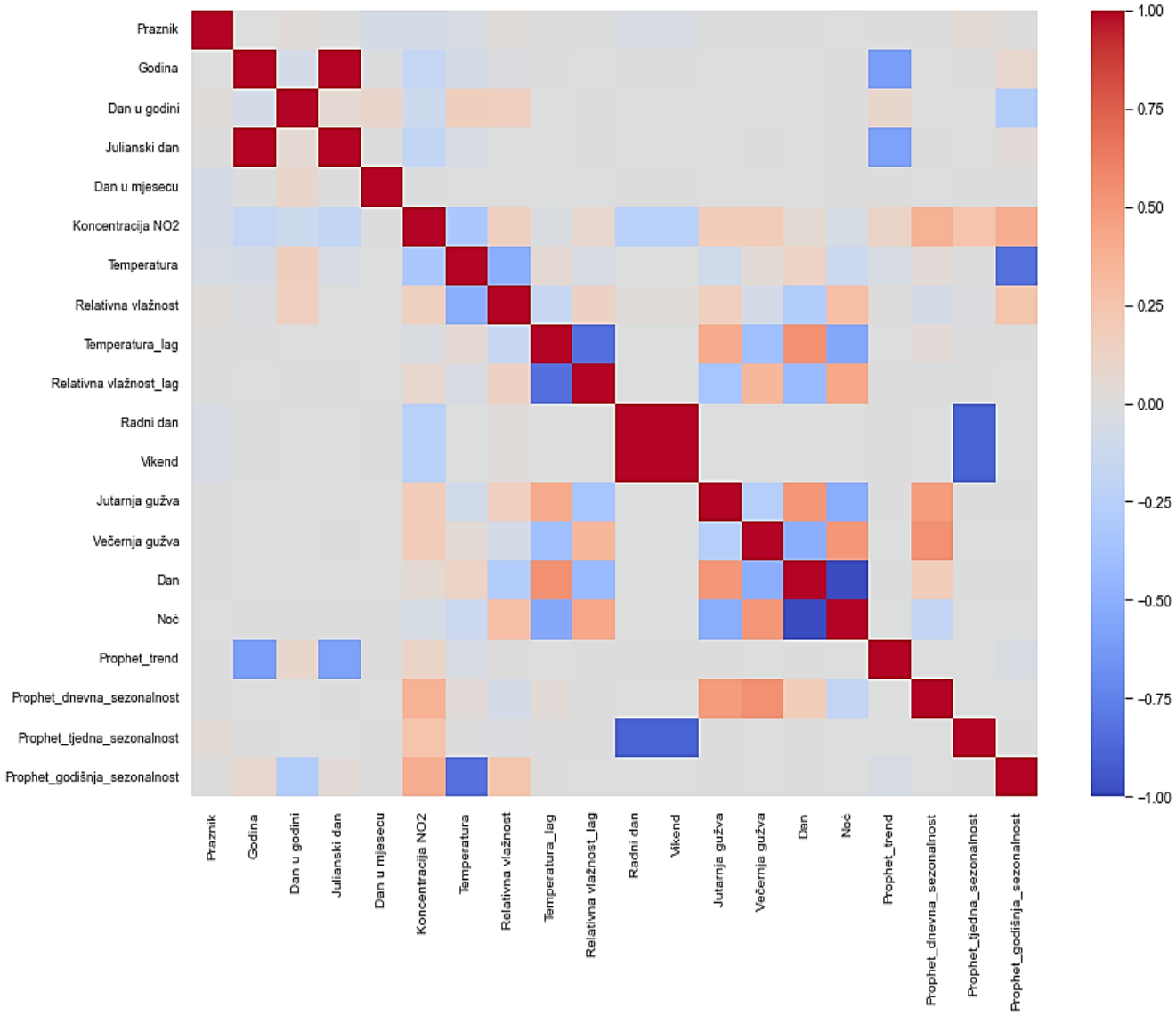
Prethodno je u radu pokazano kako postoje određene dnevne, tjedne, mjesečne, godišnje sezonalnosti te kako su temporalni podaci kao primjerice dani u tjednu, godišnja doba i mjeseci povezani s promjenom koncentracije NO_2. Zbog toga su uz mjerene meteorološke podatke uključeni i temporalni podaci u izradi *Random Forest* modela. Temporalni podaci, kao i druge dodatne značajke koje se koriste za razvoj modela, navedeni su u tablici 4. U radu su u obzir uzete i značajke imenovane kao radni dani/vikendi, jutarnje/večernje gužve, dnevno vrijeme/noćno vrijeme. Osim spomenutih značajki kreirane su i *lag* značajke pomoću metode klizećeg prozora. Na taj način raste količina podataka i olakšava modelu da točnije predviđa koncentracije NO_2. *Lag* značajke kreirane su za one meteorološke podatke koje su dostupne za svaku mjernu postaju.

Tablica 4. Značajke koje se uz meteorološke podatke koriste za razvoj *Random Forest* modela.

TIP ZNAČAJKE	ZNAČAJKA	
Temporalne značajke	Praznik	Mjesec
	Godina	Godišnje doba
	Dan u godini	Radni dan
	Julianski dan	Vikend
	Dan u mjesecu	Jutarnja gužva
	Dan u tjednu	Večernja gužva
	Noć	Dan
Značajke vremenske prognoze	Temperatura_lag	Oborine_lag
	Relativna vlažnost_lag	Brzina vjetra_lag
	Tlak_lag	Smjer vjetra_lag
	Radijacija_lag	Maksimalni naleti brzine vjetra_ lag
Prophet značajke	Trend	Tjedna sezonalnost
	Dnevna sezonalnost	Godišnja sezonalnost

Na slici 28. prikazana je toplinska karta međusobnih korelacija značajki korištenih za razvoj *Random Forest* modela za postaju Don Bosco. Crvena boja pokazuje maksimalnu pozitivnu korelaciju (+1.0), a plava boja maksimalnu negativnu korelaciju (-1.0). Smanjenje intenziteta boje označava smanjenje vrijednosti korelacija, odnosno vrijednosti korelacija približavaju se nuli. Budući da je u radu koncentracija NO_2 varijabla od najvećeg interesa jer je prediktivna varijabla, najviše se pozornosti posvećuje korelacijama značajki s koncentracijom NO_2. Na toplinskoj karti nisu

navedeni koeficijenti korelacija iz razloga što ima puno značajki pa bi karta bila nepregledna. . Ipak, pomoću funkcije *df.corrwith(df['Koncentracija_NO2']* dobivene su numeričke vrijednosti korelacija svih značajki s koncentracijom NO_2, a u nastavku su navedene one značajke koje imaju najveće vrijednosti korelacija s koncentracijom NO_2. Temperatura je značajka koja je obrnuto proporcionalna s koncentracijom NO_2 te njena korelacijska vrijednost iznosi -0.331. Relativna vlažnost je pozitivno korelirana s koncentracijom NO_2, a njena vrijednost korelacije iznosi +0.152. Osim meteoroloških značajki, od značajnog utjecaja su i temporalne značajke poput radnog dana/vikenda čija korelacijska vrijednost s koncentracijom NO_2 iznosi -0.246 te jutarnje/večernje gužve s vrijednostima korelacija u odnosu na koncentraciju NO_2 u pozitivnom iznosu +0.200. Vrlo bitan utjecaj na predviđanje koncentracije NO_2 imaju *Prophet* značajke. Sve *Prophet* značajke pokazuju pozitivne korelacije, a iznose redom za trend, dnevnu, tjednu i godišnju sezonalnost: +0.115, +0.373, +0.256, +0.394. Godišnja sezonalnost je značajka koja pokazuje najveću korelaciju s koncentracijom NO_2 stoga ona u ovom radu u svakom razvijenom modelu koji sadrži *Prophet* značajke najviše pridonosi predviđanju koncentracija NO_2.



Slika 28. Toplinska karta međusobnih korelacija značajki korištenih za razvoj *Random Forest* modela za postaju Don Bosco (preostale postaje dane su u dodatku 1).

4.1.3. *Random Forest* modeli

Izrada *Random Forest* modela započinje učitavanjem potrebnih biblioteka i podataka koji sadrže koncentracije NO_2, meteorološke podatke i značajke navedene u tablici 4. Skup podataka se dijeli na skup za obučavanje (1. siječnja 2014. – 15. ožujka 2019.) i validiranje (15. ožujka 2019. – 15. ožujka 2020.). U sljedećem koraku postavlja se hipoteza modela odnosno definira se mreža hiperparametara i njihovih vrijednosti (slika 29). Metodom *GridSearchCV* odabire se kombinacija najboljih hiperparametara na temelju kojih se model dalje uklapa pozivanjem *model.fit*

funkcije. Metoda *permutation_importance* izabire značajke koje najviše pridonose modelu. Nakon dobivenih izabranih značajki opet slijedi korištenje metode *GridSearchCV* kako bi se dobila najbolja kombinacija parametara uzimajući samo u obzir prethodno izabrane značajke. Nadalje, model se uklapa prema dobivenim najboljim parameterima izabranih značajki i konačno se dobivaju predviđene vrijednosti koncentracija NO_2 u periodu od 15.03.2019.g. do 15.03.2020.g.

```
param_mreža = {'max_features': ['auto', 'sqrt', 'log2'],
               'ccp_alpha': [0.1,  0.01,  0.001],
               'max_depth' : [6, 7, 8],
               'min_samples_split' : [2, 3, 4],
               'min_samples_leaf' : [3, 4, 5],
               'n_estimators' : [100, 150, 200],
               'random_state': [42]}
```

Slika 29. Prikaz mreže hiperparametara i njihovih potencijalnih vrijednosti.

Na opisani način razvijena su četiri tipa *Random Forest* modela za svaku postaju kako bi se modeli međusobno usporedili s ciljem dobivanja točnijih prediktivnih vrijednosti, tj. koncentracija NO_2. Tipovi *Random Forest* modela razvijeni u ovom radu:

1. *Random Forest* model s uključenim ekstremnim vrijednostima i bez značajki *Prophet* modela
2. *Random Forest* model sa isključenim ekstremnim vrijednostima i bez značajki *Prophet* modela
3. *Random Forest* model s uključenim ekstremnim vrijednostima i sa značajkama *Prophet* modela
4. *Random Forest* model sa isključenim ekstremnim vrijednostima i sa značajkama *Prophet* modela

Kako bi se prikazale uspješnosti predviđanja modela korišteni su statistički pokazatelji poput koeficijenta determinacije (R^2), prosječne apsolutne pogreške (*MAE*) i pogreška srednjeg kvadrata (*RMSE*). Formule navedenih pokazatelja prikazane su matematičkim izrazima *(2), (3), (4).* Model je uspješniji što je vrijednost R^2 bliža jedinici te što je manja pogreška odnosno što su manje vrijednosti MAE i RMSE.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

n – broj predviđenih vrijednosti

$\hat{y}_i$ – predviđena vrijednost

$\bar{y}$ – stvarna prosječna vrijednost

y_i – stvarna vrijednost

5. REZULTATI I RASPRAVA

Baza podataka sadrži 71928 uzoraka mjerenih od 1. siječnja 2014. do 17. ožujka 2022., međutim modeliranje se razvija na podacima do 15. ožujka 2020. odnosno analizirana baza podataka sadrži 54360 uzoraka. Provjerom vrijednosti prikupljenih podataka uočen je nedostatak određenog broja podataka, točnije 16737 podataka. Umjesto odbacivanja takvih uzoraka, nedostatak podataka popunjen je interpolacijom prethodne i iduće u seriji validne vrijednosti.

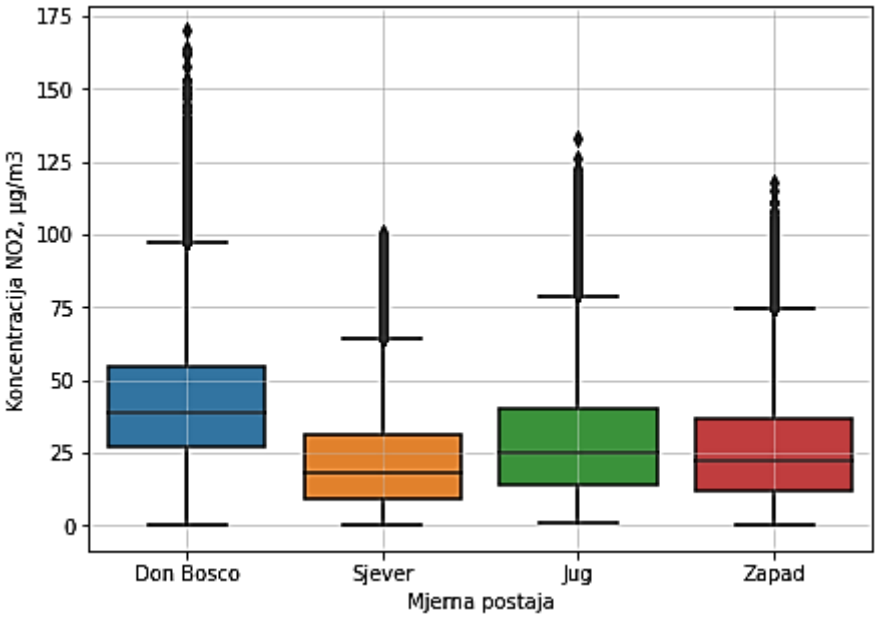
5.1. Generalna statistika mjerenih koncentracija NO_2

Dinamički opseg je razlika između maksimalne i minimalne vrijednosti podataka. Iskazan je za koncentraciju NO_2 budući da je ona ciljna varijabla, a iznosi za postaju Don Bosco 170 $\mu g/m^3$, 101 $\mu g/m^3$ za Sjever, 133 $\mu g/m^3$ za Jug i 118 $\mu g/m^3$ za Zapad. Interkvartilni opseg predstavlja razliku 25. i 75. percentila, a za postaje Don Bosco, Sjever, Jug, Zapad iznosi redom 28 $\mu g/m^3$, 22 $\mu g/m^3$, 27 $\mu g/m^3$, 25 $\mu g/m^3$. Gornju granicu interkvartilnog opsega označava vrijednost 75. percentila odnosno vrijednost iznad koje se nalazi 75% podataka, a donju granicu označava vrijednost 25. percentila odnosno vrijednost ispod koje se nalazi 25% podataka. Prosječne vrijednosti koncentracija NO_2 za promatrano razdoblje na postajama Don Bosco, Sjever, Jug, Zapad iznose redom 42,56 $\mu g/m^3$, 21,97 $\mu g/m^3$, 29,22 $\mu g/m^3$, 26,11 $\mu g/m^3$. Prethodno opisane vrijednosti koncentracija NO_2 prikazane su u tablici 5., a njihova vizualizacija grafički je prikazana pomoću *boxplot* dijagrama (slika 30). Uočeno je i postojanje ekstremnih vrijednosti (engl. *outliers*) koji se ne ubrajaju u predviđene opsege. Na slici 30. vidljivo je kako mjerna postaja Don Bosco ima najveći dinamički opseg vrijednosti koncentracija NO_2, a Sjever ima najmanji dinamički i interkvartilni opseg vrijednosti. Mjerna postaja Don Bosco pokazuje najviše vrijednosti koncentracija NO_2 zbog čega je najzagađenije mjerno mjesto u Grazu. Tomu pridonose velike količine prometa, osobito tijekom jutarnjih i večernjih gužvi. Radnim danima povećane su koncentracije NO_2 u odnosu na vikend čime se dodatno potvrđuje kako veći protok prometa utječe na povećanje koncentracija NO_2. Osim prometnog utjecaja, emisije iz obližnjeg industrijskog pogona također uvelike utječu na zagađenje Don Bosca. Mjerna postaja Jug nalazi se na sekundarnom cestovnom segmentu, ali također bilježi veće koncentracije onečišćujućih tvari zbog industrijskog kompleksa u blizini. Mjerne postaje Sjever i Zapad klasificirane su kao

urbana mjesta iz predgrađa i nalaze se u blizini manjih cesta bez posebnih doprinosa emisija u neposrednoj blizini pa stoga pokazuju niže koncentracije NO_2 u odnosu na prve dvije mjerne postaje.

Tablica 5. Prikaz srednjih vrijednosti, standardnih devijacija, minimalnih i maksimalnih vrijednosti, 25. percentila, medijana i 75. percentila koncentracija NO_2 za svaku mjernu postaju.

	DON BOSCO, µg/m^3	SJEVER, µg/m^3	JUG, µg/m^3	ZAPAD, µg/m^3
Broj uzoraka	54360	54360	54360	54360
Prosječna vrijednost	42,56	21,97	29,22	26,11
Standardna devijacija	20,81	15,82	19,01	17,37
Minimalna vrijednost	0	0	0	0
25%	27	9	14	12
50%	39	18	25	22
75%	55	31	40	37
Maksimalna vrijednost	170	101	133	118



Slika 30. *Boxplot* dijagrami koncentracija NO_2 u zraku promatranog razdoblja po mjernim postajama.

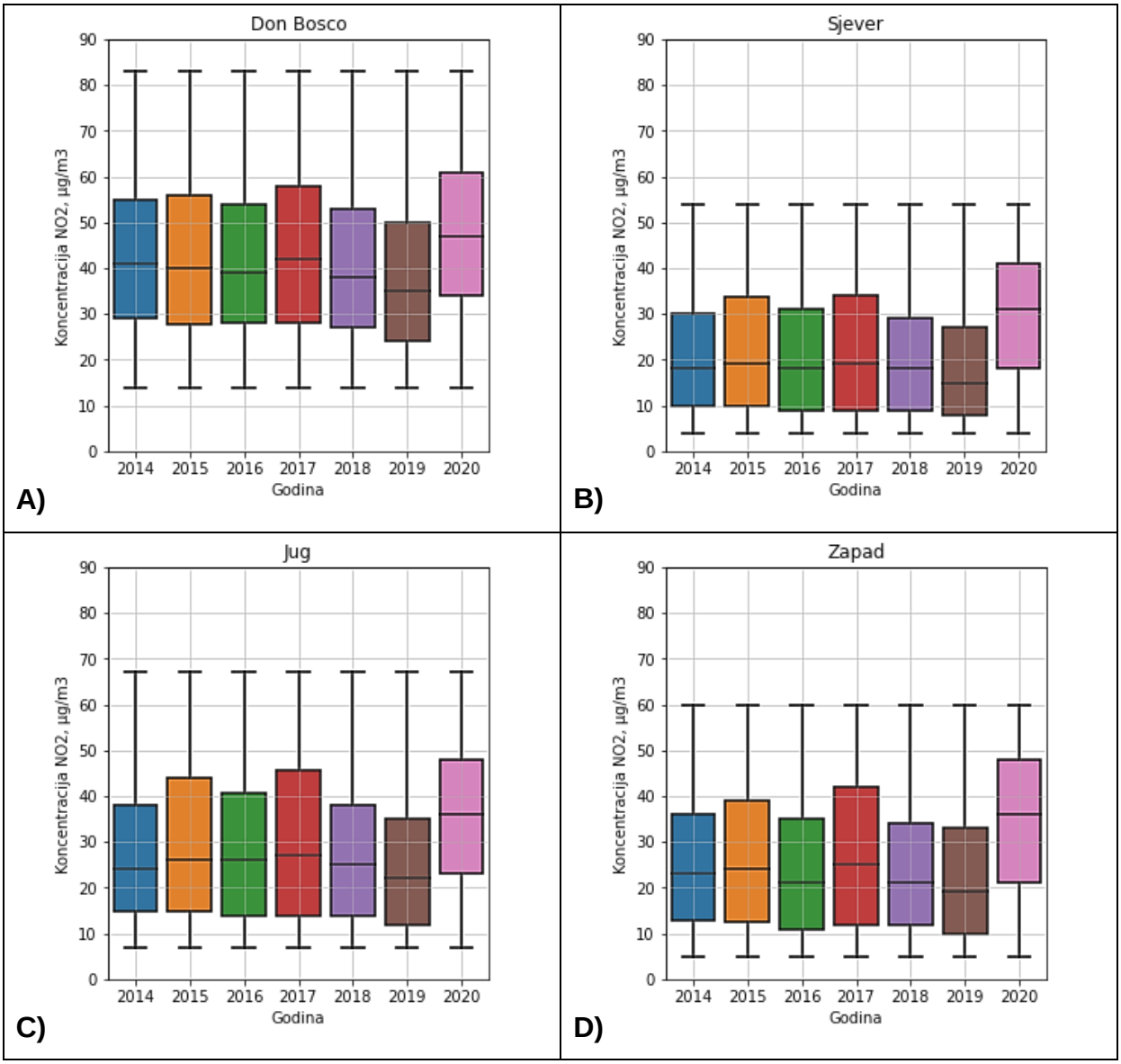
5.2. Pregled vrijednosti koncentracija NO_2 u određenim periodima

Na slici 31. uočava se razlika opsega vrijednosti koncentracija NO_2 2020. godine u odnosu na prethodne godine. To je zbog promatranog razdoblja koje završava u ožujku pa se u obzir uzimaju koncentracije NO_2 samo do ožujka 2020. godine, a time se dobiva porast koncentracije u 2020. godini (slika 32) jer generalno tijekom siječnja, veljače i ožujka koncentracije NO_2 poprimaju visoke vrijednosti. U tablici 6. navedene su vrijednosti koncentracija NO_2 agregirane po godinama promatranog razdoblja za sve četiri mjerne postaje. Iz tablice se uočava kako na svim postajama nakon 2018. godine slijedi pad koncentracije NO_2 (izuzev 2020.godine koja je prethodno objašnjena zašto pokazuje najviše vrijednosti koncentracija NO_2). Mjerna postaja Don Bosco pokazuje najviše vrijednosti koncentracija NO_2 tijekom cjelokupnog promatranog razdoblja zbog čega je najzagađenije mjerno mjesto u Grazu. Tomu pridonose velike količine prometa, osobito tijekom jutarnjih i večernjih gužvi. Osim prometnog utjecaja, emisije iz obližnjeg industrijskog pogona također uvelike utječu na zagađenje Don Bosca. Mjerna postaja Jug nalazi se na sekundarnom cestovnom segmentu, ali također bilježi veće koncentracije onečišćujućih tvari zbog industrijskog kompleksa u blizini. Mjerne postaje Sjever i Zapad klasificirane su kao urbana mjesta iz predgrađa i nalaze se u blizini manjih cesta bez posebnih doprinosa emisija u neposrednoj blizini pa stoga pokazuju niže koncentracije NO_2 u odnosu na prve dvije mjerne postaje. Prethodno je u radu pokazano kako tijekom 2017. i 2018. godine na mjernoj postaji Sjever općenito dolazi do pada vrijednosti relativne vlažnosti zraka, pada tlaka zraka i pada radijacije te do porasta količine padalina. Na slici 31.B) uočava se kako su vrijednosti koncentracija na postaji Sjever tijekom 2017. i 2018. godine smanjene što se može dovesti u direktnu vezu s padom vrijednosti relativne vlažnosti, tlaka zraka i radijacije te porastom količine padalina tijekom istog razdoblja što je pokazano prethodno u radu (*3.1.2. Meteorološki podaci*).

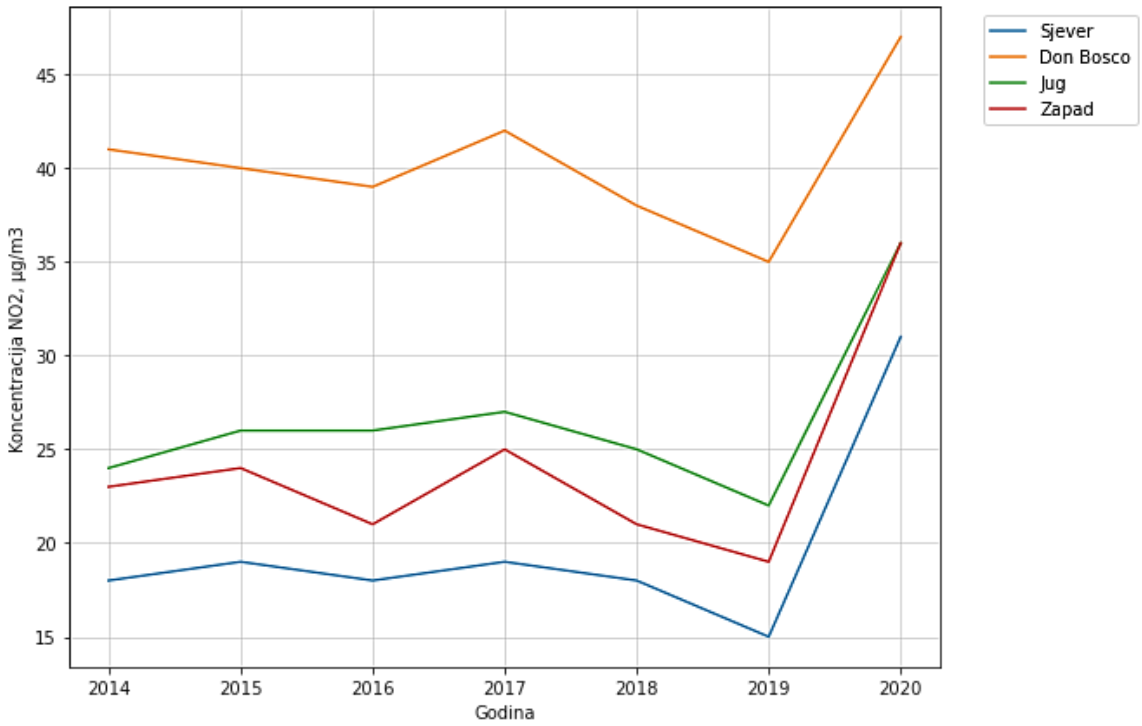
Tablica 6. Vrijednosti koncentracija NO_2 grupirane po godinama za promatrano razdoblje na sve četiri mjerne postaje.

GODINA	DON BOSCO, µg/m^3	SJEVER, µg/m^3	JUG, µg/m^3	ZAPAD, µg/m^3
2014.	41	18	24	23
2015.	40	19	26	24
2016.	39	18	26	21

2017.	42	19	27	25
2018.	38	18	25	21
2019.	35	15	22	19
2020.	47	31	36	36



Slika 31. *Boxplot* dijagrami godišnjih koncentracija NO_2 za sve četiri mjerne postaje: A) Don Bosco, B) Sjever, C) Jug, D) Zapad.

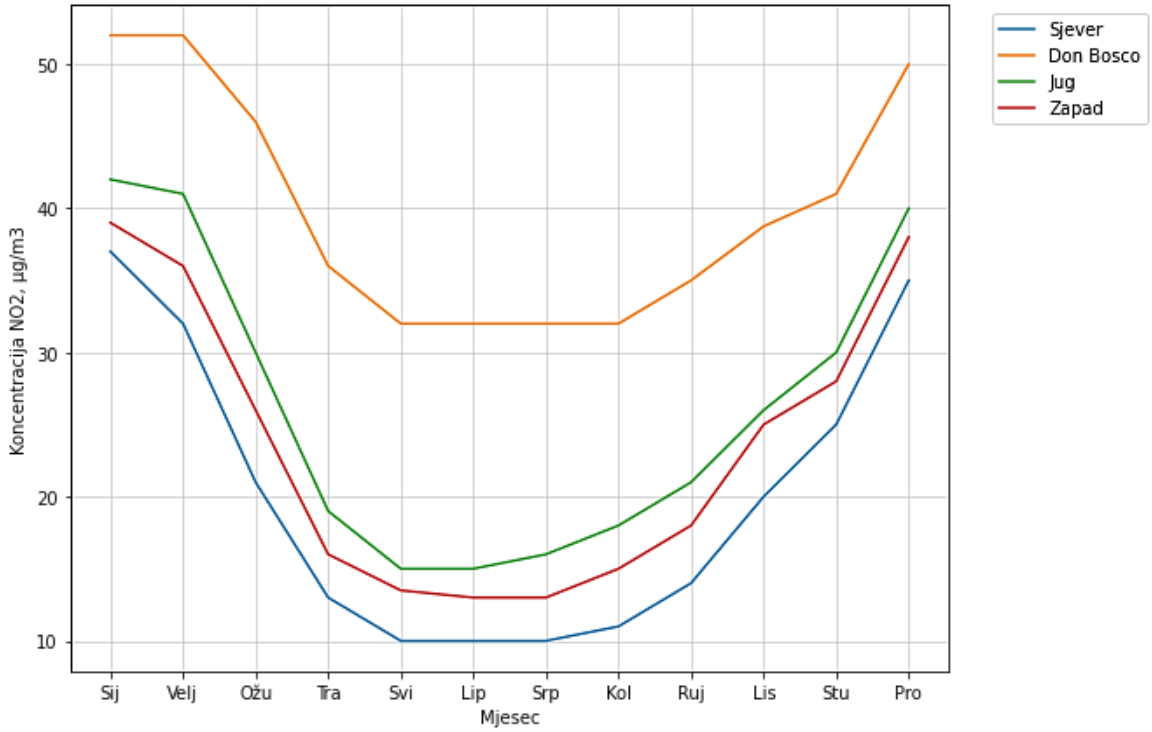


Slika 32. Koncentracije NO_2 grupirane po godinama promatranog razdoblja za sve četiri mjerne postaje.

Gledajući vrijednosti u tablici 7. primjećuje se da su tijekom lipnja koncentracije NO_2 najniže, a najviše tijekom siječnja za sve četiri mjerne postaje. Vrijednosti iz tablice 7. grafički su prikazane na slikama 33. i 34. gdje se jasnije uočava kako su prosječne mjesečne koncentracije NO_2 veće tijekom zimskih mjeseci, a to se može pripisati hladnijem vremenu zbog kojeg u kućanstvima i različitim ustanovama pojačano radi grijanje što je jedan od izvora onečišćenja zraka.

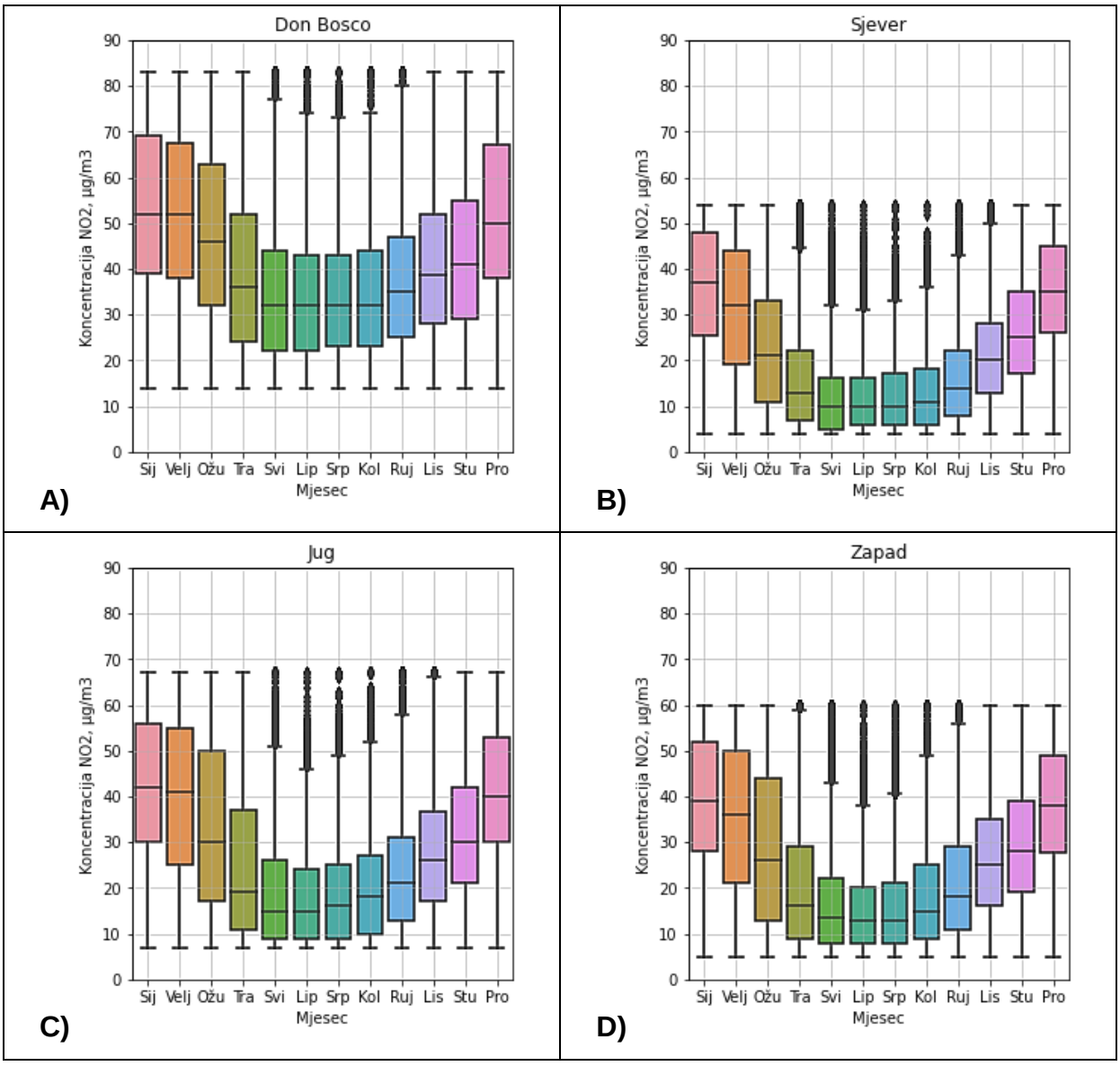
Tablica 7. Vrijednosti koncentracija NO_2 grupirane po mjesecima za promatrano razdoblje na sve četiri mjerne postaje.

MJESEC	DON BOSCO,µg/m^3	SJEVER, µg/m^3	JUG, µg/m^3	ZAPAD, µg/m^3
Siječanj	52	37	42	39
Veljača	52	32	41	36
Ožujak	46	21	30	26
Travanj	36	13	19	16
Svibanj	32	10	15	14
Lipanj	32	10	15	13
Srpanj	32	10	16	13
Kolovoz	32	11	18	15
Rujan	35	14	21	18
Listopad	38	20	26	25
Studeni	41	25	30	38
Prosinac	50	35	40	38



Slika 33. Koncentracije NO_2 grupirane po mjesecima promatranog razdoblja za sve četiri mjerne postaje.

Na postaji Sjever je tijekom ljetnih mjeseci tlak zraka niži u odnosu na ostale mjesece, a količina padalina je veća. Niski tlak zraka donosi promjenu vremena. Topli zrak se diže u vis i ostavlja iza sebe područje niskog tlaka. Zrak struji od rubova prema središtu, počinje se hladiti, vodena para se kondenzira, a kao posljedica nastaju oborine. Oborine padanjem na tlo „ispiru“ onečišćeni zrak, a čestice onečišćujućih tvari iz zraka padaju na tlo u obliku kiselih kiša. Time se koncentracija NO_2 u zraku smanjuje, ali nastaje novi problem jer se zagađuje tlo, biljni i životinjski svijet.

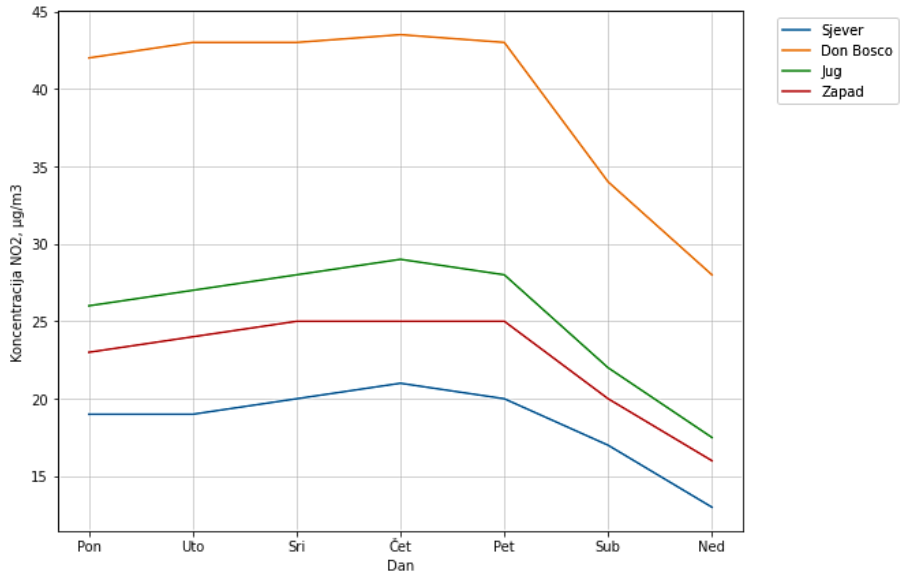


Slika 34. *Boxplot* dijagrami mjesečnih koncentracija NO_2 za mjerne postaje: A) Don Bosco, B) Sjever, C) Jug, D) Zapad.

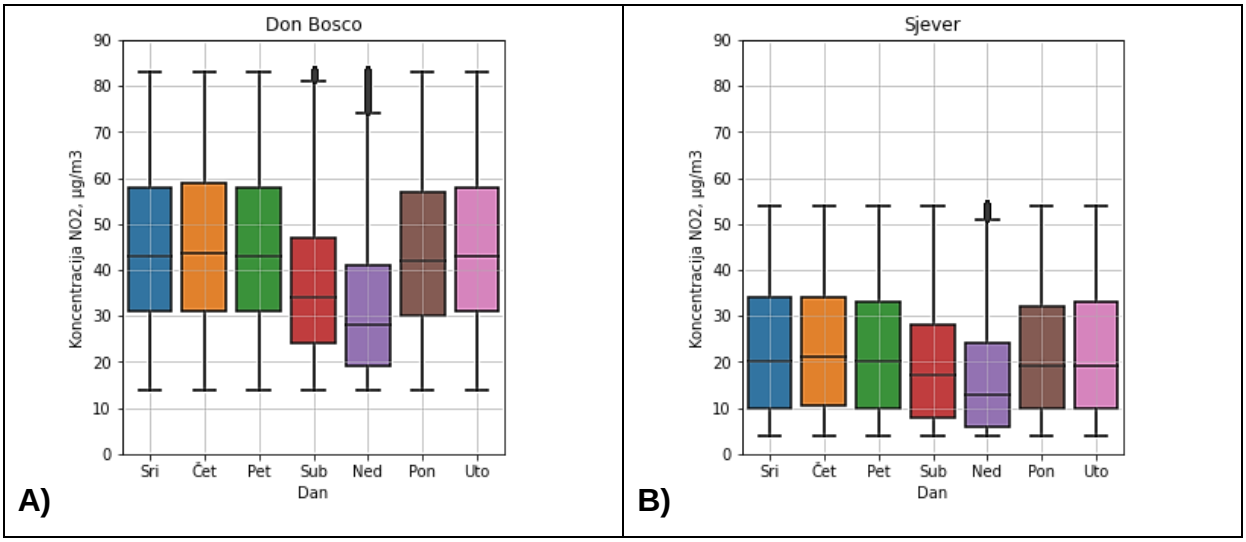
Ljudske aktivnosti obično prate sedmodnevni, tj. tjedni ciklus. Smanjenje industrijskih aktivnosti kao i obujam prometa tijekom vikenda dovodi do nižih razina emisija NO_2. Na temelju vrijednosti u tablici 8. primjećeno je da su tijekom nedjelje koncentracije NO_2 najniže, a najviše tijekom četvrtka za sve četiri mjerne postaje. Vrijednosti iz tablice 8. grafički su prikazane na slikama 35. i 36. gdje su očekivano prosječne tjedne koncentracije NO_2 veće tijekom radnog tjedna, a manje tijekom vikenda. Snažnije smanjenje koncentracije NO_2 tijekom vikenda može se primjetiti u industrijski, prometno i populacijski gustom području kao što je Don Bosco.

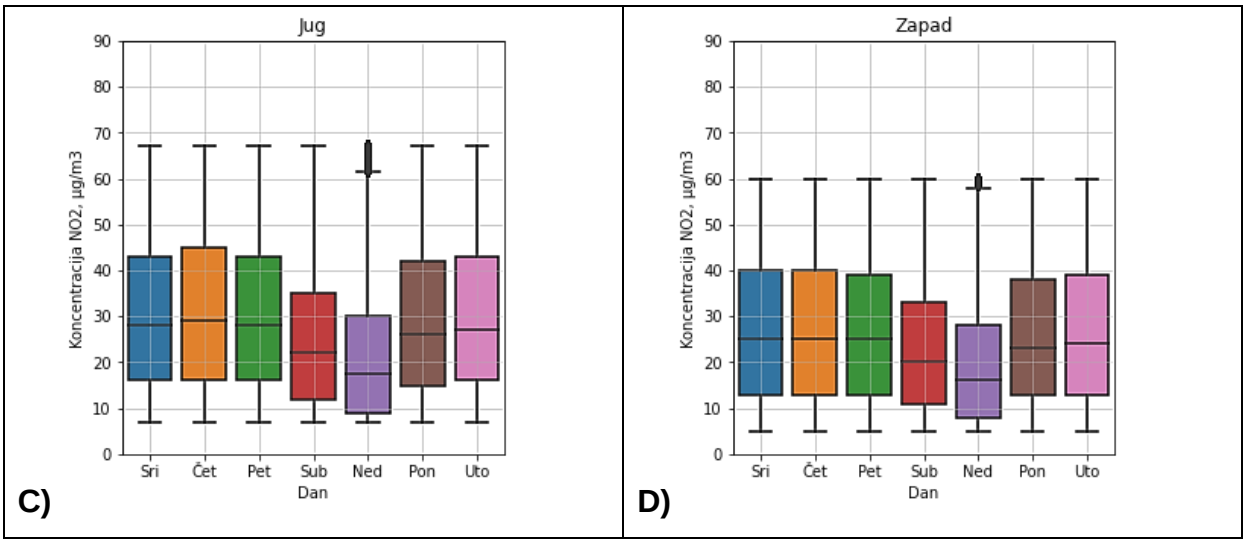
Tablica 8. Vrijednosti koncentracija NO_2 grupirane po danima u tjednu za promatrano razdoblje na sve četiri mjerne postaje.

DAN U TJEDNU	DON BOSCO, µg/m^3	SJEVER, µg/m^3	JUG, µg/m^3	ZAPAD, µg/m^3
Ponedjeljak	42	19	26	23
Utorak	43	19	27	24
Srijeda	43	20	28	25
Četvrtak	44	21	29	25
Petak	43	20	28	25
Subota	44	17	22	20
Nedjelja	28	13	18	16



Slika 35. Koncentracije NO_2 grupirane po danima u tjednu promatranog razdoblja za sve četiri mjerne postaje.



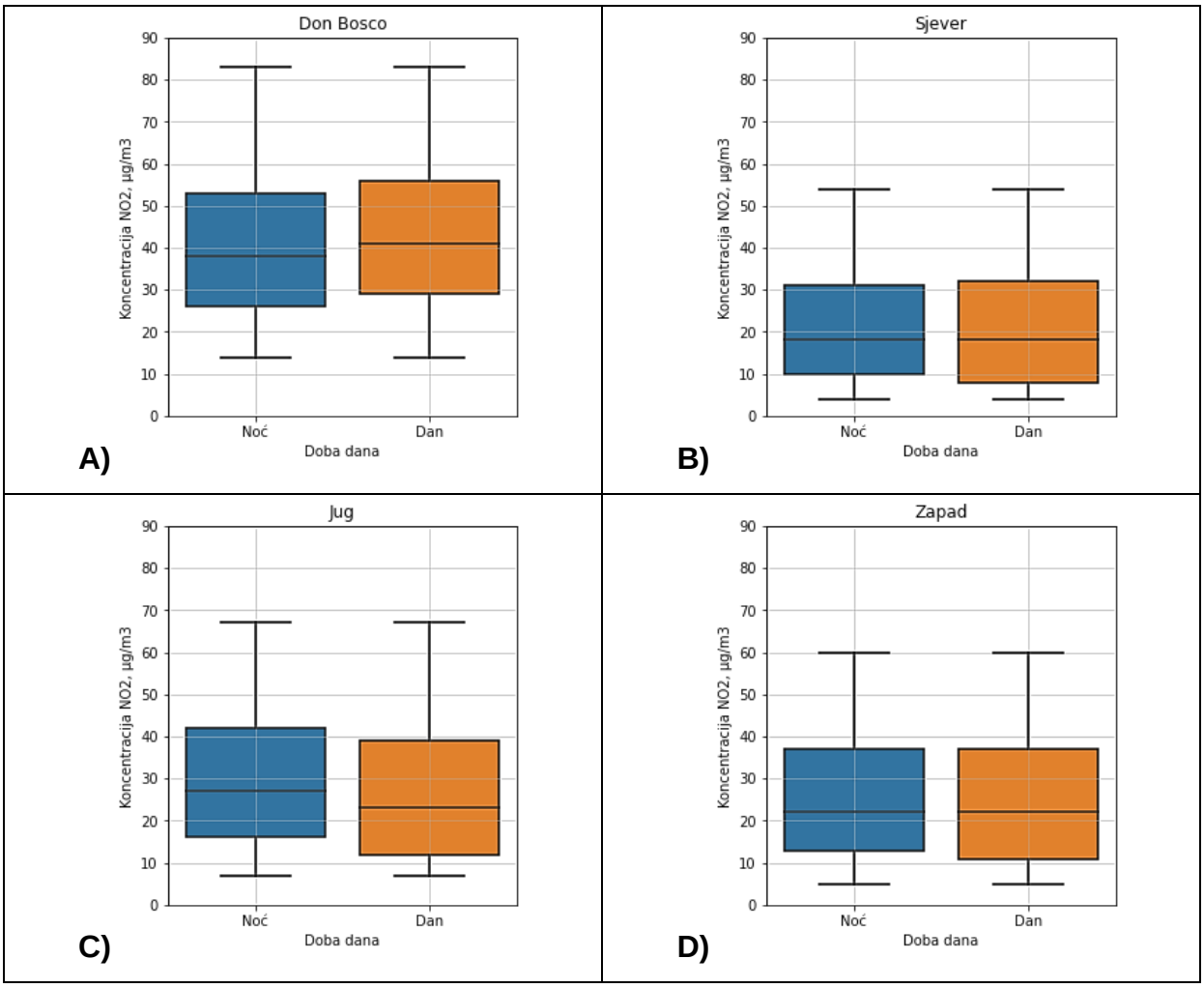


Slika 36. *Boxplot* dijagrami tjednih koncentracija NO_2 za mjerne postaje: A) Don Bosco, B) Sjever, C) Jug, D) Zapad.

Iz tablice 9. primjećuje se kako se koncentracije NO_2 ne razlikuju značajno tijekom noći i dana u promatranom razdoblju. Ipak, može se reći kako su mjerene koncentracije NO_2 na postaji Don Bosco malo veće tijekom dana u odnosu na noć, na postaji Jug su nezamjetno veće tijekom noći, a na preostale dvije postaje nema razlike u koncentracijama NO_2 noću i danju. Odnos koncentracija NO_2 tijekom dana i noći jasnije se uočava na slici 37.

Tablica 9. Vrijednosti koncentracija NO_2 grupirane po dobu dana za promatrano razdoblje na sve četiri mjerne postaje.

DOBA DANA	DON BOSCO, µg/m^3	SJEVER, µg/m^3	JUG, µg/m^3	ZAPAD, µg/m^3
Dan	41	18	23	22
Noć	38	18	27	22



Slika 37. *Boxplot* dijagrami koncentracija NO_2 po dobu dana za mjerne postaje: A) Don Bosco, B) Sjever, C) Jug, D) Zapad.

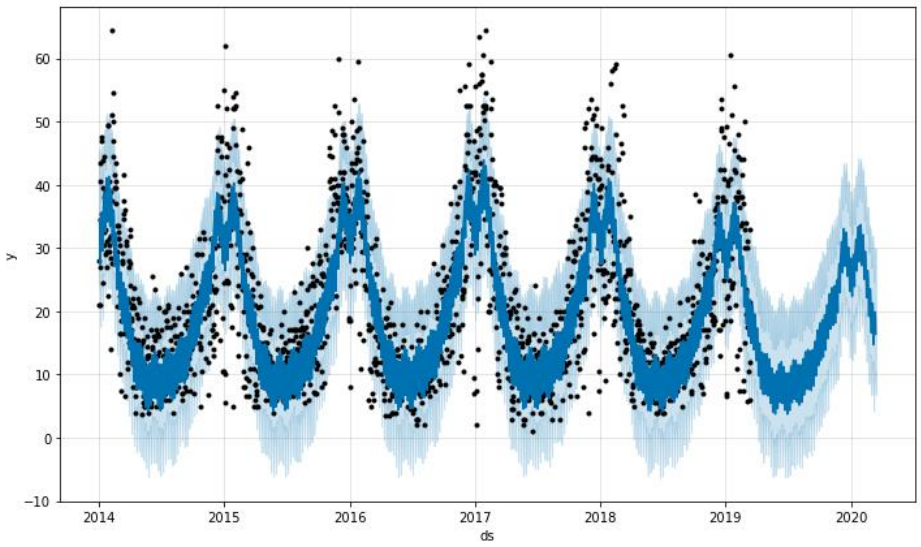
Budući da je prema prethodnim grafičkim prikazima *boxplot* dijagrama uočeno postojanje ekstremnih vrijednosti korišten je i skup podataka ograničen na vrijednosti koncentracije NO_2 između 5. i 95. percentila. Vrijednosti 5. i 95. percentila za svaku postaju navedeni su u tablici 10. Vrijednosti 5. i 95. percentila izračunati su samo na skupu podataka za obučavanje modela, odnosno do 15. ožujka 2019. jer bi se u suprotnom, kada bi se u obzir uzeli i podaci za validiranje modela, izgubila relevantnost modela. Ukupan postotak ekstremnih vrijednosti koncentracija NO_2 u skupu podataka za obučavanje modela kreće se između 8% i 10% na svakoj postaji. Skup podataka bez ekstremnih vrijednosti jednako je kvalitetan kao i sa ekstremnim vrijednostima jer se ne odbacuje veliki udio podataka. One vrijednosti koncentracija NO_2 koje su iznad 95. percentila zamjenjuju se vrijednošću 95. percentila koje su navedene u tablici 10. za svaku postaju, a vrijednosti ispod 5. percentila zamjenjuju se vrijednošću 5. percentila koje su također navedene u tablici 10.

Tablica 10. Vrijednosti 5. i 95. percentila za sve četiri mjerne postaje.

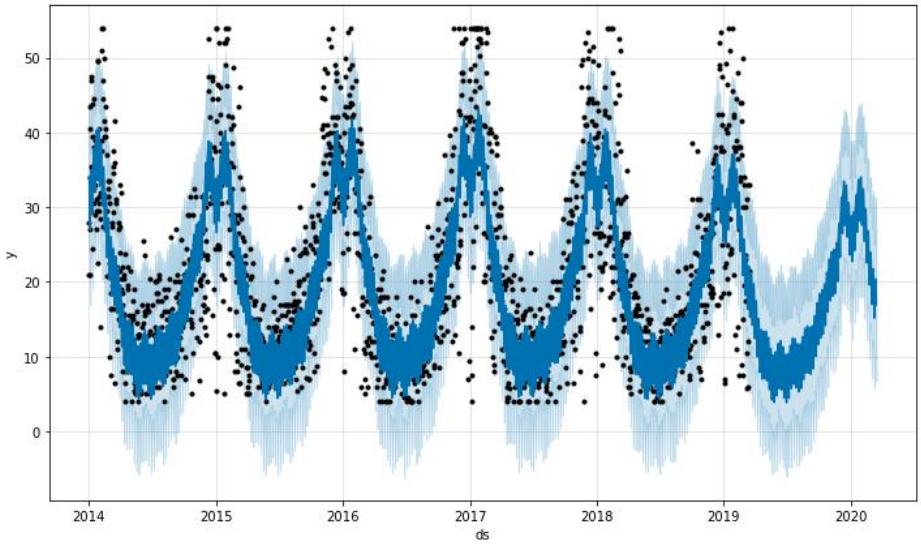
	DON BOSCO $\mu g/m^3$	SJEVER, $\mu g/m^3$	JUG $\mu g/m^3$	ZAPAD $\mu g/m^3$
5. percentil	14	4	7	5
95. percentil	83	54	67	60
Broj uzoraka < 5. percentil	1988	2051	2157	1552
Broj uzoraka > 95. percentil	2228	2234	2251	2277
Ukupan broj ekstremnih vrijednosti	4216	4285	4408	3829
Ukupno ekstremnih vrijednosti u skupu podataka za obučavanje (%)	9,25	9,40	9,67	8,40

5.3. *Prophet* modeli

Grafički prikaz stvarne i predviđene vrijednosti koncentracije NO_2 kao i minimalne i maksimalne granice predviđenih vrijednosti koncentracija NO_2 (intervale nesigurnosti) za postaju Sjever nalazi se na slici 38. za *Prophet* model obučavan na podacima s uključenim ekstremnim vrijednostima te na slici 39. za *Prophet* model obučavan na podacima sa isključenim ekstremnim vrijednostima. Crne točkice označuju stvarne mjerene vrijednosti koncentracija NO_2, plavom bojom označene su predviđene vrijednosti koncentracija NO_2, a svjetloplava nijansa označava intervale nesigurnosti. Predviđene koncentracije NO_2 za 2020.g. nižih su vrijednosti u odnosu na prethodne godine što je i očekivano budući da je i komponenta trenda padajuća. Na grafičkim prikazima je također jasno uočljivo kako postoji godišnja sezonalnost te da su koncentracije niže tijekom sredine godine (ljetni mjeseci), a povišene tijekom početka i kraja godine (zimski mjeseci).



Slika 38. Predviđanje *Prophet* modela po danima s uključenim ekstremnim vrijednostima na primjeru mjerne postaje Sjever (preostale postaje dane su u dodaktu 1).



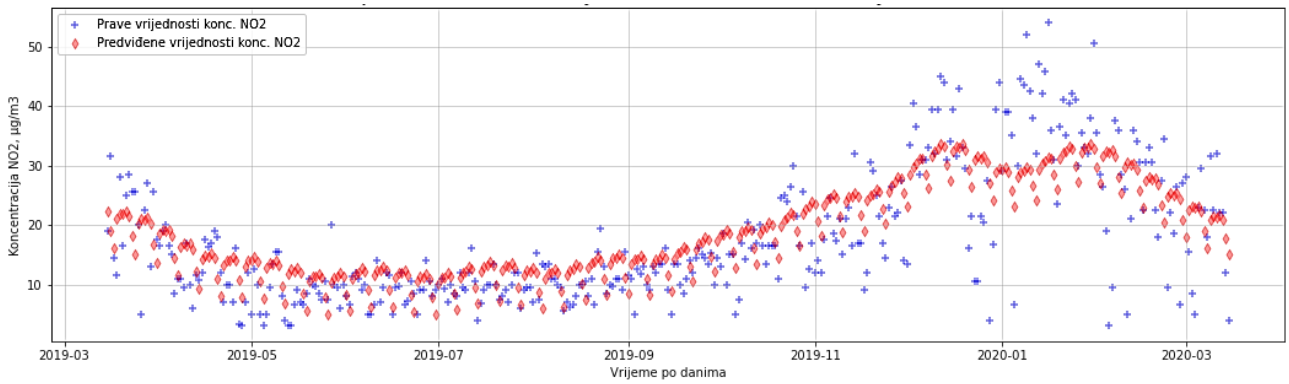
Slika 39. Predviđanje *Prophet* modela po danima sa isključenim ekstremnim vrijednostima na primjeru mjerne postaje Sjever (preostale postaje dane su u dodaktu 1).

Tablica 11. Dobivene vrijednosti statističkih pokazatelja za oba *Prophet* modela.

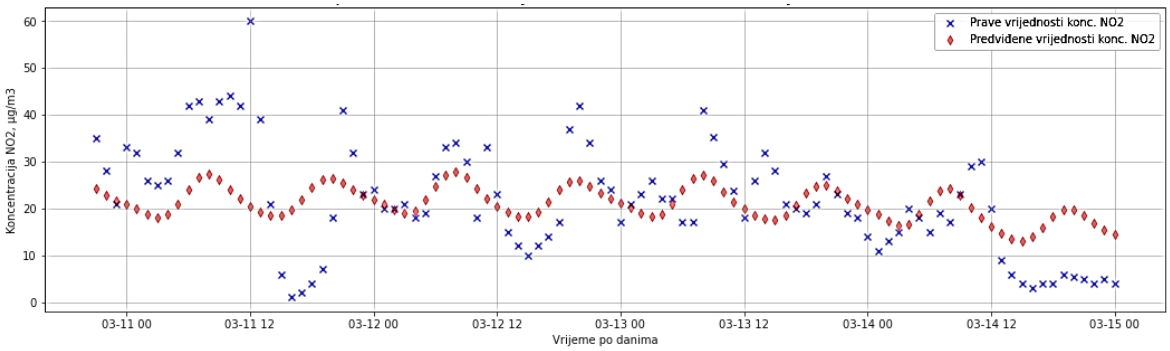
	DON BOSCO		SJEVER		JUG		ZAPAD	
	Prophet s ekstremnim vrijednostima	Prophet bez ekstremnih vrijednosti	Prophet s ekstremnim vrijednostima	Prophet bez ekstremnih vrijednosti	Prophet s ekstremnim vrijednostima	Prophet bez ekstremnih vrijednosti	Prophet s ekstremnim vrijednostima	Prophet bez ekstremnih vrijednosti
R^2	0,35	0,36	0,41	0,41	0,37	0,38	0,38	0,39
MAE	11,62	11,49	7,87	7,88	10,07	10,04	9,80	9,77
RMSE	216,56	214,04	110,56	110,48	168,44	167,63	162,40	161,74

Budući da se prema vrijednostima u tablici 11. oba *Prophet* modela bitno ne razlikuju jer su im vrijednosti R^2 gotovo iste, *RMSE* vrijednosti se također minimalno razlikuju, potvrđeno je kako je *Prophet* model robustan na podatke koji nedostaju te kako dobro podnosi skup podataka s ekstremnim vrijednostima. Ipak, čak i mala poboljšanja koja pokazuje *Prophet* model bez ekstremnih vrijednosti vodi ka tome da se za razvoj *Random Forest* modela također uzmu u obzir podaci s uključenim i isključenim ekstremnim vrijednostima. Usporedbe stvarnih i predviđenih dnevnih vrijednosti koncentracija NO_2 dobivenih na temelju *Prophet* modela za postaju Sjever prikazane su na slici 40. Za lakšu vizualnu usporedbu podudaranja stvarnih i predviđenih vrijednosti grafički je prikazano posljednjih sto vrijednosti promatranog razdoblja (slika 41). *Prophet* model uglavnom predviđa koncentracije NO_2 koje su lokalizirane oko prosječnih vrijednosti u danom periodu, ali nema sposobnost predviđanja nižih ili viših vrijednosti od tih lokaliziranih. Slika 42. pokazuje raspršenost predviđenih vrijednosti koncentracija NO_2 u odnosu na idealni regresijski pravac (R^2 = 1) za sve mjerne postaje. Crvene točke označavaju predviđene vrijednosti *Prophet* modela s uključenim ekstremnim vrijednostima, a zelene točke predviđene vrijednosti koncentracija NO_2 *Prophet* modela sa isključenim ekstremnim vrijednostima. *Prophet* model sa isključenim ekstremnim vrijednostima ima malo

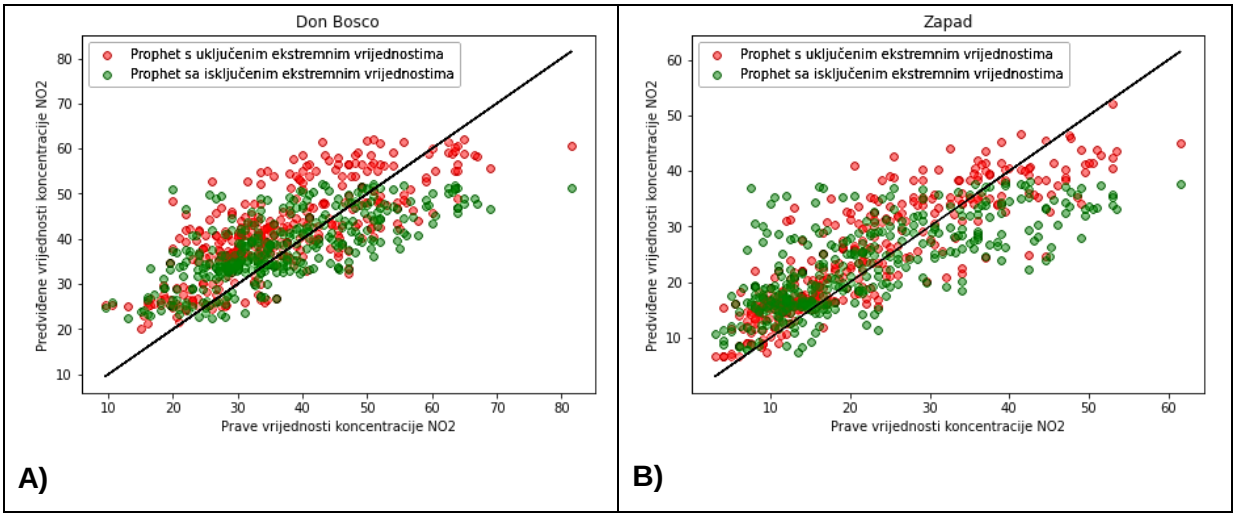
bolje, tj. uže raspršene podatke oko idealnog regresijskog pravca pa je stoga i bolji model u odnosu na *Prophet* model s uključenim ekstremnim vrijednostima.

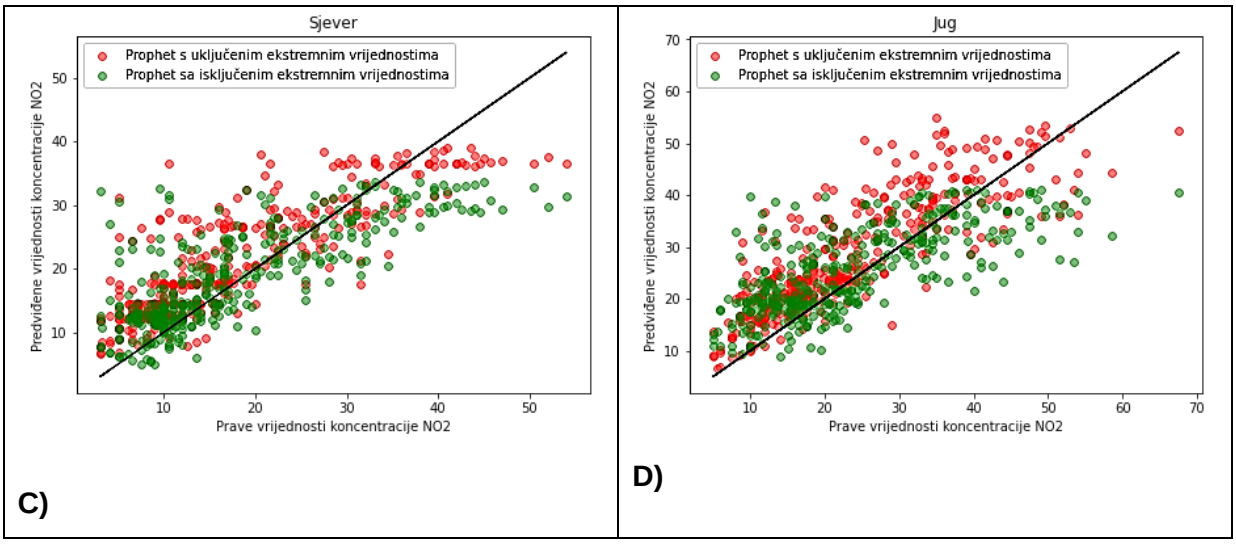


Slika 40. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 agregiranih po danu i dobivenih na temelju *Prophet* modela za postaju Sjever (preostale postaje dane su u dodatku 1).



Slika 41. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 agregiranih po danima i dobivenih na temelju *Prophet* modela za postaju Sjever (preostale postaje dane su u dodatku 1).





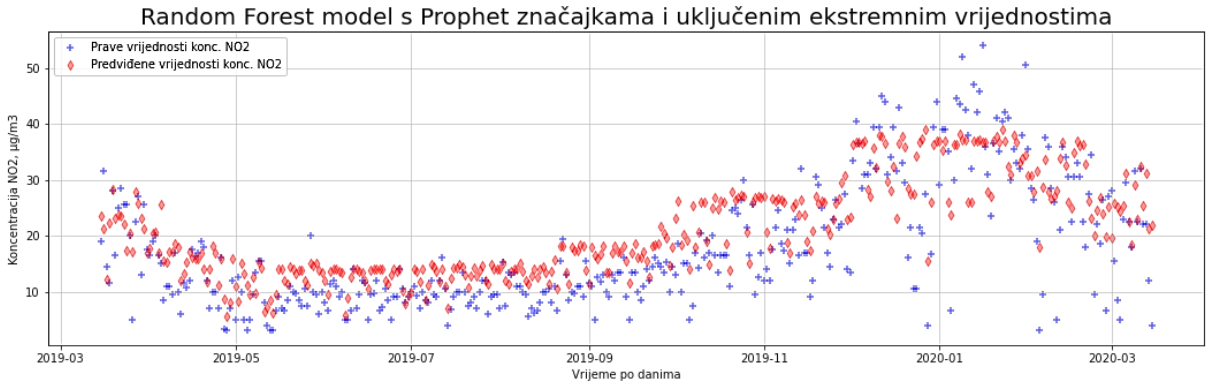
Slika 42. Raspršenost predviđenih vrijednosti koncentracija NO_2 u odnosu na idealni regresijski pravac (R^2 = 1) za mjerne postaje: A) Don Bosco, B) Zapad, C) Sjever, D) Jug.

5.4. *Random Forest* modeli

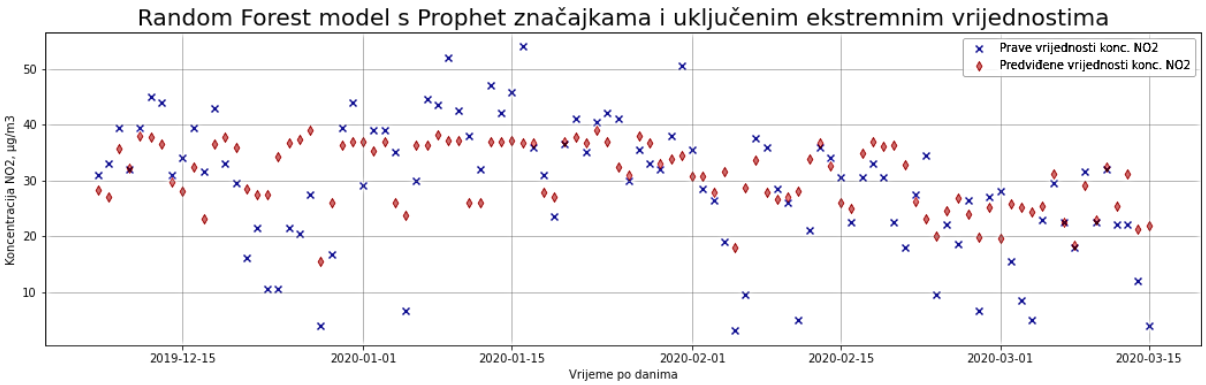
Tablica 12. Dobivene vrijednosti statističkih pokazatelja za četiri tipa *Random Forest* modela.

MODELI	MJERNE POSTAJE	R^2	*MAE*	*RMSE*
Random Forest **model s ekstremnim vrijednostima**	**Don Bosco**	0,45	10,54	182,05
	Sjever	0,56	7,01	82,15
	Jug	0,59	8,09	110,18
	Zapad	0,63	7,29	93,65
Random Forest **model bez ekstremnih vrijednosti**	**Don Bosco**	0,47	10,44	176,60
	Sjever	0,56	7,17	83,22
	Jug	0,61	7,93	104,70
	Zapad	0,64	7,49	96,20
Random Forest **model s ekstremnim vrijednostima + *Prophet* značajke**	**Don Bosco**	0,50	10,05	166,53
	Sjever	0,58	6,66	79,93
	Jug	0,62	7,68	100,64
	Zapad	0,65	7,15	93,45
Random Forest **model bez ekstremnih vrijednosti + *Prophet* značajke**	**Don Bosco**	0,49	10,11	169,20
	Sjever	0,57	6,70	80,49
	Jug	0,62	7,70	101,34
	Zapad	0,64	7,30	94,05

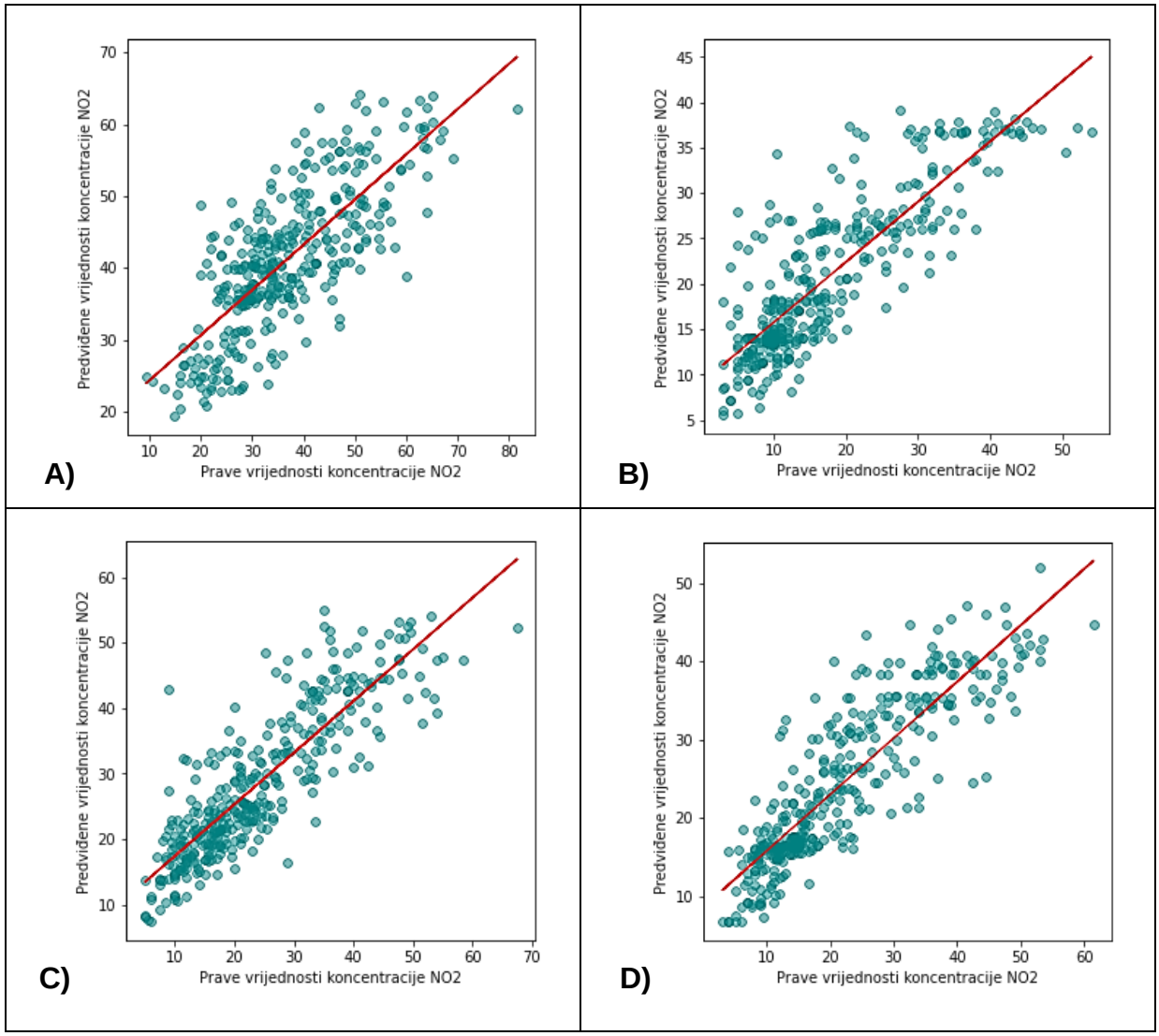
Usporedbe stvarnih i predviđenih dnevnih vrijednosti koncentracija NO_2 dobivenih na temelju najboljeg tipa *Random Forest* modela za postaju Sjever prikazane su na slici 43. Za lakšu vizualnu usporedbu podudaranja stvarnih i predviđenih vrijednosti grafički je prikazano posljednjih sto vrijednosti promatranog razdoblja (slika 44). Uspoređujući sa slikama 40. i 41. (*Prophet* modelima) jasno je vidljivo kako je *Random Forest* model s *Prophet* značajkama i uključenim ekstremnim vrijednostima bolji u odnosu na oba *Prophet* modela, a to je dokazano i boljim vrijednostima R^2, *RMSE* i *MAE* koje za najbolji tip *Random Forest* modela iznose: R^2 = 0,58, *MAE* = 6,66 i *RMSE* = 79,93. Prema tome, kombinacija *Random Forest* modela s *Prophet* značajkama pokazala se najboljim odabirom za razvoj modela predviđanja koncentracija NO_2. Slika 45. prikazuje raspršenost predviđenih vrijednosti koncentracija NO_2 za sve mjerne postaje za najbolji tip *Random Forest* modela. Za skup podataka na mjernoj postaji Zapad, *Random Forest* model s *Prophet* značajkama i uključenim ekstremnim vrijednostima pokazao se najboljim u odnosu na podatke ostalih postaja (R^2 = 0,65), a najlošiji za postaju Don Bosco (R^2 = 0,50). Razlog tomu je nedostatak značajki koje opisuju meteorološke podatke, a koji na postaji Don Bosco nisu mjereni.



Slika 43. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 agregiranih po danu i dobivenih na temelju najboljeg *Random Forest* modela za postaju Sjever (preostale postaje dane su u dodatku 1).



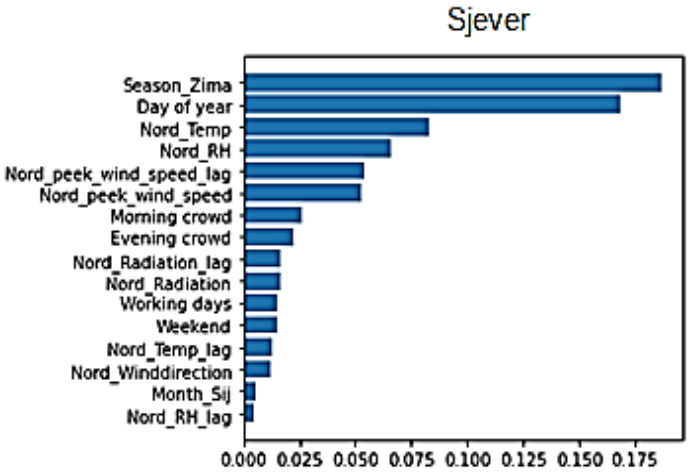
Slika 44. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 agregiranih po danu i dobivenih na temelju najboljeg *Random Forest* modela za postaju Sjever (preostale postaje dane su u dodatku 1).



Slika 45. Prikaz regresijskog pravca i raspršenosti predviđenih vrijednosti koncentracija NO_2 za najbolji tip *Random Forest* modela za mjerne postaje: A) Don Bosco, B) Sjever, C) Zapad, D) Jug.

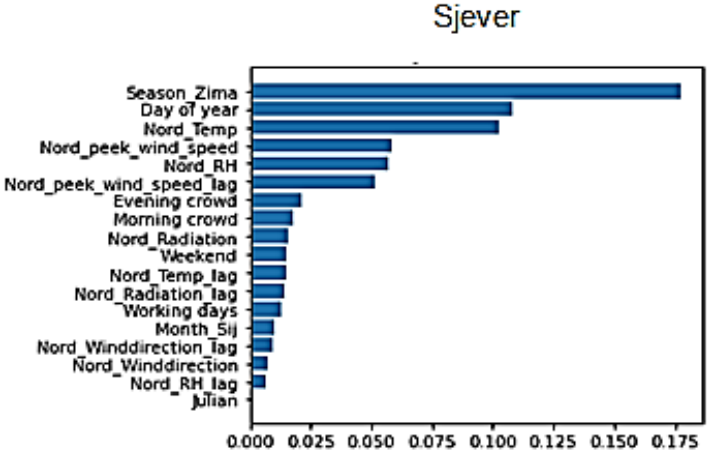
5.4.1. Izabrane značajke

5.4.1.1. *Random Forest* model s uključenim ekstremnim vrijednostima i bez značajki *Prophet* modela



Slika 46. Izabrane značajke koje najviše utječu na rezultate predviđanja koncentracija NO_2 prikazane padajućim vrijednostima za postaju Sjever.

5.4.1.2. Random Forest model sa isključenim ekstremnim vrijednostima i bez značajki Prophet modela



Slika 47. Izabrane značajke koje najviše utječu na rezultate predviđanja koncentracija NO_2 prikazane padajućim vrijednostima za postaju Sjever.

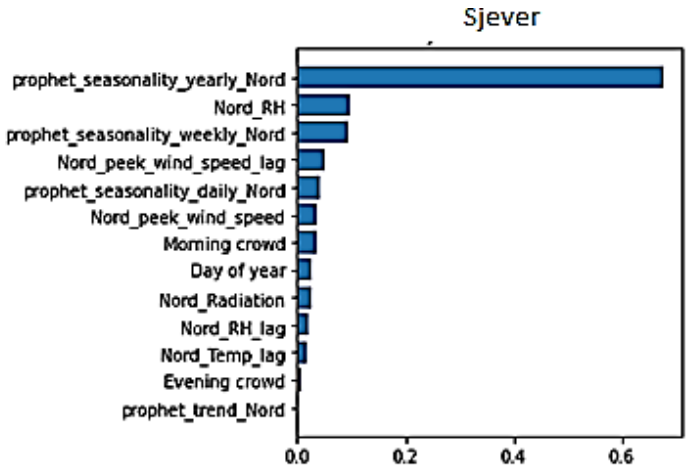
Prema slikama 46. i 47. značajke koje najviše utječu na *Random Forest* modele bez *Prophet* značajki na postaji Sjever su temporalne značajke koje pokazuju sezonalnost, a to su godišnje doba (zima) i dan u godini. Zatim slijede meteorološki podaci poput temperature, relativne vlažnosti i maksimalnih naleta brzine vjetra. Tijekom hladnijih mjeseci, kada su temperature niže, povećanje koncentracija NO_2

može se pripisati povećanom radu toplana i grijanja u kućanstvima što je jedan od glavnih izvora onečišćenja zraka dušikovim dioksidom. Visoka vlažnost zraka može smanjiti koncentraciju NO_2 što bi moglo biti posljedica činjenice da se onečišćujuće tvari jače talože na vlažnim površinama. Što se tiče relativne vlažnosti, u okolišu je dušična kiselina (glavna sastavnica kiselih kiša) u ravnoteži s NO_2, NO i H_2O prema jednadžbi *(5):*

$$2\ HNO_{2(aq)} \longleftrightarrow NO_{(g)} + NO_{2\ (g)} + H_2O_{(l)} \qquad (5)$$

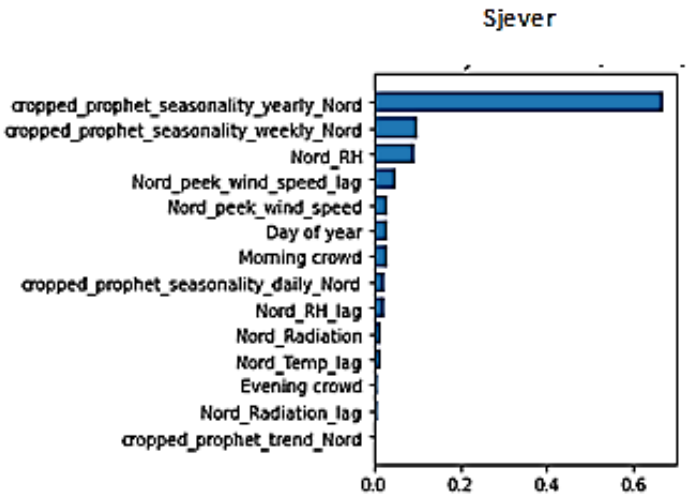
Smjer ove reakcije ovisi o klimatskim uvjetima; visoka vlažnost potiče nastanak HNO_2 jer višak H_2O pomiče ravnotežu u lijevo odnosno prema nastajanju HNO_2. Da bi se ravnoteža uspostavila, s vodom mora reagirati određena količina dušikovih oksida NO i NO_2 čime se gubi određena količina NO_2 u zraku jer prelazi u HNO_2. To rezultira smanjenjem koncentracije NO_2 u zraku. Jači vjetar pogoduje transportu i difuziji onečišćujućih tvari u zraku pa su zbog toga maksimalni naleti brzine vjetra jedne od bitnih značajki u predviđanju koncentracija NO_2 u zraku. Utjecaj jutarnjih i večernjih gužvi također je bitan budući da je u tom periodu povećan promet zbog čega su i povećane razine emisija NO_2 iz motornih vozila, a to utječe na povećanje koncentracije NO_2 u zraku. Vikend predstavlja značajku koja doprinosi predviđanju smanjenja koncentracija u tom periodu zbog smanjenih industrijskih i prometnih aktivnosti.

5.4.1.3. *Random Forest* model sa uključenim ekstremnim vrijednostima i značajkama *Prophet* modela



Slika 48. Izabrane značajke koje najviše utječu na rezultate predviđanja koncentracija NO_2 prikazane padajućim vrijednostima za postaju Sjever.

5.4.1.4. *Random Forest* model sa značajkama *Prophet* modela i isključenim ekstremnim vrijednostima



Slika 49. Izabrane značajke koje najviše utječu na rezultate predviđanja koncentracija NO_2 prikazane padajućim vrijednostima za postaju Sjever.

Prema slikama 48. i 49. značajke koje najviše utječu na *Random Forest* modele sa *Prophet* značajkama na postaji Sjever su *Prophet* značajke koje pokazuju sezonalnost. Najviše utječe godišnja sezonalnost kao što je to i pokazano toplinskom kartom korelacija. *Prophet* značajke objedinjuju temporalne značajke koje govore o sezonalnosti kao što su godišnja doba, dan u godini, tjednu. Na taj način model ima objedinjeno više značajki u jednu značajku pa time smanjuje mogući broj grananja odluka, smanjuje vjerojatnost pogreške i poboljšava predviđanja. Ostale meteorološke značajke koje utječu na predviđanje koncentracija NO_2 jednake su kao i za *Random Forest* model bez *Prophet* značajki.

6. ZAKLJUČAK

U ovom radu analizirani su prikupljeni podaci koncentracija NO_2 (od 1. siječnja 2014.g. do 15. ožujka 2020.g.) i razvijeni su modeli za procjenu koncentracije NO_2 u zraku u razdoblju od 15. ožujka 2019. do 15. ožujka 2020.g u Grazu. Analiza prikupljenih podataka pokazuje kako je Don Bosco, najzagađenije mjerno područje zbog najvećih koncentracija NO_2 u zraku, a postaja Sjever najmanje zagađeno mjerno područje. Velike koncentracije NO_2 u zraku na mjernoj postaji Don Bosco posljedica su velikih cestovnih prometnica i emisija iz obližnjih industrijskih pogona dok je postaja Sjever klasificirana kao urbano mjesto iz predgrađa s manjim cestovnim prometnicama zbog čega bilježi manje koncentracije NO_2. Modeli su razvijeni na skupovima podataka s uključenim i isključenim ekstremnim vrijednostima koncentracija NO_2. Ekstremne vrijednosti zamijenjene su vrijednostima 5. i 95. percentila ovisno o tome jesu li ekstremne vrijednosti niže ili više od 5. i 95. percentila. Sveukupan broj ekstremnih vrijednosti koncentracija NO_2 za postaje Don Bosco, Sjever, Jug i Zapad iznosi redom 4216, 4285, 4408, 3829 od ukupnih 45575 podataka koji čine skup za učenje modela. Prema dobivenim vrijednostima koeficijenata determinacije razvijeni modeli dali su zadovoljavajuće rezultate. Modeli su se pokazali robusnim i kada podaci sadrže ekstremne vrijednosti s obzirom da svi modeli razvijeni na oba skupa podataka pokazuju vrlo slične vrijednosti koeficijenata determinacije. Kombinacija *Prophet* značajki s *Random Forest* modelom daje najbolja predviđanja koncentracija NO_2 u zraku na svim postajama. Najbolji rezultati ostvareni su na mjernoj postaji Zapad gdje vrijednost koeficijenta determinacije R^2 iznosi 0,65, a najlošiji na postaji Don Bosco gdje R^2 iznosi 0,50. Razlog tomu je nedostatak meteoroloških podataka kao što su brzina vjetra, smjer vjetra i maksimalni naleti brzine vjetra koji nisu mjereni na postaji Don Bosco dok su na preostalim postajama mjereni. *Prophet* modelima pokazano je kako u skupu podataka postoji sezonalnost, a *Random Forest* modelima pokazano je kako je sezonalnost najvažnija značajka pri procjeni vrijednosti koncentracija NO_2 te da najveći utjecaj ima godišnja sezonalnost. Svi *Random Forest* modeli pokazali su kako lošije procjenjuju niske i visoke koncentracije NO_2, a to je zbog toga što su one rijetke pa model nema veliki broj podataka na kojima bi mogao učiti. Zbog toga postoji mogućnost poboljšanja *Random Forest* modela na način da se cijeli skup podataka podijeli na dva skupa podataka, odnosno na skup podataka iznad vrijednosti

medijana i na skup podataka ispod vrijednosti medijana zbog čega bi lakše mogao predviđati visoke i niske koncentracije. *Random Forest* modeli bi se mogli poboljšati i na način da se pri odabiru hiperparametara modela u programu kreira petlja kojom bi se odabrala najbolja kombinacija parametara u zadanom velikom intervalu potencijalnih vrijednosti. Međutim, većim prostorom pretraživanja mogućih parametara modela proračun razvoja modela bi se mogao značajno produljiti.

7. LITERATURA

[1] Honda, T., Pun, V.C., Manjourides, J., Suh, H., 2017. Associations between long-term exposure to air pollution, glycosylated hemoglobin and diabetes. Int. J. Hyg. Environ. Health 220, 1124–1132. https://doi.org/10.1016/J.IJHEH.2017.06.004

[2] Honda, T., Pun, V.C., Manjourides, J., Suh, H., 2017. Associations between long-term exposure to air pollution, glycosylated hemoglobin and diabetes. Int. J. Hyg. Environ. Health 220, 1124–1132. https://doi.org/10.1016/J.IJHEH.2017.06.004

[3] WHO, 2018. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action, 9 out 10 people Worldw. breathe polluted air, but more Ctries. are Tak. action.

[4] Lovrić, M., Pavlović, K., Vuković, M., Grange, S.K., Haberl, M., Kern, R., 2021. Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. Environ. Pollut. 274.

[5] WHO's Fifth WHO Air Quality Database of over 6000 Cities Updated April 2022, n.d. URL https://www.who.int/news-room/questions-and-answers/item/who-s-fifth-who-air-quality-database-of-over-6000-cities-updated-april-2022 (pristup 20.7.2022.).

[6] Honda, T., Pun, V.C., Manjourides, J., Suh, H., 2017. Associations between long-term exposure to air pollution, glycosylated hemoglobin and diabetes. Int. J. Hyg. Environ. Health 220, 1124–1132. https://doi.org/10.1016/J.IJHEH.2017.06.004

[7] Ljudske aktivnosti i onečišćenje zraka – Airq, n.d. URL https://www.airq.hr/ljudske-aktivnosti-i-oneciscenje-zraka/ (pristup 15.7.2022.)

[8] NAAQS Table | US EPA, n.d. URL https://www.epa.gov/criteria-air-pollutants/naaqs-table (pristup 15.7.2022.)

[9] Basic Information about Carbon Monoxide (CO) Outdoor Air Pollution | US EPA, n.d. URL https://www.epa.gov/co-pollution/basic-information-about-carbon-monoxide-co-outdoor-air-pollution#What is CO (pristup 19.7.2022.)

[10] Basic Information about Lead Air Pollution | US EPA, n.d. URL https://www.epa.gov/lead-air-pollution/basic-information-about-lead-air-pollution#how (pristup 15.7.2022.).

[11] WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide, 2006.

[12] Sulfur Dioxide Basics | US EPA, n.d. URL https://www.epa.gov/so2-pollution/sulfur-dioxide-basics#what is so2 (pristup 19.7.2022.)

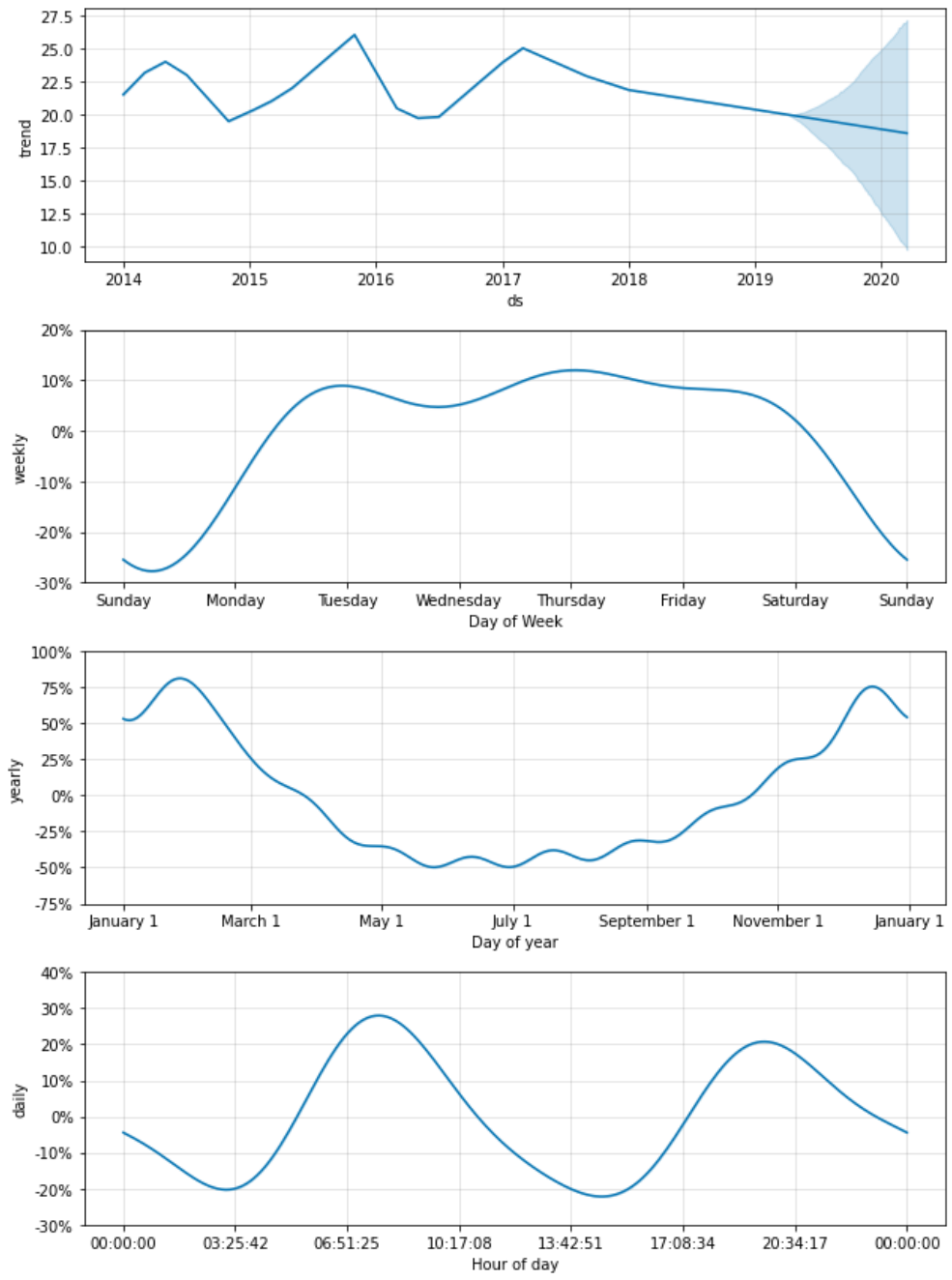
[13] Basic Information about NO2 | US EPA , n.d. URL https://www.epa.gov/no2-pollution/basic-information-about-no2#What is NO2 (pristup 19.7.2022.)

[14] Filipović-Lipanović, Opća i anorganska kemija II, Školska knjiga, Zagreb, 1995.

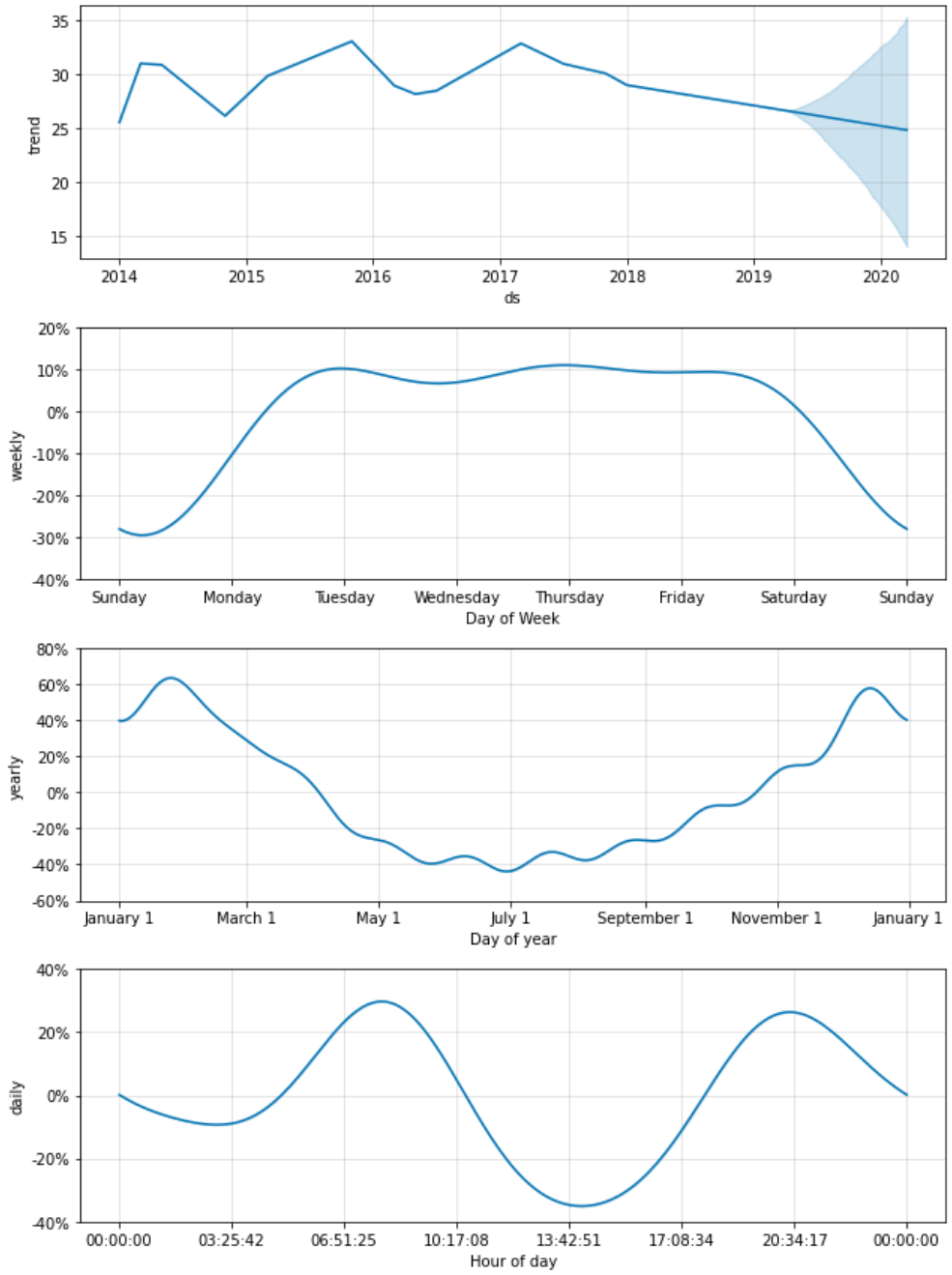
[15] Nitric Oxide, Nitrogen Dioxide and Nitrous Oxide, Laughing Gas Stock Vector - Illustration of nitrous, isolated: 169431487 [WWW Document], n.d. URL https://www.dreamstime.com/nitric-oxide-nitrogen-dioxide-nitrous-oxide-laughing-gas-nitric-oxide-no-nitrogen-dioxide-no-nitrous-oxide-n-o-laughing-image169431487 (pristup 19.7.2022.)

[16] Review of Nitrous Oxide (N2O) Emissions from Motor Vehicles on JSTOR, n.d. URL https://www.jstor.org/stable/27034495?seq=3#metadata_info_tab_contents (pristup 23.7.2022.)
[17] Babić, V., Rad, Z., n.d. SVEUČILIŠTE U ZAGREBU FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE SVEUČILIŠNI PREDDIPLOMSKI STUDIJ.
[18] Hrvatska enciklopedija, n.d.URL https://www.enciklopedija.hr/ (pristup 4.8.2022.)
[19] Honda, T., Pun, V.C., Manjourides, J., Suh, H., 2017. Associations between long-term exposure to air pollution, glycosylated hemoglobin and diabetes. Int. J. Hyg. Environ. Health 220, 1124–1132.
[20] Utjecaj onečišćenja zraka na zdravlje — Europska agencija za okoliš, n.d. URL https://www.eea.europa.eu/hr/signals/signals-2013/graficki-informacije/utjecaj-oneciscenja-zraka-na-zdravlje-2/view (pristup 23.7.2022.)
[21] Svaki naš udisaj — Europska agencija za okoliš, n.d. URL https://www.eea.europa.eu/hr/signals/signals-2013/clanci/svaki-nas-udisaj (pristup 23.7.2022.)
[22] Bolf, N., Zagrebu, S.U., n.d. OSVJEŽIMO ZNANJE Strojno učenje.
[23] J.D. Kelleher, B. Tierney, (2021), Znanost o podacima: Uvod u strojno učenje, Zagreb, MATE d.o.o.
[24] Hyperparameter Tuning the Random Forest in Python | by Will Koehrsen | Towards Data Science, n.d. URL https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74 (pristup 22.7.2022.)
[25] Python - How to plot grid seach layout and random search layout - Stack Overflow, n.d. URL https://stackoverflow.com/questions/65682419/how-to-plot-grid-seach-layout-and-random-search-layout (pristup 23.7.2022.)
[26] Multivariate Time Series | Vector Auto Regression (VAR), n.d. URL https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/? (pristup 10.8.2022.)
[27]Time Series as Features | Kaggle, n.d. URL https://www.kaggle.com/code/ryanholbrook/time-series-as-features (pristup 10.8.2022.)
[28] Tutorial: Time Series Forecasting with Prophet | Kaggle, n.d. URL https://www.kaggle.com/code/prashant111/tutorial-time-series-forecasting-with-prophet/notebook (pristup 10.8.2022.)
[29] Taylor, S.J., Letham, B., 2018. Forecasting at Scale. Am. Stat. 72, 37–45.
[30] Graz | Austria | Britannica, n.d. URL https://www.britannica.com/place/Graz (pristup 10.8.2022.)

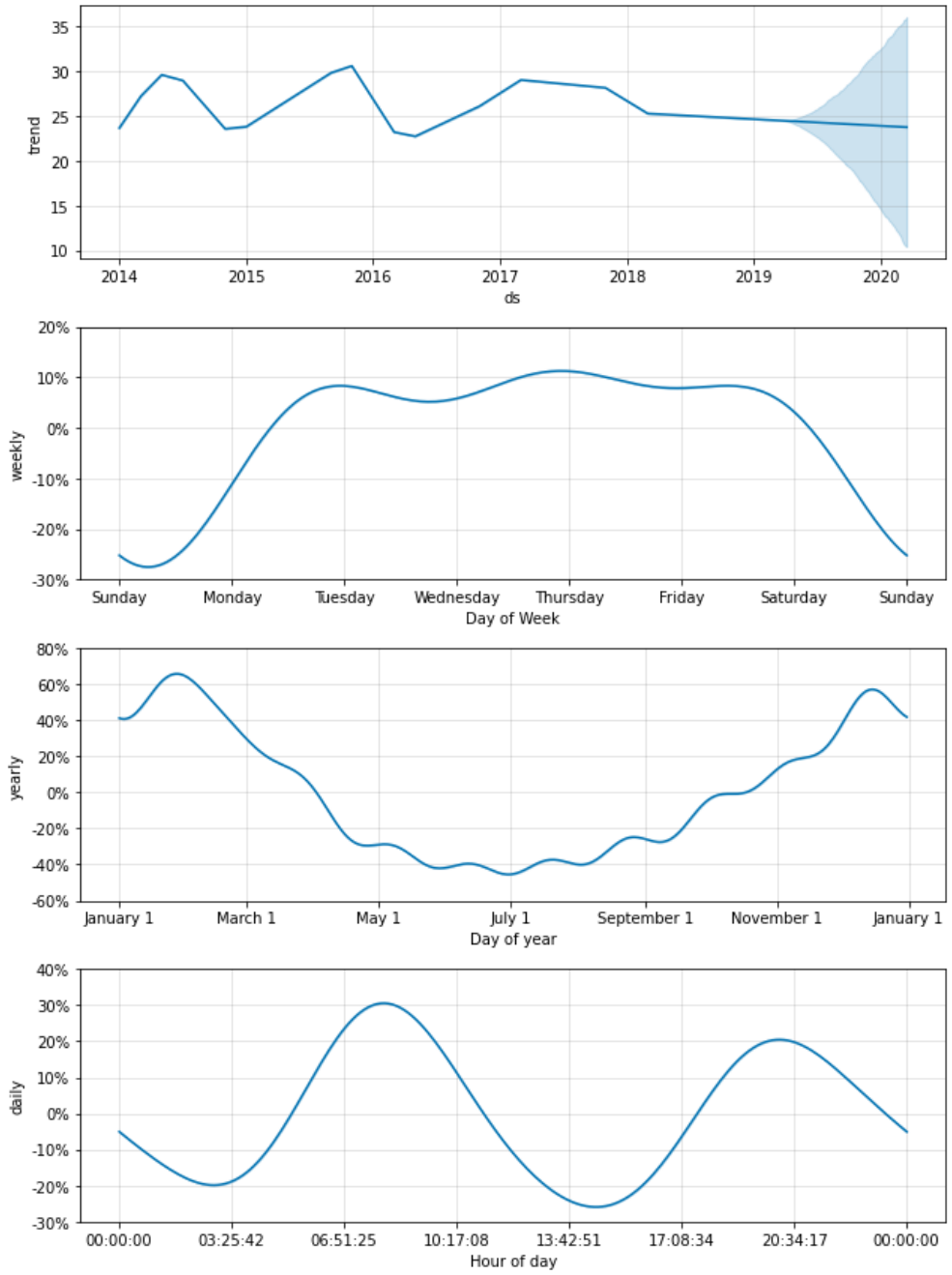
DODATAK 1



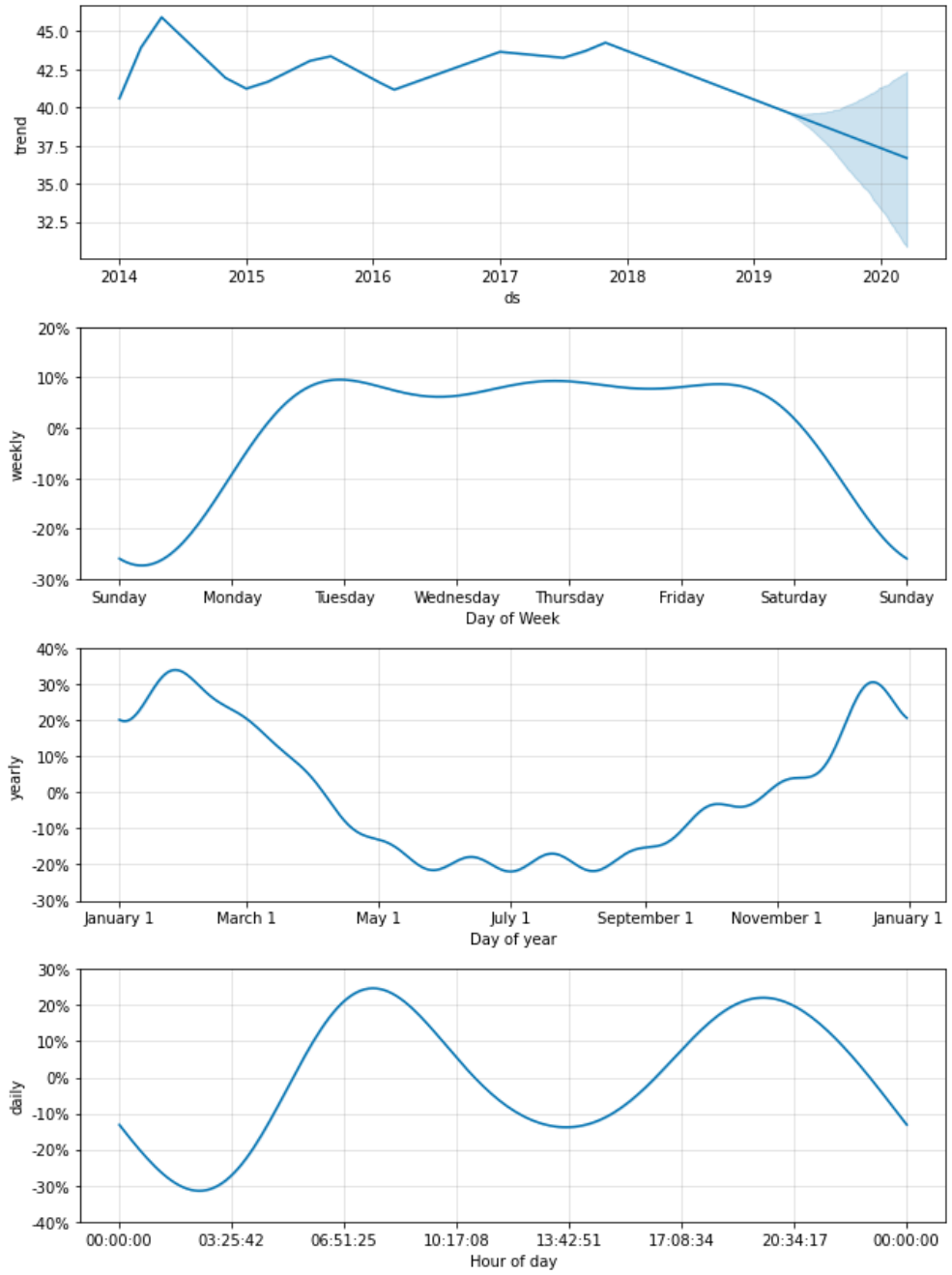
Slika 1. Komponente predviđanja *Prophet* modela s uključenim ekstremnim vrijednostima za Sjever.



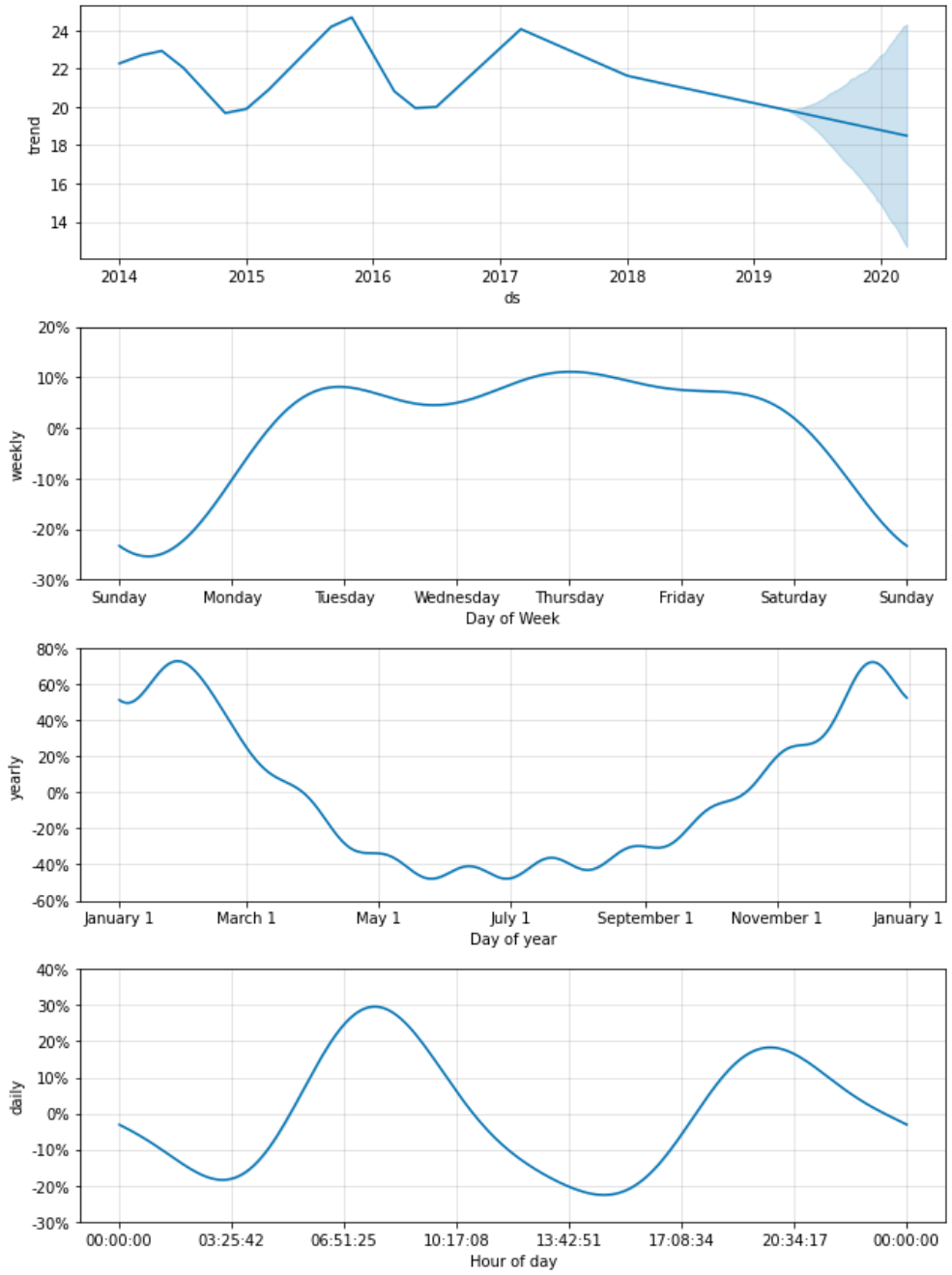
Slika 2. Komponente predviđanja *Prophet* modela s uključenim ekstremnim vrijednostima za Jug.



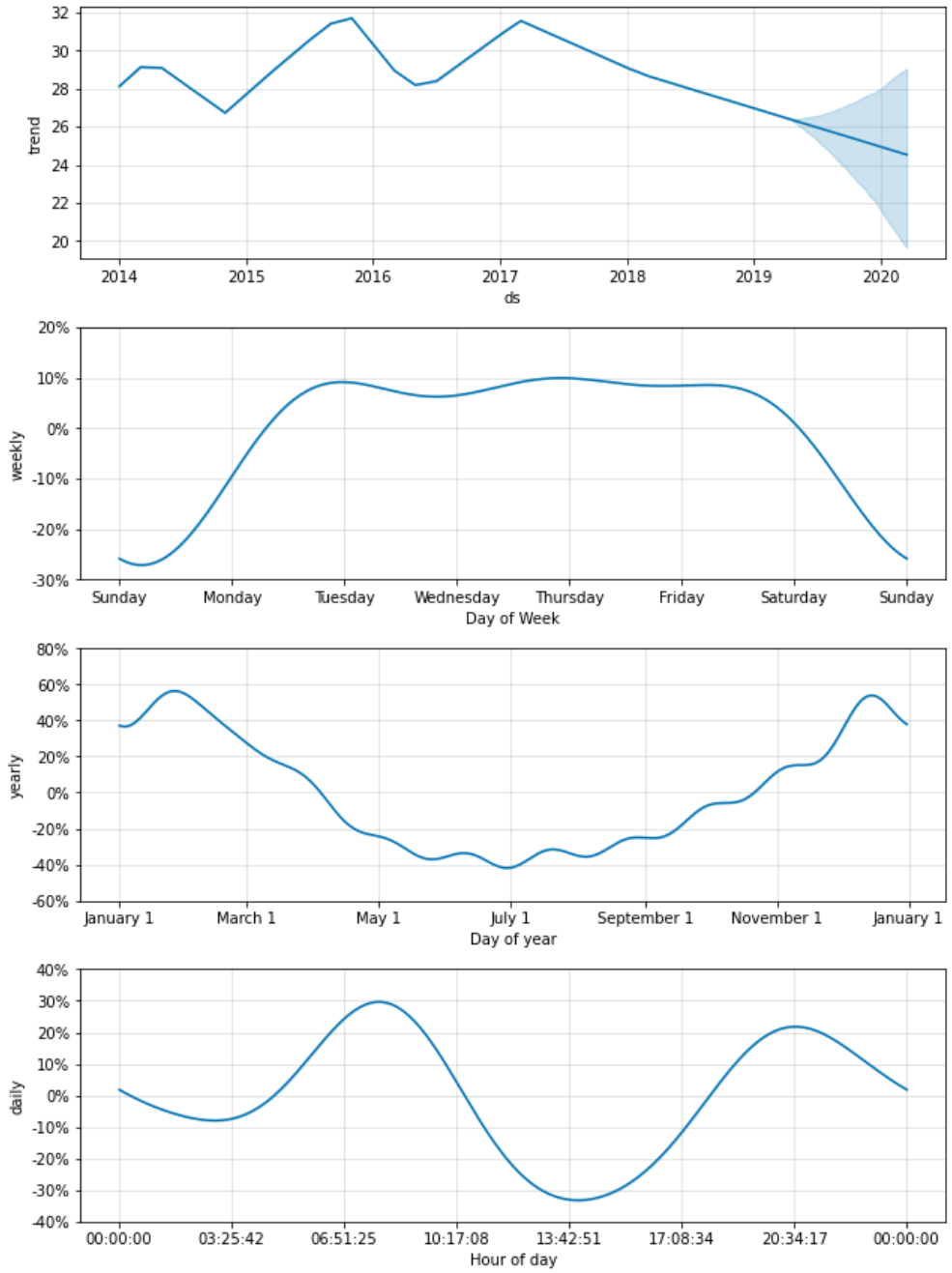
Slika 3. Komponente predviđanja *Prophet* modela s uključenim ekstremnim vrijednostima za Zapad.



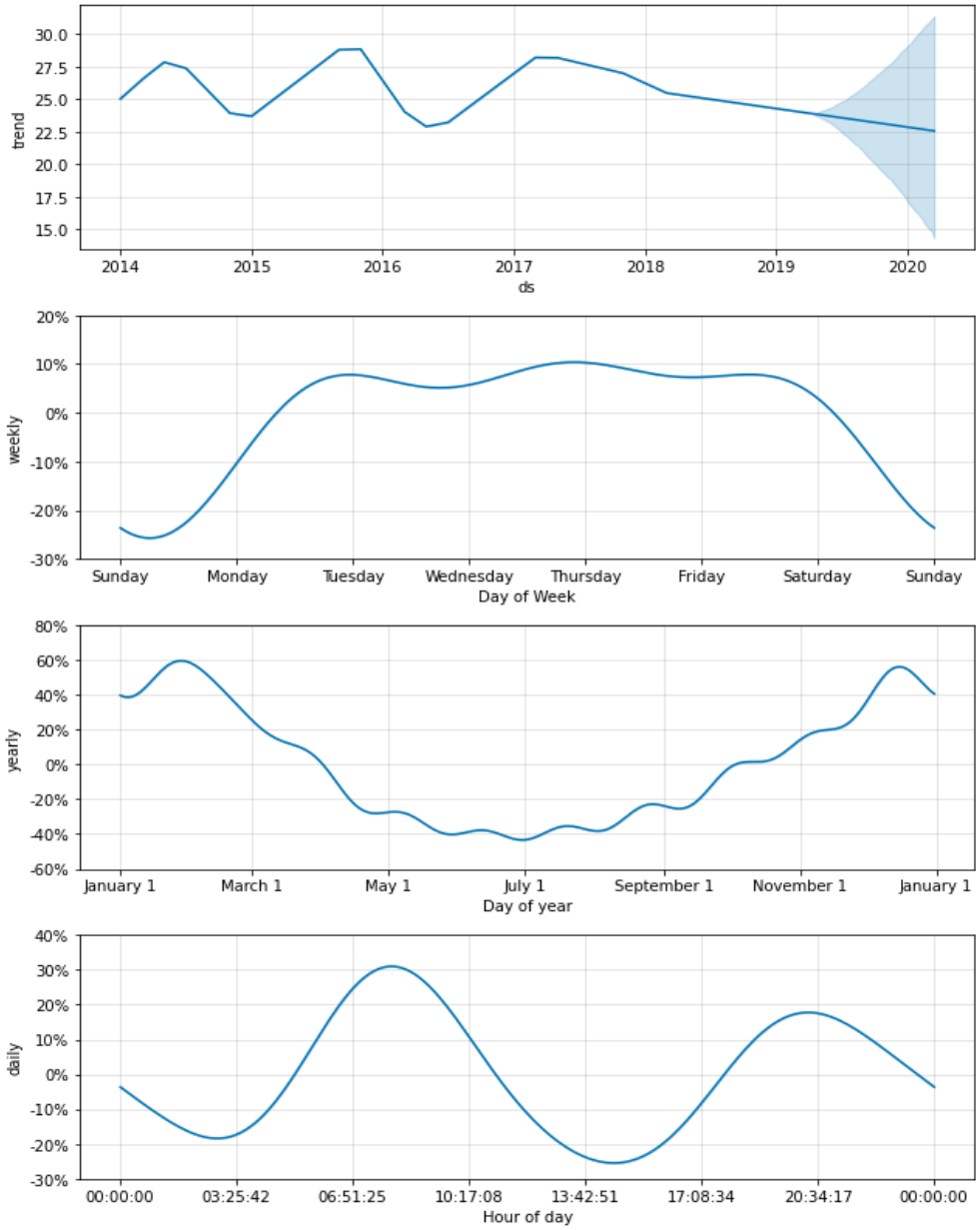
Slika 4. Komponente predviđanja *Prophet* modela sa isključenim ekstremnim vrijednostima za Don Bosco.



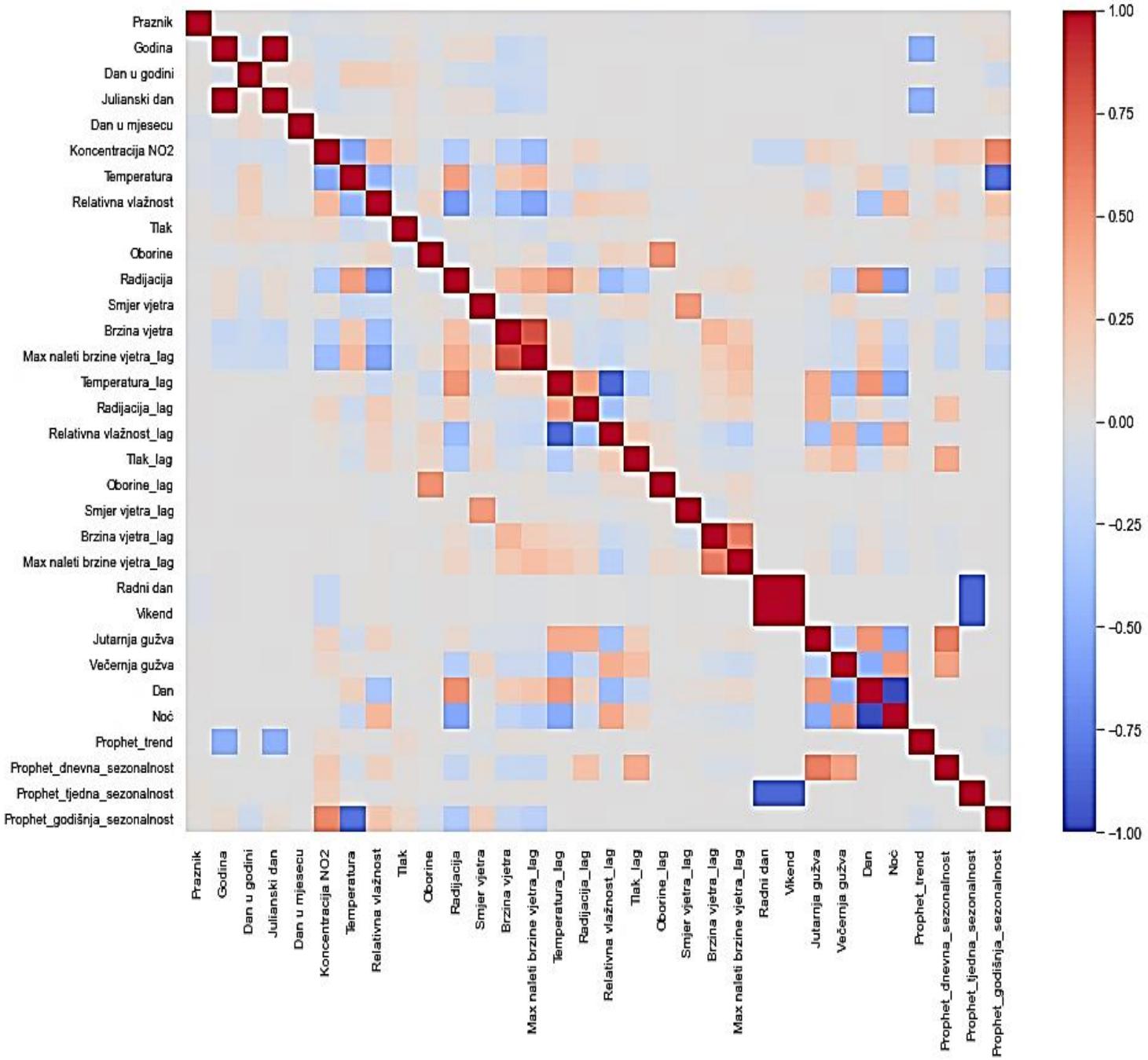
Slika 5. Komponente predviđanja *Prophet* modela sa isključenim ekstremnim vrijednostima za Sjever.



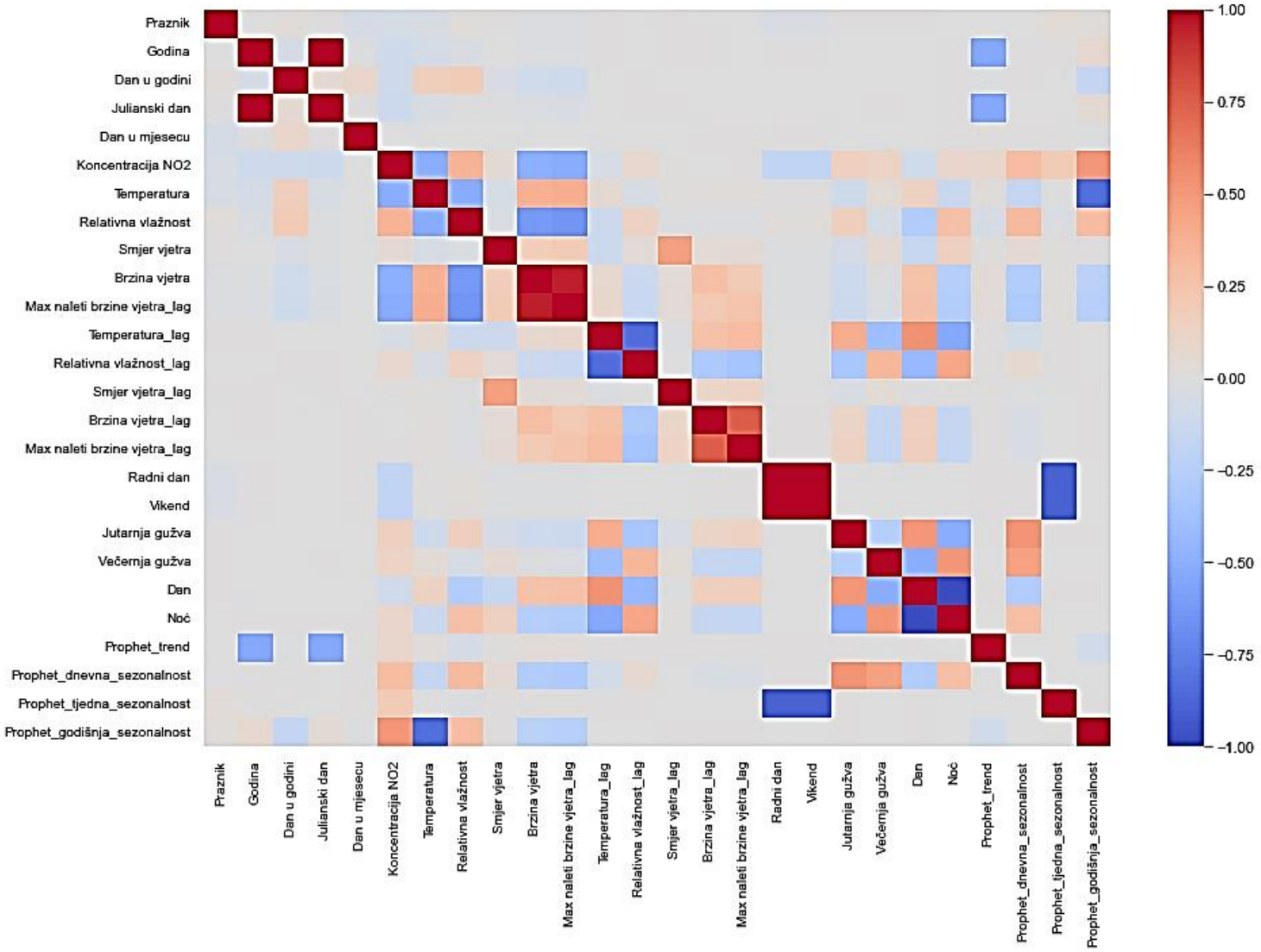
Slika 6. Komponente predviđanja *Prophet* modela sa isključenim ekstremnim vrijednostima za Jug.



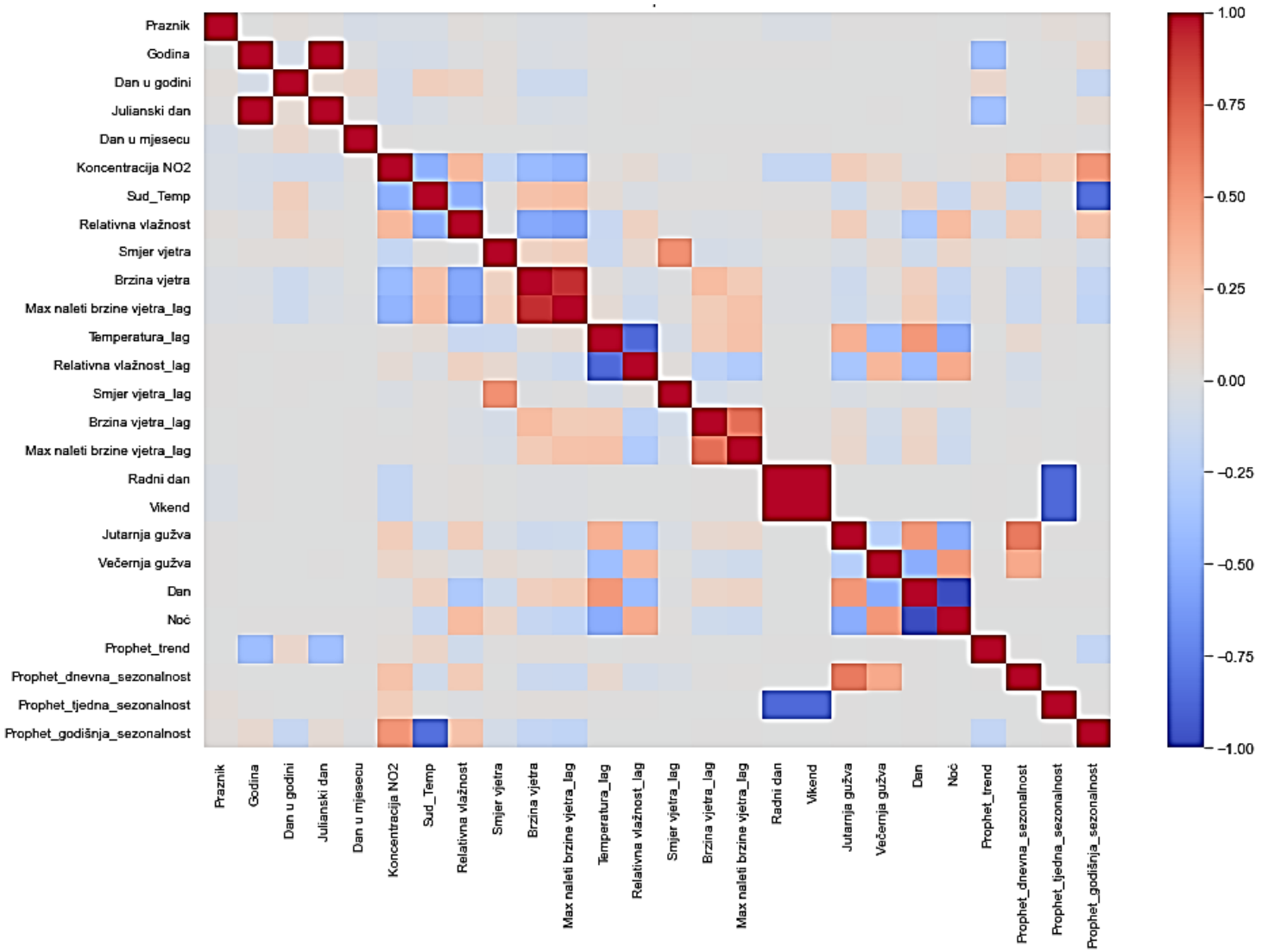
Slika 7. Komponente predviđanja *Prophet* modela sa isključenim ekstremnim vrijednostima za Zapad.



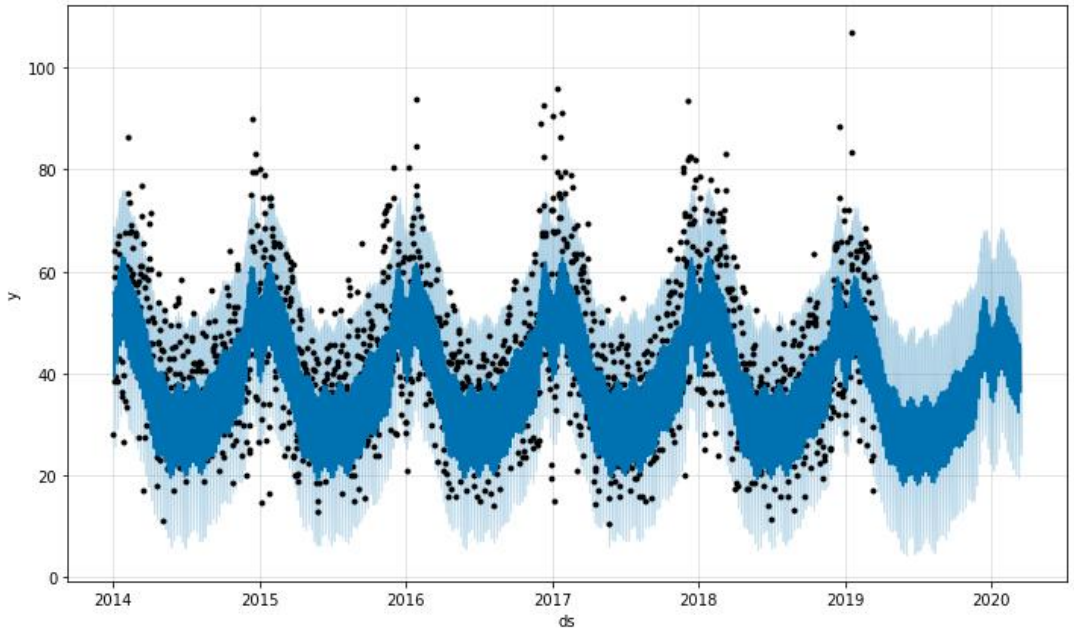
Slika 8. Toplinska karta međusobnih korelacija značajki korištenih za razvoj *Random Forest* modela (Sjever).



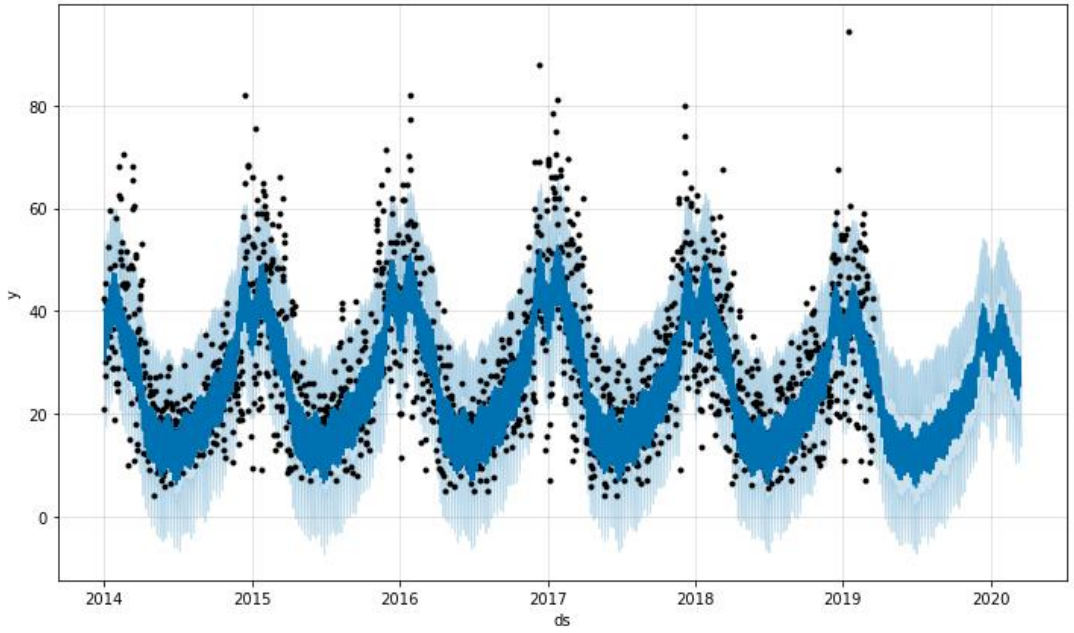
Slika 9. Toplinska karta međusobnih korelacija značajki korištenih za razvoj *Random Forest* modela (Jug).



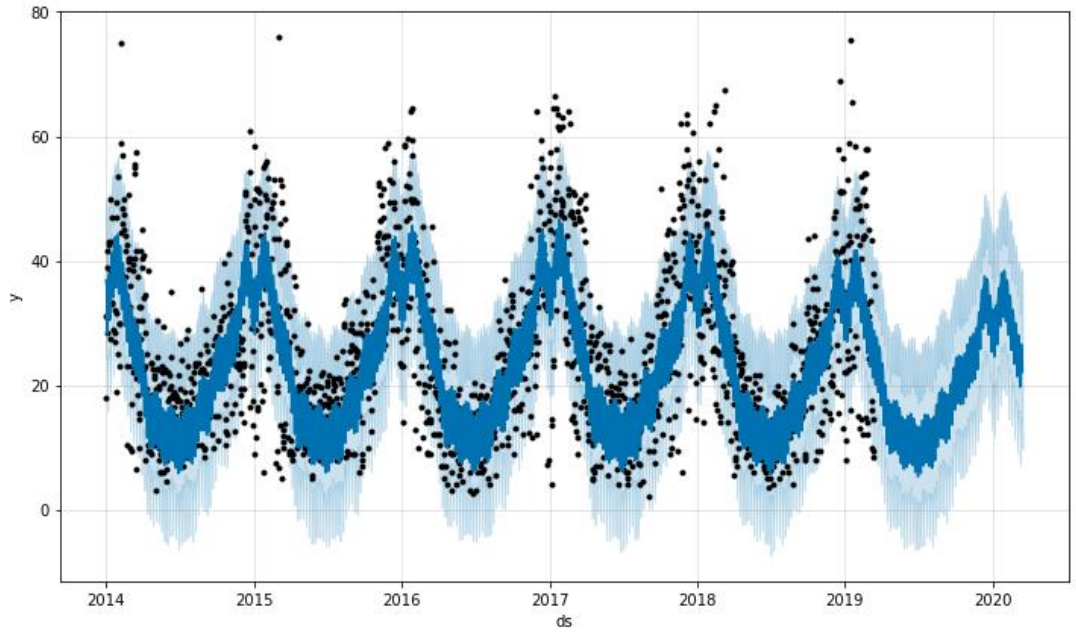
Slika 10. Toplinska karta međusobnih korelacija značajki korištenih za razvoj *Random Forest* modela (Zapad).



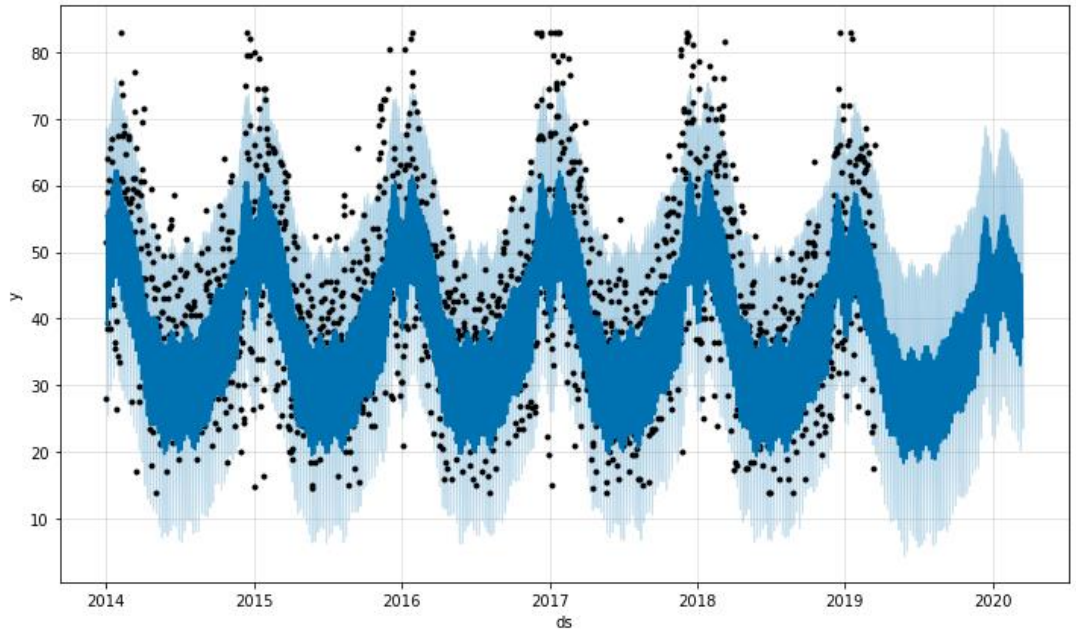
Slika 11. Predviđanje *Prophet* modela po danima s uključenim ekstremnim vrijednostima za mjernu postaju Don Bosco.



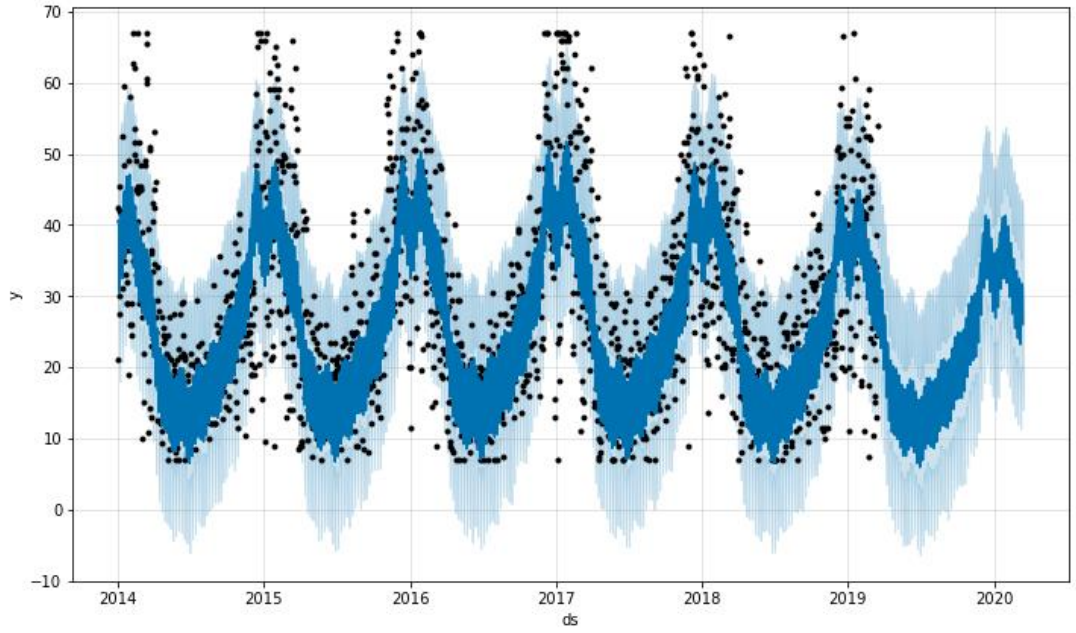
Slika 12. Predviđanje *Prophet* modela po danima s uključenim ekstremnim vrijednostima za mjernu postaju Jug.



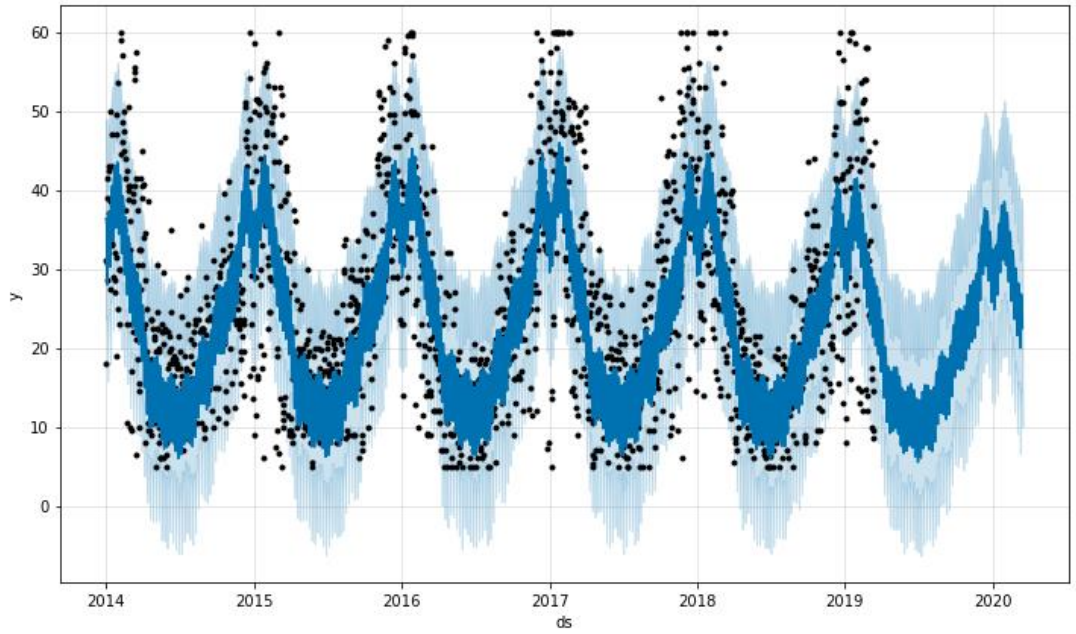
Slika 13. Predviđanje *Prophet* modela po danima s uključenim ekstremnim vrijednostima za mjernu postaju Zapad.



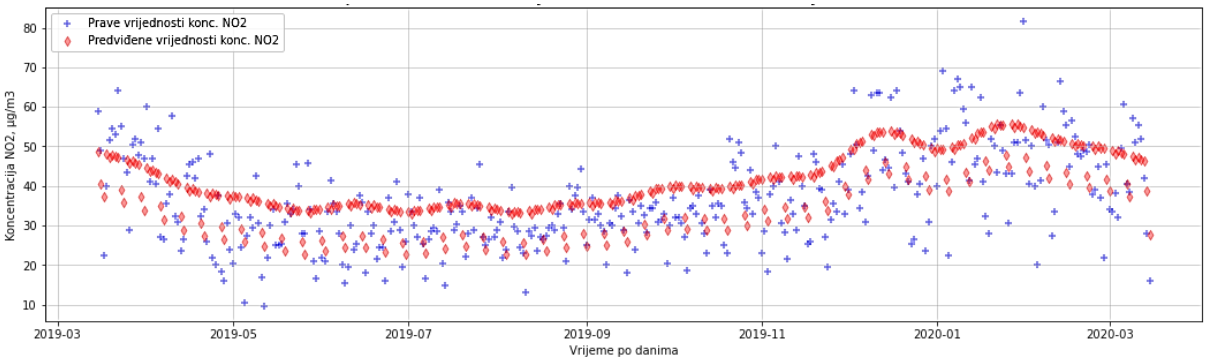
Slika 14. Predviđanje *Prophet* modela po danima sa isključenim ekstremnim vrijednostima za mjernu postaju Don Bosco.



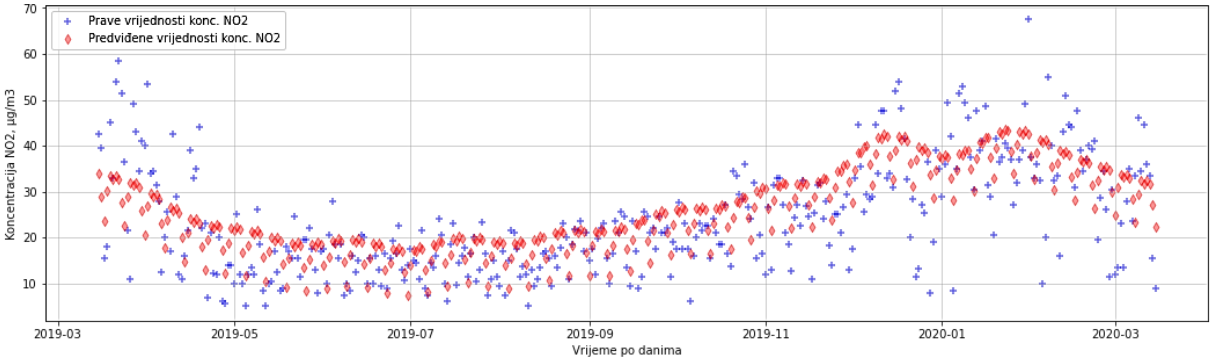
Slika 15. Predviđanje *Prophet* modela po danima sa isključenim ekstremnim vrijednostima za mjernu postaju Jug.



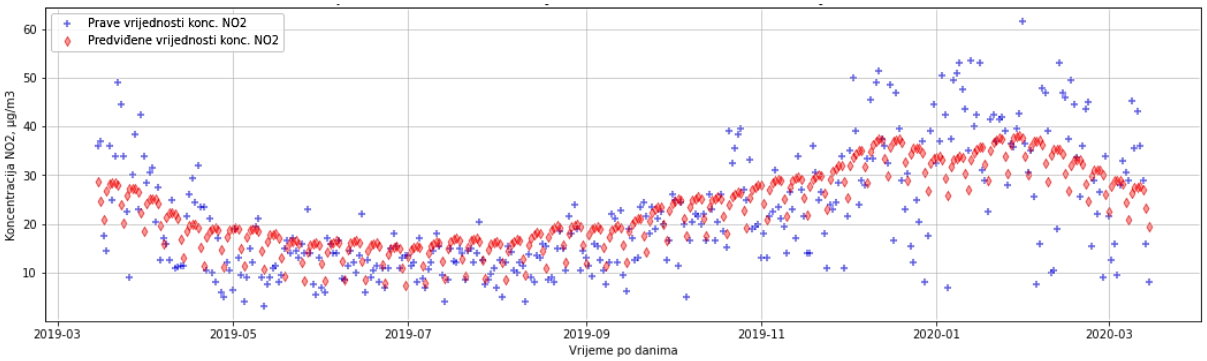
Slika 16. Predviđanje *Prophet* modela po danima sa isključenim ekstremnim vrijednostima za mjernu postaju Zapad.



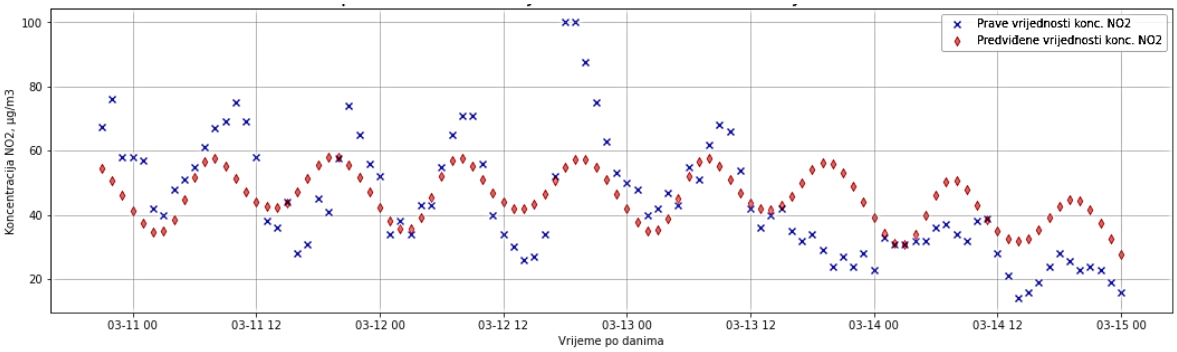
Slika 17. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 agregiranih po danu i dobivenih na temelju *Prophet* modela za postaju Don Bosco.



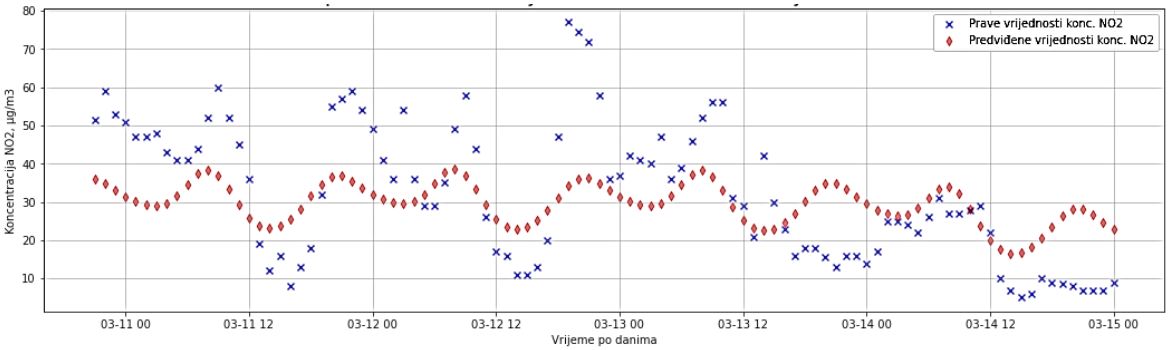
Slika 18. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 agregiranih po danu i dobivenih na temelju *Prophet* modela za postaju Jug.



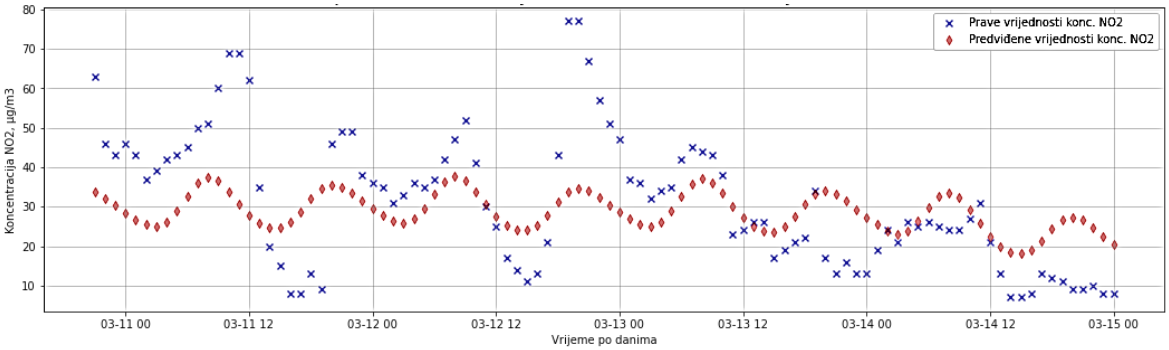
Slika 19. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 agregiranih po danu i dobivenih na temelju *Prophet* modela za postaju Zapad.



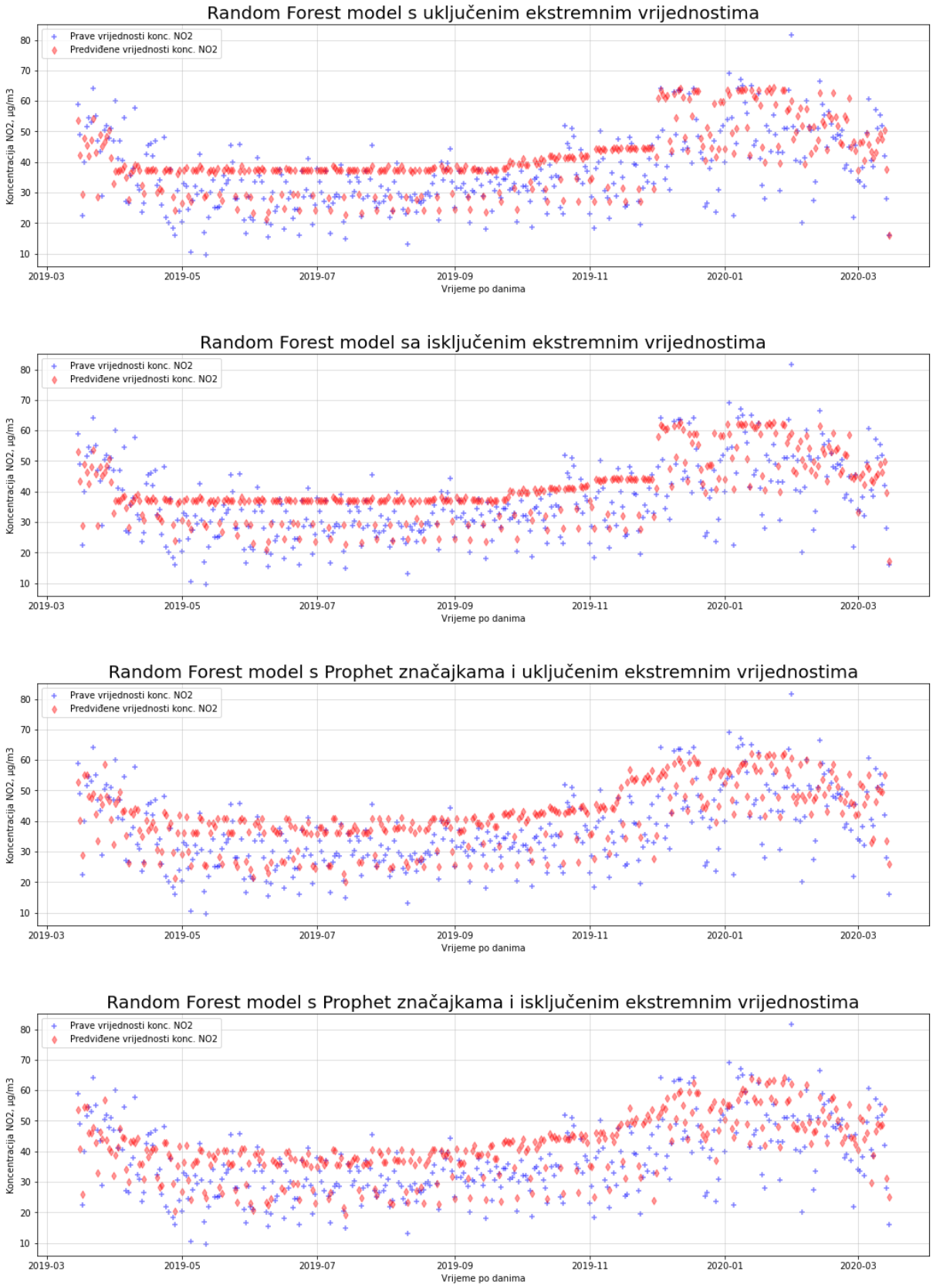
Slika 20. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 agregiranih po danima i dobivenih na temelju *Prophet* modela za postaju Don Bosco.



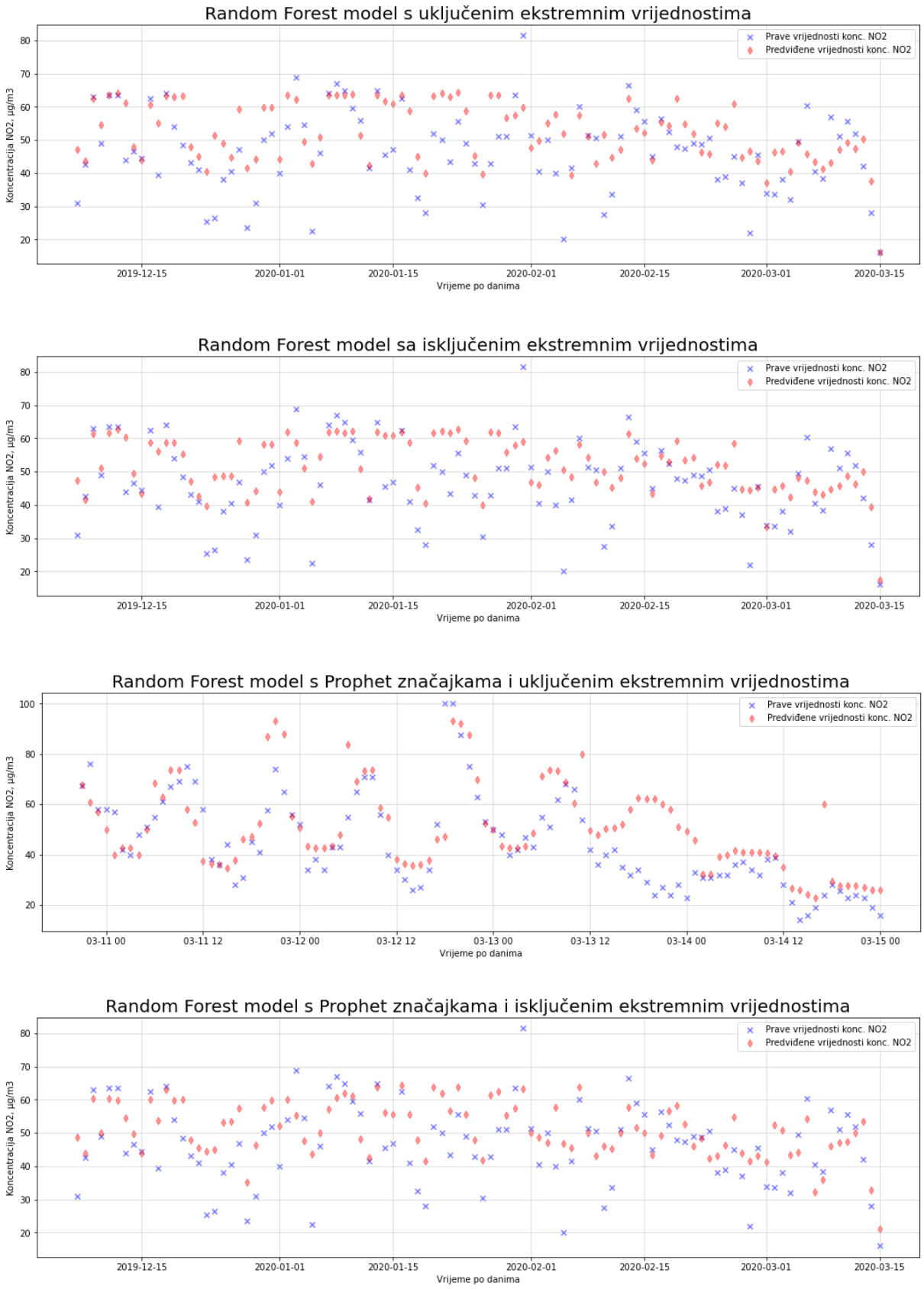
Slika 21. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 agregiranih po danima i dobivenih na temelju *Prophet* modela za postaju Jug.



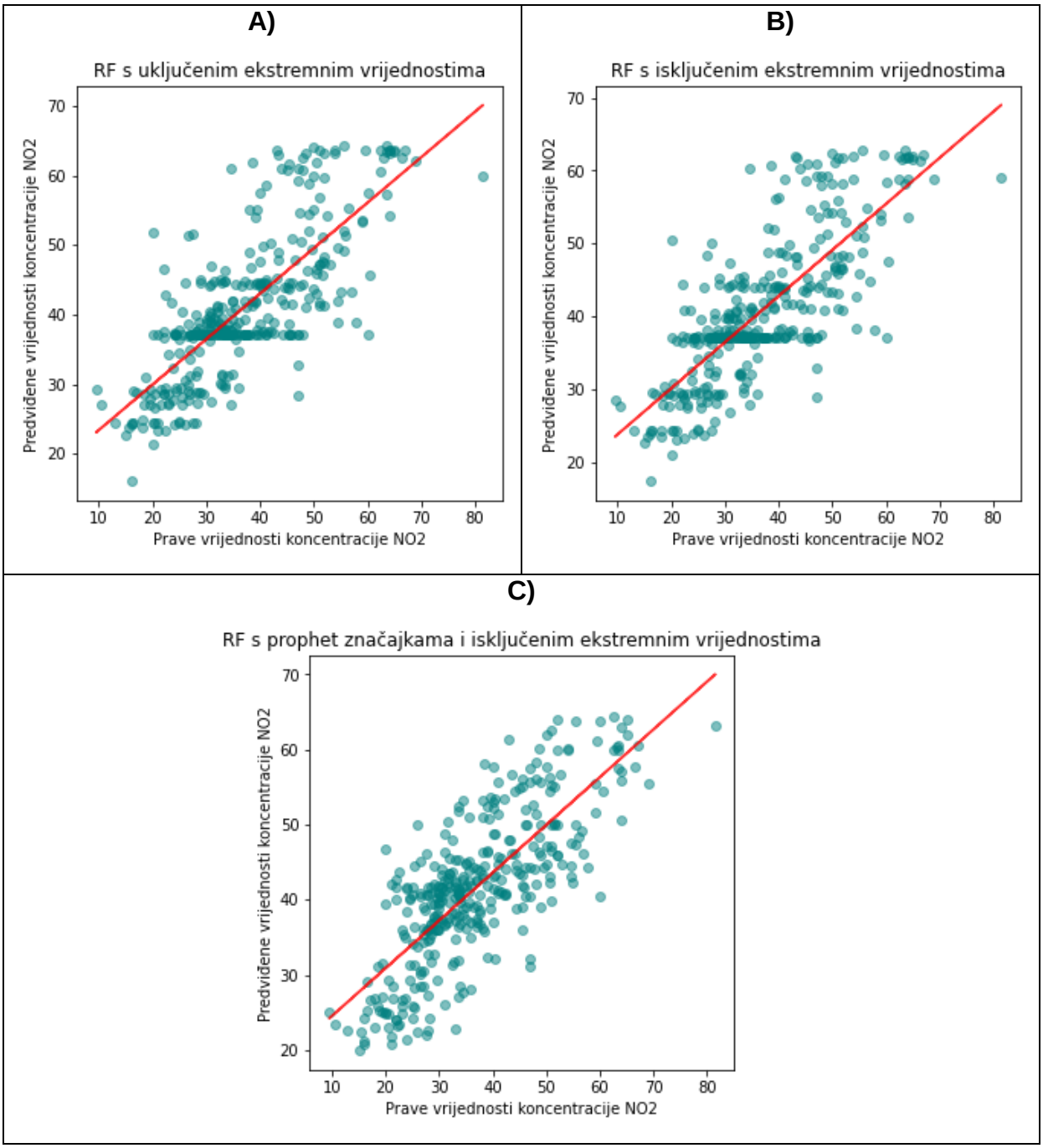
Slika 22. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 agregiranih po danima i dobivenih na temelju *Prophet* modela za postaju Zapad.



Slika 23. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 dobivenih na temelju četiri *Random Forest* modela za Don Bosco.



Slika 24. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 dobivenih na temelju četiri *Random Forest* modela za Don Bosco.

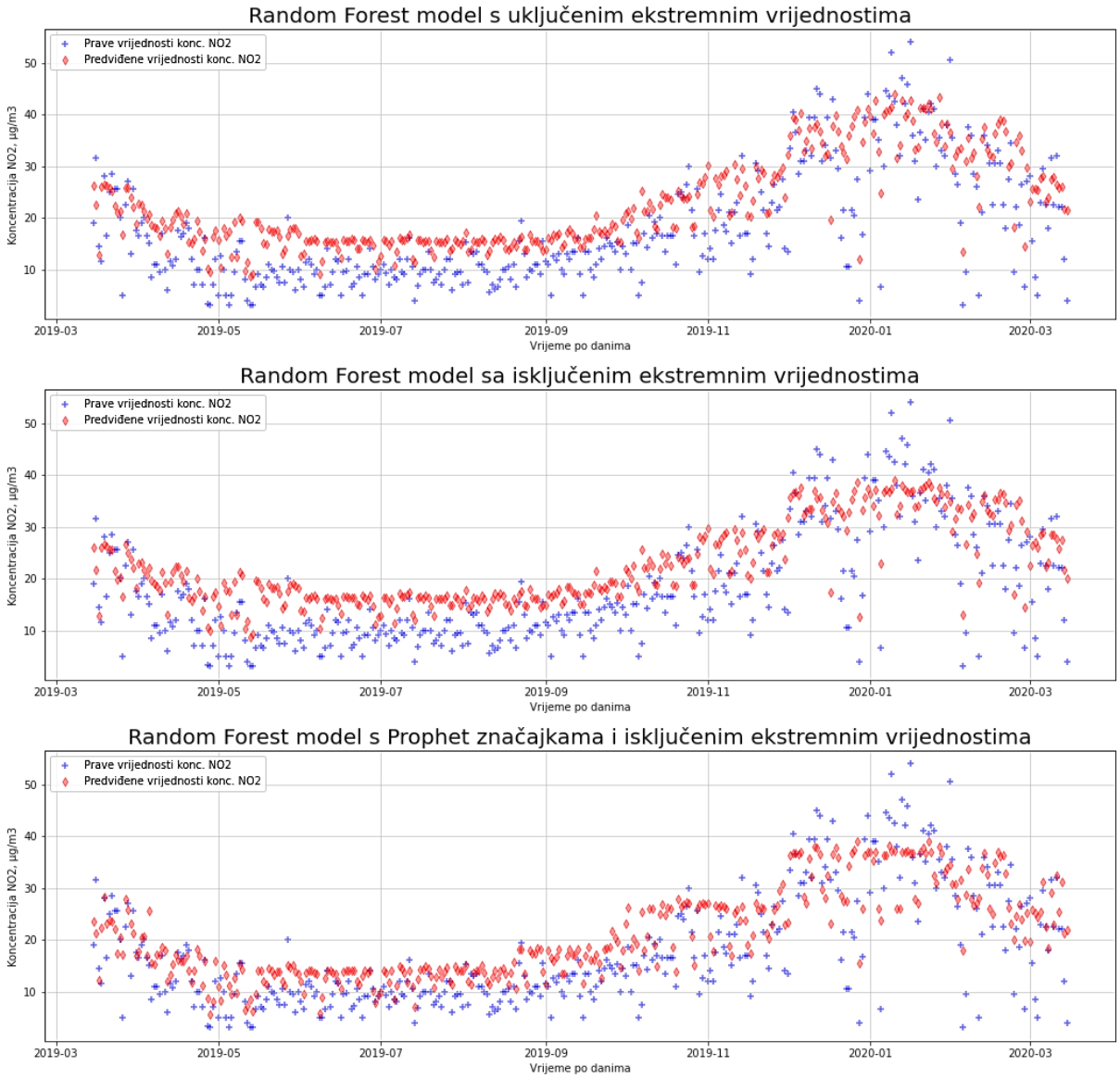


Slika 25. Prikaz regresijskog pravca i raspršenosti predviđenih vrijednosti koncentracija NO_2 za tri tipa *Random Forest* modela za Don Bosco:

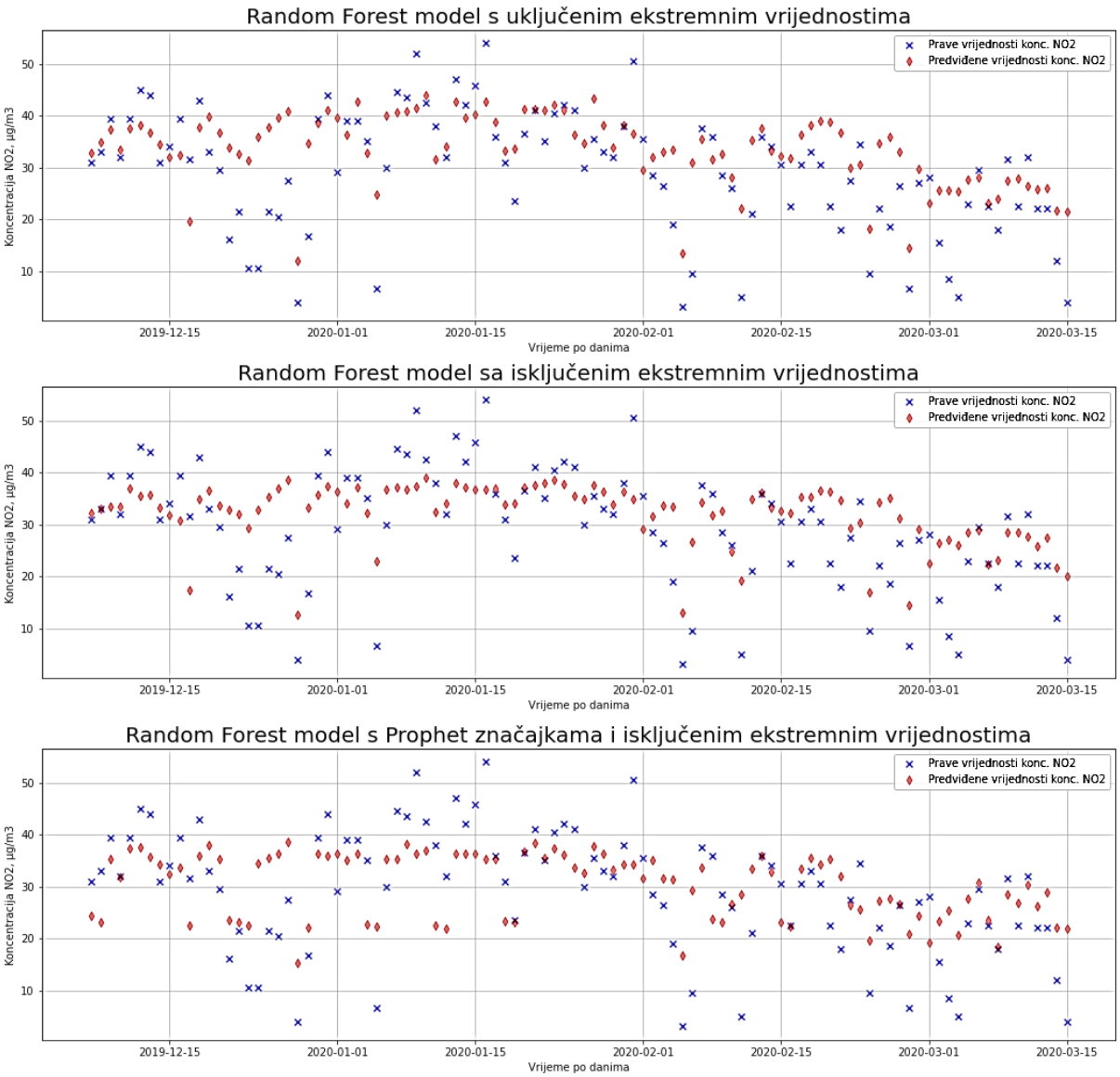
A) *Random Forest* model s uključenim ekstremnim vrijednostima

B) *Random Forest* model sa isključenim ekstremnim vrijednostima

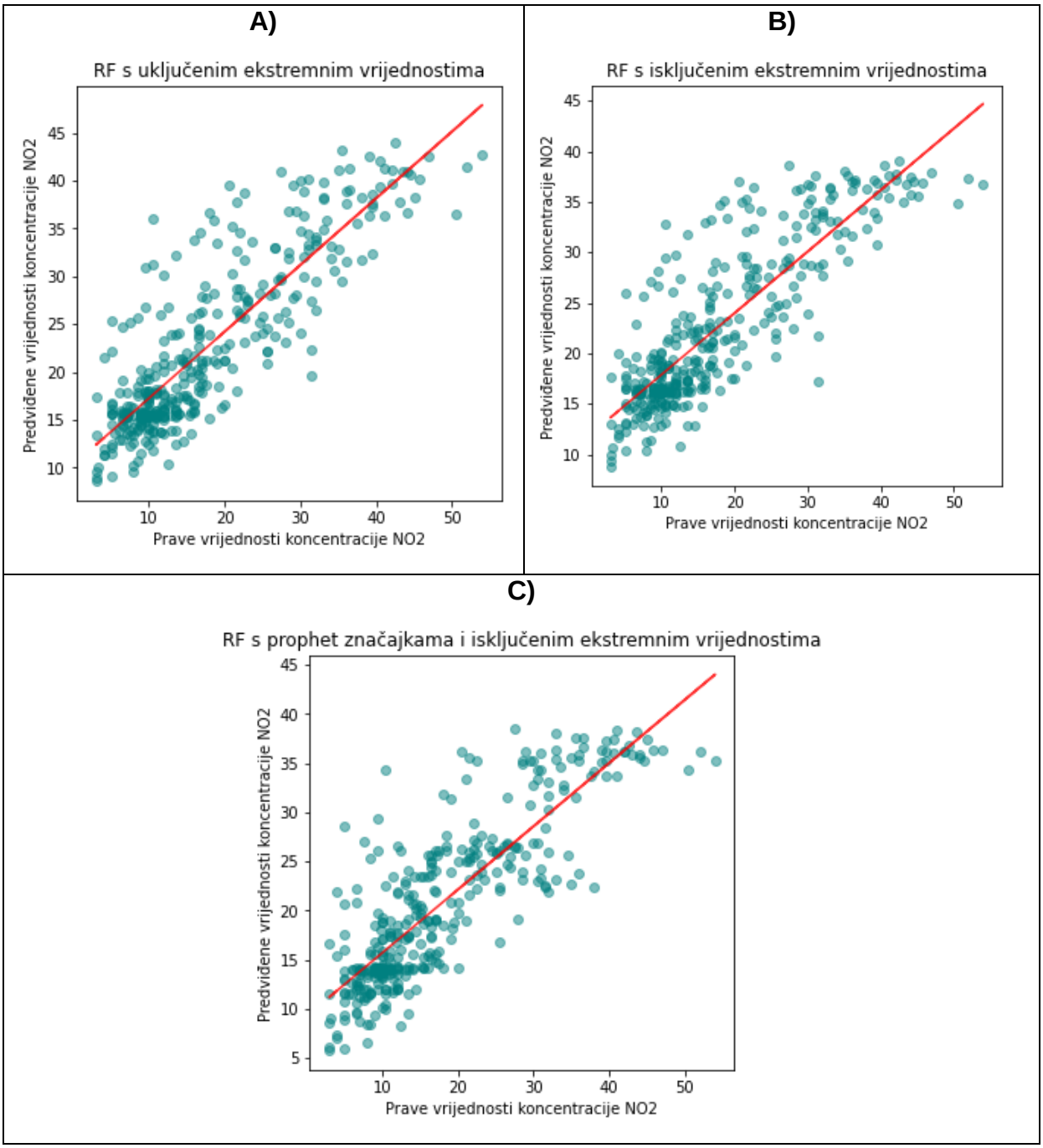
C) *Random Forest* model s *Prophet* značajkama i isključenim ekstremnim vrijednostima.



Slika 26. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 dobivenih na temelju preostala tri *Random Forest* modela za Sjever.



Slika 27. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 dobivenih na temelju preostala tri *Random Forest* modela za Sjever.

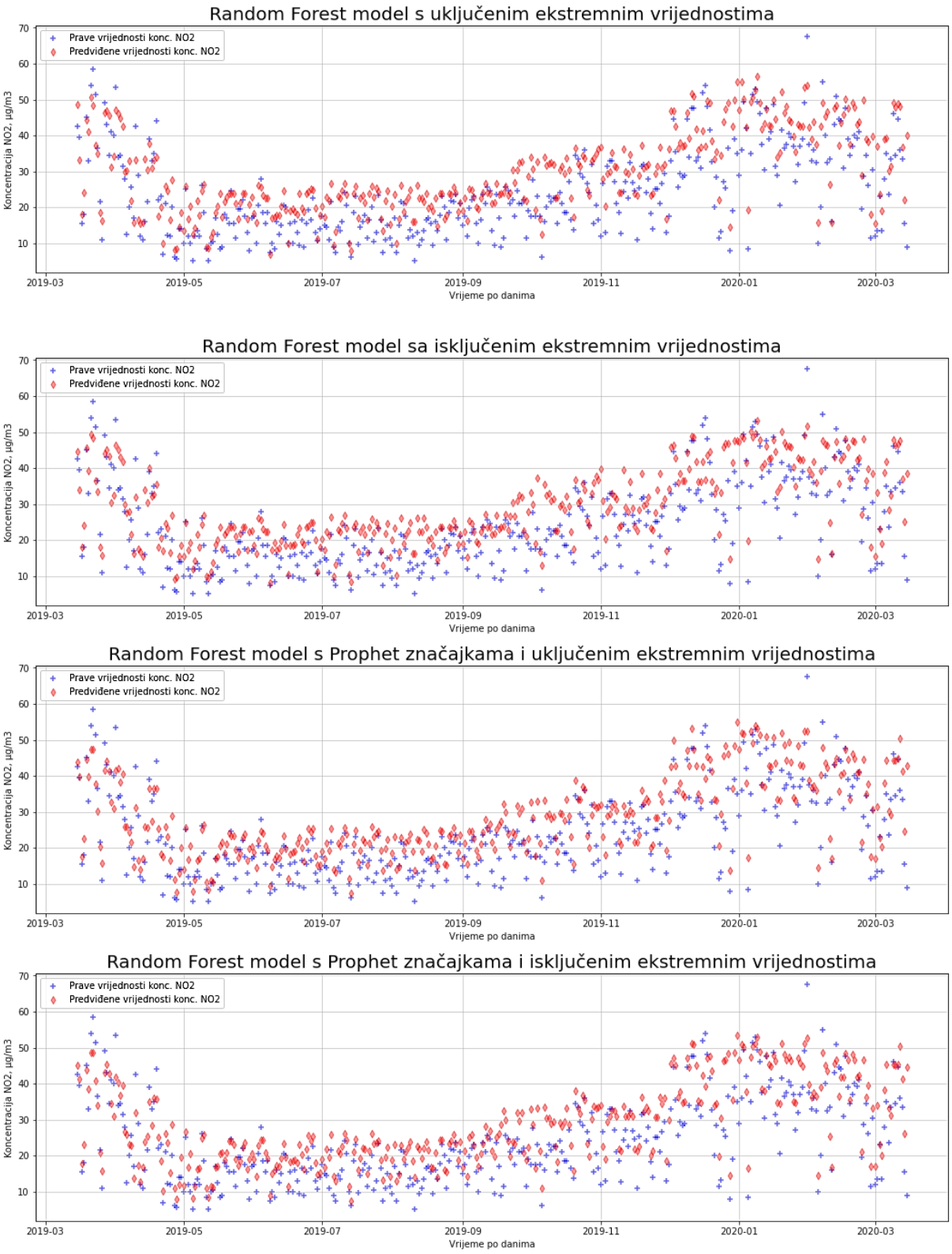


Slika 28. Prikaz regresijskog pravca i raspršenosti predviđenih vrijednosti koncentracija NO_2 za tri tipa *Random Forest* modela za Sjever:

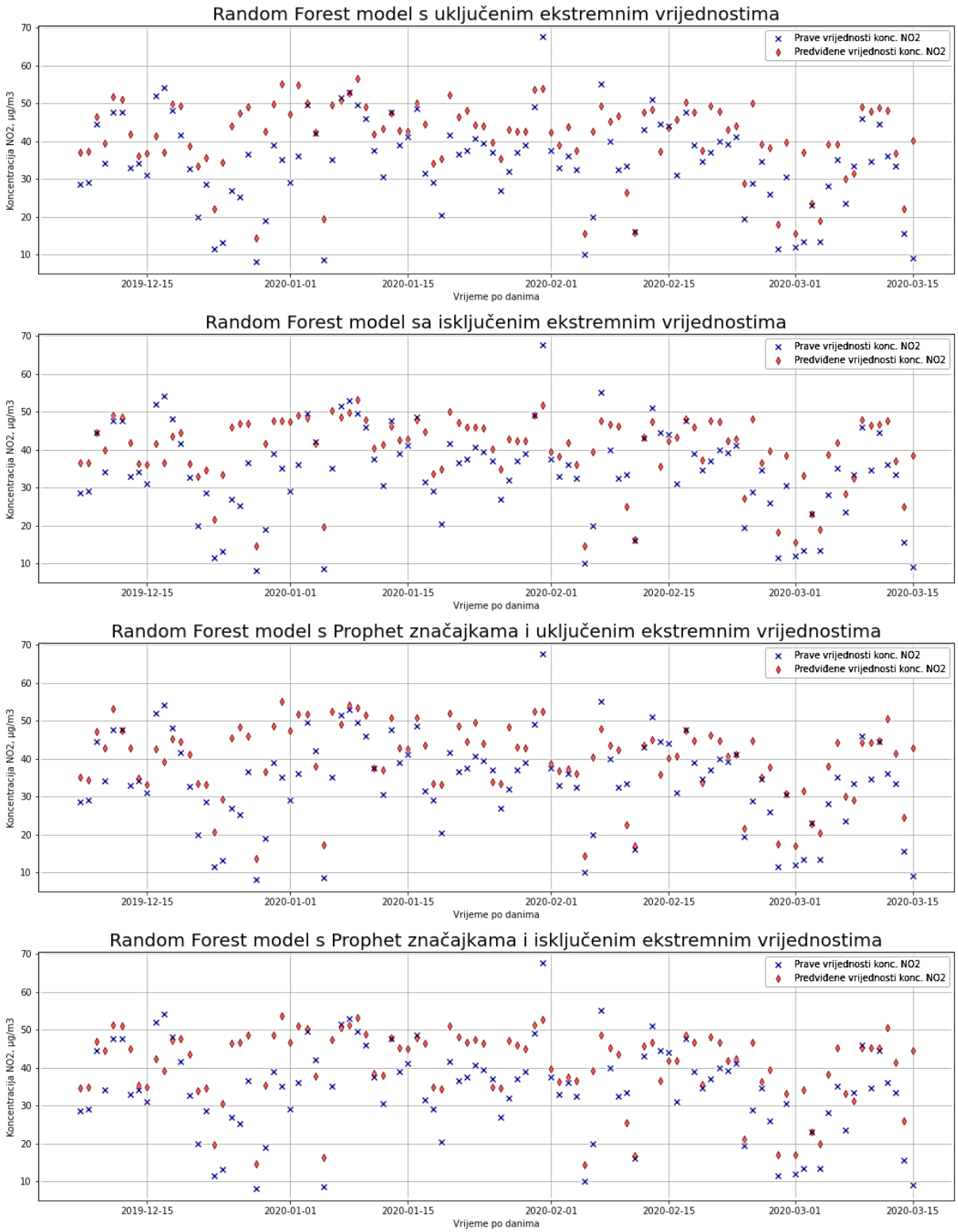
A) *Random Forest* model s uključenim ekstremnim vrijednostima

B) *Random Forest* model sa isključenim ekstremnim vrijednostima

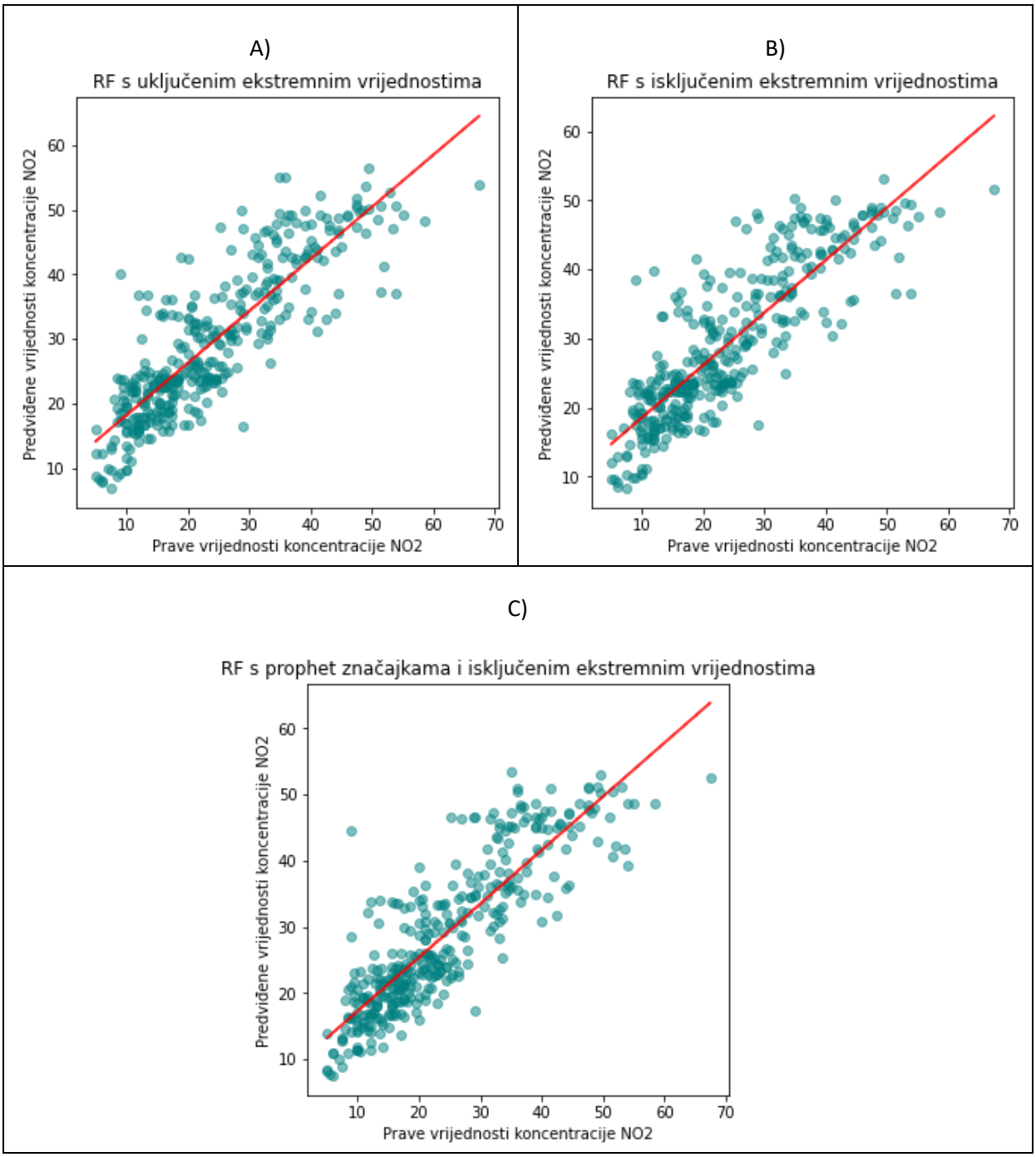
C) *Random Forest* model s *Prophet* značajkama i isključenim ekstremnim vrijednostima.



Slika 29. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 dobivenih na temelju četiri *Random Forest* modela za Jug.



Slika 30. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 dobivenih na temelju četiri *Random Forest* modela za Jug.

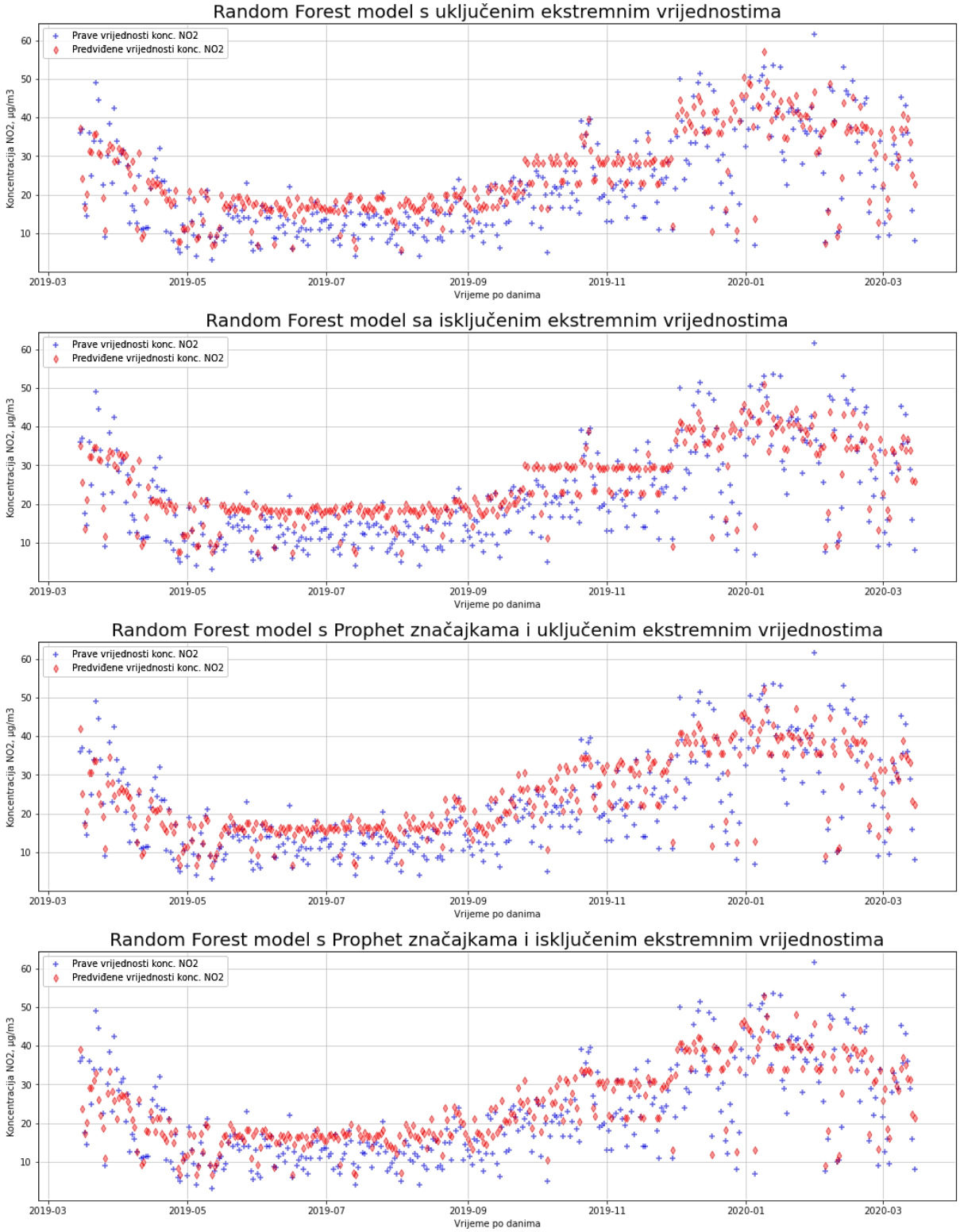


Slika 31. Prikaz regresijskog pravca i raspršenosti predviđenih vrijednosti koncentracija NO_2 za tri tipa *Random Forest* modela za Jug:

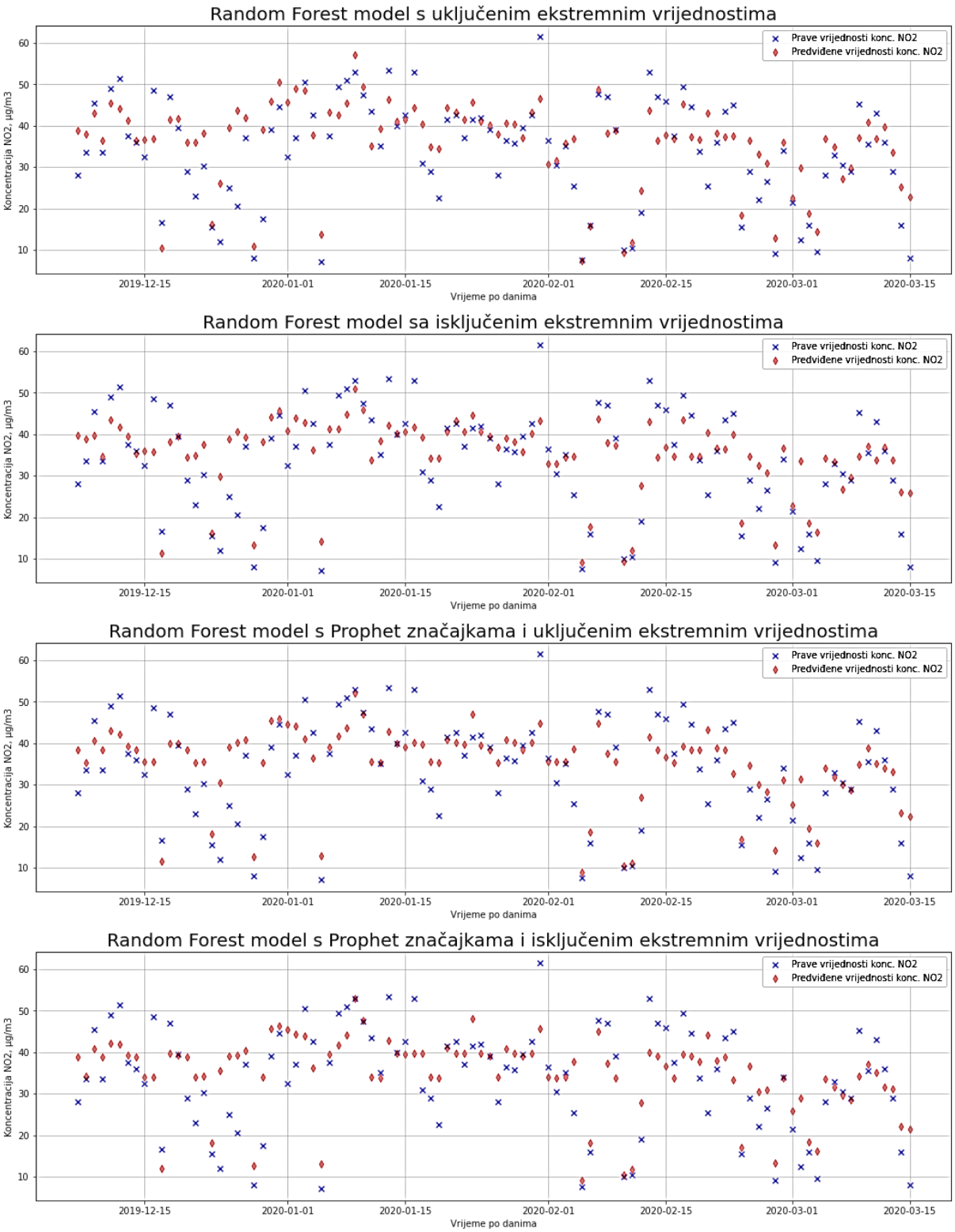
A) *Random Forest* model s uključenim ekstremnim vrijednostima

B) *Random Forest* model sa isključenim ekstremnim vrijednostima

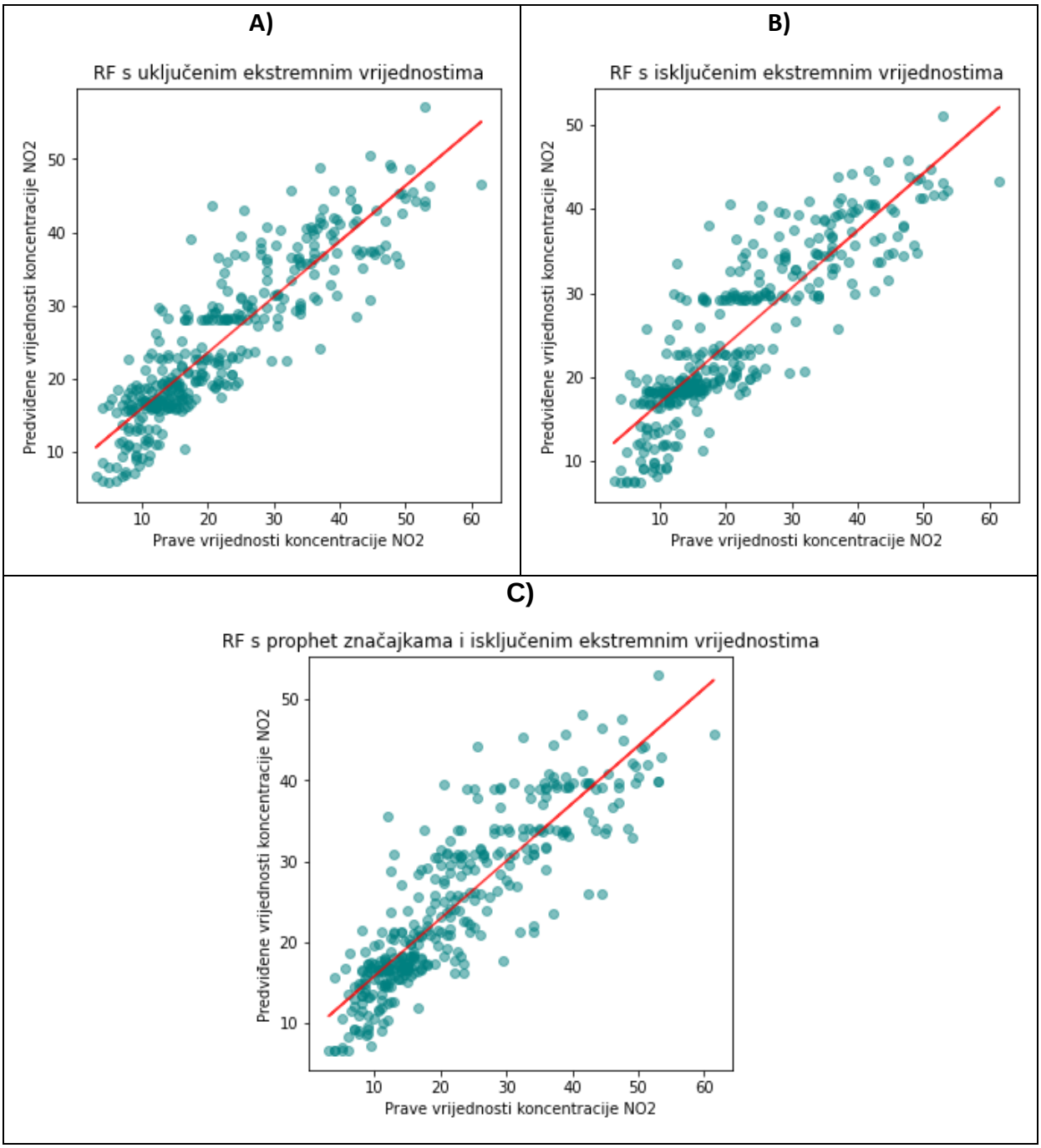
C) *Random Forest* model s *Prophet* značajkama i isključenim ekstremnim vrijednostima.



Slika 32. Usporedbe stvarnih i predviđenih vrijednosti koncentracija NO_2 dobivenih na temelju četiri *Random Forest* modela za Zapad.



Slika 33. Usporedbe posljednjih sto vrijednosti stvarnih i predviđenih koncentracija NO_2 dobivenih na temelju četiri *Random Forest* modela za Zapad.

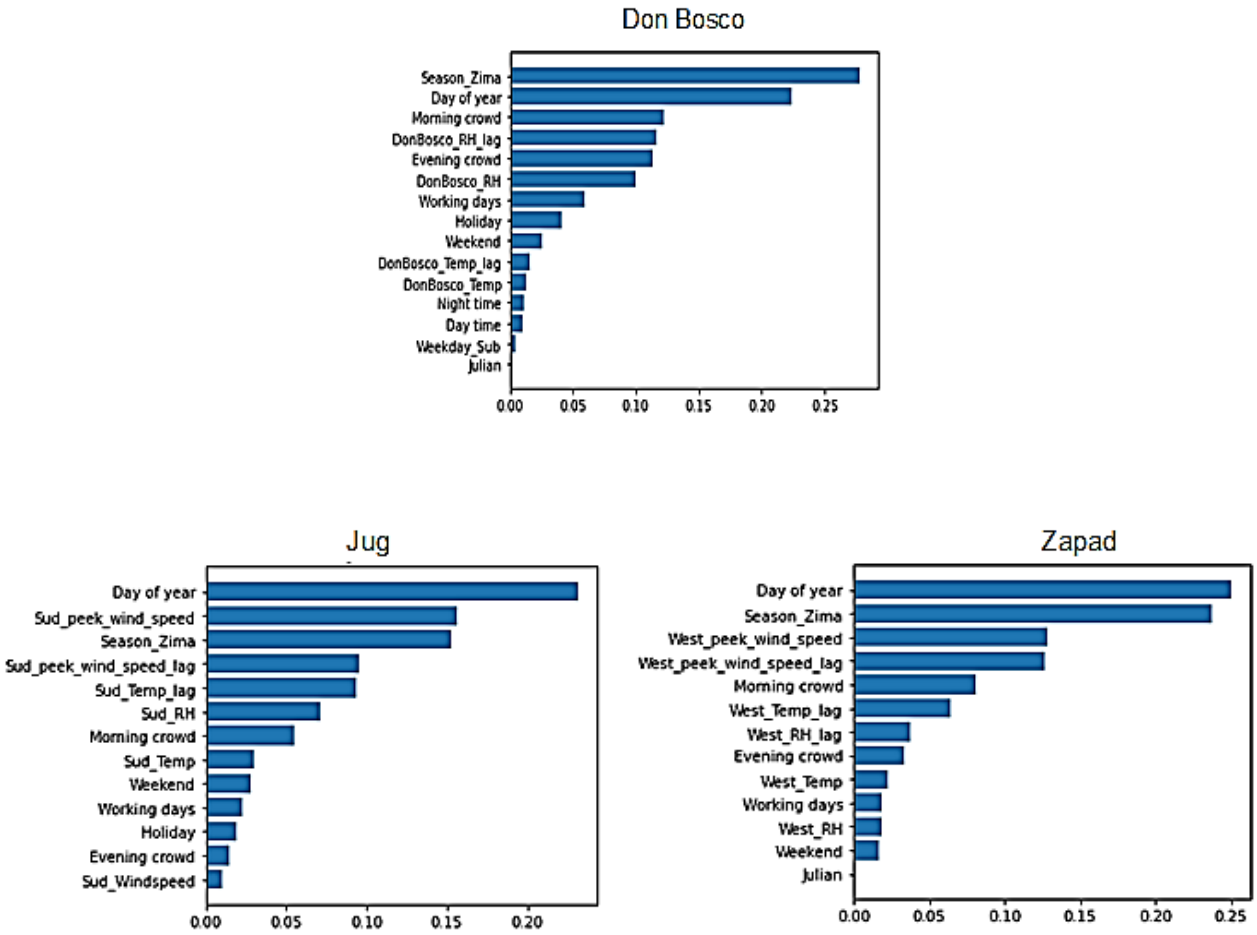


Slika 34. Prikaz regresijskog pravca i raspršenosti predviđenih vrijednosti koncentracija NO_2 za tri tipa *Random Forest* modela za Zapad:

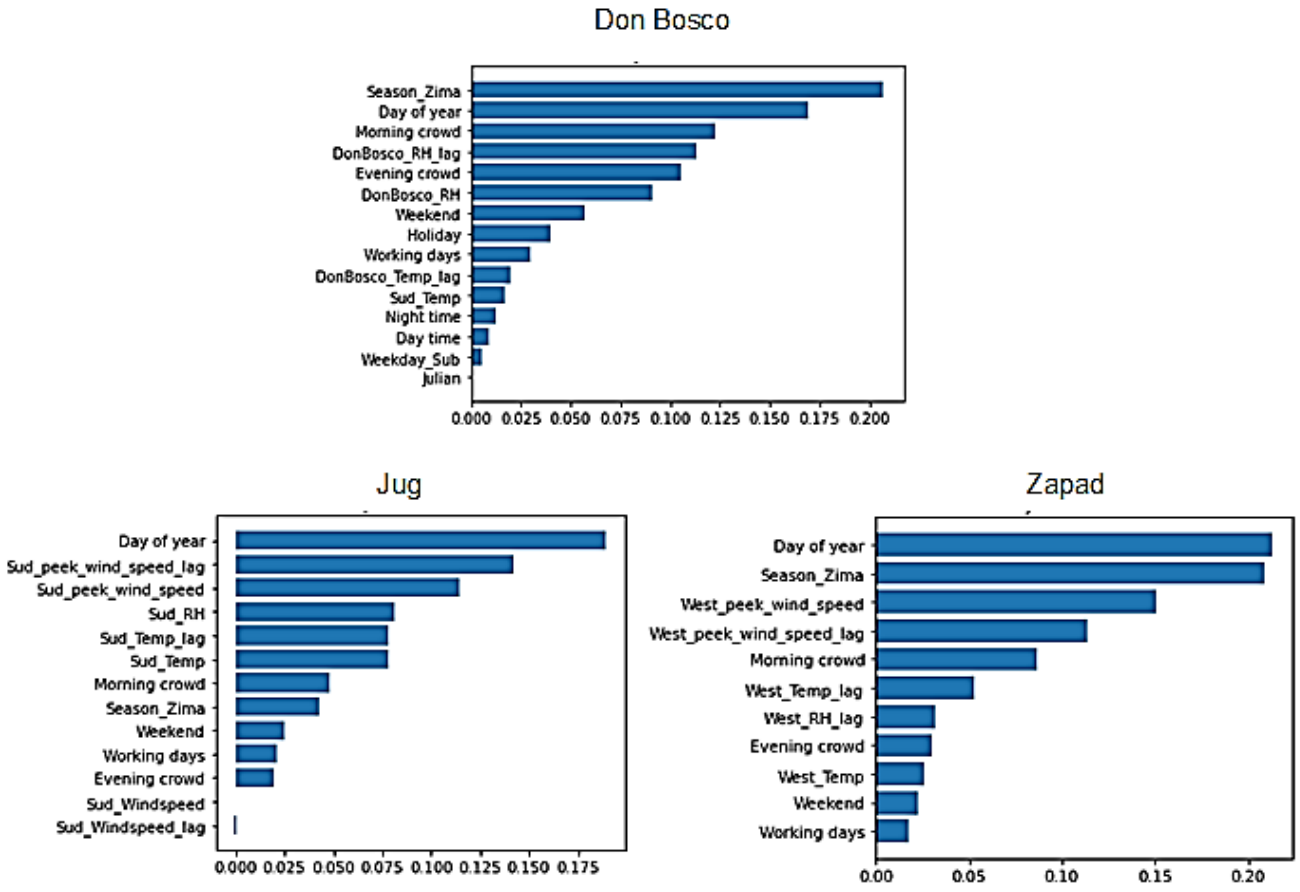
A) *Random Forest* model s uključenim ekstremnim vrijednostima

B) *Random Forest* model sa isključenim ekstremnim vrijednostima

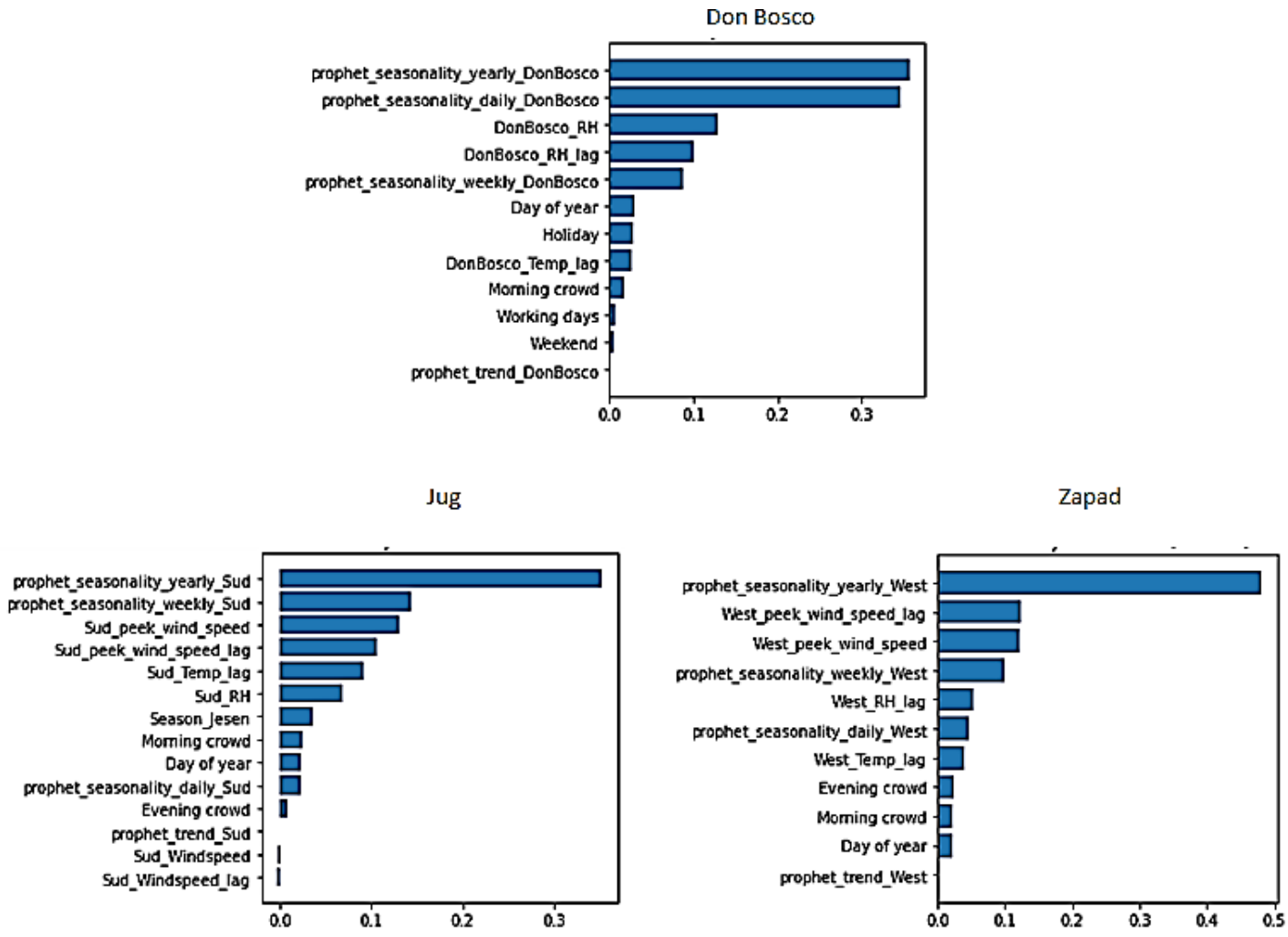
C) *Random Forest* model s *Prophet* značajkama i isključenim ekstremnim vrijednostima.



Slika 35. Izabrane značajke koje najviše utječu na rezultate predviđanja koncentracija NO_2 prikazane za *Random Forest* model s uključenim ekstremnim vrijednostima i bez značajki *Prophet* modela.

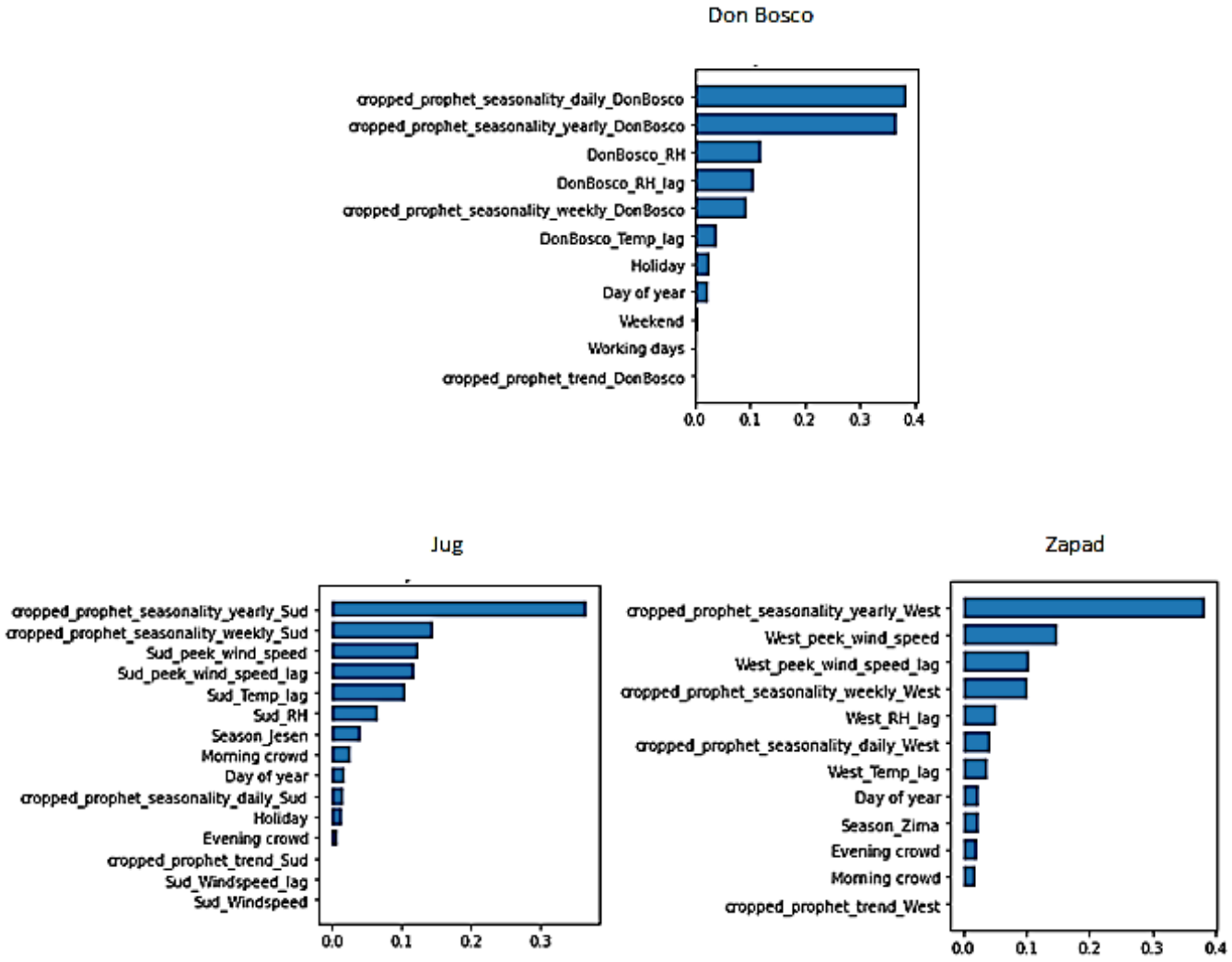


Slika 36. Izabrane značajke koje najviše utječu na rezultate predviđanja koncentracija NO_2 prikazane za *Random Forest* model sa isključenim ekstremnim vrijednostima i bez značajki *Prophet* modela.



Slika 37. Izabrane značajke koje najviše utječu na rezultate predviđanja koncentracija NO_2 prikazane za *Random Forest* model s uključenim ekstremnim vrijednostima i značajkama *Prophet* modela.

.



Slika 38. Izabrane značajke koje najviše utječu na rezultate predviđanja koncentracija NO_2 prikazane za *Random Forest* model sa isključenim ekstremnim vrijednostima i značajkama *Prophet* modela.

ŽIVOTOPIS

Ivana Kučinić [redacted] Školovanje započinje 2004. godine u osnovnoj školi „Mahično“. Po završetku osnovne škole, upisuje Gimnaziju Karlovac, opći smjer. Nakon položene državne mature, 2016. godine upisuje preddiplomski studij Kemija i inženjerstvo materijala na Fakultetu kemijskog inženjerstva i tehnologije Sveučilišta u Zagrebu. U rujnu 2020. godine, obranom završnog rada na temu „Primjena i uvećanje procesa mehanokemijske sinteze“ pod mentorstvom izv. prof. dr. sc. Igora Dejanovića, stječe naziv sveučilišne prvostupnice.

Osim školovanja, prioritet joj je i sport. Neprekidnih 17 godina aktivna je rukometašica u rukometnom klubu „RK Škola rukometa Karlovac“, a dolaskom na fakultet postaje članica rukometne selekcije FKIT-a s kojom 2016. godine sudjeluje na međunarodnom sveučilišnom natjecanju Euroijada u Pragu osvojivši drugo mjesto. Osim rukometa, dolaskom na fakultet postaje i članica ženske futsal selekcije s kojom 2017. i 2018. godine sudjeluje na znanstveno – sportskom natjecanju „Tehnologijada“.