

**Univerzitet u Beogradu
Interdisciplinarne studije**

Ognjen S. Milićević

**Računarsko modelovanje i mogućnosti
predikcije reproduktionog broja COVID-19
pomoću sredinskih i sociodemografskih
parametara**

DOKTORSKA DISERTACIJA

Beograd, 2026

**University of Belgrade
Interdisciplinary studies**

Ognjen S. Milićević

Computational modeling and prediction
capabilities of the COVID-19 reproduction
number using environmental and
sociodemographic parameters

DOCTORAL DISSERTATION

Belgrade, 2026

Mentori:

prof. dr Marko Đorđević, redovni profesor,
Biološki fakultet, Univerzitet u Beogradu

dr Igor Salom, naučni savetnik,
Institut za fiziku u Beogradu, Univerzitet u
Beogradu

Komisija:

dr Bojana Ilić, viši naučni saradnik,
Institut za fiziku u Beogradu, Univerzitet u Beogradu

dr Sofija Marković, naučni saradnik,
Biološki fakultet, Univerzitet u Beogradu

doc. dr Vedrana Pavlović, docent,
Medicinski fakultet, Univerzitet u Beogradu

Datum odbrane:

Zahvalnice

Doktorske studije, naročito one sa protrahovanim tokom kao što su to bile ove, prožmu život na jedan suptilan način. U tom smislu je lako zaboraviti i uspehe i prepreke, a naročito ljude koji su direktno i indirektno doprineli.

Svakako treba početi sa mentorima Markom Đorđevićem i Igorom Salomom, njihovom sjajnom istraživačkom grupom, i njihovim naizgled beskrajnim strpljenjem. Podržavali su mene i jedni druge i zaista napravili značajan naučni doprinos u koji su uključili mene. Uz zahvalnicu im želim da imaju strpljenja i vremena za još mnoge mlade (i ne tako mlade) doktorande.

Hvala profesorkama Aleksandri Knežević i Maji Stanojević što su me uključili u svet virusologije i omogućili mi drugu stranu uvida u kompleksni svetski problem pandemije COVID-19. Hvala mojim drugaricama i profesorkama Ani Banko i Danijeli Miljanović na mozganju o lokalnim okolnostima ovog svetskog događaja, nadam se da ćemo odraditi još koji krug po naučnom podijumu sa nekom vedrijom temom. Hvala Marku Jankoviću koji je u svoja pametovanja uključio moje matematikovanje i napravio nauku prijatnijom i zabavnijom.

Hvala kolegama sa katedre za Statistiku, po azbučnom redu Aleksi, Kseniji, Nataši, Nini i Vedrani. Drugari iz rovova su uvek tu jedni za druge. Hvala što ste me dočekali i sačekali u Aleji doktora nauka i specijalista.

Moram da se ponovo obratim mojim dvema rukama u životu Darku i Mileni za dug i zabavan život koji je preduslov ovog uspeha.

Hvala revidiranom Berlinskom kongresu na intelektualnom i emotivnom paritetu, na ljubavi i jednostavnoj i nezahtevnoj, na podršci u komplikovanju i rešavanju.

Hvala dobronamernim uzgrednim pitačima „a kada će kraj“ – kumovima, Šilji, Mladenu, Luki i drugima.

Mama Mileni dugujem svoje postojanje i sve svoje veštine, a u kontekstu ove teze njeno autentično pitanje „da li je do tebe“ će mi uvek vratiti perspektivu, osmeh i elan.

Zahvaljujem se mojoj Maci, boljoj i lepšoj polovini Ognjene Marije, koja je na samom početku našeg suživota uzela na sebe navigaciju ka završetku ove teze. Dva goluba će proći svaki kišni dan i naći svoju strehu i svoje sunce.

Računarsko modelovanje i mogućnosti predikcije reproduktionog broja COVID-19 pomoću sredinskih i sociodemografskih parametara

Ova doktorska disertacija ispituje ulogu faktora životne sredine i sociodemografskih faktora u transmisiji SARS-CoV-2 tokom rane faze pandemije COVID-19. Kao primarnu meru prenosivosti virusa, istraživanje koristi osnovni reproduktioni broj (R_0), procenjen iz inicijalne faze eksponencijalnog rasta epidemije, čime se umanjuju pristrasnosti uzrokovane razlikama u politikama testiranja i izveštavanja između regiona.

Istraživanje je strukturirano kroz dve komplementarne empirijske studije. Prva studija daje analizu na globalnoj skali, obuhvatajući 118 zemalja i integriše višedomenski skup podataka. Sezonalnost se pojavljuje kao modulator transmisije, sa jasnim hemisferski zavisnim sinusoidalnim obrascima – viši R_0 tokom zimskih, a niži tokom letnjih meseci – koji su formalno kvantifikovani Bajesijanskom MCMC analizom. Za reprezentativne severnohemisferske zemlje (Srbija, Litvanija) procenjene srednje vrednosti R_0 iznose 3,36–3,37, dok za Argentinu kao južnohemisfersku zemlju iznosi 3,70, sa invertovanom sezonskom dinamikom.

Druga studija analizira podatke za sve savezne države Sjedinjenih Američkih Država i, primenom regularizovanih regresionih metoda (Lasso, Elastic Net) i nelinearnih ansambl algoritama (Random Forest, Gradient Boosting), identifikuje fine čestične materije ($PM_{2.5}$) kao pojedinačno najznačajniji prediktor R_0 . Procenjuje se da varijacije u koncentracijama $PM_{2.5}$, u opsegu tipično zabeleženom u SAD, mogu izmeniti virusnu prenosivost za 25–40%, što uspostavlja kvantitativnu vezu između kvaliteta vazduha i otpornosti javnog zdravlja na respiratorne patogene.

Pored epidemioloških nalaza, disertacija naglašava značaj mogućnosti interpretacije modela, koristeći analizu važnosti prediktora unutar različitih algoritama. Informisan pristup je značajan za precizno usmeravanje javnozdravstvenih intervencija i strateško planiranje u uslovima visoke kompleksnosti epidemioloških podataka.

Ključne reči: COVID-19, SARS-CoV-2, osnovni reproduktioni broj, $PM_{2.5}$, zagađenje vazduha, sezonalnost, mašinsko učenje, epidemiološko modelovanje, javno zdravlje

Naučna oblast: Medicinske nauke

Uža naučna oblast: Statističko modelovanje

Computational modeling and prediction capabilities of the COVID-19 reproduction number using environmental and sociodemographic parameters

This doctoral dissertation examines the role of environmental and sociodemographic factors in the transmission of SARS-CoV-2 during the early phase of the COVID-19 pandemic. As the primary measure of viral transmissibility, the research utilizes the basic reproduction number (R_0), estimated from the initial phase of the epidemic's exponential growth, thereby mitigating biases caused by differences in testing and reporting policies across regions.

The research is structured through two complementary empirical studies. The first study provides a global-scale analysis, encompassing 118 countries and integrating a multi-domain dataset. Seasonality emerges as a transmission modulator, with clear hemisphere-dependent sinusoidal patterns – higher R_0 during the winter and lower during the summer months – which are formally quantified using Bayesian MCMC analysis. For representative Northern Hemisphere countries (Serbia, Lithuania), the estimated mean R_0 values are 3.36–3.37, while for Argentina, as a Southern Hemisphere country, it is 3.70, exhibiting inverted seasonal dynamics.

The second study analyzes data for all states of the United States of America and, by applying regularized regression methods (Lasso, Elastic Net) and nonlinear ensemble algorithms (Random Forest, Gradient Boosting), identifies fine particulate matter ($PM_{2.5}$) as the single most significant predictor of R_0 . It is estimated that variations in $PM_{2.5}$ concentrations, within the range typically recorded in the USA, can alter viral transmissibility by 25–40%, establishing a quantitative link between air quality and public health resilience to respiratory pathogens.

In addition to epidemiological findings, the dissertation highlights the importance of model interpretability, utilizing predictor importance analysis across different algorithms. An informed approach is crucial for precisely targeting public health interventions and strategic planning under conditions of highly complex epidemiological data.

Keywords: COVID-19, SARS-CoV-2, basic reproduction number, $PM_{2.5}$, air pollution, seasonality, machine learning, epidemiological modeling, public health

Scientific field: Medical sciences

Scientific subfield: Statistical modeling

Sadržaj

ZAHVALNICE.....	4
RAČUNARSKO MODELOVANJE I MOGUĆNOSTI PREDIKCIJE REPRODUKCIONOG BROJA COVID-19 POMOĆU SREDINSKIH I SOCIODEMOGRAFSKIH PARAMETARA.....	5
COMPUTATIONAL MODELING AND PREDICTION CAPABILITIES OF THE COVID-19 REPRODUCTION NUMBER USING ENVIRONMENTAL AND SOCIODEMOGRAPHIC PARAMETERS.....	6
SADRŽAJ.....	7
LISTA SLIKA.....	11
LISTA TABELA.....	13
LISTA SKRAĆENICA.....	14
1 UVOD.....	18
1.1 PANDEMIJA COVID-19: GLOBALNI NAUČNI IZAZOV	18
1.2 ULOGA FAKTORA ŽIVOTNE SREDINE I SOCIODEMOGRAFSKIH FAKTORA U VIRUSNOJ TRANSMISIJI	18
1.3 METODOLOŠKE PREPREKE U EPIDEMIOLOŠKOM MODELOVANJU.....	19
1.4 POJAVA MAŠINSKOG UČENJA U EPIDEMIOLOGIJI.....	19
1.5 CILJEVI I ZADACI ISTRAŽIVANJA.....	20
1.6 STRUKTURA TEZE	20
2 OSNOVE I PREGLED LITERATURE.....	21
2.1 MEDICINSKE OSNOVE COVID-19.....	21
2.1.1 VIRUSOLOGIJA SARS-CoV-2	21
2.1.2 PATOFIZIOLOGIJA COVID-19.....	21
2.1.3 NAČINI TRANSMISIJE	22
2.1.4 EPIDEMIOLOGIJA I TERET BOLESTI	22
2.1.5 VREMENSKI TOK GLOBALNE PANDEMIJE.....	22
2.2 MATEMATIČKO MODELOVANJE ZARAZNIH BOLESTI.....	23
2.2.1 KOMPARTMENTALNI MODELI (SEIR)	23
2.2.2 OSNOVNI REPRODUKCIONI BROJ (R_0): PROCENA I ZNAČAJ.....	23

2.2.3	LINEARNO MODELOVANJE I PREDIKCIJA REPRODUKCIJNOG BROJA.....	23
2.2.4	ANSAMBL METODE I NELINEARNI MODELI U EPIDEMIOLOŠKIM ISTRAŽIVANJIMA.....	26
2.3	ULOGA FAKTORA ŽIVOTNE SREDINE U TRANSMISIJI RESPIRATORNIH VIRUSA	29
2.3.1	ULOGA ZAGAĐENJA VAZDUHA ($PM_{2.5}$, PM_{10} , NO_2)	29
2.3.2	METEOROLOŠKI FAKTORI I SEZONALNOST	30
2.4	SOCIODEMOGRAFSKI I EKONOMSKI FAKTORI U PANDEMIJAMA.....	30

3 OPŠTA METODOLOGIJA..... 32

3.1	PROCENA OSNOVNOG REPRODUKCIJNOG BROJA (R_0).....	33
3.1.1	IZBOR VREMENSKOG INTERVALA ZA PROCENU R_0	34
3.1.2	PROCENA EKSPONENCIJALNE STOPE RASTA EPIDEMIJE	35
3.2	PRIKUPLJANJE PODATAKA, PRETPROCESIRANJE I INŽENJERING OBELEŽJA.....	37
3.2.1	IZVORI PODATAKA.....	37
3.2.2	PRETPROCESIRANJE PODATAKA	38
3.3	MODELOVANJE PRIMENOM METODA MAŠINSKOG UČENJA.....	40
3.3.1	IZBOR ALGORITAMA MAŠINSKOG UČENJA.....	41
3.3.2	VALIDACIJA MODELA I OPTIMIZACIJA HIPERPARAMETARA.....	42
3.3.3	ANALIZA ZNAČAJA PREDIKTORA I INTERPRETACIJA MODELA.....	44

4 GLOBALNA ANALIZA POKRETAČA TRANSMISIJE SARS-COV-2: PREDIKTIVNO MODELOVANJE I KVANTIFIKACIJA SEZONSKOG FORSIRANJA U 118 ZEMALJA..... 46

4.1	UVOD	46
4.2	METODOLOGIJA.....	46
4.2.1	PROCENA OSNOVNOG REPRODUKCIJNOG BROJA R_0	46
4.2.2	PRIKUPLJANJE I HARMONIZACIJA PODATAKA.....	47
4.2.3	PRETPROCESIRANJE PODATAKA	52
4.2.4	PREDIKTIVNO MODELOVANJE	53
4.2.5	BAJESIJSKA KARAKTERIZACIJA SEZONSKE DINAMIKE	54
4.3	REZULTATI	55
4.3.1	KARAKTERISTIKE FINALNOG SKUPA PODATAKA I INICIJALNA EKSPLOKACIJA.....	55
4.3.2	PRELIMINARNA GLOBALNA KORELACIONA ANALIZA I SELEKCIJA OBELEŽJA	57
4.3.3	PERFORMANSE MODELA MAŠINSKOG UČENJA	61
4.3.4	GODIŠNJI OBRASCI PREDVIĐENOG R_0 I HEMISFERSKA ZAVISNOST	63
4.3.5	BAJESIJSKA KARAKTERIZACIJA SEZONSKOG FORSIRANJA.....	66
4.4	DISKUSIJA	71
4.4.1	SEZONALNOST KAO DOMINANTNI GLOBALNI MODULATOR TRANSMISIJE	71
4.4.2	IZAZOVI GLOBALNOG INŽENJERINGA PODATAKA	72

4.4.3	OGRANIČENJA STUDIJE	73
4.5	ZAKLJUČAK	75

5 UTICAJ ZAGAĐENJA VAZDUHA NA PRENOSIVOST COVID-19 U SJEDINJENIM AMERIČKIM DRŽAVAMA..... 77

5.1	UVOD	77
5.2	METODOLOGIJA.....	78
5.2.1	PROCENA OSNOVNOG REPRODUKCIONOG BROJA.....	78
5.2.2	PODACI O ZAGAĐENJU VAZDUHA	79
5.2.3	METEOROLOŠKI PODACI	80
5.2.4	SOCIODEMOGRAFSKI I ZDRAVSTVENI PODACI.....	80
5.2.5	OBRADA PODATAKA I TRANSFORMACIJE.....	80
5.2.6	REDUKCIJA DIMENZIONALNOSTI PUTEM PCA.....	81
5.2.7	LASSO REGRESIJA.....	82
5.2.8	ELASTIC NET REGRESIJA.....	82
5.2.9	RANDOM FOREST I GRADIENT BOOST.....	82
5.3	REZULTATI	83
5.3.1	INICIJALNA KORELACIONA ANALIZA: PROMINENCIJA $PM_{2.5}$	83
5.3.2	SELEKCIJA OBELEŽJA PUTEM REGULARIZOVANE REGRESIJE.....	85
5.3.3	ZNAČAJ VARIJABLI U NELINEARNIM ANSAMBL MODELIMA	87
5.3.4	KVANTIFIKACIJA UTICAJA ZAGAĐENJA VAZDUHA NA R_0	89
5.4	DISKUSIJA I ZAKLJUČNA RAZMATRANJA.....	91
5.4.1	EKOLOŠKO-EPIDEMIOLOŠKI OKVIR TRANSMISIJE COVID-19	91
5.4.2	METODOLOŠKI DOPRINOS I ROBUSTNOST PRISTUPA	93
5.4.3	OGRANIČENJA STUDIJE I MERNI GREŠKA	93
5.4.4	PRAVCI BUDUĆIH ISTRAŽIVANJA.....	94
5.4.5	ZAKLJUČAK: ŠIRE IMPLIKACIJE ZA JAVNO ZDRAVLJE.....	94

6 OPŠTA DISKUSIJA..... 95

6.1	SINTEZA KLJUČNIH NALAZA.....	95
6.1.1	SVEOBUHVAJNI UTICAJ SEZONALNOSTI NA GLOBALNOJ SKALI	95
6.1.2	ZAGAĐENJE VAZDUHA KAO KONZISTENTAN POKRETAČ TRANSMISIJE.....	95
6.1.3	POREĐENJE LOKALNIH (SAD) I GLOBALNIH POKRETAČA PANDEMIJE.....	96
6.2	NALAZI U KONTEKSTU EVOLUIRAJUĆEG PANDEMIJSKOG PEJZAŽA	96
6.3	METODOLOŠKI DOPRINOSI I INOVACIJE.....	96
6.4	POZICIONIRANJE NALAZA ISTRAŽIVANJA U ODNOSU NA NAUČNU LITERATURU	97
6.5	IMPLIKACIJE ZA POLITIKU JAVNOG ZDRAVLJA I BUDUĆU PRIPRAVNOST ZA PANDEMIJE.....	97

6.6	OGRANIČENJA ISTRAŽIVANJA.....	98
6.7	PRAVCI BUDUĆIH ISTRAŽIVANJA	98
<u>7</u>	<u>ZAKLJUČAK</u>	<u>99</u>
7.1	SAŽETAK DOPRINOSA.....	99
7.2	ZAVRŠNE NAPOMENE.....	99
<u>8</u>	<u>BIBLIOGRAFIJA.....</u>	<u>101</u>
<u>9</u>	<u>APENDIKS.....</u>	<u>112</u>

Lista slika

ILUSTRACIJA 1: GRAFIKON REZIDUALA NASPRAM PREDVIĐANJA REGRESIONOG MODELA – UOČENE EKSTREMNE VREDNOSTI (BANGLADEŠ, BRAZIL, JAPAN). IZLOMLJENA CRVENA LINIJA PREDSTAVLJA LOESS PROCENU TRENDJA KOJA NIJE INDIKOVALA FENOMENE KOJI ZAHTEVAJU DALJE ISPITIVANJE.	56
ILUSTRACIJA 2: GRAFIKON REZIDUALA I UTICAJA – IRAN KAO ZNAČAJNA EKSTREMNA VREDNOST U POZITIVNOM SMERU. KUKOVA DISTANCA OD 0.5 I 1 KAO UOBIČAJENE GRANICE UTICAJNIH OPSERVACIJA DATA JE ISPREKIDANOM CRVENOM LINIJOM. LOESS TREND KAO NA PROŠLOJ SLICI.	56
ILUSTRACIJA 3: A) DIJAGRAM RASTURANJA KAŠNJENJA POČETKA EPIDEMIJE I R_0 POKAZUJE ZNAČAJNU UMERENU KORELACIJU B) PIRSONOVA KORELACIJA R_0 SA POLUTANTIMA GDE NO_2 I SO_2 POKAZUJU STATISTIČKI ZNAČAJNU POZITIVNU KORELACIJU SA R_0 C) DIJAGRAM RASTURANJA R_0 I PADAVINA KAO ZNAČAJNO KORELISANOG PREDIKTORA D) PIRSONOVA KORELACIJA R_0 SA OSTALIM METEOROLOŠKIM PARAMETRIMA (SA LEVA NA DESNO): T—TEMPERATURA, H—SPECIFIČNA VLAŽNOST, UV—INDEKS ULTRALJUBIČASTOG ZRAČENJA, P—VAZDUŠNI PRITISAK, WS—BRZINA VETRA. SEZONALNO VARIRAJUĆI METEOROLOŠKI PARAMETRI SU BILI STATISTIČKI ZNAČAJNI (T, H, UV).	58
ILUSTRACIJA 4: ANALIZA GLAVNIH KOMONENTI ZA METEOROLOŠKE PODATKE. (A) KUMULATIVNA OBJAŠNJENA VARIJANSA. PRVE TRI KOMONENTE IMAJU KUMULATIVNU VARIJANSU OD 95% (B) VARIJABLE KOJE SU U NAJBOLJOJ KORELACIJI SA METEOROLOŠKIM GLAVNIM KOMONENTAMA (PC). PRVA GLAVNA METEOROLOŠKA KOMONENTA JE SKORO SAVRŠENO KORELISANA SA TEMPERATUROM. (C) DIJAGRAM RASTURANJA METEOROLOŠKE PC1 NASPRAM TEMPERATURE U STANDARDIZOVANIM VREDNOSTIMA. (D) KORELACIJA METEOROLOŠKIH VARIJABLI SA TEMPERATUROM. TEMPERATURA, VLAŽNOST VAZDUHA, PADAVINE I UV ZRAČENJE SU VISOKO KORELISANI I PREDSTAVLJAJU SEZONALNOST. T—TEMPERATURA, WS—BRZINA VETRA, P—VAZDUŠNI PRITISAK, H—SPECIFIČNA VLAŽNOST, UV—INDEKS ULTRALJUBIČASTOG ZRAČENJA, PR—PADAVINE.	60
ILUSTRACIJA 5: TEST-MSE ZA 27 KOMBINACIJA ALGORITAM-SKUP PREDIKTORA SA INTERVALIMA POVERENJA IZ 100 PONOVLJENIH PODELA. ISPREKIDANA LINIJA OZNAČAVA PRAG PRIHVATLJIVOSTI; 17 KOMBINACIJA JE ISPOD PRAGA, A 12 NJIH ČIJI INTERVALI POVERENJA NE SEKU 1,0 ČINE FINALNI SKUP	62
ILUSTRACIJA 6: PRIMER SEZONSKE VARIJACIJE PREDVIĐENOG R_0 ZA NEMAČKU – POREĐENJE ALGORITAMA PO SKUPOVIMA PREDIKTORA (LASSO, RF, UNION)	64
ILUSTRACIJA 7: DNEVNE PREDIKCIJE R_0 TOKOM 2021. GODINE ZA 20 ZEMALJA RASPOREĐENIH PO GEOGRAFSKOJ ŠIRINI – SEVERNA, EKVATORIJALNA I JUŽNA HEMISFERA (RIDGE REGRESIJA, UNION SKUP PREDIKTORA)	65
ILUSTRACIJA 8: BAJESIJANSKI KORNER PLOT ZA SRBIJU – POSTERIORNA RASPODELA PARAMETARA SEZONSKOG MODELA.....	67
ILUSTRACIJA 9: BAJESIJANSKI KORNER PLOT ZA LITVANIJU – POSTERIORNA RASPODELA SREDNJEG R_0 , AMPLITUDE SEZONSKOG FORSIRANJA (ASF) I VREMENA MINIMALNOG R_0 (TMIN)	68
ILUSTRACIJA 10: BAJESIJANSKI KORNER PLOT ZA ARGENTINU – OBRNUTA SEZONSKA DINAMIKA U ODNOSU NA SEVERNU HEMISFERU, TMIN POMEREN ZA ~182 DANA.....	69
ILUSTRACIJA 11: ANALIZA SEZONSKIH PARAMETARA R_0 NA NIVOU 118 ZEMALJA. (A) AMPLITUDA SEZONSKOG FORSIRANJA ASF U FUNKCIJI APSOLUTNE GEOGRAFSKE ŠIRINE, SA IZRAZITOM POZITIVNOM LINEARNOM ZAVISNOŠĆU ($r = 0.85$). (B) RASPODELA ASF MEĐU 77 ZEMALJA UMERENOG POJASA ($ GEOGRAFSKA ŠIRINA > 23.5^\circ$) SA FITOVANOM GAUSOVOM RASPODELOM (SREDNJA VREDNOST = 0.064, SD = 0.015). (C) L1 RASTOJANJE U DANIMA IZMEĐU PIKA R_0 I MINIMUMA UV ZRAČENJA (OPENUV) ZA ZEMLJE IZNAD 10° GEOGRAFSKE ŠIRINE, FITOVANO GAMA RASPODELOM. (D) L1 RASTOJANJE U DANIMA IZMEĐU PIKA R_0 I MINIMUMA TEMPERATURE (T2M_MIN), FITOVANO GAMA RASPODELOM. ŠIRA RASPODELA U POREĐENJU SA PANELOM (C) SUGERIŠE DA JE UV ZRAČENJE PROKSIMALNIJI POKRETAČ SEZONSKE TRANSMISIJE OD TEMPERATURE. (E) MEDIJANA GODIŠNJEG R_0 U FUNKCIJI APSOLUTNE GEOGRAFSKE ŠIRINE, SA POZITIVNIM LINEARNIM TRENDOM ($r = 0.59$).	70
ILUSTRACIJA 12: EKSPONENCIJALNI REŽIM RASTA I EKSTRAKCIJA R_0 . PRIKAZANE SU LOG-TRANSFORMISANE KRIVE KUMULATIVNIH DETEKTOVANIH SLUČAJEVA ZA REPREZENTATIVNE SAVEZNE DRŽAVE SAD U IDENTIFIKOVANOM VREMENSKOM PROZORU	

EKSPONENCIJALNE FAZE. LINEARNOST KRIVIH POTVRĐUJE VALIDNOST EKSPONENCIJALNOG MODELA, A NAGIB SVAKE KRIVE DIREKTNO ODREĐUJE PROCENJENI R_0	79
ILUSTRACIJA 13: KORELACIONA ANALIZA IZMEĐU KONCENTRACIJE $PM_{2.5}$ I OSNOVNOG REPRODUKCIJNOG BROJA R_0 NA NIVOU SAVEZNIH DRŽAVA SAD. POZITIVAN NAGIB REGRESIJE LINIJE KONZISTENTAN JE KROZ CELOKUPAN DISTRIBUCIONI RASPON UZORKA.	84
ILUSTRACIJA 14: VAŽNOST PREDIKTORA U ANSAMBL MODELIMA MAŠINSKOG UČENJA. PRIKAZANO JE RANGIRANJE PREDIKTORA PREMA DOPRINOSU SMANJENJU PREDIKCIJE GREŠKE U RANDOM FOREST I GRADIENT BOOSTING ALGORITMIMA. $PM_{2.5}$ ZAUZIMA DOMINANTNU POZICIJU U OBA MODELA, SA JASNOJ SEPARACIJOM OD SEKUNDARNIH PREDIKTORA KOJI UKLJUČUJU KOMPONENTE HRONIČNIH BOLESTI, OKSIDE AZOTA I DEMOGRAFSKE FAKTORE.	86
ILUSTRACIJA 15: DISTRIBUCIJA PROCENJENE RELATIVNE PROMENE OSNOVNOG REPRODUKCIJNOG BROJA ($\Delta R_0/R_0$) PO SAVEZNIM DRŽAVAMA SAD, DOBIJENA KONTRAFAKTUALNOM SIMULACIJOM U RANDOM FOREST I GRADIENT BOOSTING MODELIMA. REALISTIČNE RAZLIKE U $PM_{2.5}$ KONCENTRACIJAMA PROIZVODE PROMENU R_0 U RASPONU OD 25 DO 40%, KONZISTENTNO IZMEĐU OBA MODELA.	90

Lista tabela

TABELA 1: PREGLED IMENA I ZNAČENJA TEMPERATURNIH PROMENLJIVIH U KONAČNOM SKUPU	48
TABELA 2: GRUPISANJE VARIJABLI I ODGOVARAJUĆE PCA KOMPONENTE	81
TABELA 3: PREGLED SVIH VARIJABLI KONAČNOG SKUPA PODATAKA PODELJENIH U PET GRUPA. VARIJABLE VEZANE ZA IZRAČUNAVANJE OSNOVNOG REPRODUKCIONOG BROJA NISU KORIŠĆENE ZA PREDIKCIJU, KAO NI DUPLIKATI METEOROLOŠKIH PODATAKA IZ WAQI BAZE, PA JE UKUPAN BROJ RAZMATRANIH PREDIKTORA 96.....	112
TABELA 4: ODABRANI PREDIKTORI ZA SVAKI OD TESTIRANIH SETOVA ZA 118 DRŽAVA.....	116
TABELA 5: PREGLED VARIJABLI — ANALIZA SAVEZNIH DRŽAVA SAD (70 VARIJABLI U 4 KATEGORIJE + 4 IZVEDENE). VREDNOSTI SU POPULACIONO PONDERISANE I VREMENSKI POMERENE ZA 5 DANA UNAZAD U ODNOSU NA EPIDEMIOLOŠKE PODATKE. U ANALIZI SU KORIŠĆENE I 4 IZVEDENE VARIJABLE: dPol2017 (RAZLIKA PM _{2.5} IZMEĐU 2017. I PROSEKA 2013–2017), KAO I PC KOMPONENTE TEMPERATURE, VLAŽNOSTI I RADIJACIJE DOBIJENE PCA TRANSFORMACIJOM (TABELA 2).	116

Lista skraćenica

ACE2 – Angiotenzin-konvertujući enzim 2 (Angiotensin-Converting Enzyme 2)

AI – Veštačka inteligencija (Artificial Intelligence)

AIC – Akaike informacioni kriterijum (Akaike information criterion)

AQI – Indeks kvaliteta vazduha (Air Quality Index)

ARDS – Akutni sindrom respiratorne insuficijencije (Acute Respiratory Distress Syndrome)

BCG – Bacillus Calmette–Guérin, vakcina protiv tuberkuloze

BMI – Indeks telesne mase (Body Mass Index)

BOD – Teret bolesti (Burden of Disease)

BP – Krvni pritisak (Blood Pressure)

CFR – Stopa smrtnosti slučajeva (Case Fatality Rate)

CHE – Tekući izdaci za zdravstvo (Current Health Expenditure)

CHOL – Holesterol (Cholesterol)

COVID-19 – Korona virusna bolest 2019 (Coronavirus Disease 2019)

EDGAR – Baza podataka o emisijama za globalna atmosferska istraživanja (Emissions Database for Global Atmospheric Research)

EHR – Elektronski zdravstveni karton (Electronic Health Record)

GDP – Bruto domaći proizvod (Gross Domestic Product)

GHED – Globalna baza podataka o zdravstvenim izdacima (Global Health Expenditure Database)

GHO – Globalna zdravstvena opservatorija (Global Health Observatory)

GLUC – Glukoza (Glucose)

HMC – Hamiltonijanski Monte Karlo (Hamiltonian Monte Carlo)

HWF – Zdravstvena radna snaga (Health Workforce)

IFR – Stopa smrtnosti infekcija (Infection Fatality Rate)

IRLS – Iterativno reponderisani najmanji kvadrati (Iteratively Reweighted Least Squares)

MCMC – Markovljev lanac Monte Karlo (Markov Chain Monte Carlo)

MDG – Milenijumski razvojni ciljevi (Millennium Development Goals)

MERS-CoV – Koronavirus sindroma respiratorne insuficijencije Bliskog istoka (Middle East Respiratory Syndrome Coronavirus)

MH – Mentalno zdravlje (Mental Health)

ML – Mašinsko učenje (Machine Learning)

NASA POWER – NASA baza meteoroloških i solarnih podataka (Prediction of Worldwide Energy Resources)

NCD – Nezarazne bolesti (Non-Communicable Diseases)

NO₂ – Azot-dioksid (Nitrogen Dioxide)

OD – Otvorena defekacija (Open Defecation)

OLS – Obična regresija najmanjih kvadrata (Ordinary Least Squares)

PAA – Fizička aktivnost, starosno standardizovana (Physical Activity, Age-standardized)

PAC – Fizička aktivnost, uporediva (Physical Activity, Comparable)

PASC – Post-akutne posledice COVID-19 / „Dugi COVID“ (Post-Acute Sequelae of COVID-19)

PCA – Analiza glavnih komponenti (Principal Component Analysis)

PCR – Regresija glavnih komponenti (Principal Component Regression)

PM₁₀ – Čestične materije aerodinamičnog prečnika manjeg od 10 µm

PM_{2.5} – Fine čestične materije aerodinamičnog prečnika manjeg od 2,5 µm

R₀ – Osnovni reprodukциони broj (Basic Reproduction Number)

RBD – Domen za vezivanje receptora (Receptor Binding Domain)

RMSEP – Kvadratni koren srednje kvadratne greške predikcije (Root Mean Square Error of Prediction)

RNK – ribonukleinska kiselina

SA – Zloupotreba supstanci (Substance Abuse)

SAIC – Postepeni (stepwise) AIC

SARS-CoV – Koronavirus teškog akutnog respiratornog sindroma (Severe Acute Respiratory Syndrome Coronavirus)

SARS-CoV-2 – Koronavirus teškog akutnog respiratornog sindroma 2 (Severe Acute Respiratory Syndrome Coronavirus 2)

SDG – Ciljevi održivog razvoja (Sustainable Development Goals)

SEIR – Kompartmentalni epidemiološki model (Susceptible-Exposed-Infected-Recovered)

SHA – Sistem zdravstvenih računa (System of Health Accounts)

SIR – Kompartmentalni epidemiološki model (Susceptible-Infected-Recovered)

SO₂ – Sumpor-dioksid (Sulfur Dioxide)

SVM – Mašina potpornih vektora (Support Vector Machine)

WAQI – Globalni indeks kvaliteta vazduha (World Air Quality Index)

WASH/WSH – Vodosnabdevanje, sanitacija i higijena (Water, Sanitation and Hygiene)

WHO – Svetska zdravstvena organizacija (World Health Organization)

WHOSIS – Statistički informacioni sistem WHO (WHO Statistical Information System)

WHS – Svetska zdravstvena statistika (World Health Statistics)

XGBoost – Ekstremno pojačavanje gradijentom (Extreme Gradient Boosting)

1 Uvod

1.1 Pandemija COVID-19: Globalni naučni izazov

Pojava novog koronavirusa SARS-CoV-2 krajem 2019. godine, i njegovo naknadno proglašenje za globalnu pandemiju od strane Svetske zdravstvene organizacije (WHO) 11. marta 2020. godine, označilo je početak krize javnog zdravlja bez presedana (World Health Organization, 2020). Pandemija COVID-19 je od tada imala dubok uticaj na društva širom sveta, dovodeći do ogromnih gubitaka života, dovodeći zdravstvene sisteme do tačke pucanja, i izazivajući široku ekonomsku i društvenu disrupciju.

Suočena sa ovim izazovom, globalna naučna zajednica se mobilisala vanrednom brzinom. Međutim, trajektorija pandemije je bila okarakterisana značajnom heterogenošću. Različiti regioni, pa čak i različite populacije unutar istog regiona, doživeli su veoma različite ishode u pogledu stopa infekcije, težine bolesti i mortaliteta. Ova disproporcija je bila evidentna već na početku, pošto su neke zemlje izveštavale o uspehu sa umerenim merama javnog zdravlja, dok su druge nametnule stroga zatvaranja i ipak su pretrpele teške epidemije. Razumevanje faktora koji pokreću ove varijacije postalo je centralno i hitno pitanje za istraživače različitih disciplina.

1.2 Uloga faktora životne sredine i sociodemografskih faktora u virusnoj transmisiji

Dok su intervencije javnog zdravlja, poput fizičkog distanciranja, obaveznog nošenja maski i vakcinacionih kampanja, poznate kao kritične u kontroli širenja virusa, jasno je da je inherentna prenosivost virusa poput SARS-CoV-2 takođe modulirana kompleksnom međuigrom eksternih faktora. Uočene varijacije u napredovanju pandemije snažno sugerišu da lokalna sredina i karakteristike populacije igraju značajnu ulogu u pogoršavanju ili ublažavanju njegovog širenja.

Rano u pandemiji, počeli su da se pojavljuju dokazi koji ukazuju na potencijalnu vezu između **faktora životne sredine** i transmisije COVID-19. Ovo uključuje meteorološke uslove poput temperature i vlažnosti, za koje je poznato da utiču na preživljavanje i transmisiju drugih respiratornih virusa (Sarkodie & Owusu, 2020). Od posebnog interesa bila je uloga zagađenja vazduha u okolini. Nekoliko studija je predložilo mehanizme pomoću kojih čestične materije, posebno fine čestične materije (PM_{2.5}), mogu olakšati virusnu transmisiju delujući kao nosilac kapi sa virusom ili povećavajući podložnost populacije kroz ugroženo respiratorno i imuno zdravlje (Comunian et al., 2020; Contini & Costabile, 2020).

Uz faktore životne sredine, širok spektar **sociodemografskih i ekonomskih faktora** se smatraju faktorima koji utiču na ranjivost populacije. Tu uključujemo gustinu stanovništva, starosnu strukturu, prevalenciju komorbiditeta (kao što su kardiovaskularne bolesti i

dijabetes), zdravstvenu infrastrukturu, ekonomski prosperitet i disparitete u javnom zdravlju (Bontempi et al., 2020). Ovi faktori kolektivno oblikuju pejzaž na kome se epidemija odvija, utičući na sve, od individualnog rizika od izloženosti do kapaciteta sistema javnog zdravlja da efikasno reaguje.

1.3 Metodološke prepreke u epidemiološkom modelovanju

Istraživanje uticaja ove ogromne i međusobno povezane mreže faktora predstavlja značajne metodološke izazove. Sprovođenje robusnog kauzalnog rasuđivanja na kompleksnom globalnom sistemu gde postoji bezbroj korelisanih parametara je ogroman zadatak. Nekoliko ključnih prepreka komplikuje analizu:

1. Kontrola dinamike infekcije: Jednostavno poređenje sirovih brojeva slučajeva ili stopa mortaliteta između regiona je problematičan pristup. Ovi brojevi su pod jakim uticajem lokalnih politika testiranja, standarda izveštavanja i specifične faze epidemijske krive. Potrebna je robusnija metrika koja zahvata inherentnu prenosivost virusa, nezavisno od ovih pristrasnosti. **Osnovni reprodukциони broj (R_0)**, koji meri prenosivost u potpuno podložnoj populaciji, nudi stabilniju i uporediviju meru.

2. Konfundirajuće varijable i multikolinearnost: Potencijalni pokretači transmisije su brojni i često visoko korelisani. Na primer, visoki nivoi zagađenja vazduha često koegzistiraju sa visokom gustinom stanovništva i socioekonomskom deprivacijom. Ova multikolinearnost otežava tradicionalnim statističkim modelima da razdvoje nezavisni efekat bilo koje pojedinačne varijable.

3. Nelinearne relacije: Odnos između datog faktora i virusne transmisije nije uvek linearan. Na primer, efekat temperature može biti optimalan za transmisiju samo unutar određenog opsega. Tradicionalni linearni regresioni modeli nisu sposobni da uhvate takve složene, nelinearne zavisnosti.

4. Visokodimenzionalni podaci: Veliki broj potencijalno relevantnih varijabli—koje obuhvataju faktore životne sredine, demografske, medicinske i ekonomske domene—često prevazilazi broj opservacionih jedinica (npr. zemlje ili države). Ova "kletva dimenzionalnosti" predstavlja veliki rizik od preprilagođavanja modela i lažnih korelacija.

1.4 Pojava mašinskog učenja u epidemiologiji

Metodološke prepreke navedene gore zahtevaju upotrebu naprednijih analitičkih tehnika sposobnih da se nose sa visokodimenzionalnim, složenim i korelisanim podacima. Mašinsko učenje (ML) nudi moćan skup alata za rešavanje ovih izazova. ML modeli, kao što su regularizovane regresije (npr. Lasso) i ansambl metode (npr. Random Forest, Gradient Boosting), dizajnirani su da rukuju velikim brojem prediktorskih varijabli, izvode automatsku selekciju obeležja kako bi identifikovali najvažnije pokretače, i da uoče složene nelinearne relacije bez jakih prethodnih pretpostavki.

1.5 Ciljevi i zadaci istraživanja

Ova teza ima za cilj da rasplete složenu mrežu faktora koji utiču na rano širenje pandemije COVID-19 koristeći napredne tehnike mašinskog učenja. Primarni ciljevi ovog istraživanja su:

1. Razviti robusnu i reproduktivnu metodologiju za procenu osnovnog broja reprodukcije (R_0) SARS-CoV-2 i sastaviti sveobuhvatan, multi-domenski skup podataka za prediktivno modelovanje na nacionalnoj (Sjedinjene Američke Države) i globalnoj skali.
2. Sistematski istražiti i identifikovati ključne faktore životne sredine, sociodemografske i medicinske faktore koji pokreću prenosivost SARS-CoV-2.
3. Specifično kvantifikovati uticaj zagađenja vazduha u okolini, posebno $PM_{2.5}$, i meteorološke sezonalnosti na širenje COVID-19.
4. Izgraditi, validirati i upoređivati skup modela mašinskog učenja po njihovoj sposobnosti da predvide virusnu prenosivost.

1.6 Struktura teze

Ova teza je strukturirana u sedam poglavlja. **Poglavlje 2** pruža pregled relevantne literature, pokrivajući epidemiologiju SARS-CoV-2, matematičke i računske tehnike modelovanja, i ustanovljenu ulogu faktora životne sredine i sociodemografskih faktora u transmisiji respiratornih bolesti. **Poglavlje 3** detaljno opisuje opštu metodologiju koja se koristi kroz tezu, uključujući pristup za procenu R_0 , prikupljanje i inženjering podataka, i tok mašinskog učenja.

Centralni empirijski rad je predstavljen kroz dve studije. **Poglavlje 4** daje analizu na globalnoj skali koristeći podatke iz 118 zemalja da bi identifikovao univerzalne pokretače transmisije i kvantifikovao uticaj sezonalnosti. **Poglavlje 5** predstavlja fokusiranu analizu Sjedinjenih Američkih Država, istražujući ulogu zagađenja vazduha kao primarnog prediktora prenosivosti COVID-19 između država.

Poglavlje 6 pruža opštu diskusiju koja sintetiše nalaze iz obe studije, razmatra metodološke doprinose i razmatra šire implikacije za politiku javnog zdravlja i buduću pripravnost za pandemije. Konačno, **Poglavlje 7** zaključuje tezu rezimeom njenih ključnih doprinosa i završnim napomenama.

2 Osnove i pregled literature

2.1 Medicinske osnove COVID-19

U ovom poglavlju je dat kratak pregled medicinskih aspekata virusa i posledične pandemije. Navedeni aspekti su neophodni za puno razumevanje uočenih zavisnosti i daju osnovu za diskusiju i ponuđena objašnjenja istih.

2.1.1 Virusologija SARS-CoV-2

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) je novi, omotan, pozitivno-usmereni, jednolančani RNK virus koji pripada rodu *Betacoronavirus* porodice *Coronaviridae* (N. Zhu et al., 2020). Genomski, deli značajnu homologiju sa virusom odgovornim za epidemiju teškog akutnog respiratornog sindroma iz 2003. godine (*Severe Acute Respiratory Syndrome* - SARS-CoV) i koronavirusom sindroma respiratorne insuficijencije Srednjeg istoka (*Middle East Respiratory Syndrome* - MERS-CoV). Viralna čestica je okarakterisana velikim glikoproteinskim "šiljcima" koji štrče sa njene površine, što virusu daje izgled nalik na krunu (latinski "corona") pod elektronskim mikroskopom. Ovaj spajk (od engleskog *spike* – *šiljak*, *bodlja*) protein je od kritične važnosti, jer posreduje u ulasku virusa u ćelije domaćina. Specifično, domen za vezivanje receptora (*receptor binding domain* - RBD) spajk proteina se vezuje sa visokim afinitetom za receptor angiotenzin-konvertujućeg enzima 2 (ACE2) na površini ljudskih ćelija (Hoffmann et al., 2020). Visoka stopa mutacija njegovog RNK genoma dovela je do pojave brojnih varijanti tokom pandemije, od kojih su neke pokazale povećanu prenosivost, izbegavanje imuniteta ili virulenciju.

2.1.2 Patofiziologija COVID-19

Kada SARS-CoV-2 uđe u telo, tipično preko respiratornog trakta, koristi svoj spajk protein da se veže za ACE2 receptore, koji su prisutni u izobilju na epitelnim ćelijama u plućima, ali su takođe prisutni u srcu, bubrezima i crevima. Nakon vezivanja za receptor, virus se spaja sa membranom ćelije domaćina i oslobađa svoj RNK genom, preuzimajući biohemijsku mašineriju ćelije da se replicira i proizvodi nove viralne čestice.

Rezultujuća bolest, COVID-19, se manifestuje širokim kliničkim spektrom. Kod mnogih pojedinaca, infekcija je ograničena na gornji respiratorni trakt, uzrokujući blage do umerene simptome kao što su groznica, kašalj i umor. Međutim, u težim slučajevima, virus može napredovati do donjeg respiratornog trakta, dovodeći do virusne pneumonije. Kritična karakteristika teške forme COVID-19 je preteran zapaljenjski odgovor, često nazivan "citokinskom olujom", koji može dovesti do akutnog sindroma respiratorne insuficijencije (ARDS), multiorganske insuficijencije i smrti (Mehta et al., 2020). Štaviše, značajan podskup pojedinaca, čak i onih sa blagom inicijalnom bolešću, doživljava perzistentne simptome nedeljama ili mesecima, stanje poznato kao Post-Acute Sequelae of COVID-19 (PASC), ili "dugi COVID".

2.1.3 Načini transmisije

Primarni način transmisije SARS-CoV-2 je kroz inhalaciju respiratornih tečnosti koje sadrže virus (CDC, 2021). Ove tečnosti se izbacuju kada zaražena osoba diše, govori, kašlje ili kija i mogu postojati kao veće **kapljice**, koje imaju tendenciju da brzo padnu na zemlju na kratkim rastojanjima, ili kao manji **aerosoli**, koji mogu ostati suspendirani u vazduhu duže vreme i putovati dalje, posebno u slabo ventiliranim zatvorenim prostorima. Ovaj vazdušni put transmisije je ključni razlog za visoku prenosivost virusa. Ključna karakteristika koja olakšava njegovo brzo širenje je značajna uloga **asimptomatske i pre-simptomatske transmisije**, gde pojedinci mogu biti zarazni i širiti virus pre nego što postanu svesni da su bolesni (He et al., 2020). Iako transmisija može takođe da se desi kontaktom sa kontaminiranim površinama (fomitima), ovo se smatra manje značajnim putem u poređenju sa respiratornom transmisijom.

2.1.4 Epidemiologija i teret bolesti

Prvi slučajevi nove pneumonije su identifikovani u Vuhanu, Kina, u decembru 2019. godine. Uzročnik je ubrzo identifikovan kao SARS-CoV-2, a visoka prenosivost virusa dovela je do njegove brze globalne diseminacije. Teret bolesti je varirao dramatično između različitih populacija i vremenskih perioda, pod uticajem faktora kao što su demografija populacije, kapacitet zdravstvene zaštite, odgovori javnog zdravlja i cirkulacija različitih virusnih varijanti. Ključne epidemiološke metrike kao što su stopa smrtnosti slučajeva (CFR) i stopa smrtnosti infekcije (IFR) su pokazale jaku zavisnost od starosti, sa rizikom od smrti koji eksponencijalno raste sa godinama (Levin et al., 2020; O'Driscoll et al., 2021; Verity et al., 2020).

2.1.5 Vremenski tok globalne pandemije

Tok pandemije COVID-19 može biti okarakterisan kroz nekoliko ključnih faza i događaja:

- **Decembar 2019. - Januar 2020.:** Pojava klastera slučajeva pneumonije nepoznatog porekla u Vuhanu, Kina. Identifikovan je novi koronavirus.
- **11. mart 2020.:** Svetska zdravstvena organizacija (SZO) proglašava epidemiju COVID-19 globalnom pandemijom.
- **Kasna 2020.:** Prve COVID-19 vakcine (zasnovane na mRNA) dobijaju odobrenje za hitnu upotrebu, označavajući prekretnicu u pandemijskom odgovoru.
- **2021.:** Pojava i globalno širenje visoko prenosivih varijanti visokog rizika, naročito **Alfa** (B.1.1.7) i **Delta** (B.1.617.2) varijanti, dovode do velikih talasa infekcije širom sveta.
- **Kasna 2021. - Rana 2022.:** **Omikron** (B.1.1.529) varijanta se pojavljuje, okarakterisana većom prenosivošću i značajnim izbegavanjem imuniteta, izazivajući do tada najveći porast slučajeva globalno.
- **2023. - Danas:** SZO proglašava kraj globalne zdravstvene hitnosti za COVID-19 u maju 2023. godine. Virus prelazi iz pandemijske faze u endemsku, nastavljajući da cirkuliše i evoluira, ali sa imunitetom populacije iz vakcinacije i prethodnih infekcija koji značajno smanjuje stope teške bolesti i smrti.

2.2 Matematičko modelovanje zaraznih bolesti

Matematičko modelovanje je nezamenljiv alat u epidemiologiji za razumevanje dinamike zaraznih bolesti, predikciju njihove trajektorije i evaluaciju potencijalnog uticaja intervencija javnog zdravlja.

2.2.1 Kompartmentalni modeli (SEIR)

Među najšire korišćenim okvirima za modelovanje epidemija su kompartmentalni modeli. Ovi modeli dele populaciju na odvojene kompartmane koji predstavljaju različite stadijume bolesti. Model Susceptible-Infected-Recovered (SIR) (Kermack & McKendrick, 1927) je najjednostavniji oblik, ali za bolesti sa značajnim inkubacionim periodom, kao što je COVID-19, Susceptible-Exposed-Infected-Recovered (SEIR) model pruža tačniju reprezentaciju (Maier & Brockmann, 2020). U ovom okviru, pojedinci se kreću od toga da budu **Podložni** (Susceptible) do **Izloženi** (Exposed) (zaraženi ali još uvek nisu zarazni), zatim do **Zaraznih** (Infectious), i konačno do **Oporavljenih** (Recovered) (ili uklonjenih). Dinamika ovog toka je opisana sistemom diferencijalnih jednačina, koji zavise od ključnih epidemioloških parametara kao što su stope transmisije, inkubacije i oporavka.

2.2.2 Osnovni reprodukциони broj (R_0): Procena i značaj

Centralni koncept u epidemiologiji izveden iz ovih modela je **osnovni reprodukциони broj (R_0)**. Definisan je kao prosečan broj sekundarnih infekcija koje proizvodi jedan zaraženi pojedinac u potpuno podložnoj populaciji, u odsustvu bilo kakvih intervencija (Keeling & Rohani, 2008). R_0 je parametar sa kritičnim pragom: ako je $R_0 > 1$, očekuje se da će epidemija rasti, dok ako je $R_0 < 1$, očekuje se da će opadati i na kraju nestati.

Procena R_0 je primarni cilj u ranim fazama epidemije, jer kvantifikuje inherentnu prenosivost patogena. Uobičajena metoda za njegovu procenu, i ona koja se koristi u ovoj tezi, je analiza početne faze eksponencijalnog rasta epidemije. Tokom ovog perioda, broj novih slučajeva raste eksponencijalno po stopi rasta r (Jednačina (1)), koja može biti ekstraktovana iz podataka o broju slučajeva log-linearnom regresijom. Ova stopa rasta je direktno povezana sa osnovnim parametrima modela, omogućavajući izračunavanje R_0 (Salom et al., 2021). Ovaj pristup ima značajnu prednost što je manje osetljiv na razlike u politikama testiranja i pristrasnostima u izveštavanju, pružajući standardizovanu meru prenosivosti koja može biti upoređivana između različitih regiona.

2.2.3 Linearno modelovanje i predikcija reprodukcionog broja

2.2.3.1 Osnovne L2 i L1 regresije

Obična regresija metodom najmanjih kvadrata (Ordinary Least Squares — OLS) predstavlja bazični alat statističkog modelovanja (Draper & Smith, 1998) i polaznu tačku za većinu prediktivnih analiza u epidemiologiji (Tibshirani, 1996). OLS procenjuje regresione

koeficijente minimizacijom funkcije gubitka kao sume kvadrata reziduala — razlika između opaženih i predviđenih vrednosti zavisne promenljive. Ovaj princip minimizacije L2 norme obezbeđuje matematički jedinstven i analitički rešiv estimator sa poznatim optimalnim svojstvima: pod pretpostavkama Gauss-Markov teoreme — linearnost, homoskedastičnost, nezavisnost grešaka i odsustvo savršene multikolinearnosti — OLS estimator je najbolji linearni nepristrasni estimator (Best Linear Unbiased Estimator, BLUE), odnosno ima najmanju varijansu među svim nepristrasnim linearnim estimatorima (Hayashi, 2011). Ova teorijska garancija čini OLS prirodnim prvim korakom u svakom modelovanju, jer uspostavlja referentne performanse sa kojom se složenije metode mogu upoređivati.

Međutim, optimalna svojstva OLS estimatora počivaju na pretpostavkama koje su u epidemiološkoj praksi retko u potpunosti zadovoljene. Posebno problematična je osetljivost OLS procedure na ekstremne vrednosti što je direktna posledica kvadriranja reziduala. U kontekstu globalnog epidemiološkog modelovanja, anomalne opservacije sa ekstremnim vrednostima su očekivana pojava a ne izuzetak. Robusna regresija adresira ovaj problem zamenom funkcije gubitka takvom koja ograničava uticaj ekstremnih reziduala. Umesto minimizacije sume kvadrata (L2 norme), robusni estimatori koriste alternativne penalizacije koje rastu sporije za velike reziduale. Najšire korišćeni pristup je M-estimacija (Huber, 1964), koja minimizuje sumu generalizovane funkcije gubitka umesto sume kvadrata reziduala. Huberova funkcija gubitka se ponaša kao kvadratna funkcija za male reziduale zadržavajući efikasnost OLS-a u centralnom delu distribucije, ali prelazi na linearnu penalizaciju (L1) za velike reziduale čime se drastično redukuje uticaj anomalnih opservacija. Parametar k (obično postavljen na 1,345 standardnih devijacija reziduala, te i u ovom istraživanju) kontroliše prag prelaska i optimizuje kompromis između smanjenja pristrasnosti i robusnosti (Huber, 1964; Venables & Ripley, 1997).

Praktični značaj robusne regresije u transnacionalnim epidemiološkim studijama ogleda se u njenoj sposobnosti da identifikuje konzistentan signal u inherentno heterogenim podacima. Kada se modeluje osnovni reprodukcioni broj na osnovu ekoloških i sociodemografskih prediktora za 118 zemalja, izvesno je da će pojedine zemlje odstupati od globalnog obrasca - bilo usled sistemskih razlika u kapacitetu zdravstvenog nadzora, specifičnih kulturnih praksi, ili geografske izolovanosti. OLS model bi dozvolio da ove anomalije deformišu procenjene koeficijente za sve zemlje, dok robusna regresija obezbeđuje da procenjeni odnosi odražavaju obrazac prisutan u većini podataka, istovremeno signalizirajući koje opservacije ne slede taj obrazac kroz dodeljene pondere (Rousseeuw & Leroy, 1987).

2.2.3.2 Principal Component Regression (PCR): Rasplitanje kompleksnog uticaja izloženosti životnoj sredini

Epidemiološka istraživanja koja nastoje da kvantifikuju zdravstvene efekte izloženosti životnoj sredini se često suočavaju sa jedinstvenim analitičkim dilemama. Ove dileme proizlaze iz inherentne korelisane strukture ekoloških podataka. Pretpostavimo jednostavnu studiju koja ima za cilj da razdvoji uticaje različitih zagađivača vazduha—kao što su ozon (O_3), čestične materije (PM_{10} i $PM_{2.5}$), azot-dioksid (NO_2) i sumpor-dioksid (SO_2)—na zdravstvene ishode kao što su respiratorne hospitalizacije ili kardiovaskularna

smrt. Ovi zagađivači ne postoje izolovano; oni su emitovani iz sličnih izvora (saobraćaj, industrijski objekti, sagorevanje fosilnih goriva), disperzovani istim meteorološkim procesima, i, kao rezultat toga, njihove koncentracije u okolini su snažno međusobno korelisane (Billionnet et al., 2012; Roberts & Martin, 2006).

Pokušaj da se direktno uključe svi ovi visoko korelisani zagađivači kao odvojeni prediktori u tradicionalnom regresionom modelu vodi numeričkoj nestabilnosti i problemu poznatom kao multikolinearnost. Koeficijenti postaju nepouzdati, standardne greške podležu inflaciji, a model može proizvesti kontraintuitivne ili čak nemoguće rezultate (npr. sugerišući da zagađivač ima zaštitni efekat). Ovo ne odražava odsustvo kauzalnog efekta već odražava neuspeh metode da razdvoji snažno isprepletene signale.

Principal Component Regression (PCR) nudi robusno rešenje za ovu multikolinearnost. Umesto direktnog modelovanja individualnih, korelisanih zagađivača, PCR prvo transformiše skup korelisanih prediktorskih varijabli u novi skup nekorelisanih varijabli nazvanih **glavne komponente (principal components - PC)** koristeći tehniku Principal Component Analysis (PCA) (Jolliffe, 2002). Svaka PC je linearna kombinacija originalnih varijabli i dizajnirana je da uhvati maksimalnu količinu varijanse u podacima. Prvih nekoliko glavnih komponenti tipično objašnjava veliku većinu ukupne varijanse u originalnim prediktorima. Ove komponente se zatim koriste umesto originalnih zagađivača u regresionoj analizi ishoda.

U kontekstu zagađenja vazduha, prva PC može predstavljati "opšti faktor zagađenja" ili "urbanog zagađenja", koji reflektuje ukupan nivo zagađenja sa visokim opterećenjima na svim zagađivačima. Druga PC bi mogla razlikovati izvore, kao što je "saobraćajno naspram industrijskog zagađenja", sa različitim kombinacijama emisija (Laden et al., 2000; Thurston & Ito, 2001). Važna poenta je da su ove PC ortogonalne (nekorelisane), rešavajući problem multikolinearnosti i omogućavajući modelu da stabilno proceni njihov efekat na zdravlje (Roberts & Martin, 2006).

Ključna snaga PCR pristupa, demonstrirana u citiranim studijama je njegova sposobnost da obezbedi transparentan i matematički robusan način rukovanja korelisanim ekološkim izloženostima. Umesto da isključuje varijable proizvoljno ili da naivno ignoriše multikolinearnost, on eksplicitno modeluje osnovnu strukturu korelacije, izvlači nepovezane dimenzije varijacije, i pruža stabilne ocene efekata koji mogu biti tačno mapirani nazad na originalne mere izloženosti. Ovaj metod omogućava istraživačima da odgovore na kritična pitanja javnog zdravlja, kao što je da li bi smanjenje specifične vrste zagađenja vazduha dovelo do merljivih poboljšanja zdravlja, čak i kada više zagađivača kovarira u realnim uslovima.

2.2.3.3 Elastic Net: Selekcija varijabli u visokodimenzionalnim genomskim podacima

Revolucija genomike je transformisala medicinu i javno zdravlje, nudeći nov uvid u genetske osnove bolesti, odgovor na tretman i ranjivost populacije (Green et al., 2011). Međutim, ovaj napredak je došao sa novim analitičkim izazovom ogromne razmere: problem " $p \gg n$ ". U studijama genetskih asocijacija, broj potencijalnih prediktorskih varijabli (p) – individualni genetski markeri poput polimorfizama pojedinačnih nukleotida (*single nucleotide*

polymorphisms - SNPs) – rutinski dostiže stotine hiljada ili čak milione, daleko prevazilažeći broj subjekta studije (n) (K. L. Ayers & Cordell, 2010). Ova ogromna dimenzionalnost čini tradicionalne regresione metode numerički nemogućim i matematički nedefinisanim bez jedinstvenog rešenja. Štaviše, SNP-ovi unutar istog gena ili hromozomskog regiona su često u stanju linkage disequilibrium (LD), što znači da nisu nezavisno nasleđeni već su statistički povezani (K. L. Ayers & Cordell, 2010). Ovo uvodi dodatni sloj složenosti: ne samo da postoje previše varijabli, već su mnoge od tih varijabli visoko korelisane. Modeli koji ne uzimaju u obzir ovu strukturu rizikuju da propuste važne genetske signale ili da proizvedu nestabilne, nereproducibilne rezultate.

Elastic Net regresija je metoda penalizovane (ili regularizovane) regresije koja je izuzetno pogodna za ovo visokodimenzionalno, korelisano okruženje podataka (Zou & Hastie, 2005). Ona postiže napredak kombinujući dve različite kaznene funkcije—Lasso (L1) penalizaciju i Ridge (L2) penalizaciju—da bi postigla i selekciju varijabli i robusno procesiranje korelisanih prediktora. Lasso komponenta penalizacije radi tako što redukuje koeficijente uz mnoge prediktore na nulu, efektivno ih uklanjajući iz modela (L. Ayers & Cordell, 2010; Tibshirani, 1996). Ovo proizvodi prostiji i interpretabilniji model, što je vrlo poželjno kada se počinje sa milionima potencijalnih prediktora. Međutim, kada se suoči sa grupom visoko korelisanih prediktora (poput SNP-ova u LD), Lasso ima tendenciju da proizvoljno selektuje samo jednu varijablu iz grupe i odbaci ostale. Ovde Ridge komponenta Elastic Net-a postaje kritična. Ridge penal ohrabruje model da smanji koeficijente onih prediktora koji su korelisani jedni prema drugima, efektivno dozvoljavajući im da budu uključeni ili isključeni iz modela kao grupa (Zou & Hastie, 2005). Mešanjem ove dve penalizacije, Elastic Net zadržava selekcionišuće osobine Lasso-a dok pozajmljuje stabilnost i sposobnost rukovanja grupama od Ridge regresije. Ovo je čini moćnim i široko korišćenim alatom u genomici za zadatke kao što su identifikacija genetskih markera osetljivosti na lekove ili otkrivanje SNP-ova povezanih sa složenim bolestima, pružajući bolju ravnotežu između lažno pozitivnih i lažno negativnih nego bilo Lasso ili Ridge sami (Cho et al., 2010; Zou & Hastie, 2005).

Ove napredne regresione metode demonstriraju da izbor prediktivnog modela nije odluka jednog alata za sve slučajeve već zahteva nijansirano razumevanje specifičnog tipa složenosti podataka. „No free lunch” teorija iz mašinskog učenja govori upravo o ovome (Wolpert & Macready, 1997). Robusna regresija se bavi *složenošću vrednosti* (anomalne opservacije), PCR se bavi *strukturnom složenošću* (međusobnim odnosima među prediktorima), a Elastic Net se bavi *dimenzionalnom složenošću* (ogromnim brojem prediktora). Svaka služi kao specijalizovani alat za prevazilaženje različitih, ali uobičajenih izazova u analizi podataka javnog zdravlja.

2.2.4 Ansambl metode i nelinearni modeli u epidemiološkim istraživanjima

Dok su regresione tehnike koje smo prethodno diskutovali moćne, one su primarno dizajnirane da modeluju linearne odnose ili odnose koji mogu biti transformisani u linearnu skalu. Međutim, mnogi kauzalni putevi u javnom zdravlju su inherentno složeni, nelinearni i okarakterisani zamršenim interakcijama između više faktora. Na primer, odnos između prihoda i zdravlja nije prava linija; zdravstvene koristi dodatnih hiljadu dolara prihoda su daleko veće za pojedinca sa niskim prihodima nego za osobu sa visokim prihodima.

Tradicionalni regresioni modeli mogu biti primorani da uhvate takve nelinearnosti kroz ručno uključivanje polinomskih ili interakcionih članova, ali to zahteva apriorno znanje o prirodi ovih odnosa. Metode mašinskog učenja, posebno ansambl i nelinearni modeli, nude fleksibilniji pristup vođen podacima (Wiemken & Kelley, 2020). Ovi algoritmi su dizajnirani da uče složene obrasce i interakcije direktno iz podataka, bez potrebe za unapred specifikovanim pretpostavkama o funkcionalnom obliku odnosa, čineći ih izuzetno pogodnim za multifasetnu prirodu epidemioloških istraživanja (Breiman, 2001; Wiemken & Kelley, 2020).

2.2.4.1 *Random Forest: Razlučivanje društvenih determinanti zdravlja*

Centralni izazov u modernom javnom zdravlju je razumeti i adresirati društvene determinante zdravlja—složenu mrežu društvenih, ekonomskih, bihejvioralnih i ekoloških faktora koji oblikuju zdravstvene ishode. Istraživanje stanja kao što je depresija, na primer, zahteva pomeranje izvan čisto biomedicinskih okvira da bi se razmotrio uticaj faktora kao što su prihodi, obrazovanje, socijalni kapital, diskriminacija i porodična istorija (Remes et al., 2021). Ove varijable su često brojne, visoko međusobno korelisane (npr. obrazovanje i prihod), i njihov uticaj na mentalno zdravlje je verovatno nelinearan i interaktivan.

Random Forest algoritam - metoda ansambl učenja, je posebno pogodna za ovu vrstu problema javnog zdravlja (Breiman, 2001; Wiemken & Kelley, 2020). Random Forest funkcioniše tako što konstruiše mnoštvo individualnih stabala odlučivanja tokom treniranja. Svako stablo je izgrađeno koristeći nasumični podskup podataka (tehnika zvana bagging) i, na svakom razdvajanju u stablu, razmatra se samo nasumični podskup prediktorskih varijabli. Da bi napravio predviđanje, model objedinjuje glasove svih individualnih stabala u "šumi" (Breiman, 2001). Ova dvostruka strategija randomizacije ima nekoliko prednosti za analizu društvenih determinanti. Prvo, ona čini model inherentno robusnim prema multikolinearnosti koja je uobičajena u društvenim podacima, pošto korelisani prediktori verovatno neće biti izabrani zajedno u istom razdvajanju stabla (Breiman, 2001; Wiemken & Kelley, 2020). Drugo, usrednjavanjem predikcija mnogih raznolikih stabala, model može uhvatiti visoko složene, nelinearne odnose i interakcione efekte bez prilagođavanja na trening podacima.

Možda najvažnije za društvenu epidemiologiju, Random Forest modeli pružaju moćan, ugrađeni mehanizam za interpretaciju: skorove značajnosti varijabli (Breiman, 2001; Suarez-Lledo & Alvarez-Galvez, 2019; Wiemken & Kelley, 2020). Merenjem koliko se tačnost predikcije modela smanjuje kada se određena varijabla nasumično izmeša između opservacija, algoritam može rangirati sve prediktore po njihovom relativnom doprinosu modelu. Ova karakteristika pomaže da se model transformiše iz "crne kutije" u opservabilan analitički alat. U studiji o društvenim determinantama depresije, ova sposobnost je omogućila istraživačima da identifikuju i rangiraju ključne pokretače depresije unutar različitih podgrupa populacije; pristup vođen podacima pruža komplement tradicionalnom, teorijski vođenom sociološkom istraživanju i ističe potencijalne mete za javnozdravstvene intervencije (Breiman, 2001; Wiemken & Kelley, 2020).

2.2.4.2 Support Vector Machines (SVM): Klasifikacija visokog rizika za dijagnozu i stratifikaciju rizika

U mnogim scenarijima javnog zdravlja i kliničke prakse, prediktivni zadatak nije procena vrednosti neke kontinualne varijable već donošenje visoko rizične klasifikacione odluke. Primeri uključuju dijagnostikovanje tumora kao benignog ili malignog na osnovu karakteristika ekstrahovanih iz medicinske slike, stratifikaciju pacijenata u grupe visokog ili niskog rizika za postoperativne komplikacije, ili predikcija da li će doći do epidemije bolesti na osnovu podataka nadzora (Almotairi et al., 2023). U ovim primenama, cena pogrešne klasifikacije može biti ekstremno visoka, što zahteva model sa visokom preciznošću i robusnošću.

Support Vector Machines (SVM) su klasa algoritama mašinskog učenja koji se ističu u takvim klasifikacionim zadacima (Almotairi et al., 2023). Osnovna ideja SVM-a je da pronađe optimalnu hiperravan tj. granicu odlučivanja koja najbolje razdvaja tačke podataka različitih klasa u visokodimenzionalnom prostoru ulaznih priomenljivih. "Optimalna" hiperravan je definisana kao ona koja ima maksimalnu marginu, ili rastojanje, između sebe i najbližih tačaka podataka bilo koje klase ("support vectors") (Cortes & Vapnik, 1995). Ovaj princip maksimalne margine čini model robusnim i manje podložnim uticaju malih varijacija u podacima. Štaviše, kroz upotrebu tehnike nazvane "kernel trick" (Aizerman et al., 2019), SVM-ovi mogu efikasno pronaći nelinearne granice odlučivanja, omogućavajući im da klasifikuju podatke koji nisu linearno razdvojivi.

Zahvaljujući svojoj visokoj tačnosti i pouzdanosti, SVM-ovi su široko primenjeni u zdravstvu. Korišćeni su za postizanje vrhunskih performansi u dijagnozi raka iz histopatoloških i genomskih podataka, u predikciji kardiovaskularnih bolesti iz elektronskih zdravstvenih kartona (EHR), i u personalizovanoj medicini (Guido et al., 2024). Njihova svestranost se takođe proteže na nadzor zaraznih bolesti, gde mogu biti korišćeni za predikciju epidemija obradom kombinacije strukturiranih podataka (npr. EHR) i nestrukturiranih podataka (npr. tekst sa društvenih medija ili izveštaji sa vesti) da bi detektovali rane znake epidemije (Ekundayo, 2024; Villanueva-Miranda et al., 2025).

2.2.4.3 Extreme Gradient Boosting (XGBoost): Visokoperformantna predikcija epidemija zaraznih bolesti

Razvoj blagovremenih i tačnih sistema ranog upozorenja za epidemije zaraznih bolesti je kritična funkcija javnog zdravlja. Predikcija trajektorije bolesti poput gripa ili denga groznice zahteva modele koji mogu uhvatiti složenu vremensku dinamiku, sezonalnost i uticaj spoljašnjih pokretača kao što su klimatske varijable (temperatura, padavine), mobilnost populacije i obrasci vazdušnog saobraćaja (Long et al., 2025). Odnosi između ovih faktora i incidence bolesti su često nelinearni i uključuju vremenski zakasnele efekte.

Extreme Gradient Boosting (XGBoost) je napredna i visoko optimizovana implementacija gradient boosting algoritma, ansambl tehnike koja je postala dominantna metoda za prediktivno modelovanje na tabularnim podacima (T. Chen & Guestrin, 2016). Kao i Random Forest, to je ansambl stabala odlučivanja. Međutim, umesto izgradnje nezavisnih stabala, gradient boosting ih gradi sekvencijalno. Svako novo stablo je trenirano da ispravi greške (rezidualne greške) koje je napravilo ansambl svih prethodnih stabala (Friedman, 2001).

Ovaj iterativni, proces korekcije grešaka, kombinovan sa naprednim tehnikama regularizacije za sprečavanje prenaučavanja, čini XGBoost izuzetno moćnim i tačnim (T. Chen & Guestrin, 2016).

XGBoost je poznat po svojoj superiornoj prediktivnoj performansi i računskoj efikasnosti, često osvajajući takmičenja mašinskog učenja i nadmašujući druge modele u realnim zadacima predikcije (T. Chen & Guestrin, 2016; Uddin & Lu, 2024). U kontekstu javnog zdravlja, uspešno je primenjen za predikciju incidencije gripa nekoliko nedelja unapred koristeći samo istorijske podatke epidemiološkog nadzora (Q. Chen et al., 2024). U studijama o denga groznici, XGBoost modeli su demonstrirali visoku tačnost u predikciji epidemija identifikujući najvažnije prediktore iz mešavine epidemioloških i meteoroloških podataka, kao što su incidencija dengue groznice u zemljama porekla avionskih putnika ili specifični zakasneli efekti oblačnosti i padavina (Q. Chen et al., 2024; Guo et al., 2017). Brzina i tačnost XGBoost-a ga čine neprocenjivim alatom za operativne sisteme nadzora, gde brze i pouzdane prognoze mogu direktno usmeravati javnozdravstvene akcije kao što su kampanje kontrole vektora ili planovi pripravnosti bolnica.

Izbor među ovim moćnim modelima mašinskog učenja često odražava fundamentalnu tenziju između prediktivne moći i direktne interpretabilnosti, odluka koja je u krajnjoj liniji vođena specifičnim ciljem javnog zdravlja. Kada je cilj razumeti osnovne kauzalne faktore da bi se informisala dugoročna politika, kao sa društvenim determinantama depresije, model kao Random Forest koji nudi jasne metrike značaja varijabli je veoma vredan (Gohari et al., 2024). Kada je zadatak visoko rizična klasifikacija sa ozbiljnim posledicama za grešku, kao dijagnoza raka, model kao SVM poznat po svojim robusnim i preciznim granicama odlučivanja je često preferiran (Kourou et al., 2015). Suprotno, kada je primarni cilj generisati najtačniju i najblagovremeniju predikciju moguću da se pokrene neposredan javnozdravstveni odgovor, kao u predikciji neizbežne epidemije bolesti, visokoperformantni model kao XGBoost može biti optimalan izbor, čak i ako su njegova unutrašnja funkcionisanja neopservabilnija (Keshavamurthy et al., 2025). "Najbolji" model je tako definisan ne jednom statističkom metrikom, već njegovom prikladnošću za specifičnu svrhu javnog zdravlja, bilo da je ta svrha objašnjenje, klasifikacija ili predikcija.

2.3 Uloga faktora životne sredine u transmisiji respiratornih virusa

Transmisija respiratornih virusa nije isključivo biološki proces već je takođe pod jakim uticajem okruženja. Značajan korpus literature (Comunian et al., 2020; Copat et al., 2020; Yao et al., 2021) je istraživao ulogu zagađenja vazduha i meteoroloških uslova u moduliranju širenja virusa poput gripa i, odskora, SARS-CoV-2, što će biti opisano u ovoj sekciji.

2.3.1 Uloga zagađenja vazduha ($PM_{2.5}$, PM_{10} , NO_2)

Zagađenje vazduha u okolini je dugo povezivano sa povećanim morbiditetom i mortalitetom od respiratornih infekcija. Nekoliko studija je istraživalo njegovu ulogu u pandemiji COVID-19, sa posebnim fokusom na fine čestične materije ($PM_{2.5}$), azot-dioksid (NO_2) i ozon (O_3)

(Copat et al., 2020; Yao et al., 2021). Predložena su dva primarna mehanizma za objašnjenje ove veze:

1. Zagađivači kao virusni nosači: Čestične materije mogu delovati kao fizički vektor za virus. Respiratorne kapi sa virusom mogu se zakačiti za PM čestice, potencijalno produžavajući njihovo vreme suspenzije u vazduhu i olakšavajući njihov transport na veće udaljenosti (Contini & Costabile, 2020). Ovaj mehanizam može povećati efektivni domet i dozu izloženosti.

2. Povećana podložnost domaćina: Hronična izloženost zagađivačima vazduha može ugroziti funkciju respiratornog i imunog sistema. Na primer, zagađivači mogu izazvati upalu u disajnim putevima i stimulirati prekomernu ekspresiju ACE2 receptora, upravo receptora koje SARS-CoV-2 koristi za ulazak u ćeliju, time povećavajući podložnost pojedinca na infekciju (Paital & Agrawal, 2021; Sagawa et al., 2021).

Brojne ekološke studije su pronašle jake korelacije između dugotrajne izloženosti zagađivačima, posebno PM_{2.5}, i viših stopa zaraze i smrti usled COVID-19 (Comunian et al., 2020; Stieb et al., 2020). Istraživanje predstavljeno u ovoj tezi se direktno oslanja na zaključke ovog rada, uzimajući u obzir prostornu i vremensku dinamiku da bi potvrdilo i kvantifikovalo ulogu PM_{2.5} kao važnog prediktora virusne prenosivosti.

2.3.2 Meteorološki faktori i sezonalnost

Transmisija mnogih respiratornih virusa, kao što je grip, pokazuje jake sezonske obrasce, sa vrhuncima koji se tipično javljaju u hladnijim, suvljim mesecima zime. Ovo je dovelo do intenzivnog istraživanja da li COVID-19 sledi sličan sezonski trend. Meteorološki faktori kao što su temperatura, vlažnost i UV zračenje mogu uticati na stabilnost virusa na površinama i u aerosolima, kao i na ljudsko ponašanje (npr. vreme provedeno u zatvorenom prostoru) i imune odgovore (Byun et al., 2021). Dok su neke studije prijavile inverznu asocijaciju između temperature i vlažnosti i transmisije COVID-19, dokazi su bili mešoviti, posebno u ranim fazama pandemije kada je imunitet populacije bio nizak. Globalna analiza u ovoj tezi pruža robusne argumente za jak, hemisferski zavisani sezonski signal, sugerišući da kako pandemija postaje endemska, sezonalnost će verovatno biti dominantan pokretač njene dinamike.

2.4 Sociodemografski i ekonomski faktori u pandemijama

Pandemije nisu samo biološki događaji; one su društvene pojave. Širenje i uticaj bolesti su duboko oblikovani društvenim, demografskim i ekonomskim karakteristikama populacije. Faktori kao što je gustina stanovništva mogu uticati na stope kontakta, iako je njena uloga u COVID-19 bila predmet debate, sa nekim studijama koje sugerišu da su povezanost i mobilnost važniji od jednostavne gustine (Hamidi et al., 2020; Jo et al., 2021). Starosna struktura populacije je kritična, pošto je težina COVID-19 snažno zavisna od starosti (Santesmasses et al., 2020). Prevalenca komorbiditeta kao što su dijabetes, hipertenzija i gojaznost takođe značajno povećava rizik od teških ishoda (Williamson et al., 2020).

Štaviše, socioekonomski faktori igraju krucijalnu ulogu. Ekonomski prosperitet, kapacitet zdravstvenog sistema i dispariteti u javnom zdravlju mogu stvoriti ogromne razlike u ranjivosti i ishodima. Populacije sa nižim socioekonomskim statusom i manjinske grupe su često bile neproporcionalno pogođene COVID-19, tragična realnost povezana sa faktorima kao što su uslovi stanovanja, profesionalna izloženost i nejednak pristup zdravstvenoj zaštiti (Chakraborty, 2021). Ove varijable su esencijalni konfundirajući faktori koji moraju biti uzeti u obzir u bilo kom modelu koji nastoji da izoluje efekte faktora životne sredine.

3 Opšta metodologija

Ovo poglavlje opisuje sveobuhvatan metodološki okvir razvijen za istraživanje determinanti prenosivosti virusa **SARS-CoV-2** na globalnom nivou. Predloženi metodološki pristup zasnovan je na integraciji epidemioloških principa, savremenih statističkih metoda i tehnika mašinskog učenja, sa ciljem da se identifikuju ključni faktori koji utiču na dinamiku širenja respiratornih infekcija u različitim geografskim i društveno-ekonomskim kontekstima. Poseban fokus stavljen je na analizu prostorne heterogenosti transmisije i na kvantifikaciju odnosa između epidemioloških, demografskih, klimatskih i socioekonomskih determinanti i intenziteta prenosa infekcije.

Razvijeni pristup koncipiran je tako da bude metodološki robustan, statistički validan i reproduktivan, i uzme u obzir složenost i visok stepen dimenzionalnosti podataka koji karakterišu globalne epidemiološke studije. U istraživanju su korišćeni podaci prikupljeni iz više međunarodnih javno dostupnih izvora, uključujući epidemiološke baze podataka, meteorološke i ekološke skupove podataka, kao i indikatore socioekonomskog razvoja. Integracija ovih heterogenih izvora podataka zahtevala je sistematski pristup prikupljanju, standardizaciji i transformaciji podataka kako bi se obezbedila njihova međusobna uporedivost i statistička konzistentnost. Posebna pažnja posvećena je kontroli kvaliteta podataka, identifikaciji i obradi nedostajućih vrednosti, kao i redukciji potencijalnog uticaja sistematskih grešaka i pristrasnosti koje mogu nastati u velikim međunarodnim bazama podataka.

Metodološki okvir ove studije zasniva se na kvantitativnoj analizi odnosa između determinanti prenosa infekcije i epidemioloških indikatora širenja bolesti. U tom kontekstu, primarni epidemiološki indikator korišćen u analizi predstavlja **osnovni reprodukcion broj (R_0)**, koji se u epidemiologiji zaraznih bolesti definiše kao prosečan broj sekundarnih infekcija koje nastaju od jednog infektivnog slučaja u potpuno osetljivoj populaciji. R_0 predstavlja ključni parametar za razumevanje potencijala širenja infekcije, jer vrednosti veće od jedan ukazuju na mogućnost eksponencijalnog rasta epidemije, dok vrednosti manje od jedan sugerišu postepeno gašenje transmisije.

Ukupna metodologija istraživanja organizovana je u tri međusobno povezana metodološka segmenta. Prvi segment obuhvata postupak procene primarne zavisne varijable, odnosno osnovnog broja reprodukcije (R_0), korišćenjem epidemioloških podataka o incidenciji bolesti i odgovarajućih statističkih modela za procenu transmisije infekcije. Drugi segment odnosi se na proces prikupljanja podataka, njihovu obradu i transformaciju, uključujući postupke čišćenja podataka, normalizacije promenljivih, kao i inženjeringa obeležja (feature engineering) radi konstruisanja informativnih prediktora relevantnih za modelovanje širenja infekcije. Treći segment metodologije obuhvata primenu naprednih tehnika mašinskog učenja u cilju izgradnje prediktivnih modela koji omogućavaju identifikaciju najznačajnijih faktora povezanih sa varijabilnošću R_0 između različitih geografskih regiona.

Ovakav integrisani pristup omogućava istovremeno razmatranje velikog broja potencijalnih determinanti transmisije, kao i procenu njihovog relativnog doprinosa u objašnjenju

prostorne i vremenske varijabilnosti prenosivosti SARS-CoV-2. Kombinacija klasičnih epidemioloških metoda i savremenih algoritama mašinskog učenja omogućava dublje razumevanje kompleksnih obrazaca širenja respiratornih infekcija i predstavlja osnovu za razvoj preciznijih modela predikcije epidemijskih trendova, što može imati značajne implikacije za planiranje javnozdravstvenih intervencija i strategija kontrole budućih pandemija.

3.1 Procena osnovnog reproduktionog broja (R_0)

Osnovni reprodukcioni broj (R_0) predstavlja jedan od ključnih epidemioloških parametara koji kvantifikuje potencijal širenja infektivnih bolesti u populaciji. U kontekstu respiratornih infekcija, uključujući infekciju izazvanu virusom SARS-CoV-2, R_0 se definiše kao prosečan broj sekundarnih slučajeva infekcije koji nastaju kao posledica jednog primarnog slučaja u potpuno osetljivoj populaciji, u odsustvu intervencija javnog zdravlja. Vrednost ovog parametra direktno odražava intenzitet transmisije patogena i predstavlja osnovu za razumevanje dinamike epidemijskog procesa.

U epidemiološkim modelima širenja zaraznih bolesti, R_0 se može konceptualno izraziti kao proizvod tri osnovne komponente: verovatnoće prenosa infekcije po kontaktu (p), stope kontakata između osetljivih i infektivnih pojedinaca (k), kao i trajanja infektivnog perioda (D). Formalno, ovaj odnos može se predstaviti sledećim izrazom:

$$R_0 = p * k * D = \beta * D \quad (1)$$

gde je β efektivna stopa transmisije. U realnim epidemiološkim uslovima, procena ovih parametara često nije direktno dostupna, te se R_0 najčešće procenjuje na osnovu podataka o incidenciji bolesti tokom ranih faza epidemije, kada je većina populacije i dalje osetljiva na infekciju. U ovom istraživanju procena R_0 zasniva se na analizi rane eksponencijalne faze epidemijskog rasta, tokom koje se broj novoregistrovanih slučajeva može aproksimirati eksponencijalnom funkcijom oblika:

$$I(t) = I_0 e^{rt} \quad (2)$$

gde $I(t)$ predstavlja broj slučajeva zaražavanja u vremenu t , I_0 početni broj slučajeva, dok r označava eksponencijalnu stopu rasta epidemije. Parametar r (dan^{-1}) predstavlja trenutnu eksponencijalnu stopu rasta broja slučajeva u početnoj, nekontrolisanoj fazi epidemije i direktno se procenjuje kao nagib log-linearnom regresijom $\log(D(t))$ na t (Jednačina (11)). Procena parametra r omogućava indirektno određivanje vrednosti osnovnog reproduktionog broja kroz odnos između stope rasta epidemije i distribucije serijskog intervala infekcije. Serijski interval definisan je kao vremenski razmak između pojave simptoma kod primarnog slučaja i pojave simptoma kod sekundarnog slučaja koji je zaražen od strane primarnog slučaja. U skladu sa prethodnim epidemiološkim studijama SARS-CoV-2 infekcije, distribucija serijskog intervala modelovana je korišćenjem gama distribucije sa parametrima procenjenim na osnovu objavljenih epidemioloških analiza (Nishiura et al.,

2020). Kombinacijom procene eksponencijalne stope rasta epidemije i distribucije serijskog intervala moguće je dobiti pouzdanu procenu osnovnog reproduktionog broja.

U cilju povećanja robusnosti procena, analiza je sprovedena na nivou pojedinačnih zemalja i regiona, koristeći agregirane podatke o dnevnom broju potvrđenih slučajeva infekcije. Za svaku posmatranu teritoriju identifikovan je početni vremenski interval epidemijskog rasta u kojem se dinamika širenja bolesti može smatrati približno eksponencijalnom. Ovaj vremenski prozor definisan je na osnovu kriterijuma stabilnog rasta incidencije i minimalnog uticaja nefarmaceutskih intervencija, kao što su mere fizičkog distanciranja, ograničenja mobilnosti i javnozdravstvene restrikcije.

Procene R_0 sprovedene su korišćenjem statističkih metoda implementiranih u programskom okruženju R, uz primenu specijalizovanih epidemioloških paketa za analizu transmisije zaraznih bolesti. Posebna pažnja posvećena je proceni intervala pouzdanosti, kako bi se kvantifikovala nesigurnost procena koja proizlazi iz varijabilnosti epidemioloških podataka i ograničenja dostupnih informacija tokom ranih faza pandemije.

Dobijene vrednosti osnovnog reproduktionog broja predstavljaju primarnu zavisnu varijablu u daljoj analizi i korišćene su kao ciljna promenljiva u modelima mašinskog učenja razvijenim u okviru ove studije. Na taj način omogućeno je identifikovanje ključnih determinanti koje doprinose prostornoj i vremenskoj varijabilnosti prenosivosti SARS-CoV-2 na globalnom nivou.

3.1.1 Izbor vremenskog intervala za procenu R_0

Precizna procena osnovnog reproduktionog broja (R_0) u velikoj meri zavisi od adekvatnog izbora vremenskog intervala u kojem se analiziraju epidemiološki podaci. U teorijskom smislu, R_0 se definiše u uslovima potpuno osetljive populacije i odsustva intervencija javnog zdravlja, zbog čega se njegova procena u realnim epidemiološkim uslovima najčešće zasniva na analizi **rane faze epidemijskog rasta**, kada je uticaj kontrolnih mera minimalan, a transmisija patogena približno neometana.

U ovom istraživanju posebna pažnja posvećena je identifikaciji vremenskog prozora koji najbolje odgovara pretpostavkama eksponencijalnog rasta epidemije. Tokom početnih faza širenja SARS-CoV-2 infekcije, broj novoregistrovanih slučajeva često pokazuje obrazac približno eksponencijalnog povećanja, što omogućava primenu matematičkih modela zasnovanih na eksponencijalnoj funkciji rasta. U tom periodu može se pretpostaviti da većina populacije još uvek nije stekla imunitet, bilo prirodnom infekcijom ili vakcinacijom, čime su zadovoljene osnovne pretpostavke za procenu R_0 .

Proces identifikacije odgovarajućeg vremenskog intervala sproveden je kroz nekoliko uzastopnih analitičkih koraka. Najpre su analizirani vremenski nizovi dnevne incidencije potvrđenih slučajeva COVID-19 za svaku posmatranu zemlju ili region. Nakon toga je izvršena preliminarne vizuelna i statistička analiza dinamike rasta broja slučajeva kako bi se identifikovao period stabilnog eksponencijalnog povećanja incidencije. Ovaj interval je definisan kao vremenski segment u kojem logaritamski transformisan broj slučajeva

pokazuje približno linearan trend u funkciji vremena, što predstavlja indikator eksponencijalnog rasta epidemije.

Dodatni kriterijumi za izbor vremenskog intervala uključivali su i procenu uticaja nefarmaceutskih intervencija (engl. *non-pharmaceutical interventions*, NPI), kao što su zatvaranje obrazovnih institucija, ograničenja mobilnosti stanovništva, zabrane javnih okupljanja i druge javnozdravstvene mere uvedene sa ciljem suzbijanja transmisije virusa. Periodi tokom kojih su ove intervencije imale značajan uticaj na dinamiku epidemije nisu uključivani u procenu R_0 , jer bi takve mere mogle dovesti do potcenjivanja stvarnog potencijala transmisije patogena.

Kako bi se dodatno smanjio uticaj statističkog šuma i dnevnih fluktuacija u prijavljenim slučajevima, primenjene su metode glaćenja vremenskih nizova (Hyndman & Khandakar, 2008), uključujući pokretne proseke i druge standardne postupke filtriranja epidemioloških podataka. Ovakav pristup omogućio je stabilniju procenu eksponencijalne stope rasta epidemije i smanjio uticaj nepravilnosti koje mogu nastati usled kašnjenja u prijavljivanju slučajeva ili promena u testiranju populacije.

Na osnovu definisanog vremenskog intervala rasta epidemije procenjena je eksponencijalna stopa rasta r , koja predstavlja osnovu za dalju procenu osnovnog reproduktionog broja. Ovakav pristup omogućava metodološki konzistentnu procenu R_0 između različitih geografskih regiona i predstavlja važan korak u obezbeđivanju uporedivosti rezultata u globalnim epidemiološkim analizama.

Na kraju, dobijeni vremenski intervali korišćeni su kao osnova za procenu parametara transmisije u narednim fazama metodološkog okvira, pri čemu su procenjene vrednosti R_0 integrisane u dalju analizu determinanti širenja infekcije korišćenjem statističkih i mašinskih metoda modelovanja.

3.1.2 Procena eksponencijalne stope rasta epidemije

U ovoj sekciji je prikazan sistem diferencijalnih jednačina koji modeluje širenje infekcije, a zatim i njegovo rešenje korišćeno za procenu osnovnog reproduktionog broja.

3.1.2.1 Matematički okvir

Za opis dinamike korišćen je modifikovani SEIR model (Keeling & Rohani, 2011), koji populaciju deli na pet kompartmana:

- $S(t)$ – osetljivi pojedinci
- $E(t)$ – izloženi (zaraženi ali još neinfektivni)
- $I(t)$ – infektivni
- $R(t)$ – oporavljeni / uklonjeni
- $D(t)$ – detektovani slučajevi

Ukupna veličina populacije označena je sa $N = S + E + I + R$, pri čemu se pretpostavlja da je N konstantna tokom posmatranog perioda (zanemaruju se migracije i prirodni priraštaj). Dinamika se opisuje sistemom diferencijalnih jednačina:

$$dS/dt = -\beta SI/N \quad (3)$$

$$dE/dt = \beta SI/N - \sigma E \quad (4)$$

$$dI/dt = \sigma E - \gamma I \quad (5)$$

$$dR/dt = \gamma I \quad (6)$$

$$dD/dt = \varepsilon \delta I \quad (7)$$

gde β predstavlja stopu prenosa infekcije, parametri $\sigma = 0,2 \text{ dan}^{-1}$ i $\gamma = 0,4 \text{ dan}^{-1}$ odgovaraju recipročnim vrednostima prosečnog latentnog perioda ($1/\sigma = 5$ dana) i prosečnog infektivnog perioda ($1/\gamma = 2,5$ dana), a ε i δ redom predstavljaju efikasnost i stopu detekcije. Vrednost latentnog perioda od približno 5 dana usaglašena je sa ranim globalnim procenama prosečnog perioda inkubacije (Lauer et al., 2020; Li et al., 2020). S druge strane, kratak efektivni infektivni period od 2,5 dana odražava visoku vršnu infektivnost karakterističnu za SARS-CoV-2 u danima neposredno oko pojave simptoma, usled kratkog serijskog intervala (Prem et al., 2020). Ova specifična kombinacija matematičkih parametara pokazala se kao visoko robusna za modelovanje stope eksponencijalnog rasta u inicijalnoj, nekontrolisanoj fazi pandemije (Salom et al., 2021).

3.1.2.2 Eksponencijalni režim

U ranoj fazi epidemije važi aproksimacija:

$$\frac{S}{N} \approx 1 \quad (8)$$

jer je gotovo cela populacija osetljiva. U tom slučaju sistem diferencijalnih jednačina (3)–(7) se linearizuje i svodi na sistem za E i I koji ima eksponencijalna rešenja oblika $e^{\lambda t}$. Dominantna (najveća realna) sopstvena vrednost ovog linearizovanog sistema označena je sa λ^+ i predstavlja merenu stopu eksponencijalnog rasta epidemije — identičnu parametru r iz Jednačine (2). Broj infektivnih i detektovanih slučajeva tada raste kao:

$$I(t) \sim e^{\lambda_+ t} \quad (9)$$

što implicira:

$$D(t) \sim e^{\lambda_+ t} \quad (10)$$

Logaritmovanjem dobijamo:

$$\log(D(t)) = \lambda_+ t + C \quad (11)$$

Dakle, nagib log-linearne regresije direktno daje λ^+ . Analitička veza između λ^+ i R_0 je:

$$R_0 = 1 + \frac{[\lambda_+(\gamma + \sigma) + \lambda_+^2]}{\gamma\sigma} \quad (12)$$

Ključna prednost ovog pristupa je što je R_0 procenjen na ovaj način manje osetljiv na apsolutni obim testiranja. Sem toga, čak i ako je stopa detekcije konstantno pristrasna, nagib eksponencijalnog rasta ostaje nepromenjen. Takođe, fokus isključivo na inicijalnu fazu epidemije znači da rezultati nisu pod uticajem kasnijih talasa, pojave novih sojeva, niti mera socijalne distance čiji se efekti u krivoj slučajeva pojavljuju sa kašnjenjem od ~10 dana usled inkubacionog perioda.

3.2 Prikupljanje podataka, pretprocesiranje i inženjering obeležja

U cilju analize determinanti prenosivosti respiratornih infekcija i identifikovanja faktora koji utiču na varijabilnost osnovnog reproduccionog broja (R_0), u ovom istraživanju korišćen je integrisani skup podataka koji obuhvata epidemiološke, klimatske, demografske, socioekonomske i ekološke varijable. S obzirom na kompleksnost fenomena transmisije zaraznih bolesti, bilo je neophodno integrisati podatke iz više međunarodnih izvora kako bi se obuhvatile različite dimenzije koje mogu uticati na dinamiku širenja infekcije.

3.2.1 Izvori podataka

Prikupljanje podataka sprovedeno je korišćenjem javno dostupnih međunarodnih baza podataka koje obezbeđuju standardizovane i uporedive indikatore na globalnom nivou. Epidemiološki podaci o broju potvrđenih slučajeva COVID-19 preuzeti su iz međunarodnih epidemioloških repozitorijuma (Dong et al., 2020) koji prate dnevnu incidenciju infekcije po državama i regionima. Ovi podaci korišćeni su kao osnov za procenu R_0 koji predstavlja primarnu zavisnu varijablu u daljoj analizi.

Pored epidemioloških podataka, u analizu su uključeni i klimatski i meteorološki indikatori koji potencijalno utiču na stabilnost virusa u spoljašnjoj sredini i na obrasce transmisije respiratornih infekcija. Ovi podaci obuhvataju varijable kao što su srednja temperatura vazduha, relativna vlažnost, količina padavina, brzina vetra i solarna radijacija. Meteorološki parametri preuzeti su iz globalnih klimatskih baza podataka (NASA Langley Research Center, 2020) koje omogućavaju prostorno i vremenski standardizovane meteorološke procene za veliki broj geografskih lokacija.

U istraživanje su takođe uključeni i indikatori kvaliteta vazduha i zagađenja životne sredine (Crippa et al., 2020), s obzirom na to da prethodne studije ukazuju na potencijalnu povezanost između nivoa zagađenja i težine respiratornih infekcija (Comunian et al., 2020; Copat et al., 2020). Varijable koje opisuju koncentracije suspendovanih čestica ($PM_{2.5}$ i PM_{10}), kao i koncentracije gasovitih polutanata kao što su azotni dioksid (NO_2), sumpor-

dioksid (SO_2) i ozon (O_3), korišćene su kao potencijalni prediktori u modelima mašinskog učenja.

Pored ekoloških i klimatskih determinanti, u analizu su uključeni i demografski i socioekonomski indikatori koji mogu uticati na intenzitet prenosa infekcije. Ove varijable obuhvataju gustinu naseljenosti, strukturu stanovništva prema starosti, nivo urbanizacije, bruto domaći proizvod po stanovniku, kao i različite indikatore socijalne povezanosti i mobilnosti stanovništva. Demografski i socioekonomski podaci preuzeti su iz međunarodnih statističkih baza podataka koje obezbeđuju uporedive indikatore za veliki broj zemalja (Bailey et al., 2018; European Commission Global Human Settlement, 2020).

S obzirom na to da su prikupljeni podaci poticali iz heterogenih izvora i različitih formata, sproveden je opsežan proces pretprocesiranja podataka kako bi se obezbedila njihova konzistentnost i pogodnost za statističku analizu. Ovaj proces uključivao je standardizaciju naziva geografskih jedinica, usklađivanje vremenskih serija, kao i harmonizaciju mernih jedinica različitih varijabli. Posebna pažnja posvećena je identifikaciji i obradi nedostajućih vrednosti, pri čemu su primenjene odgovarajuće statističke metode imputacije ili filtriranja podataka u zavisnosti od prirode i obima nedostajućih informacija.

Nakon završetka procesa čišćenja i standardizacije podataka sproveden je postupak inženjeringa obeležja (engl. *feature engineering*). Ovaj korak obuhvata transformaciju postojećih varijabli i konstrukciju novih prediktora koji mogu biti informativniji za modelovanje epidemioloških procesa. U okviru ovog postupka izvršene su različite transformacije podataka, uključujući logaritamsku transformaciju varijabli sa izrazito asimetričnom distribucijom, standardizaciju i normalizaciju numeričkih varijabli, kao i agregaciju pojedinih indikatora u kompozitne indekse.

Konačni skup podataka organizovan je u obliku analitičke matrice u kojoj svaki red predstavlja posmatranu geografsku jedinicu (zemlju ili region), dok kolone odgovaraju pojedinačnim prediktorima i zavisnoj promenljivoj R_0 . Ovakva struktura podataka omogućava primenu naprednih statističkih metoda i algoritama mašinskog učenja u cilju identifikacije ključnih determinanti prenosivosti SARS-CoV-2 i analize njihovog relativnog doprinosa u objašnjenju prostorne varijabilnosti epidemijskog širenja.

3.2.2 Pretprocesiranje podataka

S obzirom na heterogenost i kompleksnost prikupljenih skupova podataka, pre njihove statističke analize bilo je neophodno sprovesti sistematski proces pretprocesiranja kako bi se obezbedila konzistentnost, uporedivost i statistička validnost podataka. Pretprocesiranje podataka predstavlja ključnu fazu u analizi velikih epidemioloških skupova podataka, jer omogućava identifikaciju i korekciju potencijalnih nepravilnosti koje mogu nastati usled različitih metoda prikupljanja podataka, administrativnih kašnjenja u prijavljivanju slučajeva ili nedoslednosti u definisanju pojedinih indikatora.

Prvi korak u procesu pretprocesiranja obuhvatao je harmonizaciju geografskih jedinica i standardizaciju naziva država i regiona u svim prikupljenim skupovima podataka. Budući da

su podaci poticali iz različitih međunarodnih baza, često su korišćene različite nomenklature za iste geografske entitete. Kako bi se omogućilo pravilno povezivanje podataka iz različitih izvora, primenjena je standardizacija prema međunarodno prihvaćenim kodovima država (ISO standardi), čime je obezbeđena konzistentnost u identifikaciji geografskih jedinica.

Nakon harmonizacije geografskih identifikatora sprovedeno je usklađivanje vremenskih serija epidemioloških i pratećih varijabli. S obzirom na to da su pojedini skupovi podataka bili dostupni u različitim vremenskim rezolucijama (dnevnom, mesečnim ili godišnjim), bilo je potrebno izvršiti odgovarajuću agregaciju ili interpolaciju podataka kako bi se obezbedila njihova međusobna kompatibilnost. Granularni podaci su direktno ili ponderisano sabirani do nivoa željene administrativne jedinice (npr. države), a vremenski asinhroni podaci su interpolirani na najgušću vremensku skalu svih pronađenih setova. Ovaj proces omogućio je usklađivanje vremenskog okvira svih promenljivih sa periodom koji je korišćen za procenu osnovnog reproduktionog broja.

Posebna pažnja posvećena je identifikaciji i obradi nedostajućih vrednosti, koje su česta pojava u velikim međunarodnim skupovima podataka. Nedostajuće vrednosti identifikovane su kroz deskriptivnu statističku analizu i vizuelnu inspekciju distribucije podataka. U zavisnosti od obima i strukture nedostajućih podataka primenjene su različite strategije njihove obrade. U slučajevima kada je procenat nedostajućih vrednosti bio mali, primenjivane su metode imputacije zasnovane na statističkim procenama, kao što su imputacija srednjom vrednošću ili interpolacija vremenskih nizova. U situacijama kada je nedostajući deo podataka bio značajan, odgovarajuće promenljive ili posmatranja su izostavljena iz analize kako bi se izbeglo uvođenje sistematske pristrasnosti u modele.

Dodatni korak u procesu pretprocesiranja odnosio se na detekciju i obradu ekstremnih vrednosti (outliers) koje mogu značajno uticati na stabilnost statističkih modela. Identifikacija ekstremnih vrednosti sprovedena je korišćenjem standardnih statističkih metoda, uključujući analizu distribucije podataka, box-plot dijagrame i kriterijume zasnovane na interkvartilnom opsegu. U zavisnosti od prirode identifikovanih ekstremnih vrednosti primenjene su odgovarajuće strategije njihove obrade, uključujući transformaciju promenljivih ili ograničavanje ekstremnih vrednosti na statistički prihvatljive opsege.

U cilju poboljšanja statističkih karakteristika podataka sprovedene su i transformacije pojedinih varijabli, naročito u slučajevima kada su distribucije pokazivale izraženu asimetriju ili visok stepen varijabilnosti. Logaritamske transformacije primenjivane su kod promenljivih koje karakteriše eksponencijalna raspodela vrednosti, čime je omogućena stabilnija statistička analiza i smanjen uticaj ekstremnih vrednosti. Za svaku varijablu je pokušano nekoliko transformacija (različiti korenovi, logaritam itd.) a zatim ona koja proizvodi najveću p-vrednost Shapiro-Wilk testa (Shapiro & Wilk, 1965) je odabrana i pridružena ovoj promenljivoj. Na kraju pretprocesiranja sprovedena je standardizacija numeričkih promenljivih, kako bi se obezbedila njihova uporedivost u okviru modela mašinskog učenja. Standardizacija je izvršena transformacijom promenljivih u oblik sa srednjom vrednošću nula i standardnom devijacijom jedan, čime je omogućeno da sve promenljive imaju sličan numerički opseg i da nijedna promenljiva ne dominira u procesu modelovanja zbog svoje skale merenja. Neke metode imaju bolje performanse nakon ove

transformacije (npr. obična linearna regresija), dok su druge numerički stabilne i standardizacija ne utiče na rezultate.

Rezultat ovog procesa bio je konačni, metodološki konzistentan skup podataka spreman za dalju analizu i primenu algoritama mašinskog učenja, koji su korišćeni za identifikaciju ključnih determinanti prenosivosti SARS-CoV-2 u različitim geografskim i socioekonomskim kontekstima.

3.3 Modelovanje primenom metoda mašinskog učenja

U cilju identifikacije ključnih determinanti prenosivosti SARS-CoV-2 i kvantifikacije njihovog relativnog doprinosa varijabilnosti osnovnog reproduktionog broja (R_0), u ovom istraživanju primenjen je integrisani pristup zasnovan na metodama mašinskog učenja. Za razliku od tradicionalnih statističkih metoda koje često podrazumevaju unapred definisane pretpostavke o funkcionalnom obliku odnosa između promenljivih, algoritmi mašinskog učenja omogućavaju modelovanje složenih, nelinearnih i visoko dimenzionalnih odnosa između velikog broja prediktora i ciljne promenljive.

U kontekstu epidemioloških istraživanja, primena metoda mašinskog učenja omogućava analizu velikih skupova heterogenih podataka koji obuhvataju epidemiološke, klimatske, demografske i socioekonomske varijable. Ovakav pristup omogućava identifikaciju latentnih obrazaca u podacima i otkrivanje kompleksnih interakcija između različitih determinanti transmisije infekcije koje bi mogle ostati neprimećene u okviru klasičnih regresionih modela. U ovom istraživanju modeli mašinskog učenja korišćeni su sa dva osnovna cilja. Prvi cilj bio je izgradnja prediktivnih modela koji omogućavaju procenu vrednosti osnovnog reproduktionog broja na osnovu velikog broja potencijalnih determinanti transmisije. Drugi cilj odnosio se na selekciju obeležja (feature selection) i identifikaciju najznačajnijih faktora koji doprinose prostornoj varijabilnosti prenosivosti infekcije.

Proces modelovanja započet je podelom konačnog skupa podataka na trening skup i test skup. Trening skup korišćen je za obuku modela i procenu parametara algoritama, dok je test skup korišćen za nezavisnu evaluaciju performansi modela i procenu njihove sposobnosti generalizacije na nove podatke. Ovakav pristup omogućava smanjenje rizika od prekomernog prilagođavanja modela (engl. *overfitting*), koje može nastati kada model previše precizno opisuje specifičnosti trening skupa, ali ne uspeva da adekvatno generalizuje obrasce u novim podacima.

U cilju obezbeđivanja stabilnosti i pouzdanosti procena performansi modela, primenjena je i metoda unakrsne validacije (cross-validation). Ova metoda podrazumeva višestruko deljenje trening skupa na podskupove koji se naizmenično koriste za obuku i validaciju modela, čime se dobija robusnija procena prediktivnih performansi i smanjuje zavisnost rezultata od slučajne podele podataka.

Performanse modela procenjivane su korišćenjem standardnih metrika za evaluaciju regresionih modela (Hastie et al., 2009), uključujući koeficijent determinacije (R^2), srednju

kvadratnu grešku (MSE) i srednju apsolutnu grešku (MAE). Ove metrike omogućavaju kvantitativnu procenu tačnosti predikcije modela i upoređivanje performansi različitih algoritama mašinskog učenja primenjenih u analizi.

Poseban značaj u okviru ovog istraživanja ima i analiza važnosti prediktora (feature importance), koja omogućava identifikaciju promenljivih koje imaju najveći uticaj na vrednost ciljne promenljive R_0 . Analiza važnosti prediktora sprovedena je korišćenjem metoda koje procenjuju relativni doprinos svake promenljive u procesu predikcije, čime se omogućava interpretacija rezultata modela i identifikacija ključnih epidemioloških i ekoloških determinanti transmisije infekcije.

Primena metoda mašinskog učenja u ovom istraživanju omogućila je sveobuhvatnu analizu kompleksnih odnosa između velikog broja potencijalnih determinanti i dinamike širenja respiratornih infekcija. Ovakav pristup predstavlja značajan metodološki napredak u odnosu na tradicionalne epidemiološke analize i omogućava dublje razumevanje faktora koji oblikuju obrasce transmisije SARS-CoV-2 na globalnom nivou.

3.3.1 Izbor algoritama mašinskog učenja

Izbor odgovarajućih algoritama mašinskog učenja predstavlja ključni metodološki korak u procesu modelovanja, jer različiti algoritmi imaju različite sposobnosti u pogledu identifikacije složenih obrazaca u podacima, otpornosti na šum i interpretabilnosti rezultata. U ovom istraživanju izbor algoritama zasnovan je na nekoliko kriterijuma: sposobnosti modelovanja nelinearnih odnosa između promenljivih, robusnosti u radu sa relativno velikim brojem prediktora, otpornosti na multikolinearnost između varijabli i mogućnosti procene relativne važnosti pojedinačnih prediktora.

S obzirom na kompleksnu prirodu epidemioloških procesa i potencijalno nelinearne odnose između determinanti transmisije i osnovnog reprodukcionog broja (R_0), u analizi su korišćeni algoritmi zasnovani na **metodama ansambla (ensemble learning)**, koji kombinuju više modela kako bi se postigla veća tačnost i stabilnost predikcija. Ansambl metode su posebno pogodne za analizu visoko dimenzionalnih skupova podataka jer omogućavaju identifikaciju kompleksnih interakcija između promenljivih i smanjuju rizik od prekomernog prilagođavanja modela.

Jedan od primarnih algoritama korišćenih u ovom istraživanju bio je **Random Forest** model. Random Forest predstavlja ansambl metodu zasnovanu na velikom broju stabala odlučivanja (*decision trees*), pri čemu se svako stablo trenira na slučajno odabranom podskupu podataka i podskupu prediktora. Konačna predikcija modela dobija se agregacijom rezultata pojedinačnih stabala. Prednost ovog pristupa ogleda se u njegovoj visokoj robusnosti na šum u podacima, sposobnosti modelovanja nelinearnih odnosa i mogućnosti procene važnosti pojedinačnih prediktora u procesu modelovanja. Ključni hiperparametri Random Forest modela koji se podešavaju su: maksimalni broj prediktora razmatranih pri svakom grananju (mtry ili max splits) — manji broj uvodi nasumičnost koja smanjuje korelaciju između stabala i poboljšava robusnost ansambla; i minimalna veličina terminalnog čvora (min leaf size) — minimalni broj opservacija koji mora biti prisutan u

završnom čvoru stabla, čime se kontroliše dubina stabala i sprečava preprilagođavanje. Broj stabala u ansamblu se tipično bira kao vrednost pri kojoj greška prestaje da se smanjuje.

Pored Random Forest algoritma, u analizi je primenjen i algoritam **Gradient Boosting**, koji predstavlja još jednu široko primenjivanu ansambl tehniku u okviru savremenih metoda mašinskog učenja. Za razliku od Random Forest pristupa, koji paralelno konstruiše više stabala odlučivanja, Gradient Boosting modeli stabla konstruišu sekvencijalno, pri čemu svako naredno stablo pokušava da minimizuje greške prethodnog modela. Ovakav pristup omogućava veoma precizno modelovanje kompleksnih nelinearnih odnosa između prediktora i ciljne promenljive. Ključni hiperparametri Gradient Boosting modela koji se podešavaju su: maksimalni broj podela po stablu (max splits) — kontroliše složenost pojedinačnih stabala i njihovu sposobnost hvatanja interakcija između prediktora; minimalna veličina terminalnog čvora (min leaf size) — analogna uloga kao kod Random Foresta; i stopa učenja (learning rate) — određuje koliko svako naredno stablo koriguje grešku prethodnog, pri čemu niže vrednosti usporavaju učenje ali generalno poboljšavaju generalizaciju.

U cilju dodatnog poboljšanja performansi modela, u analizi su korišćene i napredne implementacije boosting algoritama, kao što su **Extreme Gradient Boosting (XGBoost)** modeli. XGBoost predstavlja optimizovanu implementaciju boosting algoritma koja omogućava efikasno modelovanje velikih skupova podataka uz primenu regularizacije, čime se smanjuje rizik od prekomernog prilagođavanja modela i poboljšava sposobnost generalizacije.

Pored ansambl metoda, u analizi su razmatrani i jednostavniji regresioni modeli kao referentni pristup za poređenje performansi. Ovakvi modeli omogućavaju procenu u kojoj meri složeniji algoritmi mašinskog učenja doprinose poboljšanju tačnosti predikcija u odnosu na tradicionalne statističke metode.

Kombinovanjem više algoritama omogućeno je sistematsko poređenje njihovih performansi i identifikacija modela koji najpreciznije opisuje odnos između determinanti transmisije i osnovnog reproduktionog broja. Ovakav pristup obezbeđuje metodološku robusnost analize i omogućava pouzdaniju interpretaciju rezultata u kontekstu globalne epidemiologije respiratornih infekcija.

3.3.2 Validacija modela i optimizacija hiperparametara

U cilju obezbeđivanja pouzdanosti i stabilnosti prediktivnih modela razvijenih u okviru ovog istraživanja, posebna pažnja posvećena je procesu validacije modela i optimizacije njihovih hiperparametara. U kontekstu metoda mašinskog učenja, validacija modela predstavlja ključni korak kojim se procenjuje sposobnost modela da generalizuje naučene obrasce na nove, prethodno neviđene podatke. Bez adekvatne validacije postoji značajan rizik od prekomernog prilagođavanja modela (*overfitting*), pri čemu model može pokazivati visoku tačnost na trening skupu, ali slabiju performansu na nezavisnim podacima.

U ovom istraživanju skup podataka podeljen je na trening skup i test skup, pri čemu je trening skup korišćen za obuku modela i procenu optimalnih parametara algoritama, dok je test skup korišćen za konačnu procenu performansi modela. Ovakva podela omogućava nezavisnu evaluaciju modela i predstavlja standardni pristup u savremenim analizama zasnovanim na mašinskom učenju.

Kako bi se dodatno povećala pouzdanost procene performansi modela, primenjena je metoda k-struke unakrsne validacije (k-fold cross-validation) (Hastie et al., 2009). Ova metoda podrazumeva deljenje trening skupa na k podskupova približno jednake veličine. Tokom procesa validacije model se trenira na $k-1$ podskupova, dok se preostali podskup koristi za validaciju modela. Ovaj postupak se ponavlja k puta, pri čemu svaki podskup jednom služi kao validacioni skup. Konačna procena performansi modela dobija se kao prosečna vrednost evaluacionih metrika iz svih iteracija validacije. Primena ove metode omogućava stabilniju procenu performansi modela i smanjuje zavisnost rezultata od slučajne podele podataka.

OLS regresija ima dobro utemeljene principe evaluacije rezultata u cilju detekcije ekstremnih vrednosti. U tu svrhu korišćena su dva komplementarna dijagnostička alata. **Grafikon reziduala naspram predikcija („residuals vs. fitted plot“)** prikazuje rezidualne (razliku između opaženih i predviđenih vrednosti R_0) u funkciji predviđenih vrednosti modela. Idealan model pokazuje rezidualne tačke nasumično raspoređene oko horizontale na nuli, bez sistematskih obrazaca. Prisustvo LOESS krive — lokalno ponderisane regresione linije koja je robusna na ekstremne vrednosti i ne pretpostavlja unapred definisanu funkcionalnu formu — omogućava vizuelnu detekciju nelinearnih sistematskih devijacija koje standardna rezidualna analiza može propustiti. U ovoj analizi LOESS kriva korišćena je kao dijagnostički alat za identifikaciju opservacija čiji su reziduali sistematski udaljeni od nule (Cook, R. D., 1982). **Grafikon reziduali-uticaj („residual-leverage plot“)** kombinuje dve dimenzije dijagnostike: veličinu standardizovanog reziduala i polugu (leverage) — meru statističkog uticaja svake opservacije na procenu regresionih koeficijenata. Poluga kvantifikuje koliko je vrednost prediktora za datu opservaciju udaljena od centroida čitavog skupa prediktora u multidimenzionalnom prostoru; visoka poluga znači da opservacija potencijalno može neproporcionalno „povući“ regresionu ravan ka sebi. Opservacije sa istovremeno visokim rezidualima i visokom polugom su tzv. uticajne tačke — one ne samo da se slabo uklapaju u model, već i aktivno menjaju procenjene parametre. Sintetička mera koja kombinuje obe dimenzije je Kukova distanca (Cook's distance), definisana kao:

$$D_i = \frac{e_i^2}{p \cdot \text{MSE} \cdot (h_{ii} / (1 - h_{ii})^2)} \quad (13)$$

gde e_i označava rezidualni vektor, h_{ii} polugu i-te opservacije, p broj parametara modela, a MSE srednju kvadratnu grešku (Cook, R. D., 1982). Vrednosti $D_i > 4/n$ (gde je n veličina uzorka) konvencionalno se smatraju indikatorom značajnog uticaja. Na grafikonu su ucrtane i Kukove izolinije na vrednostima 0,5 i 1,0 kao vizuelni prag — opservacije koje prelaze ove linije zaslužuju posebnu pažnju pri interpretaciji rezultata. Identifikacija uticajnih tačaka nije automatski razlog za njihovo isključivanje iz analize: za ekološke studije sa heterogenim

globalnim uzorkom one često nose supstantivnu informaciju o zemljama čiji profil ne odgovara globalnom obrascu, te su same po sebi interpretativno vredne.

Pored validacije modela, značajan deo analize odnosio se na optimizaciju hiperparametara algoritama mašinskog učenja. Hiperparametri predstavljaju parametre modela koji se ne procenjuju direktno tokom procesa učenja, već se unapred definišu i značajno utiču na performanse modela. U slučaju ansambl metoda, kao što su Random Forest i Gradient Boosting modeli, hiperparametri mogu uključivati broj stabala u ansamblu, maksimalnu dubinu stabala, minimalni broj posmatranja potrebnih za grananje čvora, kao i parametre koji kontrolišu brzinu učenja modela.

Optimizacija hiperparametara sprovedena je sistematskim pretraživanjem prostora mogućih vrednosti parametara korišćenjem metoda kao što su grid search i random search. Ove metode omogućavaju evaluaciju velikog broja kombinacija hiperparametara kako bi se identifikovala konfiguracija modela koja daje najbolje performanse u okviru unakrsne validacije. Na ovaj način postiže se optimalan balans između složenosti modela i njegove sposobnosti generalizacije.

Performanse modela tokom procesa validacije procenjivane su korišćenjem standardnih evaluacionih metrika za regresione modele. Među najvažnijim metrima korišćenim u analizi nalaze se koeficijent determinacije (R^2), koji kvantifikuje udeo varijanse zavisne promenljive objašnjen modelom, kao i srednja kvadratna greška (MSE) i srednja apsolutna greška (MAE), koje predstavljaju mere odstupanja predviđenih vrednosti od stvarnih vrednosti ciljne promenljive.

Rezultati validacije i optimizacije hiperparametara omogućili su identifikaciju modela koji postižu optimalne performanse u predikciji osnovnog reprodukcionog broja. Ovakav metodološki pristup obezbeđuje veću pouzdanost dobijenih rezultata i omogućava robusnu analizu determinanti prenosivosti SARS-CoV-2 u različitim geografskim i socioekonomskim kontekstima.

3.3.3 Analiza značaja prediktora i interpretacija modela

Jedan od ključnih ciljeva primene metoda mašinskog učenja u epidemiološkim istraživanjima nije samo izgradnja modela sa visokom prediktivnom tačnošću, već i razumevanje relativnog doprinosa pojedinačnih promenljivih u objašnjenju varijabilnosti posmatranog fenomena. U okviru ovog istraživanja posebna pažnja posvećena je analizi **značaja prediktora (feature importance)** kako bi se identifikovali najvažniji faktori koji utiču na vrednost osnovnog reprodukcionog broja (R_0) i samim tim na intenzitet transmisije SARS-CoV-2 infekcije.

Analiza značaja prediktora sprovedena je korišćenjem metoda koje omogućavaju procenu relativnog doprinosa svake nezavisne promenljive u procesu predikcije. Kod modela zasnovanih na stablima odlučivanja, kao što su Random Forest i Gradient Boosting modeli, značaj promenljivih procenjuje se na osnovu njihovog doprinosa smanjenju greške modela tokom procesa grananja stabla. Promenljive koje dovode do većeg smanjenja greške ili doprinose većem poboljšanju kriterijuma razdvajanja čvorova smatraju se informativnijim i

imaju veći značaj u modelu. Kao alternativni pristup selekciji promenljivih korišćena je postupna selekcija zasnovana na Akaike informacionom kriterijumu (stepwise AIC). Ovaj metod iterativno dodaje ili uklanja prediktore iz regresionog modela, u svakom koraku birajući kombinaciju koja minimizuje AIC — informaciono-teorijsku meru koja balansira kvalitet fita modela (izražen kroz likelihood funkciju) sa penalizacijom za broj uključenih parametara, čime se smanjuje preprilagođavanje (Akaike, 1974). Za razliku od pristupa zasnovanih na stabilima odlučivanja, koji rangiraju sve prediktore po značaju, stepwise AIC procedura eksplicitno selektuje podskup promenljivih koje zajedno čine efikasan model sa maksimalnim fitom uz što manje promenljivih.

Primena ovih metoda omogućava ne samo identifikaciju najznačajnijih determinanti transmisije, već i analizu smerova njihovog uticaja. Na taj način moguće je utvrditi da li određeni faktori doprinose povećanju ili smanjenju procenjenog osnovnog reproduktionog broja, kao i u kojoj meri različite vrednosti prediktora menjaju intenzitet transmisije infekcije.

Rezultati analize značaja promenljivih predstavljaju važan element interpretacije modela, jer omogućavaju identifikaciju ključnih epidemioloških, klimatskih, demografskih i socioekonomskih determinanti koje oblikuju obrasce širenja respiratornih infekcija. Ovakav pristup omogućava prelazak sa isključivo prediktivnog modelovanja na **interpretabilno modelovanje**, koje ima poseban značaj u javnozdravstvenim istraživanjima gde je razumevanje uzročno-posledičnih odnosa između faktora i ishoda od ključnog značaja za planiranje preventivnih i kontrolnih mera.

Dobijeni rezultati analize značaja prediktora korišćeni su u daljim fazama istraživanja za interpretaciju modela i identifikaciju faktora koji imaju najveći doprinos prostornoj varijabilnosti prenosivosti SARS-CoV-2. Na taj način omogućeno je dublje razumevanje kompleksnih epidemioloških obrazaca koji karakterišu širenje respiratornih infekcija na globalnom nivou i formirana je osnova za diskusiju rezultata prikazanu u narednim poglavljima disertacije.

4 Globalna analiza pokretača transmisije SARS-CoV-2: Prediktivno modelovanje i kvantifikacija sezonskog forsiranja u 118 zemalja

4.1 Uvod

Početno pitanje za bilo koju pandemiju je koji faktori određuju ranu prenosivost virusa kada se posmatra celokupna geografska i sociodemografska raznolikost planete? Držanje analitičkog fokusa na globalnoj skali otvara niz novih naučnih mogućnosti, ali i metodoloških izazova. Na jednoj strani, varijabilnost ekoloških, klimatskih i sociodemografskih uslova između 118 zemalja obezbeđuje statistički snažan osnov za identifikaciju robusnih prediktora transmisije. Na drugoj strani, heterogenost sistema zdravstvene zaštite, razlike u politikama testiranja i izveštavanja, kao i prisustvo višestrukih konfundirajućih varijabli zahtevaju pažljiv pristup odabiru zavisne promenljive i metodologiji modelovanja.

Centralna prednost pristupa primenjenog u ovoj studiji leži u izboru zavisne promenljive. Umesto oslanjanja na grube epidemiološke ishode — broj slučajeva, hospitalizacije ili mortalitet — koji su snažno zavisni od lokalnih kapaciteta testiranja i zdravstvene infrastrukture, studija koristi osnovni reprodukcionni broj (R_0) procenjen iz inicijalne, nekontrolisane faze epidemijskog rasta. Ovaj pristup, detaljno opisan u Poglavlju 3, obezbeđuje meru koja odražava inherentni transmijski potencijal virusa u datim uslovima sredine, pre uvođenja bilo kakvih intervencija javnog zdravlja. Time se eliminišu najznačajnije sistematske pristrasnosti koje opterećuju komparativne međunacionalne studije.

Ciljevi ove studije su dvojaki. Prvo, razviti robusne modele mašinskog učenja sposobne da predvide R_0 na osnovu meteoroloških, ekoloških, sociodemografskih i zdravstvenih karakteristika svake zemlje. Drugo, primeniti Bajesijanske metode za kvantifikaciju sezonske dinamike transmisije, čime se obezbeđuje formalna statistička validacija sezonskog forsiranja kao sistemskog faktora globalne pandemijske dinamike.

4.2 Metodologija

4.2.1 Procena osnovnog reprodukcionog broja R_0

Metodološki okvir za procenu R_0 oslanja se na matematičke osnove opisane u Poglavlju 3.1 i objavljene u prethodnoj publikaciji autorskog tima (Salom et al., 2021). Ovaj postupak je primenjen na javno dostupne podatke o broju novodetektovanih slučajeva zaražavanja Sars-Cov-2 virusom za 118 zemalja, pri čemu je za svaku zemlju identifikovan vremenski interval eksponencijalnog rasta koji prethodi uvođenju značajnih epidemioloških mera. Vrednosti λ , izračunate su iz javno dostupnih podataka o broju slučajeva za svaku od 118 zemalja, a

dobijeni $R_{0,free}$ predstavlja procenu inherentne transmisibilnosti virusa pre primene bilo kakvih mera ublažavanja.

Izbor $R_{0,free}$ kao centralne zavisne promenljive nije slučajan već metodološki utemeljen. Alternativni epidemiološki ishodi koji su široko korišćeni u komparativnim međunacionalnim analizama — ukupni broj slučajeva, stope hospitalizacija ili mortalitet — podložni su sistemskim pristrasnostima koji proističu iz heterogenosti kapaciteta testiranja, različitih definicija slučajeva i nejednakog kvaliteta zdravstvene evidencije između zemalja. Procena R_0 iz inicijalne, nekontrolisane faze rasta epidemije minimizuje uticaj ovih konfundirajućih faktora jer odražava biološku i ekološku dimenziju transmisije pre nego što su administrativne i bihejvioralne intervencije izmenile dinamiku širenja virusa. Time se obezbeđuje komparabilnost između zemalja različitih zdravstvenih sistema i administrativnih kapaciteta, što je preduslov za validnu globalnu analizu prediktora transmisibilnosti.

4.2.2 Prikupljanje i harmonizacija podataka

Konstruisanje skupa podataka za globalnu analizu predstavljalo je značajan metodološki poduhvat. Finalni skup obuhvata 118 zemalja i 96 varijabli, grupisanih u pet tematskih domena: meteorološke podatke, sociodemografske podatke, zdravstvene indikatore Svetske zdravstvene organizacije, podatke o zagađenju vazduha i mere društvene povezanosti (Tabela 3). Integracija podataka iz tako heterogenih izvora zahtevala je pažljivu harmonizaciju i ujednačavanje metodologije prikupljanja kako bi se obezbedila konzistentnost merenja na nivou pojedinačnih zemalja.

4.2.2.1 Meteorološki podaci

Meteorološki podaci prikupljeni su automatizovanim Python skriptom koji je pristupao servisu NASA POWER (NASA Langley Research Center, 2020). Ovaj servis obezbeđuje dnevne vremenske serije podataka za niz meteoroloških parametara na globalnom nivou, što ga čini pogodnim izvorom za sistematsko prikupljanje klimatskih podataka u studijama koje obuhvataju veliki broj zemalja.

Kako bi se konstruisala jedinstvena reprezentativna vrednost za svaku zemlju, primenjen je pristup ponderisanja prema populaciji urbanih centara. Konkretno, za svaku zemlju izračunata je ponderisana srednja vrednost dnevnih meteoroloških merenja iz glavnih urbanih centara, pri čemu je broj stanovnika pojedinog centra korišćen kao ponder. Ovaj pristup metodološki je superioran u odnosu na jednostavno geografsko usrednjavanje jer uzima u obzir činjenicu da je većina međuljudskih kontakata koncentrisana u gusto naseljenim urbanim sredinama, koje stoga u većoj meri određuju stvarnu izloženost stanovništva datim klimatskim uslovima i, posledično, dinamiku transmisije respiratornih patogena.

Geografske informacije neophodne za implementaciju ovog postupka — koordinate urbanih centara i veličina njihovih populacija — preuzete su iz baze GHS STAT UCDB2015MT GLOBE R2019A V1.0, publikovane od strane Zajedničkog istraživačkog centra Evropske komisije (JRC) (European Commission Global Human Settlement, 2020). Ova baza podataka obezbeđuje standardizovane i međunarodno uporedive informacije o urbanim centrima na globalnom nivou, što je bio preduslov za konzistentnu primenu metodologije ponderisanja u svim analiziranim zemljama.

U finalnom skupu podataka zastupljeno je ukupno osam temperaturnih promenljivih koje međusobno visoko koreliraju.

Tabela 1: Pregled imena i značenja temperaturnih promenljivih u konačnom skupu

Ime temperaturne promenljive	Objašnjenje značenja promenljive
T2M	temperatura vazduha na 2m (srednja)
T2M_MAX	maksimalna temperatura na 2m
T2M_MIN	minimalna temperatura na 2m
T10M	temperatura vazduha na 10m (srednja)
T10M_MAX	maksimalna temperatura na 10m
T10M_MIN	minimalna temperatura na 10m
TS	temperatura površine zemljišta (skin)
temperature	WAQI jedinstvena prijavljena temperatura

Zbog problema multikolinearnosti koji ove promenljive unose u regresione modele, primenjena je analiza glavnih komponenti (PCA) na ovaj podskup varijabli. Prve dve glavne komponente zadržane su kao zamena za originalnih osam promenljivih, pri čemu ove dve komponente zajedno objašnjavaju 98,2% ukupne varijanse temperaturnih merenja. Na taj način očuvana je gotovo sva informacija sadržana u originalnim temperaturnim podacima, uz istovremeno eliminisanje multikolinearnosti koja bi ugrozila interpretabilnost i stabilnost regresionih koeficijenata.

Podaci o ultravioletnom zračenju prikupljeni su putem servisa OpenUV, koji obezbeđuje informacije o maksimalnom dnevnom UV indeksu (OpenUV, 2020). Za svaku od analiziranih zemalja konstruisana je reprezentativna vrednost primenom iste metodologije ponderisanja populacijom urbanih centara koja je korišćena za meteorološke i ostale prostorno distribuirane promenljive, čime je obezbeđena metodološka konzistentnost unutar skupa podataka. U finalnom skupu promenljivih ova kategorija je zastupljena promenljivom OpenUV_max, koja se pokazala kao jedan od robusnih prediktora u modelima zasnovanim na Random Forest-u.

Uključivanje UV zračenja u skup potencijalnih prediktora transmisibilnosti SARS-CoV-2 počiva na dva biološki utemeljena mehanizma. Prvi se odnosi na direktnu viricidnu aktivnost ultravioletnog zračenja — poznato je da UV zračenje inaktivira respiratorne viruse u aerosolima i na površinama smanjenjem infektivnosti virusnih čestica (Schuit et al., 2020), čime se potencijalno skraćuje efektivno vreme transmisije u spoljašnjim sredinama. Drugi

mehanizam deluje posredno, putem sinteze vitamina D u koži pod uticajem UV-B zračenja. Vitamin D ima dokumentovanu ulogu u modulaciji urođenog i stečenog imunološkog odgovora (Martineau et al., 2017), a njegova deficijencija, koja je učestalija u populacijama sa manjom izloženošću sunčevom zračenju, povezana je sa povećanom susceptibilnošću prema respiratornim infekcijama.

Sa geografskog stanovišta, UV zračenje u značajnoj meri kovarira sa geografskom širinom i sezonskim ciklusom, što ga čini relevantnom promenljivom u kontekstu globalne analize koja obuhvata zemlje iz obe hemisfere i različitih klimatskih zona. Ova kovarijacija sa geografskim i sezonskim faktorima istovremeno znači da UV zračenje može funkcionisati i kao proksi za širi kompleks klimatskih uslova koji moduliraju transmisiju respiratornih patogena, a ne isključivo kao nezavisan prediktor sa direktnim uzročnim delovanjem na dinamiku virusa.

4.2.2.2 Sociodemografski podaci

Sociodemografski i geografski podaci prikupljeni su pretežno iz baze Global Human Settlement — Urban Centre Database (GHS-UCDB R2019A), koju je publikovao Zajednički istraživački centar Evropske komisije (JRC) (European Commission Global Human Settlement, 2020). Ova baza podataka obezbeđuje standardizovane informacije za više od 10.000 urbanih centara širom sveta, što je čini jednim od najsveobuhvatnijih i metodološki najkonzistentnijih izvora sociodemografskih podataka na globalnom nivou.

Za svaku od 118 analiziranih zemalja iz ove baze su ekstrahovani ključni indikatori na nivou gradova. Među najznačajnijim promenljivima izdvajaju se površina izgrađenog gradskog prostora po stanovniku u 2015. godini (BUCA15), koja služi kao proksi za gustinu urbane izgradnje i stepen urbanizacije, te bruto domaći proizvod na nivou grada u 2015. godini (GDP15), koji odražava ekonomski razvoj i životni standard urbane populacije. Obe promenljive su merene na nivou individualnih urbanih centara, a zatim agregirane na nivo zemlje primenom iste metodologije ponderisanja populacijom koja je korišćena za meteorološke podatke, čime je obezbeđena metodološka konzistentnost između različitih kategorija podataka.

Geografske koordinate i veličina populacije urbanih centara, takođe preuzete iz GHS-UCDB R2019A baze, bile su neophodne ne samo za izračunavanje ponderisanih proseka već i kao osnova za sve prostorne analize u okviru studije. Važno je napomenuti da su gradske populacije u ovoj bazi definisane na osnovu ujednačene metodologije koja ne zavisi od različitih nacionalnih definicija gradskih područja, što je posebno značajno u kontekstu međunarodnih komparativnih studija gde nekonzistentnost definicija može biti izvor sistemske pristrasnosti.

Pored podataka iz GHS-UCDB R2019A, u kategoriju sociodemografskih promenljivih uključene su i dodatne varijable koje reflektuju ekonomski status na nacionalnom nivou, kao što je paritet kupovne moći po stanovniku (PPP per capita, paritet kupovne moći po glavi stanovnika). Ova promenljiva pojavljuje se kao značajan prediktor u pojedinim modelima

selekcije karakteristika, što ukazuje na to da opšti ekonomski razvoj zemlje, pored specifičnih karakteristika urbane sredine, doprinosi objašnjenju varijabilnosti u vrednostima R_0 između zemalja.

4.2.2.3 WHO podaci o zdravlju populacije

Zdravstveni indikatori na nivou populacije preuzeti su iz Global Health Observatory (GHO) baze podataka Svetske zdravstvene organizacije (WHO). GHO predstavlja centralni repozitorijum zdravstvenih statistika WHO i obezbeđuje standardizovane podatke o zdravstvenom stanju populacija za veliki broj zemalja sveta, što ga čini pogodnim izvorom za komparativne međunacionalne studije.

Odabrani su oni indikatori koji su bili dostupni za značajan broj zemalja u periodu od tri godine pre početka pandemije, čime je obezbeđena dovoljna pokrivenost uzorka za statističku analizu. Kriterijum dostupnosti podataka bio je ključan u procesu selekcije jer je jedna od centralnih metodoloških prepreka u globalnim epidemiološkim studijama upravo nejednaka pokrivenost zemalja različitim zdravstvenim indikatorima, koja može uvesti sistematsku pristrasnost ukoliko se ne uzme u obzir pri interpretaciji rezultata. Finalni skup podataka sadržao je ukupno 25 GHO indikatora koji pokrivaju različite dimenzije zdravlja populacije (Tabela 3 – sekcija 3).

Među indikatorima koji su se pokazali kao značajni prediktori u modelima selekcije karakteristika posebno se izdvajaju prevalenca gojaznosti, nivoi holesterola u krvi, konzumacija alkohola i mortalitet dece mlađe od pet godina. Prisustvo gojaznosti i povišenog holesterola u skupu relevantnih prediktora konzistentno je sa literaturom koja ukazuje na to da metabolički komorbiditet modulira imunološki odgovor i time potencijalno utiče na dinamiku transmisije na nivou populacije. Konzumacija alkohola, koja se pojavljuje kao jedan od najrobusnijih prediktora u više različitih skupova karakteristika dobijenih različitim metodama selekcije, verovatno služi kao proksi (engl. *proxy* - *zamenik*) za određene obrasce društvenog ponašanja i međuljudskih kontakata koji pogoduju transmisiji respiratornih patogena. Mortalitet dece mlađe od pet godina, kao opšti indikator kvaliteta zdravstvenog sistema i životnog standarda, reflektuje širi socioekonomski i zdravstveni kontekst koji posredno utiče na ranjivost populacije prema infektivnim bolestima.

Važno je naglasiti da su svi zdravstveni indikatori mereni pre početka pandemije i da stoga ne predstavljaju posledicu već potencijalni preduslov različite transmisibilnosti virusa u različitim populacijama. Ovaj temporalni raspored merenja obezbeđuje kauzalnu usmerenost analize i isključuje mogućnost reverzne kauzalnosti kao izvora konfundiranja.

4.2.2.4 Podaci o zagađenju vazduha

Podaci o zagađenju vazduha prikupljeni su iz dva komplementarna izvora, pri čemu je potreba za kombinovanjem izvora proistekla iz fundamentalnog ograničenja primarnog

skupa podataka — neravnomerne geografske distribucije stanica za merenje kvaliteta vazduha na globalnom nivou.

Kao primarni izvor korišćena je baza World Air Quality Index (WAQI), koja sadrži istorijske dnevne podatke o koncentracijama 10 vrsta zagađivača od početka pandemije za 500 najvećih gradova na svetu. Metodološki, ovaj izvor je konzistentan sa pristupom primenjenim za meteorološke podatke, budući da obezbeđuje merenja na nivou pojedinih gradova koja su zatim mogla biti agregirana na nivo zemlje primenom iste metodologije ponderisanja populacijom. Međutim, usled neravnomerne geografske distribucije mernih stanica, koja odražava nejednaku razvijenost infrastrukture za monitoring kvaliteta vazduha u različitim delovima sveta, većina zemalja u uzorku nije bila zastupljena u ovoj bazi. Posledica toga je da su promenljive dobijene iz WAQI baze bile izuzetno retke, sa nedostajućim vrednostima za 70 do 90 zemalja od ukupno 118 analiziranih.

Uprkos visokom stepenu nepotpunosti podataka, mnoge od ovih promenljivih zagađenja vazduha ipak su pokazale statističku značajnost u modelima, što snažno ukazuje na informacionu vrednost ovih merenja i istovremeno naglašava potrebu za proširenjem pokrivenosti podataka u budućim istraživanjima.

Kao dopuna retkim vrednostima iz WAQI baze, korišćeni su podaci iz baze Emissions Database for Global Atmospheric Research (EDGAR), koja obezbeđuje vrednosti ukupne mase emitovanih zagađivača na godišnjem nivou za svaku zemlju. Iz ove baze preuzete su vrednosti za ukupno 9 vrsta zagađivača, pri čemu je svaka vrsta zastupljena u dva oblika — kao apsolutna vrednost i kao vrednost skalirana brojem stanovnika — što daje ukupno 18 promenljivih u finalnom skupu podataka. Važno je, međutim, naglasiti metodološko ograničenje ovih promenljivih: ni apsolutna vrednost ukupnih emisija ni njena populaciono skalirana varijanta ne odražavaju tačno stvarnu ličnu izloženost zagađivačima, koja zavisi od prostorne distribucije izvora emisija, meteoroloških uslova disperzije i individualnih obrazaca kretanja i boravka. U tom smislu, merenja iz WAQI baze, koja su zasnovana na direktnom uzorkovanju vazduha u urbanim sredinama, smatraju se informativnijim izvorom, dok EDGAR podaci služe kao gruba aproksimacija u slučajevima kada direktna merenja nisu dostupna.

Kombinovanjem ova dva izvora konstruisan je skup promenljivih zagađenja vazduha koji, uprkos navedenim ograničenjima, pokriva sve analizirane zemlje i obezbeđuje dovoljan stepen varijabilnosti za statističku analizu prediktivne vrednosti izloženosti zagađivačima u kontekstu transmisije SARS-CoV-2.

4.2.2.5 Mere društvene povezanosti

Mere društvene povezanosti između zemalja operacionalizovane su putem Indeksa društvene povezanosti (Social Connectedness Index — SCI) koji je objavio Facebook (Bailey et al., 2018). SCI kvantifikuje stepen povezanosti između geografskih jedinica mereno putem prijateljstava na Facebook platformi, i predstavlja jedan od retkih dostupnih izvora podataka koji direktno odražava intenzitet međunarodnih društvenih veza na globalnom nivou.

Osnova za konstruisanje promenljivih bila je matrica SCI vrednosti na nivou parova zemalja, koja opisuje relativnu gustinu Facebook prijateljstava između svake moguće kombinacije dve zemlje u uzorku.

Iz ove matrice parova zemalja konstruisana je mrežna reprezentacija globalne društvene povezanosti, pri čemu svaka zemlja odgovara čvoru u grafu, a vrednost SCI između dve zemlje određuje težinu odgovarajuće ivice. Promenljiva za svaku pojedinačnu zemlju dobijena je primenom mera centralnosti na ovaj graf, pri čemu centralnost služi kao proksi za epidemiološku važnost određene zemlje u globalnoj mreži društvenih kontakata. Osnovna pretpostavka ovog pristupa je da zemlja koja je intenzivnije povezana sa većim brojem drugih zemalja ima veći potencijal za uvoz i izvoz infektivnih slučajeva, te da njena pozicija u globalnoj mreži doprinosi objašnjenju inicijalne transmisibilnosti virusa na njenom teritoriju.

Budući da različite mere centralnosti mogu davati međusobno veoma različite rezultate i reflektovati različite aspekte mrežne pozicije čvora, odabrane su dve komplementarne mere za svaku zemlju. Prva je jednostavna centralnost stepena, definisana kao zbir vrednosti svih ivica koje polaze iz datog čvora, i odražava ukupan intenzitet međunarodnih društvenih veza određene zemlje. Druga je centralnost protoka posredništva, koja je metodološki složenija mera i kvantifikuje u kojoj meri određena zemlja posreduje u tokovima između ostalih zemalja u mreži. Ova druga mera posebno je relevantna sa epidemiološkog stanovišta jer identifikuje zemlje koje predstavljaju ključne čvorišta u globalnom širenju infektivnih bolesti, bez obzira na ukupan obim njihovih međunarodnih veza.

Uključivanje mera društvene povezanosti u analizu metodološki je opravdano činjenicom da inicijalna faza pandemije u svakoj zemlji nije bila određena isključivo lokalnim ekološkim i sociodemografskim uslovima, već i intenzitetom međunarodnih putovanja i društvenih veza koje su omogućile uvoz virusa i uspostavile inicijalni lanac transmisije. U tom smislu, SCI-zasnovane promenljive dopunjuju ostale kategorije podataka dodavanjem dimenzije globalne međupovezanosti koja je specifična za pandemijsku dinamiku respiratornih bolesti.

4.2.3 Pretprocesiranje podataka

Pretprocesiranje podataka sprovedeno je nakon podele uzorka na skup za treniranje i skup za testiranje, čime je izbegnuto implicitno curenje informacija iz test skupa u procedure transformacije i imputacije. Iste transformacije koje su primenjene na skup za treniranje primenjene su zatim i na skup za testiranje, što je preduslov za validnu procenu generalizabilnosti modela na novim podacima.

Prvi korak u pretprocesiranju bila je redukcija promenljivih zasnovana na kriterijumu nepotpunosti podataka. Promenljive za koje je nedostajalo 100 ili više vrednosti (od ukupnih 118) isključene su iz daljnje analize, budući da bi njihovo zadržavanje unelo više šuma nego informacije u modele. Za preostale promenljive koje su sadržale manji broj nedostajućih vrednosti primenjena je imputacija medijanom kolone, čime su nedostajuće vrednosti

zamenjene centralnom tendencijom distribucije odgovarajuće promenljive. Izbor medijane umesto srednje vrednosti kao mere centralne tendencije za imputaciju metodološki je opravdan činjenicom da je medijana robusna prema ekstremnim vrednostima, koje su česte u sociodemografskim i zdravstvenim promenljivama u globalnim skupovima podataka.

Nakon imputacije, primenjena je automatizovana procedura transformacije promenljivih čiji je cilj bio minimizacija asimetričnosti raspodela. Funkcija koja je sprovedena u tu svrhu sistematski je testirala nekoliko kandidat transformacija za svaku promenljivu — kvadratni koren, kubni koren, logaritamsku transformaciju i inverznu transformaciju, uključujući i identičnu transformaciju kao referentnu opciju — i birala onu koja je rezultovala najmanjom apsolutnom asimetrijom raspodele. Ovaj pristup je primenjen uniformno na sve promenljive, čime je eliminisana potreba za subjektivnom procenom odgovarajuće transformacije za svaku individualnu promenljivu. Normalizacija oblika raspodela bila je posebno važna za one algoritme u skupu modela koji su osetljivi na devijacije od normalnosti, kao što su metode zasnovane na linearnoj regresiji.

U finalnom koraku pretprocesiranja sve promenljive su standardizovane — centrirane oduzimanjem srednje vrednosti i skalirane deljenjem sa standardnom devijacijom. Standardizacija je neophodna kako bi se obezbedila numerička stabilnost i komparabilnost koeficijenata u modelima koji su osetljivi na razlike u skalama merenja između promenljivih, kao što su Ridge regresija, Lasso regresija i metoda potpornih vektora. Bez ovog koraka, promenljive merene u velikim numeričkim rasponima neproporcionalno bi dominirale u procesima optimizacije ovih algoritama, neovisno od njihove stvarne prediktivne vrednosti.

4.2.4 Prediktivno modelovanje

Za prediktivno modelovanje vrednosti R_0 odabran je skup od devet algoritama koji pokrivaju širok spektar metodoloških pristupa, od parametarskih linearnih metoda do nelinearnih ansambl tehnika. Konkretno, primenjeni su: regresija najmanjih kvadrata (OLS), robusna regresija (RoR), regresija glavnih komponenti (PCR), Ridge regresija (regresija grebenom), Lasso regresija (laso regresija), Elastic Net (EN), Random Forest, regresija potpornih vektora (SVR) i gradijentno pojačavanje (XGBoost). Ovako širok izbor algoritama bio je metodološki nameran — različite tehnike imaju različite pretpostavke o obliku veze između prediktora i ishoda, različitu osetljivost na multikolinearnost i različitu sklonost ka prilagođavanju šumu u podacima, te njihova komparativna evaluacija obezbeđuje robusniju sliku prediktivnog potencijala skupa podataka nego što bi to bio slučaj sa primenom jednog ili malog broja algoritama.

Podešavanje hiperparametara sprovedeno je pretragom mreže vrednosti (grid search), pri čemu je za metode sa većim brojem podesivih hiperparametara pretraga vršena ili simultano (ElasticNet, Random Forest, regresija potpornih vektora) ili sekvencijalno (XGBoost). Podela podataka na skupove za treniranje, validaciju i testiranje sprovedena je u razmeri 60-20-20, a čitav postupak je ponovljen 100 puta sa različitim nasumičnim podelama uzorka. Ovaj pristup višestrukog ponavljanja bio je neophodan kako bi se dobile stabilne procene

performansi modela koje nisu artefakt jedne konkretne podele podataka, s obzirom na relativno mali ukupni uzorak od 118 zemalja. Računarski zahtevi ovako dizajniranog eksperimenta bili su značajni — kompletna procedura zahtevala je više od 2.500 procesorskih sati na namenskom računarskom klasteru. Performanse modela na skupovima za treniranje i testiranje evaluirane su korenom srednje kvadratne greške predikcije (RMSEP), skaliranim standardnom devijacijom čitavog skupa podataka radi lakše interpretacije i poređenja između skupova prediktora. Odabir modela i seta prediktora je vršen ne samo po vrednosti greške, već i po razlici između trening i test seta kao mere preprilagođavanja, te mogućnosti da kasnije proizvede godišnje sezonske predikcije. Za poslednju stavku preduslov je uključenost vremenskih promenljivih.

4.2.5 Bajesijanska karakterizacija sezonske dinamike

Kao dopuna primarnim prediktivnim modelima, sprovedena je Bajesijanska analiza (Gelman et al., 2013) u cilju formalne statističke karakterizacije sezonskih obrazaca emergentnih iz predikcija modela mašinskog učenja. Osnovna motivacija za ovu sekundarnu analizu bila je potreba da se potvrdi da godišnji obrasci R_0 generisani modelom nisu artefakti procedure mašinskog učenja, već da odražavaju stvarnu sezonsku dinamiku transmisije SARS-CoV-2 konzistentnu sa poznatim sezonskim pokretačima respiratornih bolesti.

Tehnički okvir analize zasnivao se na proceduri Markovljevog lanca Monte Karlo (MCMC) koja je korišćena za fitovanje prediktovanih dnevnih vrednosti R_0 za svaku zemlju sinusoidnom funkcijom oblika:

$$R(t) \sim \text{Normal}(R_{base} + R_{base} \times A_{SF} \times \cos(\frac{2\pi(t - t_{min})}{365}), \sigma) \quad (14)$$

gde R_{base} predstavlja bazalni godišnji reprodukcionni broj oko kojeg se odvija sezonska oscilacija, A_{SF} relativnu amplitudu sezonskog forsiranja, t_{min} dan u godini u kome temperatura ima minimalnu tj. R_0 dostiže maksimalnu vrednost, a σ rezidualnu standardnu devijaciju koja obuhvata dnevnu varijabilnost neobjašnjenu sezonskim modelom.

Model je implementiran u programskom jeziku R korišćenjem Stan okruženja za Hamiltonijansko Monte Karlo (HMC) semplovanje. Specificirani su sledeće apriorne raspodele:

- $A_{SF} \sim \text{Normal}(0,35, 0,05)$, informativna apriorna distribucija za relativnu amplitudu sezonskog forsiranja izvedena iz preliminarne procena;
- $t_{min} \sim \text{Normal}(t_{minT}, 30)$, gde t_{minT} predstavlja dan minimalne temperature za datu zemlju, čime se kodira apriorno očekivanje da se pik transmisije vremenski poklapa sa najhladnijim periodom;
- $\sigma \sim \text{LogNormal}(0, 1)$, široka distribucija za rezidualnu varijabilnost;
- $R_{base} \sim \text{Normal}(R_0, 0,35 \times R_0)$, centriran na frekventističku procenu R_0 za svaku zemlju sa koeficijentom varijacije od 0,35.

Sempler je pokrenut sa 4 nezavisna lanca od po 2000 iteracija (1000 za zagrevanje), čime je generisano 4000 posteriornih uzoraka po zemlji.

Za svaku analiziranu zemlju, MCMC procedura je procenila posteriorne raspodele četiri ključna parametra. Za zemlje kod kojih procenjeni minimum pada u blizinu granice kalendarske godine (pre 60. ili posle 300. dana), vremenski niz je pomeren za 183 dana pre fitovanja, a apriorna distribucija t_{min} je analogno korigovana, kako bi se izbegle diskontinuiteti u kosinusnoj funkciji.

Kako bi se olakšala interpretacija rezultata između hemisfera, vremenska promenljiva je transformisana tako da fazni pomeraj funkcije t_{min} direktno odgovara danu u godini u kome je temperatura najniža a reprodukcioni broj najviši. Pored Bajesijanske estimacije za odabrane zemlje, sprovedena je i komplementarna frekventistička analiza na celokupnom uzorku od 118 zemalja. Za svaku zemlju, prediktovana dnevna kriva R_0 fitovana je nelinearnim najmanji kvadratima na kosinusni model oblika $R_0(t) = R_{0_median} \times (1 + A_{SF} \times \cos(2\pi(t - t_{peak})/365))$, gde je R_{0_median} uzoračka medijana dnevnih predikcija, a A_{SF} i t_{peak} procenjene amplitude sezonskog forsiranja i vremena pika, respektivno. Korišćena su dva seta inicijalnih uslova kako bi se izbegla konvergencija ka lokalnim minimumima. Kvalitet fita je ocenjen merom $1 - RSS/TSS$, gde RSS označava rezidualnu sumu kvadrata a TSS totalnu sumu kvadrata. Dobijene estimacije parametara na nivou zemalja potom su analizirane u regresijama prema apsolutnoj geografskoj širini, a tajming pika R_0 je upoređen sa nezavisno izmerenim datumima minimuma UV zračenja i minimuma temperature za svaku zemlju.

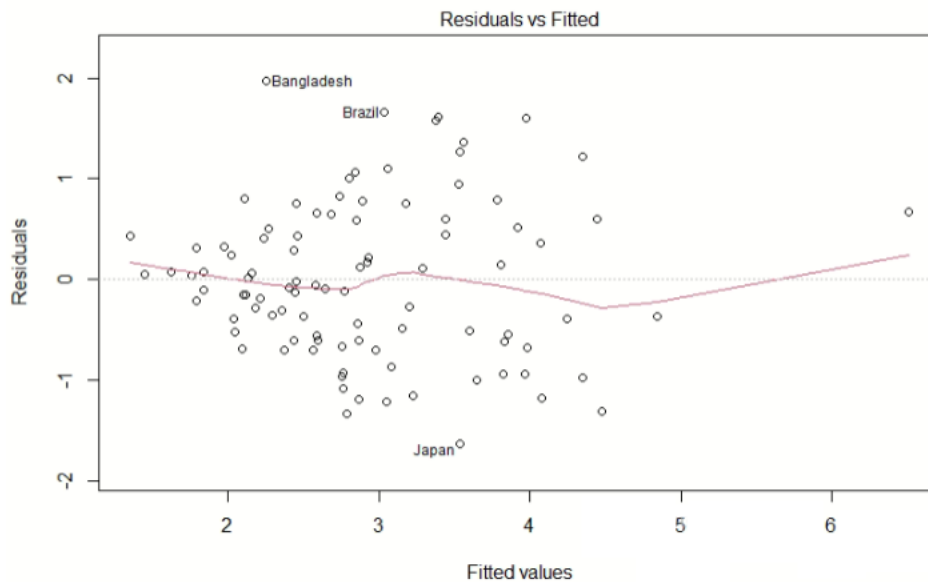
4.3 Rezultati

4.3.1 Karakteristike finalnog skupa podataka i inicijalna eksploracija

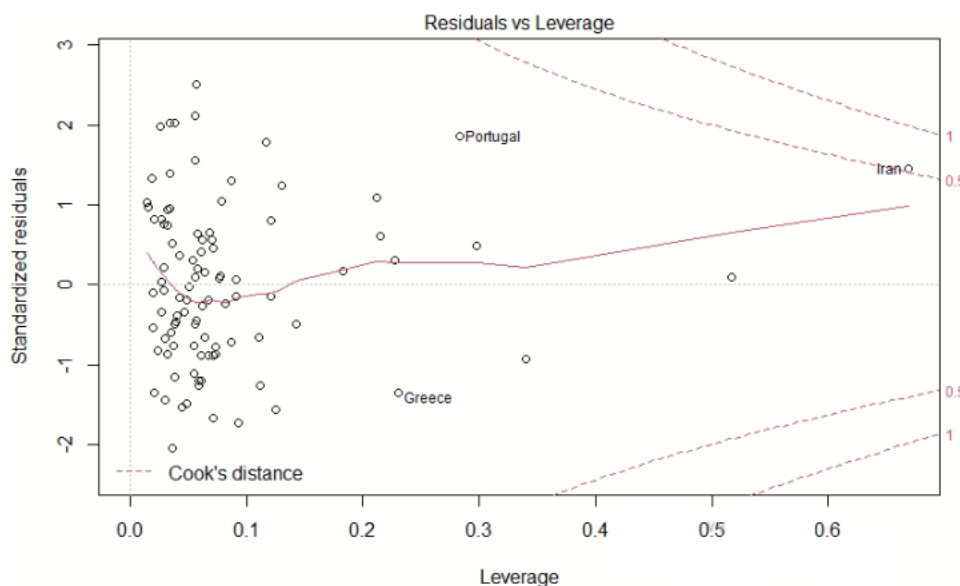
Finalni skup podataka obuhvata 118 zemalja i 96 varijabli koje pokrivaju pet tematskih domena: meteorološke podatke, sociodemografske podatke, zdravstvene indikatore WHO, podatke o zagađenju vazduha i mere društvene povezanosti. Osam temperaturnih varijabli, koje su međusobno visoko korelirale, redukovano je primenom analize glavnih komponenti na dve glavne komponente koje zajedno objašnjavaju 98,2% ukupne varijanse temperaturnih merenja, čime je problem multikolinearnosti u ovom podskupu promenljivih rešen uz minimalan gubitak informacije.

Inicijalna eksploracija skupa podataka sprovedena je metodama linearne regresije pre primene složenijih tehnika mašinskog učenja. Primarni cilj ove faze bio je identifikacija eventualnih ekstremnih vrednosti koje bi mogle neproporcionalno uticati na rezultate modelovanja i čije bi prisustvo u skupu za treniranje moglo narušiti stabilnost i generalizabilnost finalnih modela. U tu svrhu korišćena su dva komplementarna dijagnostička alata. Grafikon reziduala u odnosu na prilagođene vrednosti (Residuals vs. Fitted plot, Ilustracija 1) identifikovao je tri ekstremne vrednosti: Bangladeš i Brazil, čije su

stvarne vrednosti R_0 bile znatno više od vrednosti predviđenih modelom, te Japan, koji je pokazivao upečatljivo negativne rezidualne vrednosti. Grafikon reziduali-uticaj (Residual-Leverage plot, Ilustracija 2), koji istovremeno prikazuje veličinu reziduala i statistički uticaj pojedinih opservacija, dodatno je identifikovao Iran kao uticajnu ekstremnu vrednost u pozitivnom pravcu, sa potencijalom da neproporcionalno utiče na procenu regresionih koeficijenata. U oba slučaja trend je procenjen LOESS linijom i ne pokazuje upečatljive fenomene koji su dalje istraživani.



Ilustracija 1: Grafikon reziduala naspram predviđanja regresionog modela – uočene ekstremne vrednosti (Bangladeš, Brazil, Japan). Izlomljena crvena linija predstavlja LOESS procenu trenda koja nije indikovala fenomene koji zahtevaju dalje ispitivanje.



Ilustracija 2: Grafikon reziduala i uticaja – Iran kao značajna ekstremna vrednost u pozitivnom smeru. Kukova distanca od 0.5 i 1 kao uobičajene granice uticajnih opservacija data je isprekidanom crvenom linijom. LOESS trend kao na prošloj slici.

Prisustvo ovih ekstremnih vrednosti metodološki je zanimljivo i zaslužuje posebnu pažnju pri interpretaciji rezultata. Bangladeš i Brazil, sa vrednostima R_0 višim od očekivanih na osnovu svojih ekoloških i sociodemografskih karakteristika, mogu reflektovati specifičnosti u dinamici inicijalne faze epidemije u ovim zemljama koje nisu adekvatno zahvaćene raspoloživim skupom prediktora — na primer, specifičnu gustinu naseljenosti megagrađova, neformalne stambene uslove ili specifične obrasce kretanja i međuljudskih kontakata. Japan, sa markantno negativnim rezidualima, odnosno sa stvarnim vrednostima R_0 znatno nižim od predviđenih, potencijalno odražava specifične kulturološke obrasce ponašanja japanske populacije koji su mogli suprimirati transmisiju nezavisno od meteoroloških i sociodemografskih uslova — poput rasprostranjene prakse nošenja maski i pre uvođenja formalnih mera, visoke discipline u poštovanju higijenskih normi i specifičnih obrazaca društvenih interakcija. Iran kao statistički značajna opservacija sa visokim uticajem može delom odražavati i specifičnosti u ranom izveštavanju o broju slučajeva koje su uticale na procenu R_0 iz inicijalne faze epidemije.

Identifikacija ovih ekstremnih vrednosti u inicijalnoj eksplorativnoj fazi bila je vredna i kao metodološka provera konzistentnosti podataka i kao osnova za interpretaciju ograničenja finalnih modela. Ona ukazuje na to da, uprkos sveobuhvatnosti konstruisanog skupa prediktora, postoje dimenzije ranjivosti i transmisibilnosti koje nisu u potpunosti zahvaćene raspoloživim varijablama, i koje su specifične za pojedine zemlje ili regione.

4.3.2 Preliminarna globalna korelaciona analiza i selekcija obeležja

Primarna obrada i validacija podataka je obavljena u dva rada koji postavljaju temelje za kompletnu analizu prikazanu u ovom poglavlju.

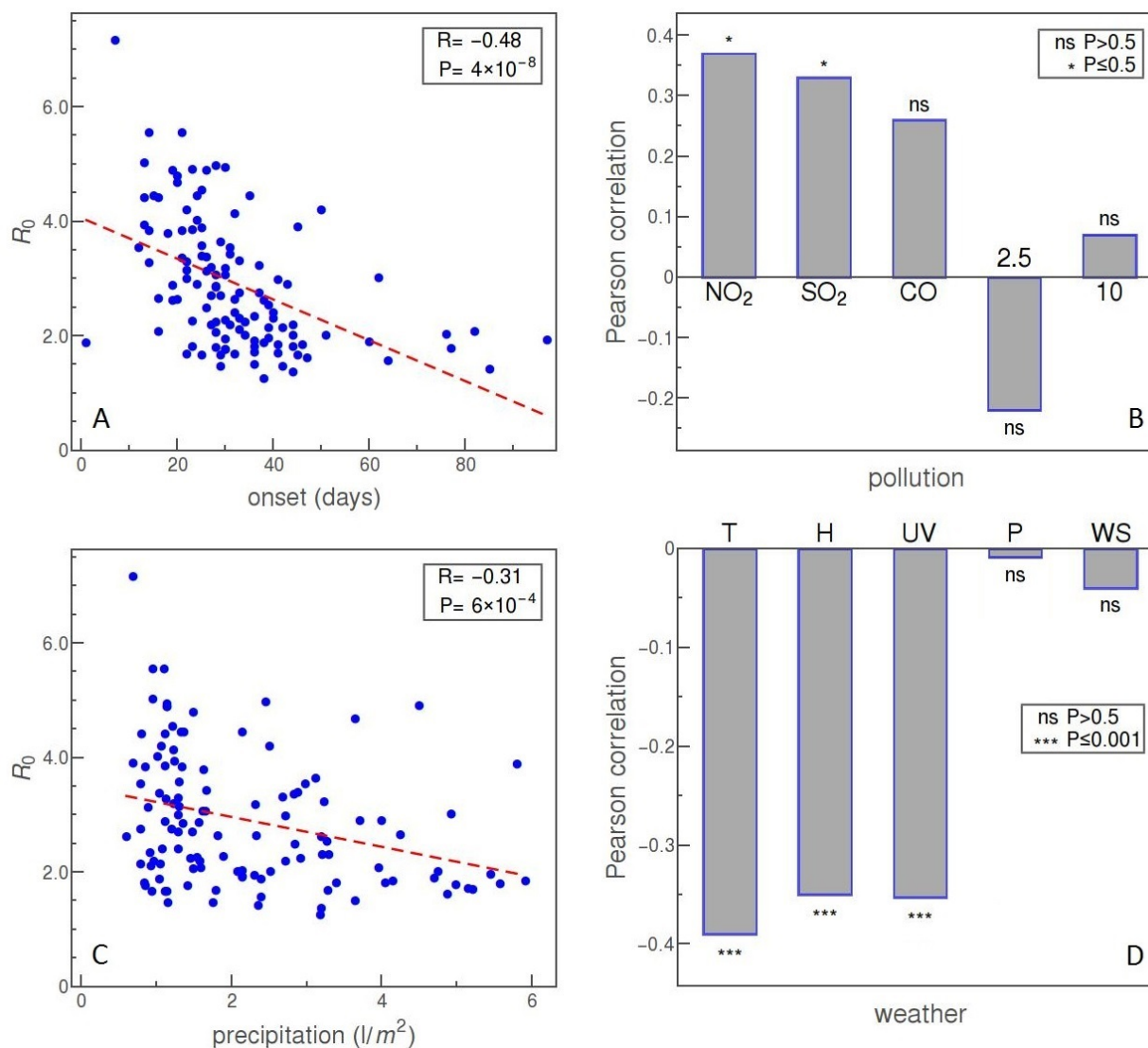
4.3.2.1 *Efekti demografskih i klimatskih parametara na COVID-19 osnovni reprodukcioni broj*

Prvi rad (Salom et al., 2021) postavlja metodološki okvir za ekstrakciju osnovnog reprodukcionog broja R_0 iz rane eksponencijalne faze epidemije u 118 zemalja. Tri važne komponente su testirane i povezane sa ishodom od interesa.

- **Meteorološki faktori:** Potvrđena je značajna negativna korelacija transmisije sa temperaturom, vlažnošću vazduha i UV zračenjem. Pronađena je i negativna korelacija sa padavinama, dok brzina vetra i vazdušni pritisak nisu pokazali uticaj na širenje virusa. Značajni faktori zagađenja su bili NO_2 i SO_2 , dok ostali zagađivači nisu pokazali statistički značajnu korelaciju sa R_0 (Ilustracija 3, B-D).
- **Sociodemografski faktori:** Utvrđena je snažna pozitivna korelacija sa indeksom humanog razvoja (HDI) i procentom urbane populacije, što ukazuje na to da su se razvijenija i povezaniya društva suočila sa bržim početnim širenjem. Zanimljivo je da gustina naseljenosti sama po sebi nije pokazala značajnu korelaciju na nivou država.

- **Zdravstveni faktori:** Transmisija je snažno pozitivno korelisana sa nivoom holesterola u krvi i konzumacijom alkohola. Analiza krvnih grupa pokazala je tendenciju ka većoj transmisiji kod Rh- negativnih fenotipova.

Meteorološke komponente koje su pokazale korelaciju sa R_0 su upravo one koje poznato imaju veliku sezonalnu varijaciju. Posebno interesantan rezultat je bila značajna umerena negativna korelacija između R_0 i vremena početka pandemije („onset“) (Ilustracija 3, Panel A). Zbog visoke korelacije između sociodemografskih indikatora razvoja i onset promenljive, nije bilo moguće tvrditi da je primećeni efekat nezavisan.

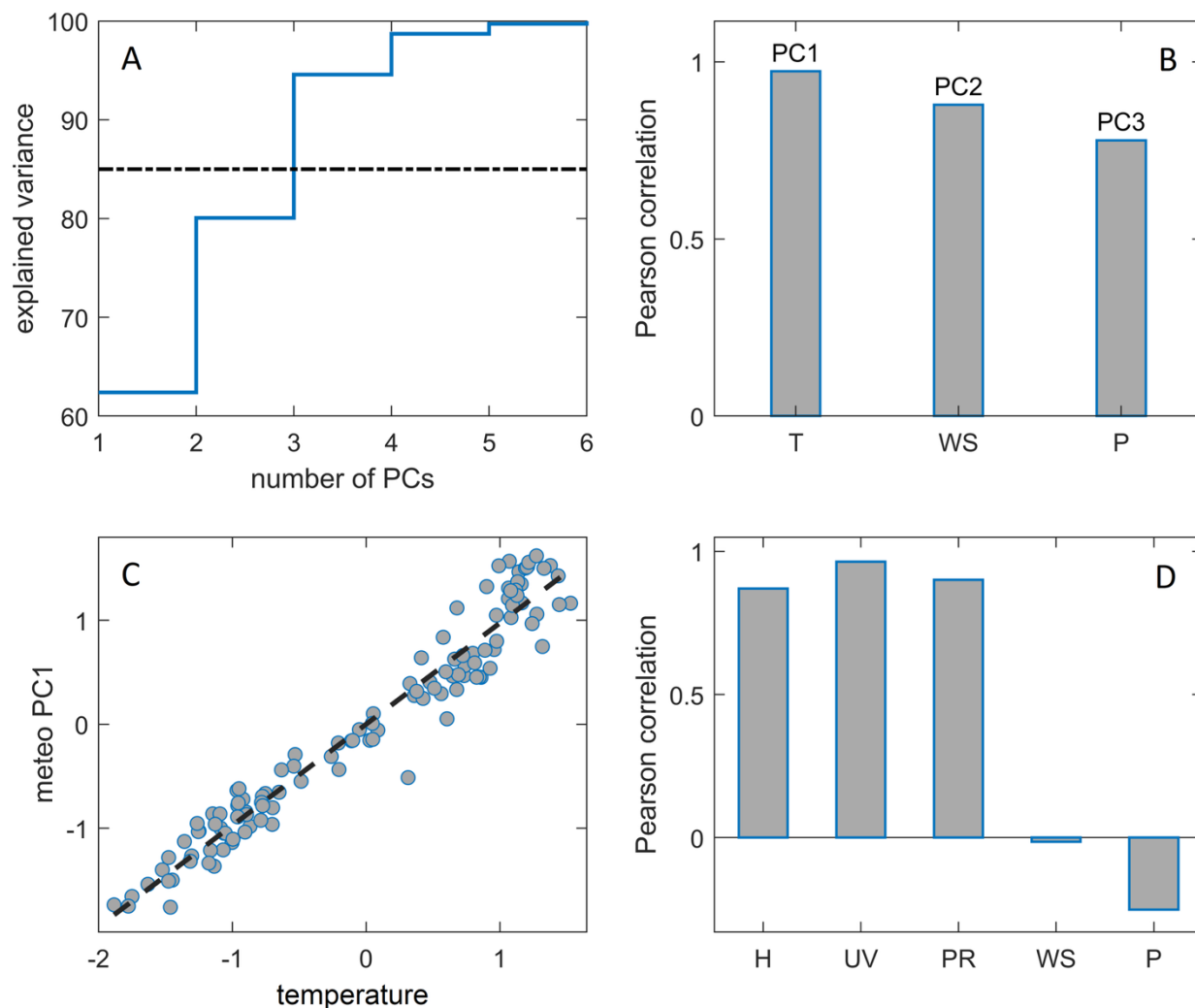


Ilustracija 3: A) Dijagram rasturanja kašnjenja početka epidemije i R_0 pokazuje značajnu umerenu korelaciju B) Pearsonova korelacija R_0 sa polutantima gde NO_2 i SO_2 pokazuju statistički značajnu pozitivnu korelaciju sa R_0 C) Dijagram rasturanja R_0 i padavina kao značajno korelisano prediktora D) Pearsonova korelacija R_0 sa ostalim meteorološkim parametrima (sa leva na desno): T—temperatura, H—specifična vlažnost, UV—indeks ultraljubičastog zračenja, P—vazdušni pritisak, WS—brzina vetra. Sezonalno varirajući meteorološki parametri su bili statistički značajni (T, H, UV).

4.3.2.2 *Inferencija glavnih vodilja transmisibilnosti putem metoda selekcije obeležja*

Drugi rad (Djordjevic et al., 2021) ide korak dalje koristeći PCA i tehnike selekcije obeležja (Lasso i Elastic Net) kako bi se identifikovali nezavisni i najvažniji prediktori R_0 . Ulazne varijable su prirodno podeljene na demografske i meteorološke podatke i PCA je nezavisno primenjena na obe grupe.

- **Primarni prediktor:** Bogatstvo i nivo razvoja zemlje (BDP po glavi stanovnika ili HDI) identifikovani su kao najvažniji globalni prediktori širenja SARS-CoV-2, verovatno kao proksi varijabla za ukupnu frekvenciju kontakata u populaciji.
- **Modulatori efekta:** Utvrđeno je da izgrađena površina po glavi stanovnika (BUAPC) usporava širenje virusa, što se interpretira kao veća dostupnost zatvorenog prostora i manja pretrpanost.
- **Stil života i okruženje:** Nezdravi životni faktori, uključujući preteranu telesnu težinu, pušenje, fizičku neaktivnost i dugotrajnu izloženost zagađenju vazduha ($PM_{2.5}$), identifikovani su kao značajni faktori koji pospešuju transmisiju.
- **Meteorološki nalazi:** Interesantno je da je najznačajnija meteorološka glavna komponenta PC1 bila u velikoj korelaciji sa temperaturom, dok su PC2 i PC3 bile u velikoj korelaciji sa brzinom vetra i vazдушnim pritiskom respektivno (Ilustracija 4). Pošto je sama temperatura visoko korelisana sa padavinama, vlažnosti vazduha i UV zračenjem, PC1 je prepoznata kao marker sezonalnosti. Lasso i Elastic Net su potvrdili uticaj sezonalnosti klimatskih uslova na R_0 nezavisno od jakih sociodemografskih signala.



Ilustracija 4: Analiza glavnih komponenti za meteorološke podatke. (a) Kumulativna objašnjena varijansa. Prve tri komponente imaju kumulativnu varijansu od 95% (b) Varijable koje su u najboljoj korelaciji sa meteorološkim glavnim komponentama (PC). Prva glavna meteorološka komponenta je skoro savršeno korelisana sa temperaturom. (c) Dijagram rasturanja meteorološke PC1 naspram temperature u standardizovanim vrednostima. (d) Korelacija meteoroloških varijabli sa temperaturom. Temperatura, vlažnost vazduha, padavine i UV zračenje su visoko korelisani i predstavljaju sezonalnost. T—temperatura, WS—brzina vetra, P—vazdušni pritisak, H—specifična vlažnost, UV—indeks ultraljubičastog zračenja, PR—padavine.

Glavne demografske komponente dobijene PCA analizom grupišu originalne varijable u nezavisne faktore koji opisuju različite aspekte posmatranih zemalja. Prva demografska komponenta (PC1) dominantno odražava ukupni društveni i ekonomski prosperitet zemlje, s obzirom na to da pokazuje izuzetno visoku korelaciju sa Indeksom ljudskog razvoja (HDI) i BDP-om po glavi stanovnika, ali i sa očekivanim životnim vekom, prosečnom starošću i procentom urbanog stanovništva. Pored nje, izdvajaju se komponente koje nose specifične, odvojene uticaje na transmisiju: PC4 obuhvata indikatore nezdravog načina života i uslova sredine (gojaznost, fizička neaktivnost, pušenje i zagađenje vazduha), dok PC7 reflektuje dostupnost zatvorenog prostora po stanovniku kroz varijablu izgrađene površine. Konačno, komponenta PC9 je gotovo isključivo određena varijablom vremena početka epidemije (onset).

Prethodne analize su pokazale da je vreme početka pandemije u značajnoj korelaciji sa HDI-jem, usled činjenice da su razvijenije i globalno povezanije zemlje prve bile pogođene pandemijom. Ipak, primena PCA u ovoj studiji nam je omogućila da delimično dekomponujemo odnos ove dve varijable. Izdvajanjem onset-a u zasebnu komponentu (PC9), nezavisnu od komponentne prosperiteta (PC1), potvrđeno je da vreme početka epidemije ima svoj nezavistan uticaj na usporavanje transmisije virusa. Ovaj nezavisni doprinos se najverovatnije može objasniti psihološkim efektom: zemlje u kojima je epidemija počela kasnije, imale su priliku da duže prate vesti o širenju bolesti u svetu. To je dovelo do spontane promene uobičajenog ponašanja stanovništva i povećane lične opreznosti, što je uticalo na smanjenje R_0 vrednosti čak i pre formalnog uvođenja strogih državnih mera intervencije.

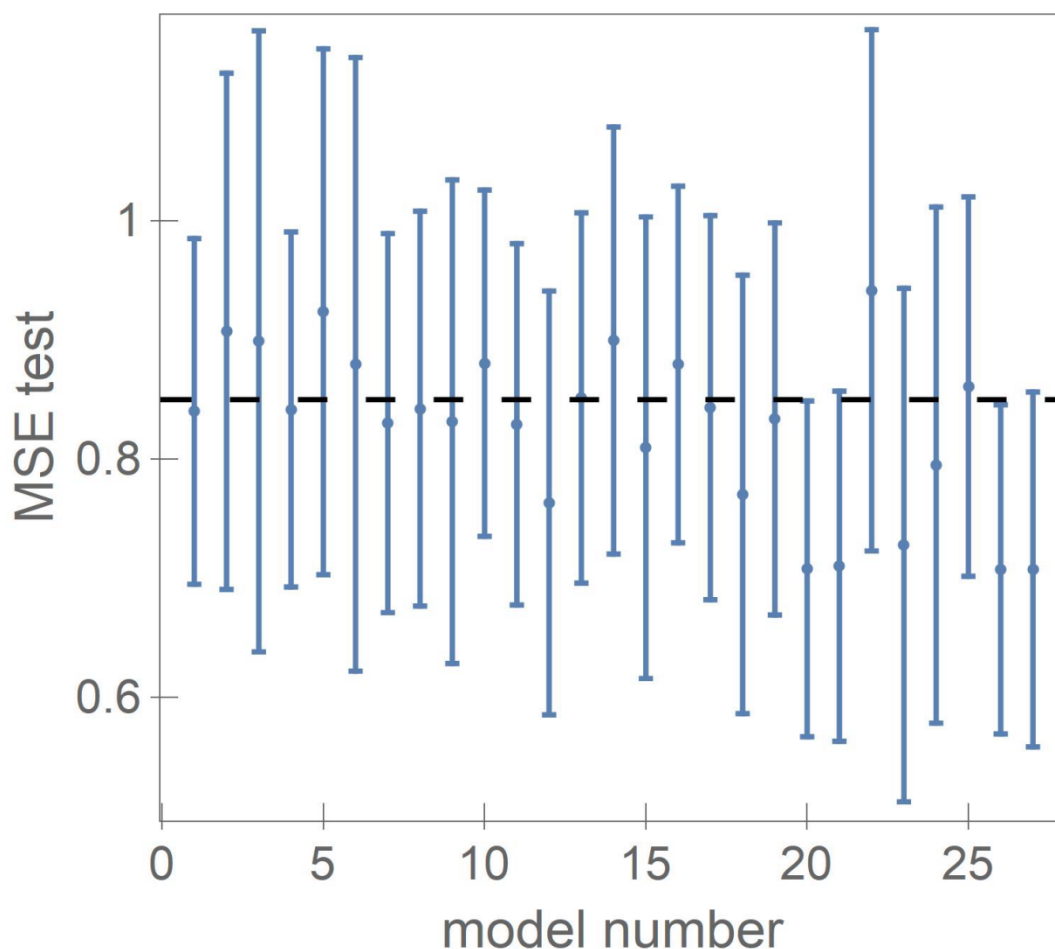
Navedeni radovi su motivisali iscrpnu prediktivnu analizu baziranu na uočenim korelacijama i izabranim obeležjima. Uočeni obrasci su u narednim poglavljima dodatno istraženi a njihova prediktivna moć kvantifikovana.

4.3.3 Performanse modela mašinskog učenja

U fazi selekcije obeležja primenjene su tri nezavisne metode koje su generisale šest različitih skupova prediktora (

Tabela 4). SAIC skup sadrži sedam varijabli odabranih postupnom selekcijom zasnovanom na Akaike informacionom kriterijumu (stepwise AIC), primenjenom na PCA-redukovani skup promenljivih. Lasso1 i Lasso2 skupovi sadrže varijable odabrane Lasso regularizacijom, pri čemu je Lasso1 primenjen na potpun skup originalnih varijabli (uključujući sirove temperaturne metrike), a Lasso2 na PCA-redukovani skup u kome su temperaturne varijable zamenjene glavnim komponentama. RF1 i RF2 skupovi sadrže varijable identifikovane permutacijskom važnošću Random Forest modela, analogno primenjenom na potpun (RF1) i PCA-redukovani (RF2) skup varijabli. Konačno, Union skup konstruisan je objedinjavanjem svih jedinstvenih varijabli koje su identifikovane kao značajne barem jednom od tri metode selekcije, čime je formiran najsveobuhvatniji skup od 21 prediktora.

Devet algoritama testirano je na šest skupova prediktora — SAIC, Lasso1, Lasso2, RF1, RF2 i Union — pri čemu su performanse evaluirane korenom srednje kvadratne greške predikcije (RMSEP) normalizovanim standardnom devijacijom celokupnog skupa podataka, što omogućava direktno poređenje između skupova prediktora različitih dimenzionalnosti i skala. Evaluirano je ukupno 54 kombinacije (9 algoritama $\times$ 6 skupova prediktora). Eliminisan su one sa test-RMSEP koji je konzistentno prelazio 1,0 ili sa izrazitom razlikom između trening i test greške koja ukazuje na neprihvatljivo preprilagođavanje. Preostalih 27 kombinacija prikazano je na Slici 5 zajedno sa intervalima poverenja dobijenim iz 100 ponovljenih podela skupa podataka. Isprekidana horizontalna linija označava unapred prag prihvatljivosti ispod 1,0 jednak minimalnoj polovini intervala poverenja: 17 kombinacija je ispod praga, dok je kod njih 12 medijana test-MSE i 95% interval poverenja konzistentno ispod 1,0. Ovih 12 modela zadržano je kao finalni skup (Ilustracija 5).



Ilustracija 5: Test-MSE za 27 kombinacija algoritam-skup prediktora sa intervalima poverenja iz 100 ponovljenih podela. Isprekidana linija označava prag prihvatljivosti; 17 kombinacija je ispod praga, a 12 njih čiji intervali poverenja ne seku 1,0 čine finalni skup

Linearne metode pokazale su umerene i međusobno relativno slične performanse. OLS regresija ostvarila je test-RMSEP vrednosti u rasponu od 1,05 do 1,44 u zavisnosti od skupa prediktora, pri čemu su najlošije performanse zabeležene na Lasso1 i Lasso2 skupovima, a nešto bolje na SAIC skupu. Robusna regresija pokazala je gotovo identične performanse kao OLS, što sugeriše da ekstremne vrednosti identifikovane u inicijalnoj eksploraciji nisu imale dramatičan destabilizujući efekat na linearne modele. Regresija glavnih komponenti (PCR) ostvarila je test-RMSEP u rasponu od 1,04 do 1,33, sa najboljim performansama na RF1 skupu prediktora. Regularizacija linearne tehnike — Ridge regresija i ElasticNet — konzistentno su nadjačale OLS regresiju, posebno na skupovima sa većim brojem prediktora, što je konzistentno sa teorijskim očekivanjem da regularizacija poboljšava generalizabilnost u uslovima kada je broj prediktora relativno velik u odnosu na veličinu uzorka. Ridge regresija ostvarila je test-RMSEP vrednosti u rasponu od 0,78 do 1,32, a ElasticNet od 0,78 do 1,39, pri čemu su obe metode postigle svoje najbolje performanse na Union skupu prediktora.

Nelinearne metode ostvarile su znatno bolje performanse od linearnih tehnika, što ukazuje na prisustvo nelinearnih odnosa između prediktora i transmisibilnosti virusa koji linearni

modeli nisu bili u stanju da adekvatno modeluju. Random forest je ostvario test-RMSEP vrednosti u rasponu od 0,70 do 0,91 u zavisnosti od skupa prediktora, sa najboljim performansama od 0,70 na SAIC skupu. XGBoost je ostvario test-RMSEP vrednosti u rasponu od 0,74 do 1,02, sa najboljim performansama od 0,74 na Lasso1 skupu i 0,77 na Union skupu. Regresija potpornih vektora pokazala je test-RMSEP vrednosti u rasponu od 0,88 do 1,25, pozicionirajući se između linearnih i ansambl metoda po prediktivnoj tačnosti.

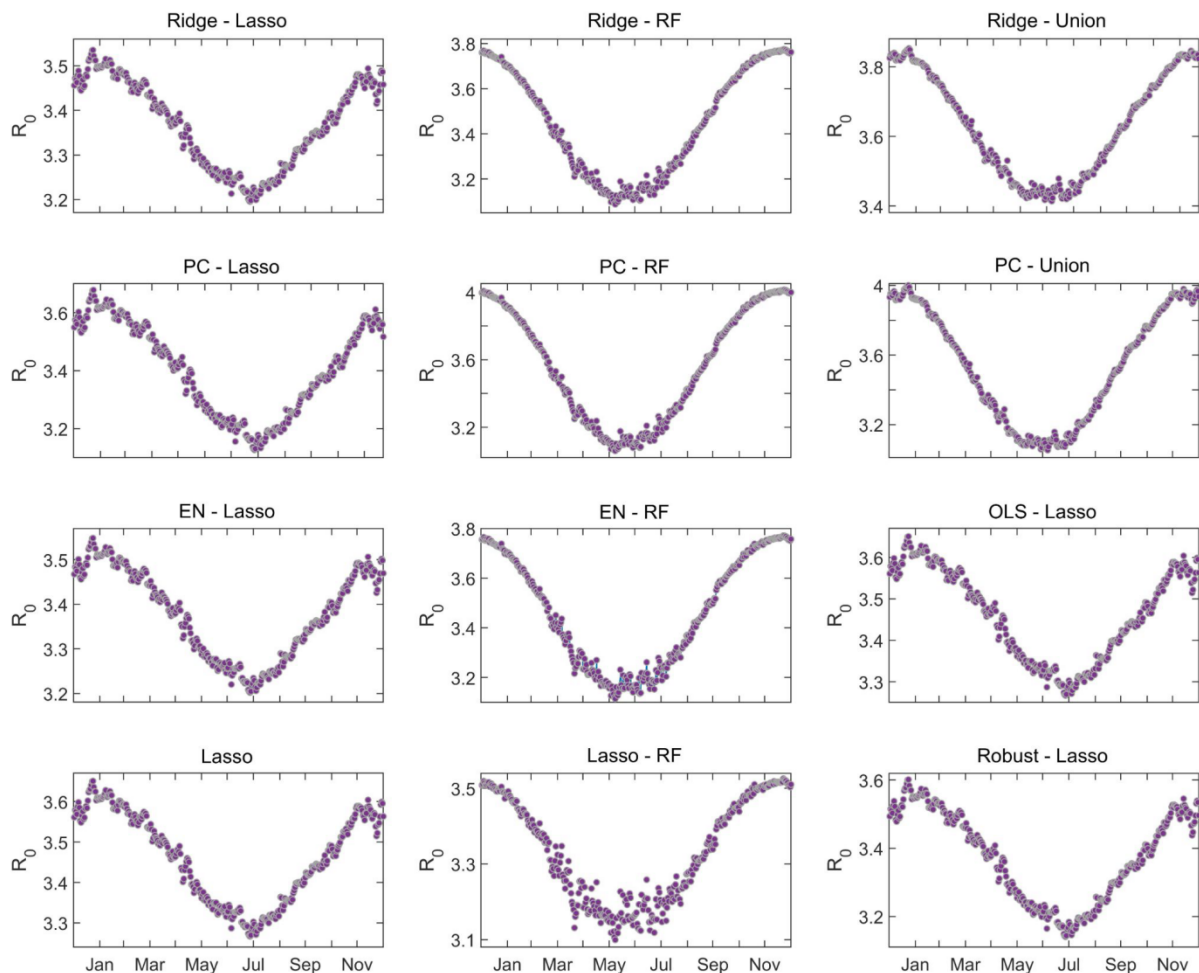
Važan metodološki aspekt evaluacije bila je analiza razlike između greške na skupu za treniranje i greške na skupu za testiranje kao indikatora stepena prilagođavanja šumu u podacima. Ansambl metode, posebno Random Forest, pokazale su izraženu tendenciju ka preprilagođavanju, sa trening-RMSEP vrednostima znatno nižim od test-RMSEP vrednosti — na primer, Random Forest na RF1 skupu ostvarila je trening-RMSEP od 0,35 nasuprot test-RMSEP od 0,91, što ukazuje na značajno preprilagođavanje. Ova diskrepanca bila je manji problem kod linearnih regularizacionih tehnika, što je bio jedan od faktora koji su uticali na konačni odabir modela.

Nakon evaluacije svih kombinacija algoritama i skupova prediktora, kao najrobusnija kombinacija identifikovana je Ridge regresija primenjena na objedinjeni Union skup prediktora, sa test-RMSEP od 0,78. Važno je napomenuti da ovo nije apsolutno najniža vrednost u čitavoj matrici rezultata — Random Forest na SAIC skupu ostvaruje test-RMSEP od 0,70, a XGBoost na Lasso1 skupu 0,74. Međutim, izbor finalnog modela bio je motivisan specifičnim zahtevom zadatka: generisanjem stabilnih godišnjih krivih R_0 za 365 dana na osnovu modela treniranog na podacima iz samo dva meseca. Za ovaj zadatak ključni su bili kriterijumi koje ansambl metode ne zadovoljavaju u istoj meri: (a) niska sklonost preprilagođavanju, merena razlikom trening i test RMSEP (0,72 vs 0,78 za Ridge/Union, nasuprot 0,26 vs 0,83 za RF/Union), što ukazuje na bolju generalizabilnost; (b) uniformna greška predikcije tokom čitave godine — Lasso skup prediktora pokazivao je povećanu dnevnu varijabilnost predikcija u hladnijim sezonama, skup prediktora Random Forest-a u toplijim sezonama, dok Union skup pokazuje ujednačeno ponašanje tokom čitave godine; i (c) sveobuhvatnost Union skupa koji integriše nalaze svih metoda selekcije, čime se smanjuje rizik od propuštanja relevantnog prediktora. Kombinacija ovih svojstava učinila je Ridge regresiju na Union skupu najprikladnijim izborom za ekstrapolaciju predikcija R_0 na čitavu kalendarsku godinu 2021., uprkos tome što nije ostvarila apsolutno najnižu tačkastu test grešku.

4.3.4 Godišnji obrasci predviđenog R_0 i hemisferska zavisnost

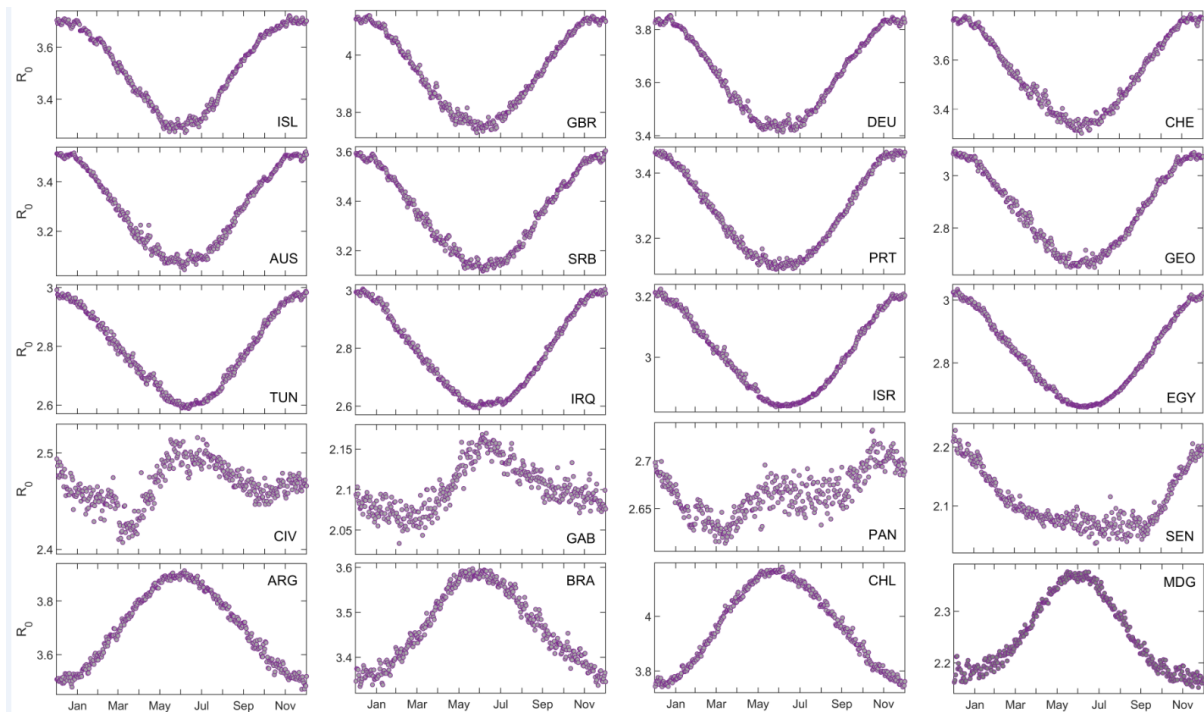
Godišnje predikcije generisane su tako što su socioekonomski prediktori za svaku zemlju zadržani na vrednostima iz perioda procene R_0 , dok su meteorološki prediktori zamenjeni stvarnim dnevnim vrednostima za svaki dan 2021. godine. Na taj način dobijene su dnevne procene R_0 za svih 12 finalnih modela. Analiza predikcija na primeru Nemačke pružila je dodatni uvid u razlike između skupova prediktora u pogledu njihovog ponašanja tokom čitave godine (Ilustracija 6). Svi algoritmi pokazali su sezonsku varijaciju u sličnom intervalu sa nižim vrednostima prema sredini godine, što potvrđuje robusnost sezonskog signala

zahvaćenog modelom. Međutim, uočene su karakteristične razlike u dnevnoj varijabilnosti između skupova prediktora — Lasso podskup karakteristika pokazivao je povećanu dnevnu varijabilnost u hladnijim sezonama sa višim vrednostima, dok je podskup karakteristika Random Foresta pokazivao povećanu dnevnu varijabilnost tokom toplijih sezona. Union skup prediktora, nasuprot tome, pokazao je uniformno ponašanje tokom čitave godine, odražavajući smanjenu varijaciju oko izgladene linije predikcije, što je bila jedna od ključnih kvalitativnih prednosti koja je opredelila njegov konačni odabir za generisanje godišnjih prognoza.



Ilustracija 6: Primer sezonske varijacije predviđenog R_0 za Nemačku – poređenje algoritama po skupovima prediktora (Lasso, RF, Union)

Primenom odabranog finalnog modela — Ridge regresije na Union skupu prediktora — generisane su dnevne predikcije R_0 tokom čitave 2021. godine za uzorak od 20 zemalja odabranih tako da pokrivaju širok opseg geografskih širina i obeju hemisfera (Ilustracija 7). Odabir ovih 20 zemalja bio je metodološki motivisan potrebom da se ispita da li model, treniran isključivo na podacima iz februara i marta 2020. godine koji ne sadrže direktan sezonski signal, uspešno reprodukuje poznate sezonske obrasce transmisije respiratornih virusa kada se primeni na čitav kalendarski ciklus.



Ilustracija 7: Dnevne predikcije R_0 tokom 2021. godine za 20 zemalja raspoređenih po geografskoj širini – severna, ekvatorijalna i južna hemisfera (Ridge regresija, Union skup prediktora)

Dobijeni godišnji obrasci pokazali su izrazitu hemisfersku zavisnost koja je u potpunosti konzistentna sa etabliranim nalazima o sezonskoj dinamici respiratornih patogena. Zemlje severne hemisfere, zastupljene u gornja tri reda slike 7, reprodukovale su opšti obrazac uočen na primeru Nemačke (Ilustracija 6) - prediktovane vrednosti R_0 bile su niže tokom proleća i leta, a više tokom jeseni i zime, odražavajući povoljniji epidemiološki kontekst toplijih meseci karakterizovan višim temperaturama, intenzivnijim UV zračenjem i smanjenom gustinom unutrašnjih prostora u kojima se odvija većina društvenih interakcija. Zemlje južne hemisfere, zastupljene u donjem redu vizualizacije, pokazale su invertovanu godišnju dinamiku u odnosu na početak kalendarske godine — maksimalne vrednosti R_0 nastupale su sredinom godine, što odgovara zimi u južnoj hemisferi, dok su minimalne vrednosti zabeležene na početku i kraju kalendarske godine, koji odgovaraju letnjim mesecima južne hemisfere.

Emergencija hemisferski zavisnih sezonskih obrazaca iz modela treniranog isključivo na prostornim varijacijama između zemalja u ranoj fazi pandemije predstavlja metodološki posebno vredan nalaz. Činjenica da model nije bio izložen nikakvim vremenskim sezonskim podacima tokom treniranja, a ipak reprodukuje temporalne sezonske obrasce konzistentne sa poznatom biologijom respiratornih virusa, snažno ukazuje na to da su meteorološke i ekološke promenljive uključene u model uspešno enkodovala sezonsku komponentu transmisibilnosti kroz svoje prostorne varijacije između zemalja različitih klimatskih zona. Ovaj nalaz istovremeno predstavlja indirektan argument za validnost čitavog analitičkog

pristupa i pruža osnov za formalnu statističku karakterizaciju sezonskog forsiranja koja je sprovedena u narednoj fazi analize.

Zemlje u blizini ekvatora, zastupljene u četvrtom redu vizualizacije, pokazale su kvalitativno drugačiji obrazac od zemalja viših geografskih širina. Umesto jasno izraženog jednogodišnjeg sinusoidnog obrasca karakterističnog za umerene klimatske zone, ekvatorijalne zemlje pokazuju multimodalne krive sa varijabilnim oblikom, što je konzistentno sa činjenicom da ove zemlje nemaju izražene godišnje temperaturne cikluse već potencijalno dva kraća kišna i sušna perioda koji različito moduliraju uslove transmisije. Posebno je metodološki značajno zapažanje da je godišnji opseg vrednosti R_0 markantno manji za ekvatorijalne zemlje nego za zemlje viših geografskih širina severne i južne hemisfere, što sugerise da je sezonska modulacija transmisibilnosti utoliko izraženija ukoliko su sezonske klimatske oscilacije intenzivnije.

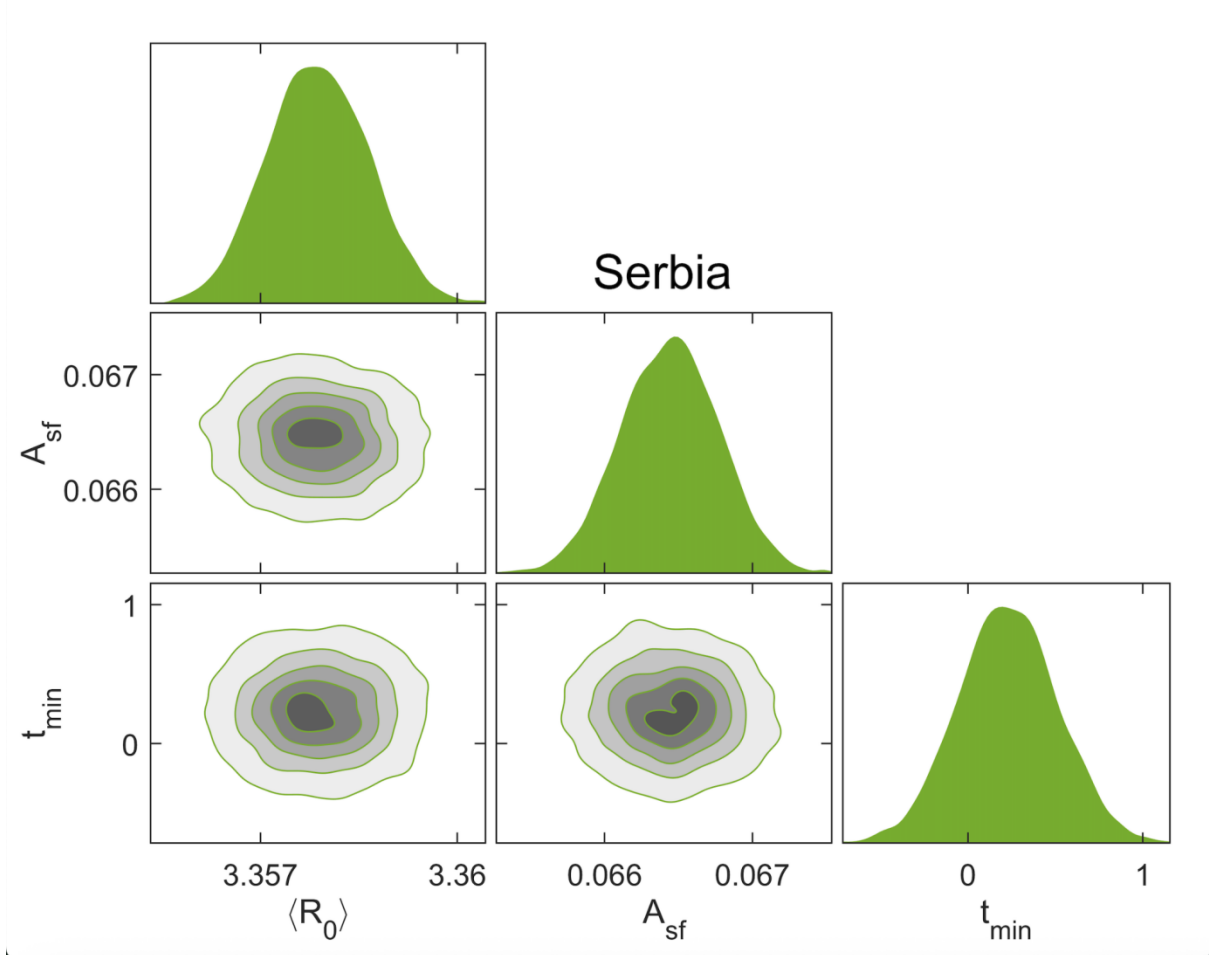
4.3.5 Bajesijanska karakterizacija sezonskog forsiranja

Emergencija hemisferski zavisnih sezonskih obrazaca u predikcijama modela mašinskog učenja zahtevala je formalnu statističku karakterizaciju kako bi se potvrdilo da uočeni obrasci nisu artefakti procedure modelovanja već odražavaju stvarnu sezonsku dinamiku transmisije SARS-CoV-2. U tu svrhu primenjena je Bajesijanska analiza zasnovana na proceduri Markovljevog lanca Monte Karlo (MCMC), kojom su fitovane prediktovane dnevne vrednosti R_0 sinusoidnoj funkciji koja matematički reprezentuje sezonsko forsiranje. Model je implementiran u programskom jeziku R korišćenjem Stan okruženja za probablističko modelovanje, koje obezbeđuje punu posteriornu distribuciju parametara umesto tačkastih procena, čime je omogućena kvantifikacija neizvesnosti u procenama sezonskih parametara.

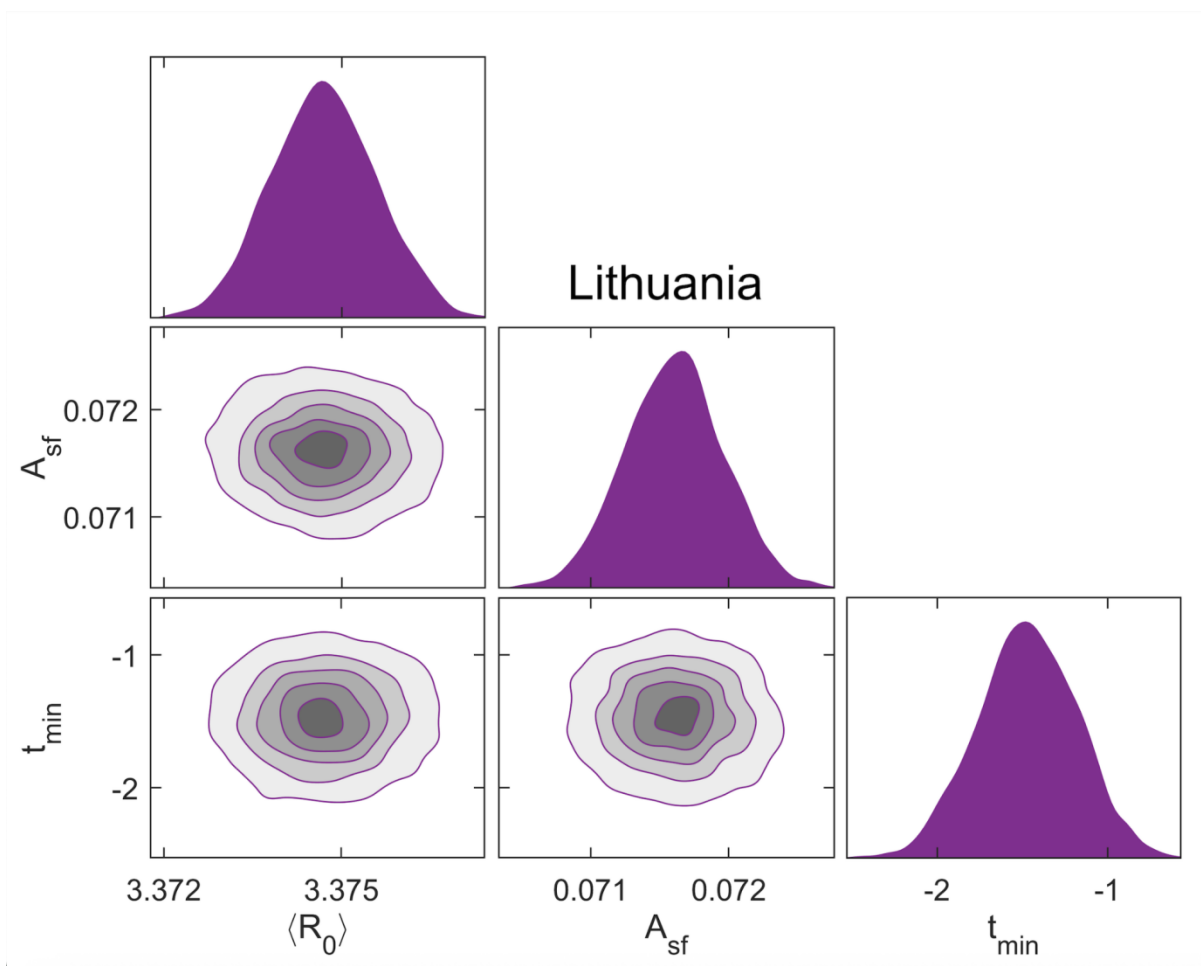
Za svaku od analiziranih zemalja MCMC procedura procenila je posteriorne distribucije četiri ključna parametra. Srednji godišnji reprodukcioni broj $\langle R_0 \rangle$ reprezentuje bazalnu transmisibilnost virusa integrisanu kroz čitav godišnji ciklus i odražava prosečnu ranjivosti populacije nezavisno od sezonskih oscilacija. Amplituda sezonskog forsiranja (A_{sf}) kvantifikuje intenzitet sezonske modulacije transmisije, odnosno veličinu oscilacije R_0 oko godišnje srednje vrednosti — veće vrednosti ovog parametra ukazuju na izraženiju sezonsku komponentu dinamike transmisije. Vreme nastupanja minimalne temperature u toku godine (t_{min}) definiše položaj sezonskog maksimuma transmisije i operacionalizovano je kao fazni pomak u danima od sredine godine, odnosno od dana 182, što omogućava direktnu i intuitivnu interpretaciju rezultata. Kako bi se olakšalo poređenje između zemalja različitih hemisfera, vremenska promenljiva je transformisana tako da vrednost parametra t_{min} direktno odgovara danu u godini u kome je reprodukcioni broj najviši. Poslednji parametar je rezidualna standardna devijacija σ koja obuhvata dnevnu varijabilnost neobjašnjenu sezonskim modelom.

Rezultati analize za odabrane severnohemisferske i južnohemisferske zemlje potvrdili su očekivane inverzne sezonske obrasce sa visokim stepenom statističke robusnosti. Za Srbiju je procenjeni srednji R_0 iznosio 3,36, pri čemu je minimum R_0 nastupio oko jednog dana

nakon sredine godine, što odgovara ranom letu u severnoj hemisferi (Ilustracija 8). Za Litvaniju je srednji R_0 procenjen na 3,37, takođe sa minimumom na danu 182+1 (Ilustracija 9), što je gotovo identično sa nalazom za Srbiju i ukazuje na konzistentnost sezonskog obrasca između severnohemisferskih zemalja sličnih geografskih širina. Vrednosti amplituda sezonskog forsiranja iznosile su 0,067 za Srbiju i 0,071 za Litvaniju, što ukazuje na umereno ali statistički robusno sezonsko forsiranje u obe zemlje.

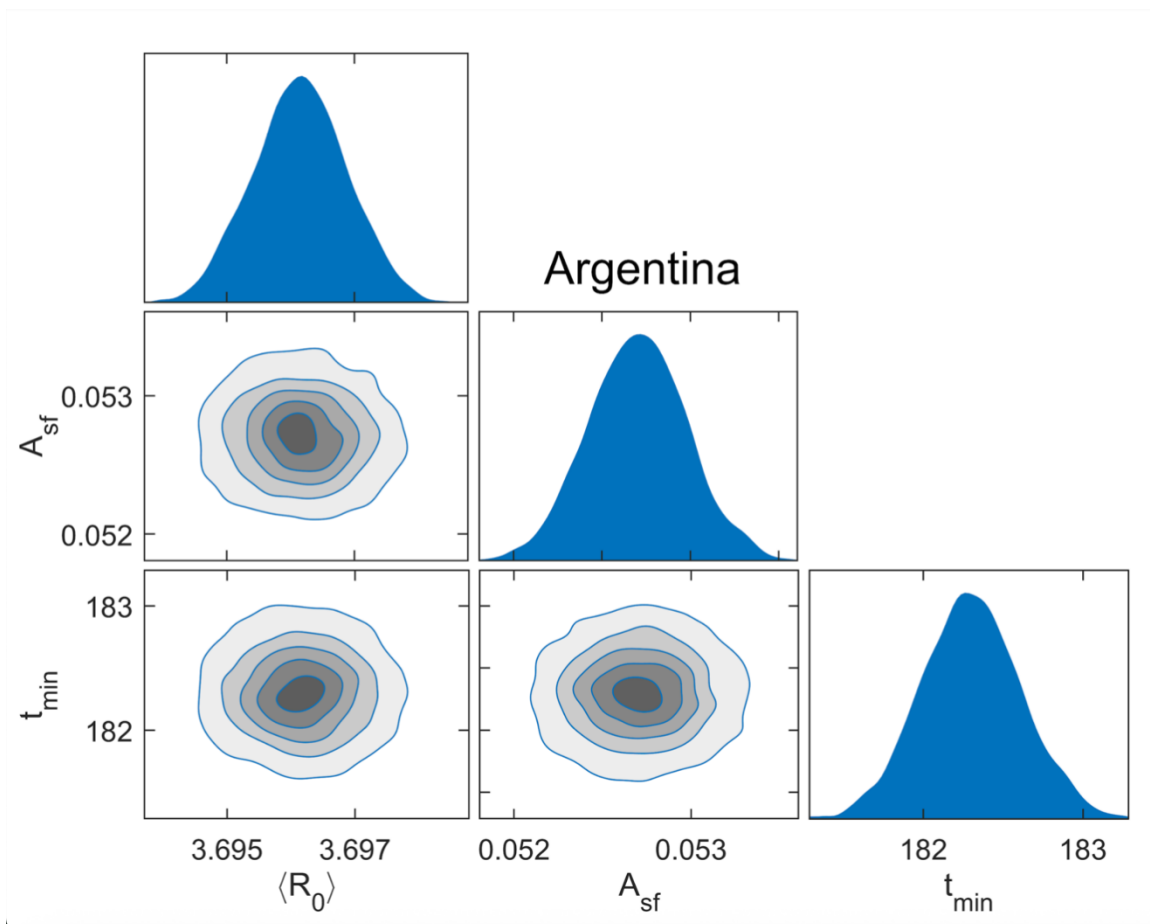


Ilustracija 8: Bajesijski korner plot za Srbiju – posteriorna raspodela parametara sezonskog modela



Ilustracija 9: Bajesijanski korner plot za Litvaniju – posteriorna raspodela srednjeg R_0 , amplitude sezonskog forsiranja (A_{sf}) i vremena minimalnog R_0 (t_{min})

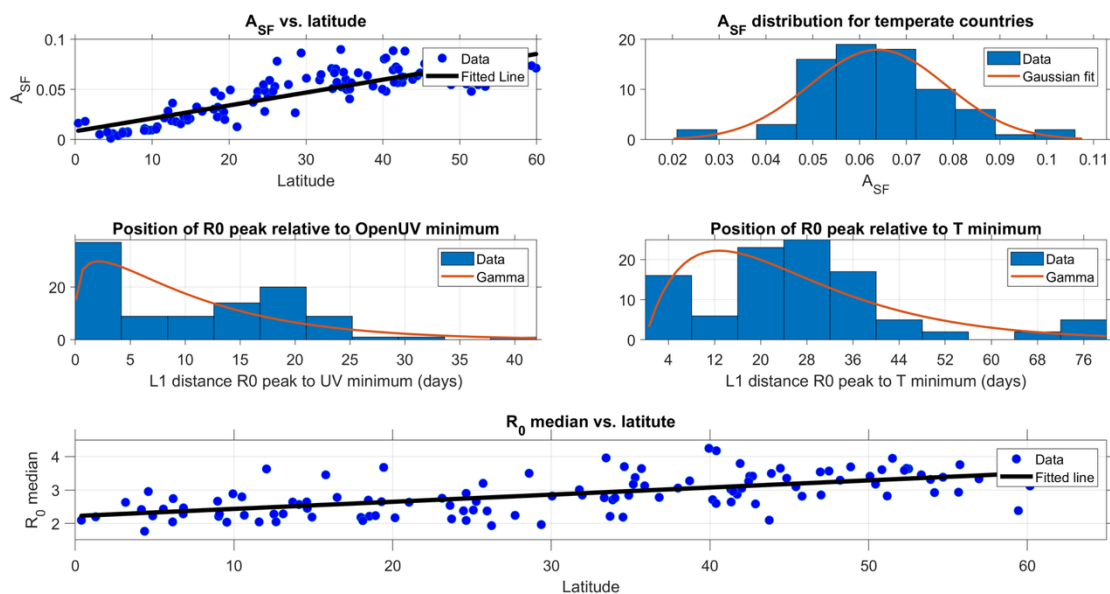
Za Argentinu, kao reprezentativnu zemlju južne hemisfere, procenjeni srednji R_0 iznosio je 3,70 (Ilustracija 10), što je nešto više od vrednosti procenjenih za severnohemisferske zemlje, ali u sličnom opsegu koji sugeriše konzistentnu bazalnu transmisibilnost virusa na globalnom nivou. Ključna razlika u odnosu na severnohemisferske zemlje ogleda se u tajmingu minimalnog R_0 — za Argentinu je minimum R_0 nastupio oko 182 dana nakon letnjeg solsticija, što odgovara letu na južnoj hemisferi i direktno potvrđuje invertovanu sezonsku dinamiku između hemisfera. Amplituda sezonskog forsiranja za Argentinu procenjena je na 0,052, što je nešto niže od vrednosti za Srbiju i Litvaniju, a može reflektovati specifičnosti klimatskog profila Argentine ili razlike u geografskoj heterogenosti ove prostorno velike zemlje.



Ilustracija 10: Bajesijanski korner plot za Argentinu – obrnuta sezonska dinamika u odnosu na severnu hemisferu, t_{min} pomeren za ~182 dana

Posteriorni ugaoni dijagrami generisani za sve tri analizirane zemlje vizualizuju združene i marginalne posteriorne distribucije za sva tri procenjena parametra i pružaju uvid u stepen statističke neizvesnosti procena kao i u korelacionu strukturu između parametara. Robusnost posteriornih distribucija, reflektovana u relativno uskim intervalima pouzdanosti za sve tri analizirane zemlje, potvrđuje da je MCMC procedura uspešno konvergirala i da su procenjeni parametri sezonskog forsiranja statistički pouzdani.

U cilju karakterizacije globalnog obrasca sezonalnosti, sinusoidalna procedura fitovanja proširena je na svih 118 zemalja korišćenjem nelinearnih najmanjih kvadrata. Rezultati ove analize na nivou čitavog uzorka sumirani su na Slici 11.



Ilustracija 11: Analiza sezonskih parametara R_0 na nivou 118 zemalja. (a) Amplituda sezonskog forsiranja A_{SF} u funkciji apsolutne geografske širine, sa izrazitom pozitivnom linearnom zavisnošću ($r = 0.85$). (b) Raspodela A_{SF} među 77 zemalja umerenog pojasa ($|\text{geografska širina}| > 23.5^\circ$) sa fitovanom Gausovom raspodelom (srednja vrednost = 0.064, SD = 0.015). (c) L1 rastojanje u danima između pika R_0 i minimuma UV zračenja (OpenUV) za zemlje iznad 10° geografske širine, fitovano gama raspodelom. (d) L1 rastojanje u danima između pika R_0 i minimuma temperature (T2M_MIN), fitovano gama raspodelom. Šira raspodela u poređenju sa panelom (c) sugerise da je UV zračenje proksimalniji pokretač sezonske transmisije od temperature. (e) Medijana godišnjeg R_0 u funkciji apsolutne geografske širine, sa pozitivnim linearnim trendom ($r = 0.59$).

Amplituda sezonskog forsiranja A_{SF} ispoljava izrazitu pozitivnu linearnu zavisnost od apsolutne geografske širine ($r = 0.85$), što ukazuje da zemlje udaljenije od ekvatora doživljavaju znatno izraženiju sezonsku modulaciju reproduktionog broja (gornji levi panel). Među 77 zemalja umerenog pojasa (geografska širina $> 23.5^\circ$), raspodela A_{SF} približno odgovara Gausovoj raspodeli sa srednjom vrednošću od 0.064 i standardnom devijacijom od 0.015 (gornji desni panel), što sugerise da je intenzitet sezonskog forsiranja relativno uniforman unutar umerenih zona, sa ograničenom varijabilnošću između zemalja. Sveukupno, Bajesijanska analiza pruža snažan kvantitativni argument za značajnu ulogu sezonalnosti u dinamici transmisije SARS-CoV-2. Nalaz da vrednosti srednjeg R_0 procenjene ovom analizom pokazuju relativnu konzistentnost između geografski, klimatski i socioekonomski raznorodnih zemalja — 3,36 za Srbiju, 3,37 za Litvaniju i 3,70 za Argentinu — sugerise da se bazalna transmisibilnost virusa kreće u sličnom opsegu između analiziranih zemalja, iako razlika između Srbije/Litvanije (3,36–3,37) i Argentine (3,70) od približno 10% ukazuje na to da geografski, demografski i socioekonomski faktori ipak moduliraju i ovu bazalnu komponentu. Sezonski uslovi zatim predvidivo amplifikuju ili suprimuju transmisiju oko ove bazalne vrednosti. Ovaj nalaz elevira sezonalnost iz statusa konfundirajuće varijable u status centralnog sistemskog faktora pandemijske dinamike i ukazuje da svako direktno poređenje efikasnosti politika javnog zdravlja ili težine epidemija između različitih regiona mora eksplicitno uzeti u obzir hemisferski zavisno sezonsko forsiranje kao pozadinski signal koji je prisutan nezavisno od primenjenih interventnih mera.

4.4 Diskusija

4.4.1 Sezonalnost kao dominantni globalni modulator transmisije

Primarni nalaz ove studije je da, kada se analiza obavi na globalnoj skali koja uključuje 118 zemalja različitih klimatskih zona, sezonalnost se pojavljuje kao snažan i predvidiv modulator transmisije SARS-CoV-2. Ovaj zaključak počiva na dva međusobno podržavajuća pristupa: emergenciji hemisferski zavisnih sezonskih obrazaca u prognozama modela mašinskog učenja, i formalnoj Bajesijanskoj kvantifikaciji parametara sezonskog forsiranja koji su sprovedeni kao nezavisni validacioni korak.

Emergencija sezonskih obrazaca iz modela treniranog isključivo na prostornim varijacijama između zemalja u ranoj fazi pandemije — dakle, bez ikakvog direktnog izlaganja vremenskim sezonskim podacima tokom treniranja — predstavlja posebno snažan argument u prilog centralnoj ulozi sezonalnosti. Kada je odabrani model, Ridge regresija na Union skupu prediktora, primenjena za generisanje dnevnih predikcija R_0 tokom čitave 2021. godine za 20 zemalja koje pokrivaju širok opseg geografskih širina, dobijeni godišnji obrasci su reprodukovali poznatu hemisfersku asimetriju sezonske dinamike respiratornih virusa. Zemlje severne hemisfere pokazale su niže vrednosti R_0 tokom proleća i leta, zemlje južne hemisfere pokazale su invertovanu godišnju dinamiku sa minimumom sredinom godine, a ekvatorijalne zemlje pokazale su multimodalne obrasce sa markantno manjim godišnjim opsegom vrednosti R_0 u poređenju sa zemljama viših geografskih širina. Činjenica da su ovi obrasci emergirali iz modela koji je uključivao isključivo prostorne razlike između zemalja snažno ukazuje na to da meteorološke i ekološke promenljive uključene u model uspešno enkoduju sezonsku komponentu transmisibilnosti kroz svoju prostornu distribuciju.

Formalna Bajesijanska karakterizacija ovih obrazaca, sprovedena fitovanjem prediktovanih godišnjih kriva R_0 sinusoidnoj funkciji putem MCMC procedure implementirane u Stan okruženju, pružila je kvantitativnu potvrdu sezonskog forsiranja sa procenjenim amplitudama od 0,067 za Srbiju, 0,071 za Litvaniju i 0,052 za Argentinu. Tajming minimalnog R_0 , definisan kao fazni pomak od sredine godine, direktno je odgovarao letnjim mesecima odgovarajuće hemisfere u svim analiziranim zemljama — minimum u ranom letu za severnohemisferske zemlje Srbiju i Litvaniju, i minimum oko 182 dana nakon letnjeg solsticija za južnohemisfersku Argentinu. Procenjene vrednosti srednjeg godišnjeg R_0 — 3,36 za Srbiju, 3,37 za Litvaniju i 3,70 za Argentinu — nalaze se u istom opsegu, ali razlika od približno 10% između severnohemisferskih i južnohemisferske zemlje ukazuje da bazalna transmisibilnost nije potpuno uniformna, već da je i ona pod uticajem lokalnih faktora koji nisu obuhvaćeni sezonskim modelom.

Ovi rezultati su u saglasnosti sa dobro utvrđenom sezonalnošću influence i drugih respiratornih virusa, za koje je poznato da dosežu transmisijsku kulminaciju u zimskim mesecima umereno klimatskih zona. Mehanizmi koji leže u osnovi sezonskog forsiranja respiratornih virusa višestruki su i međusobno isprepleteni. Na fizičkom nivou, niže temperature produžavaju preživljavanje viralnih aerosola u spoljašnjoj sredini, smanjena relativna vlažnost vazduha pogoduje transmisiji kapljica manjeg prečnika koje prodiru

dublje u respiratorni trakt, a redukovana ekspozicija UV-B zračenju ograničava viruicidno delovanje sunčevog svetla na virusne čestice u spoljašnjim sredinama. Na bihejvioralnom nivou, promene u obrascima kretanja i boravka populacije tokom hladnih meseci — veći udeo vremena provedenog u zatvorenim, slabo ventilisanim prostorima sa visokom gustinom ljudi — doprinose sezonskom povećanju efektivnog kontaktnog broja nezavisno od direktnih fizičkih efekata klimatskih faktora na viabilnost virusa. Prisustvo UV zračenja, merenog putem OpenUV_max promenljive, kao jednog od robusnih prediktora u modelima Random Foresta konzistentno je sa ovim mehanizmima i ukazuje na to da model uspešno zahvata barem deo ove biološke realnosti.

Važna epidemiološka implikacija ovog nalaza tiče se interpretacije efikasnosti javnozdravstvenih intervencija u retrospektivnim analizama pandemijskog odgovora. Svako poređenje rezultata intervencija između zemalja koje ulaze u različite faze sezonskog ciklusa metodološki je nevalidno ukoliko se ne kontroliše za sezonsku referentnu vrednost transmisije. Mere zatvaranja implementirane u zemlji koja istovremeno ulazi u period letnje minimalne transmisije nužno će pokazivati bolji epidemiološki ishod od identične intervencije sprovedene u periodu zime, bez obzira na stvarnu efikasnost primenjenih mera. Bez eksplicitne korekcije za ovu konfundirajuću varijablu, retrospektivne evaluacije pandemijskog odgovora ostaju metodološki nerobusne i podložne sistematskim greškama u atribuciji uzročno-posledičnih odnosa između politika i ishoda.

4.4.2 Izazovi globalnog inženjeringa podataka

Konstruisanje sveobuhvatnog i metodološki konzistentnog skupa podataka za globalnu epidemiološku analizu pokazalo se kao jedan od najzahtevnijih aspekata ove studije, koji je u značajnoj meri prevazišao inicijalnu procenu složenosti zadatka. Iskustvo stečeno tokom ovog procesa pruža vredne uvide u trenutno stanje globalnih infrastruktura za podatke i ukazuje na sistemske prepreke koje ograničavaju ambiciju i validnost komparativnih međunacionalnih istraživanja.

Osnovna pretpostavka sa kojom se pristupa globalnim analizama u savremenom informacionom dobu jeste da su relevantni podaci dostupni, standardizovani i komparabilni između zemalja. Ova pretpostavka pokazala se kao sistemski neosnovana čak i za naizgled jednostavne i pervazivne varijable. Populacioni podaci, koji su fundamentalni za konstruisanje ponderisanih proseka korišćenih u gotovo svim kategorijama promenljivih, prikupljaju se nekonzistentno između zemalja, sa nejasnim teritorijalnim granicama i različitim vremenskim tačkama prikupljanja. Gradske populacije, koje su bile ključne za metodologiju ponderisanja primenjenoj u ovoj studiji, definisane su naizmenično na osnovu urbanih, metropolitanskih i širokogradskih područja, pri čemu ove definicije nisu ekvivalentne ni između zemalja ni unutar iste zemlje u različitim vremenskim periodima. Ovaj naizgled tehnički detalj ima supstantivne posledice za validnost međunacionalnih poređenja jer različite definicije gradskih granica sistematski utiču na procenjenju gustinu naseljenosti i sve od nje izvedene varijable.

Otvorene baze podataka poput Wikidata nude potencijal za premošćavanje ovih praznina zbog svoje široke pokrivenosti i dostupnosti, ali podižu ozbiljna pitanja o valjanosti,

upravljanju i provenijenciji podataka. Uprkos višegodišnjim nastojanjima da se formalizuju procesi merenja kvaliteta podataka u ovakvim bazama, ovi standardi nisu široko prihvaćeni od strane globalne naučne zajednice, što otežava njihovu formalnu integraciju u istraživačke projekte koji zahtevaju stroge standarde kvalitete podataka. Komercijalne baze podataka inspirišu više poverenja u pogledu konzistentnosti i proverenosti podataka, ali dolaze sa sopstvenim skupom problema — strukturnom rigidnošću, retkim pokrićem i finansijskim barijerama pristupa koje ograničavaju reproducibilnost istraživanja.

Problem nepotpunosti podataka bio je posebno izražen u kategoriji zagađenja vazduha. WAQI baza, koja je metodološki najpogodniji izvor jer obezbeđuje direktna merenja kvaliteta vazduha u urbanim centrima konzistentna sa pristupom primenjenim za meteorološke podatke, sadrži podatke za 500 najvećih gradova na svetu. Međutim, usled neravnomerne geografske distribucije ovih gradova koja odražava nejednaku razvijenost infrastrukture za monitoring kvaliteta vazduha, pristup podacima za preostale gradove zahteva plaćenu pretplatu putem istorijskog API-ja. Paradoksalno, uprkos nedostajućim vrednostima za 70 do 90 zemalja od ukupno 118 analiziranih, promenljive zagađenja vazduha iz WAQI baze ipak su pokazale statističku značajnost u modelima selekcije karakteristika. Ovaj nalaz snažno sugeriše da bi potpuno pokriće ovim podacima potencijalno rezultovalo još snažnijim prediktorima i boljim performansama modela, te istovremeno naglašava epidemiološku vrednost investiranja u globalno proširenje mreža za monitoring kvaliteta vazduha.

Pokušaj premošćavanja nepotpunosti WAQI podataka korišćenjem EDGAR baze emisija zagađivača otkrio je suštinsko konceptualno ograničenje — ukupne godišnje emisije zagađivača na nacionalnom nivou, ni u sirovom ni u populaciono-skaliranom obliku, ne odražavaju adekvatno stvarnu individualnu izloženost zagađivačima koja zavisi od prostorne distribucije izvora emisija, meteoroloških uslova disperzije i individualnih obrazaca kretanja i boravka. Ovo iskustvo ilustruje širu metodološku lekciju o opasnostima supstitucije konceptualno adekvatnih ali nepotpunih podataka konceptualno neadekvatnim ali kompletnim podacima — kompletnost skupa podataka nije sama po sebi vrlina ukoliko se postiže na račun mernog ekvivalenta konstrukta od interesa.

Mogući put ka rešavanju problema nepotpunosti podataka o kvalitetu vazduha leži u modelu otvorene saradnje, analognom onom koji već primenjuje sama WAQI platforma agregiranjem podataka iz individualnih mernih stanica širom sveta. Volonteri sa pristupom i znanjem o lokalnim institucijama za prikupljanje podataka mogli bi doprineti proširenju pokrivenosti za gradove koji trenutno nisu zastupljeni, pri čemu bi odgovarajuće citiranje izvora bilo neophodno za održavanje kvaliteta podataka. Ovakav konzorcijski pristup bio bi, međutim, suviše obiman i složen zadatak za jedan istraživački tim, i zahtevao bi širu koordinaciju naučne zajednice.

4.4.3 Ograničenja studije

Uprkos metodološkoj sveobuhvatnosti i opsežnosti skupa podataka konstruisanog za ovu studiju, postoji niz supstantivnih ograničenja koja je neophodno eksplicitno navesti i

razmotriti pri interpretaciji rezultata, kako bi se izbeglo precenjivanje snage zaključaka i kako bi se identifikovali prioriteti za buduća istraživanja.

Prvo i najfundamentalnije ograničenje odnosi se na kvalitet i kompletnost ulaznih podataka. Kao što je detaljno razmotreno u prethodnom odeljku, konstruisani skup podataka sadrži značajan broj nedostajućih vrednosti, posebno u kategorijama zagađenja vazduha i pojedinih zdravstvenih indikatora. Premda je primenjena standardna procedura imputacije medijanom za promenljive sa manjim brojem nedostajućih vrednosti, a promenljive sa više od 100 nedostajućih vrednosti od ukupno 118 isključene iz analize, preostale nedostajuće vrednosti nisu nasumično distribuirane — one su sistematski korelisane sa nivoom razvijenosti lokalne infrastrukture, administrativnim kapacitetima i kvalitetom života, koji su sami po sebi potencijalni prediktori transmisibilnosti. Ova nasumičnost nedostajućih vrednosti znači da imputacija medijanom, koliko god bila standardna procedura, unosi sistematska pristrasnost u skup podataka čiji smer i intenzitet nije moguće precizno kvantifikovati.

Drugo ograničenje tiče se sezonske varijabilnosti koja je u modelu zahvaćena isključivo putem meteoroloških promenljivih. Značajan deo sezonske varijabilnosti u transmisiji respiratornih virusa posledica je promena u ljudskom ponašanju koje nisu direktno korelisane sa temperaturom i ostalim meteorološkim parametrima na jednostavan i predvidiv način. Na primer, gustina ljudi u spoljašnjim kafićkim baštama može biti veća u prvim danima proleća nego kasnije tokom godine, čak i kada vremenske prilike postaju povoljnije kako sezona odmiče — što odražava društvene i kulturološke obrasce koji nisu enkodovani ni u jednoj od raspoloživih promenljivih. Uključivanje novih varijabli zasnovanih na satelitskim snimcima, koje bi mogle da zahvate informacije poput intenziteta saobraćaja, zauzetosti parking mesta i gustine zagađivača izvedene iz daljinskog osmatranja, moglo bi unaprediti godišnje predikcije ne samo kroz direktne podatke o kretanju već i kroz posredne mere izloženosti zagađivačima u realnom vremenu.

Treće ograničenje odnosi se na sam proces selekcije karakteristika i konstruisanje skupova prediktora. Iako je primenjen rigorozan okvir koji kombinuje tri metodološki različita pristupa selekciji — postupnu selekciju zasnovanu na AIC kriterijumu, Lasso regularizaciju i permutacijsku važnost Random Foresta— svaki od ovih pristupa ima sopstvena implicitna pretpostavke i sklonosti koje mogu sistematski favorizovati određene tipove promenljivih. Pored toga, relativno mali uzorak od 118 zemalja u kombinaciji sa 96 polaznih promenljivih stvara uslove visoke dimenzionalnosti u kojima je rizik od slučajnih korelacija i nestabilnih procena koeficijenata povišen, a pouzdanost selekcije karakteristika ograničena. Ponovljeni eksperiment sa 100 različitih podela skupa podataka delom adresira ovaj problem obezbeđujući procene varijabilnosti performansi, ali ne eliminišu fundamentalno ograničenje veličine uzorka.

Četvrto ograničenje tiče se primenjenih algoritama mašinskog učenja i njihove tendencije ka preprilagođavanju. Kao što je pokazano u analizi performansi modela, ansambl metode poput Random Foresta pokazuju izraženu diskrepancu između trening i test greške, što ukazuje na preprilagođavanje šumu prisutnom u relativno malom skupu podataka. Premda je upravo ova karakteristika bila jedan od razloga za konačni odabir Ridge regresije kao

finalnog modela, ni linearni regularizacioni modeli nisu imuni na ovaj problem — regularizacioni parametar koji minimizuje grešku na validacionom skupu u jednoj od 100 podela skupa podataka možda nije optimalan za drugu podelu, što unosi dodatnu varijabilnost u procene finalnih performansi.

Peto ograničenje odnosi se na vremenski okvir studije i reprezentativnost podataka za kasniji tok pandemije. Vrednosti R_0 procenjene su iz inicijalne, nekontrolisane faze epidemije koja je za različite zemlje nastupila u različitim periodima između januara i aprila 2020. godine. Ove R_0 procene odražavaju transmisibilnost originalnog Vuhan soja SARS-CoV-2 u specifičnim epidemiološkim uslovima ranog perioda pandemije, pre pojave varijanti sa izmenjenom transmisibilnošću i pre sticanja imuniteta populacije. Ekstrapolacija zaključaka ove studije na kasniji tok pandemije, koji je bio obeležen sukcesijom varijanti sa različitim transmisivnim karakteristikama i progresivnim sticanjem prirodnog i vakcinalnog imuniteta, zahteva oprez i nije direktno podržana podacima korišćenim u analizi.

Konačno, važno je napomenuti da, uprkos svim navedenim ograničenjima, vrednost ove studije leži ne samo u konkretnim nalazima već i u transparentnom dokumentovanju izazova globalnog inženjeringa epidemioloških podataka koji će biti relevantni za buduća istraživanja sličnog obima i ambicije.

4.5 Zaključak

Ova studija predstavlja jednu od metodološki najsveobuhvatnijih globalnih analiza faktora koji su modulirali ranu transmisiju SARS-CoV-2. Integracijom podataka iz višestrukih izvora — NASA POWER za meteorološke podatke, WHO Global Health Observatory za zdravstvene indikatore populacije, EDGAR za emisije zagađivača, Facebook Social Connectedness Index za mere društvene povezanosti i OpenUV za podatke o ultravioletnom zračenju — za 118 zemalja, i primenom rigoroznog okvira mašinskog učenja sa ekstenzivnom ukrštenom validacijom koja je uključivala 100 ponavljanja podele skupa podataka i više od 2.500 procesorskih sati računarskog vremena, studija je postigla dva ključna cilja koji su bili eksplicitno formulisani kao istraživačka pitanja.

Prvo, demonstrirala je da su robusni modeli predikcije R_0 zasnovani na ekološkim i sociodemografskim varijablama principijelno izvodljivi. Od devet testiranih algoritama koji su pokrivali spektar od parametarskih linearnih metoda do nelinearnih ansambl tehnika, i od šest skupova prediktora konstruisanih različitim metodama selekcije karakteristika, kombinacija Ridge regresije i objedinjenog Union skupa prediktora pokazala se kao optimalna, ostvarujući test-RMSEP od 0,78 uz najstabilniji balans između prilagođavanja i generalizacije i najujednačenije ponašanje tokom čitave kalendarske godine. Ovaj rezultat potvrđuje da informacija o inherentnoj transmisibilnosti virusa nije slučajno distribuirana između zemalja već je sistematski kodovana u merljivim karakteristikama njihovog ekološkog, klimatskog i sociodemografskog profila.

Drugo, i od posebnog naučnog i javnozdravstvenog značaja, studija je formalno kvantifikovala sezonalnost kao značajan, hemisferski zavisni modulator globalne transmisibilnosti SARS-CoV-2. Ovaj zaključak počiva na dva nezavisna i međusobno podržavajuća pristupa. S jedne strane, primenom odabranog modela na generisanje dnevnih predikcija R_0 tokom čitave 2021. godine za 20 zemalja koje pokrivaju širok opseg geografskih širina, emergirali su jasni hemisferski zavisni sezonski obrasci — niže vrednosti R_0 tokom letnjih meseci za severnohemisferske zemlje, invertovana dinamika za južnohemisferske zemlje, i multimodalni obrasci sa smanjenim godišnjim opsegom za ekvatorijalne zemlje. S druge strane, Bajesijanska analiza zasnovana na MCMC proceduri implementiranoj u Stan okruženju formalno je kvantifikovala parametre sezonskog forsiranja za odabrane zemlje. Za Srbiju i Litvaniju, kao reprezentativne severnohemisferske zemlje, procenjeni srednji R_0 iznosio je 3,36 odnosno 3,37, sa minimumom R_0 neposredno nakon sredine godine i amplitudama sezonskog forsiranja od 0,067 odnosno 0,071. Za Argentinu, kao reprezentativnu južnohemisfersku zemlju, procenjeni srednji R_0 iznosio je 3,70, sa minimumom R_0 pomećenim za otprilike šest meseci u odnosu na severnohemisferske zemlje i amplitudom sezonskog forsiranja od 0,052. Procenjene vrednosti srednjeg godišnjeg R_0 kreću se u sličnom, ali ne identičnom opsegu (3,36–3,70), što sugerise da bazalna transmisibilnost virusa, iako uporediva između analiziranih zemalja, nije potpuno uniformna — ostaje pod uticajem lokalnih faktora koji prevazilaze sezonsku komponentu. Sezonski ciklus zatim predvidivo modulira transmisiju oko ove bazalne vrednosti.

Ovi zaključci imaju direktne i dalekosežne implikacije za epidemiološku metodologiju i praksu javnog zdravlja. Svaka komparativna evaluacija efikasnosti pandemijskog odgovora mora eksplicitno modelovati sezonalnu referentnu vrednost transmisije kao konfundirajuću varijablu — bez ove korekcije, poređenja između zemalja koje ulaze u različite faze sezonskog ciklusa su metodološki nevalidna i podložna sistematskim greškama u atribuciji uzročno-posledičnih odnosa između primenjenih politika i epidemioloških ishoda. Pored toga, sposobnost predikcije sezonskih vrhunaca transmisije na osnovu ekoloških i klimatskih karakteristika zemalja pruža planerima zdravstvene zaštite instrumentalni okvir za proaktivnu alokaciju resursa i strateško planiranje preventivnih kampanja u periodima povećanog transmisijskog rizika.

Konačno, emergencija sezonskih obrazaca iz modela treniranog isključivo na podacima s početka pandemije — bez ijednog opserviranog sezonskog ciklusa u skupu za treniranje, budući da su svi podaci o R_0 prikupljeni iz februara i marta 2020. godine — predstavlja snažan metodološku nalaz da ekološke varijable zahvaćene modelom nose stvarnu, biološki relevantnu informaciju o dinamici transmisije respiratornih virusa koja prevazilazi granice konkretnog vremenskog i geografskog konteksta u kome je model treniran. Ovaj nalaz postavlja empirijsku osnovu za integrisano ekološko-epidemiološko modelovanje budućih pandemijskih pretnji i sugerise da bi sistematsko prikupljanje i harmonizacija globalnih ekoloških, klimatskih i sociodemografskih podataka trebalo da bude prioritet naučne zajednice i relevantnih međunarodnih institucija u okviru priprema za buduće pandemijske scenarije.

5 Uticaj zagađenja vazduha na prenosivost COVID-19 u Sjedinjenim Američkim Državama

5.1 Uvod

Pandemija COVID-19 je globalni fenomen, ali njena dinamika nije bila homogena ni između država ni unutar njih. Posebno je izražena sub-nacionalna heterogenost u Sjedinjenim Američkim Državama, gde su pojedine države iskusile drastično različite stope ranog širenja infekcije. Ova varijabilnost predstavlja prirodni eksperiment koji omogućava razdvajanje uloge ekoloških, demografskih i zdravstvenih faktora u inicijalnoj transmisiji SARS-CoV-2. U ovoj studiji analizirano je 46 saveznih država SAD za koje je bilo moguće pouzdano identifikovati eksponencijalni režim širenja epidemije, a za svaku je konstruisan skup od 74 prediktorske varijable grupisane u četiri kategorije (Tabela 5): zagađenje vazduha (10 polutanata), meteorološki parametri, sociodemografski indikatori i zdravstvene karakteristike populacije.

U globalnoj analizi je primećeno da se parametri zagađenja pojavljuju kao statistički značajni iako su podaci dati na skali države i u velikoj meri nedostajali. Veliki broj ranih studija pokušao je da poveže zagađenje vazduha sa COVID-19 ishodima, ali su se često oslanjale na sirove brojeve slučajeva ili mortalitet, koji su snažno zavisni od testiranja, zdravstvenog sistema i političkih intervencija. Konkretno, kao što je već diskutovano, poređenje apsolutnih brojeva slučajeva između različitih regiona podložno je uticaju razlika u trenutku nastanka epidemije, primenjenih mera socijalne distance i kapaciteta testiranja — što znači da takvi pokazatelji ne mere inherentnu biološku prenosivost virusa, već kombinaciju epidemioloških i društvenih efekata (Hale et al., 2021). Nasuprot tome, osnovni reprodukcionni broj R_0 opisuje prosečan broj sekundarnih infekcija u potpuno osetljivoj populaciji u odsustvu intervencija. Budući da zavisi isključivo od nagiba eksponencijalnog rasta u ranoj fazi epidemije i time je manje osetljiv na apsolutni obim testiranja i na razlike u kontrolnim merama, R_0 omogućava daleko pouzdanije poređenje transmisibilnosti između različitih populacija (Salom et al., 2021).

Metodološki okvir korišćen u ovom poglavlju oslanja se na nelinearnu dinamiku i SEIR modelovanje inicijalnog eksponencijalnog režima širenja epidemije objašnjeno u poglavlju 3. Pošto ukupan broj varijabli (74) prevazilazi veličinu uzorka (46 država), direktna primena regresionih metoda bila bi nestabilna i sklona preprilagođavanju. Zbog toga je primenjen višemodelni pristup koji obuhvata: redukciju dimenzionalnosti putem Principal Component Analysis (PCA) unutar tematskih blokova međusobno korelisanih varijabli, regularizovane linearne regresije (Lasso i Elastic Net) za selekciju i rangiranje nezavisnih prediktora, te nelinearne ansambl metode (Random Forest i Gradient Boost) za identifikaciju nelinearnih obrazaca i kvantifikaciju veličine efekta (Milicevic et al., 2021).

Cilj ovog poglavlja je da kvantifikuje doprinos zagađenja vazduha — posebno $PM_{2.5}$ čestica — u ranom širenju COVID-19 kroz 46 država SAD, uz kontrolu velikog broja potencijalnih

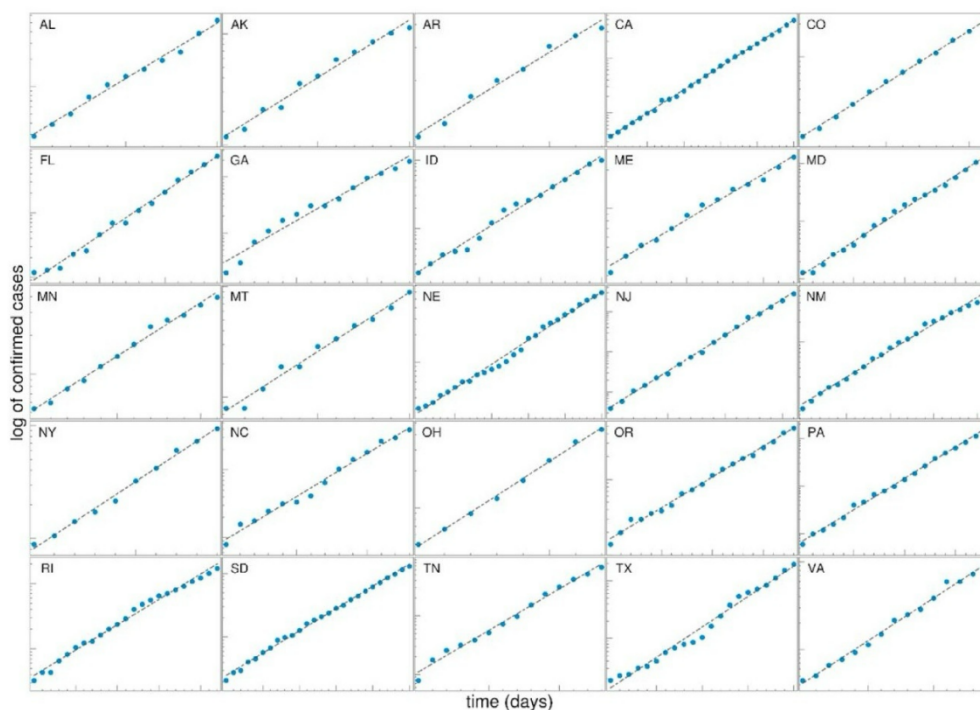
konfundera. SAD predstavljaju optimalan okvir za ovu analizu jer, za razliku od globalnih poređenja između heterogenih zemalja, obezbeđuju: dovoljno veliku varijaciju u nivoima zagađenja, relativno homogen zdravstveni sistem i standarde testiranja, kao i dostupne detaljne epidemiološke i ekološke podatke. Time se smanjuje šum koji je neizbežan u analizama koje obuhvataju administrativno i kulturološki heterogene jedinice.

5.2 Metodologija

Specifičnosti ove studije u odnosu na opštu metodologiju su izložene u ovoj sekciji.

5.2.1 Procena osnovnog reproduktionog broja

Model izložen u metodologiji omogućava preciznu procenu R_0 iz empirijskih podataka (jednačine (10)-(12)) i primenjen je na podatke o kumulativnim detektovanim slučajevima za svaku od 46 analiziranih država SAD, preuzetih iz baze Worldometer (Worldometer, 2020). Pet isključenih država su Illinois, Ajova, Indijana, Kansas i Kentaki, a dodat je Distrikt Kolumbija. Ovi podaci korišćeni su isključivo za procenu R_0 i nisu direktno ulazili kao prediktori u modele mašinskog učenja. Za svaku državu ručno je identifikovan interval čistog eksponencijalnog rasta na semi-logaritamskoj skali, primenom kriterijuma linearnosti i minimalne dužine intervala, uz izostavljanje ranih dana sa malim brojem slučajeva. Linearna regresija na log-skali daje nagib λ_+ , iz kojeg se potom direktno izračunava R_0 . Validnost eksponencijalnog režima potvrđena je za svih 46 država, što je u skladu sa prethodno objavljenim nalazima za globalni skup zemalja (Salom et al., 2021).



Ilustracija 12: Eksponecijalni režim rasta i ekstrakcija R_0 . Prikazane su log-transformisane krive kumulativnih detektovanih slučajeva za reprezentativne savezne države SAD u identifikovanom vremenskom prozoru ekspancijalne faze. Linearost krivih potvrđuje validnost ekspancijalnog modela, a nagib svake krive direktno određuje procenjeni R_0 .

5.2.2 Podaci o zagađenju vazduha

Podaci o kvalitetu vazduha preuzeti su iz EPA Air Quality System baze (US EPA, 2013). Za svaku državu prikupljeni su dnevni agregisani podaci sa svih dostupnih mernih stanica za sledeće polutante: gasovite zagađivače (O_3 , NO_2 , SO_2 , CO), čestične materije ($PM_{2.5}$ i PM_{10}) i ostale dostupne vrste, uključujući isparljiva organska jedinjenja (VOC), okside azota (NO_x) i opasne zagađivače vazduha (HAPs).

Budući da merne stanice nisu uniformno raspoređene, primenjena je populaciono ponderisana agregacija na nivou države. Vrednosti su najpre usrednjene po gradu za svaku stanicu, a zatim je državni prosek izračunat kao prosek gradskih vrednosti ponderisan brojem stanovnika, koristeći populacione podatke iz popisa:

$$X_{state} = (\sum_i P_i X_i) / (\sum_i P_i) \quad (15)$$

gde je:

X_i - vrednost na stanici i

P_i - populacija urbanog područja koje stanica pokriva

Ovom metodom sprečava se da retko naseljene oblasti disproportionalno utiču na državni prosek.

Kritično za validnost analize, vrednosti zagađenja nisu uzete kao višegodišnji proseci, već su vremenski usklađene sa periodom procene R_0 za svaku državu pojedinačno — to jest, sa identifikovanim eksponencijalnim prozorom rasta. Time se osigurava da prediktori odražavaju uslove izlaganja tokom relevantnog epidemiološkog perioda, a ne dugoročni prosek koji bi mogao maskirati kratkoročne varijacije.

5.2.3 Meteorološki podaci

Meteorološki podaci preuzeti su iz NASA POWER projekta (NASA Langley Research Center, 2020) putem API-ja, za koordinate najvećih urbanih centara svake države koji čine više od 10% državne populacije. Vrednosti su prostorno agregirane ponderisanjem populacijom gradova, i vremenski usrednjene kroz relevantni epidemiološki interval identičan onome korišćenom za zagađenje.

Uključeni parametri obuhvataju: temperature na 2 m i 10 m (srednja, maksimalna, minimalna), mere vlažnosti (temperatura vlažnog termometra, relativna vlažnost, tačka rose, specifična vlažnost), ukupne padavine, indekse insolacije (direktno i difuzno sunčevo zračenje), brzinu vetra i atmosferski pritisak. Maksimalni predviđeni UV indeks preuzet je iz OpenUV servisa. Meteorološki parametri su uključeni jer utiču na stabilnost aerosola, preživljavanje virusa u spoljašnjoj sredini i obrasce ljudskog ponašanja unutar i van zatvorenih prostora.

5.2.4 Sociodemografski i zdravstveni podaci

Demografski i zdravstveni indikatori preuzeti su iz nekoliko izvora. Demografski sastav populacije po polu, rasi i starosnoj strukturi (procenat ispod 18 i iznad 65 godina), kao i podaci o zdravstvenom osiguranju, BDP-u, očekivanom trajanju života i mortalitetu odojčadi i dece, preuzeti su sa platforme Measure of America (Social Science Research Council, 2020). Medicinski parametri — hipertenzija, holesterol, kardiovaskularne bolesti, dijabetes, rak, gojaznost, fizička neaktivnost, hronična bubrežna i hronična opstruktivna plućna bolest — preuzeti su iz America's Health Rankings baze koja hostuje CDC podatke (United Health Foundation, 2020). Procenat stranog stanovništva preuzet je sa Census Reporter platforme (Census Reporter, 2020), a subnacionalni HDI iz Global Data Lab baze (Smits & Permanyer, 2019). Gustina stanovništva, procenat urbanog stanovništva i medijana starosti preuzeti su sa US Census Bureau sajta (U.S. Census Bureau, 2020).

Ove varijable modeluju strukturne karakteristike populacije koje su relativno stabilne u kratkom vremenskom intervalu i koje se mogu smatrati populacionim faktorima rizika za transmisiju i težinu bolesti.

5.2.5 Obrada podataka i transformacije

Jedan od ključnih izazova bio je usklađivanje različitih vremenskih skala:

infekcija → inkubacija → testiranje → prijava

Da bi se to kompenzovalo, rađena je vremenska harmonizacija - ekološki podaci su pomereni unazad u odnosu na epidemiološke podatke za 5 dana. Ova korekcija osigurava da prediktori odgovaraju stvarnom vremenu izlaganja, a ne trenutku administrativne potvrde slučaja.

Kontrola nedostajućih podataka je pokazala da su nedostajuće vrednosti bile retke (<5%). Kada su postojale, zamenjene su medijanom te promenljive i potvrđeno je da imputacija ne menja distribuciju hi-kvadrat testom slaganja za značajnost od 0,05. Ovo sprečava gubitak opservacija bez uvođenja pristrasnosti.

5.2.6 Redukcija dimenzionalnosti putem PCA

Ukupan broj varijabli (74) premašuje veličinu uzorka (46 država), što predstavlja ozbiljan rizik od preprilagođavanja, naročito za nelinearne metode. Pored toga, mnoge varijable su međusobno visoko korelisane jer opisuju različite aspekte istih latentnih faktora. Da bi se smanjila dimenzionalnost bez gubitka interpretabilnosti, varijable su najpre grupisane u tematske blokove prema konceptualnoj srodnosti i očekivanoj korelisanosti, a zatim je u okviru svakog bloka primenjena PCA.

Grupisanje je sprovedeno prema dva kriterijuma: varijable moraju opisivati slične veličine kako bi interpretacija dobijenih komponenti ostala nedvosmislena, i varijable unutar grupe moraju biti međusobno visoko korelisane kako bi PCA efikasno smanjila dimenzionalnost. Grupisanje varijabli i odgovarajuće glavne komponente prikazani su ispod (Tabela 2).

Tabela 2: Grupisanje varijabli i odgovarajuće PCA komponente

PCA komponente	Varijable
PC1 temperatura	T2M, T2MMAX, T2MMIN, T10M, T10MMAX, T10MMIN, TS
PC1 vlažnost	QV2M, T2MDEW
PC1 padavine	PRECTOT, TQV
PC1 vetar	WS2M, WS10M
PC1-PC2 radijacija	CLRSKY_SFC_SW_DWN, ALLSKY_SFC_SW_DWN, ALLSKY_SFC_LW_DWN
PC1-PC2 sezonalnost	PC1 temperatura, PC1 vlažnost, PC1 padavine, PC1 radijacija, PC2 radijacija, RH2M, OpenUVmax
PC1 NO	NO ₂ , NONOxNOy
PC1-PC2 starost	Population over 65, Youth population, Medijan age
PC1-PC2 gustina	1/BUAPC, Urban population, Population density
PC1-PC4 prosperitet	Life expectancy, Infant mortality, GDP, HDI, Child mortality, Alcohol consumption, Foreign-born population
PC1-PC4 bolesti	Gojaznost, CVD deaths, Hipertenzija, Holesterol, Pušenje, KVB, Dijabetes, Rak, Hronična bubrežna bolest, HOBP, Višestruke hronične bolesti, Fizička neaktivnost
PC1-PC3 osiguranje	Zdravstveno osiguranje za svu decu, odrasle i ukupnu populaciju, po rasnim grupama (crna, starosedelačka, azijska, latinoamerička, bela)

Pre PCA sve varijable su standardizovane (oduzeta aritmetička sredina, podeljeno standardnom devijacijom) kako bi se sprečilo da varijable sa većom numeričkom skalom dominiraju komponentama. Za svaki blok zadržano je onoliko komponenti koliko je potrebno da se kumulativno objasni više od 85% varijanse unutar bloka. Provereno je da komponente u statistički dovoljnoj meri slede normalnu raspodelu, što je očekivano na osnovu primenjenih transformacija originalnih varijabli.

Neke varijable nisu zadovoljile kriterijume grupisanja i u dalju analizu su ušle samostalno: relativna vlažnost RH2M (koja nije visoko korelisana sa ostalim merama vlažnosti QV2M i T2MDEW), PM_{2.5}, PM₁₀, CO i rasne varijable.

5.2.7 Lasso regresija

Lasso (Least Absolute Shrinkage and Selection Operator) uvodi L1 regularizaciju koja dovodi do automatske selekcije varijabli skupljanjem koeficijenata nekih prediktora na tačno nulu. Sve ulazne varijable su standardizovane. Hiperparametar λ optimizovan je pretragom po eksponencijalnoj skali od numeričke nule (OLS regresija) do vrednosti koja daje model sa samo konstantom. Srednja kvadratna greška (MSE) na test skupu unakrsne validacije (200 ponavljanja, podela baze podataka 80-20 prema ukupnim opservacijama) korišćena je kao funkcija gubitka, a odabran je λ_1SE — najjednostavniji model koji je i dalje uporediv sa optimalnim. Konačni model obuhvata sve nenulte koeficijente.

5.2.8 Elastic Net regresija

Elastic Net proširuje Lasso kombinacijom L1 i L2 regularizacije, uvodeći drugi hiperparametar α koji kontroliše odnos između penalizacija. Ovaj model je posebno pogodan za situacije sa visoko koreliranim prediktorima: dok Lasso teži da zadrži samo jednu varijablu iz koreliranog klastera, Elastic Net dozvoljava ravnomerniju raspodelu težine između njih, što daje realističniju sliku u ekološkim sistemima gde faktori deluju zajedno. Dvodimenzionalna pretraga hiperparametara (λ po eksponencijalnoj skali, α linearno od 0 do 1) i unakrsna validacija sprovedeni su identično kao za Lasso. Konačni odabrani par (α , λ) je najbliži centroidu svih foldova.

5.2.9 Random Forest i Gradient Boost

Da bi se izbeglo prilagođavanje, pre primene nelinearnih metoda sprovedena je preliminarna selekcija varijabli: zadržane su samo one koje pokazuju statistički značajnu korelaciju sa R_0 ($p < 0.1$) prema bar jednoj od tri mere — Pirsonovoj, Kendalovoj ili Spirmanovoj korelaciji (Pearson, 1896; Spearman, 1904; Kendall, 1938). Pirsonova korelacija meri linearnu asocijaciju između kontinuiranih varijabli uz pretpostavku normalne raspodele, Spirmanova je njen neparametarski ekvivalent koji meri monotoni odnos rangova i robusnija je na ekstremne vrednosti i odstupanje od normalne raspodele, dok Kendalova takođe meri korelaciju rangova ali je stabilnija pri malim uzorcima i manje osetljiva na greške merenja. Ovaj kriterijum rezultovao je skupom od 13 varijabli koje su ušle u Random Forest i Gradient Boosting modele.

Za Gradient Boosting optimizovani su: maksimalni broj podela ($\{1, 2, 3, 4, 5, 8, 16\}$), minimalna veličina lista ($\{1, 2, 3, 4, 5, 8, 16, 18\}$) i stopa učenja ($\{0.1, 0.25, 0.5, 0.75, 1\}$). Za Random Forest optimizovani su maksimalni broj podela ($\{6, 12, 18, 22, 24, 26, 30, 35\}$) i minimalna veličina lista ($\{1, 2, \dots, 7\}$). Broj stabala u ansamblu biran je da minimizuje MSE na test skupu. Unakrsna validacija sprovedena je identično kao za linearne modele (200 ponavljanja, podela 80-20), a dobijeni hiperparametri korišćeni su za ponovni trening na celom skupu podataka, nakon čega je procenjena važnost prediktora za obe metode.

5.3 Rezultati

Slično prethodnom poglavlju, rezultati su dati kao inicijalna korelaciona analiza, zatim kao rigoriznija selekcija obeležja putem regularizovanih metoda, i na kraju kao konačni prediktivni modeli sa kvantifikovanom greškom i zaključcima koji proizilaze iz njih.

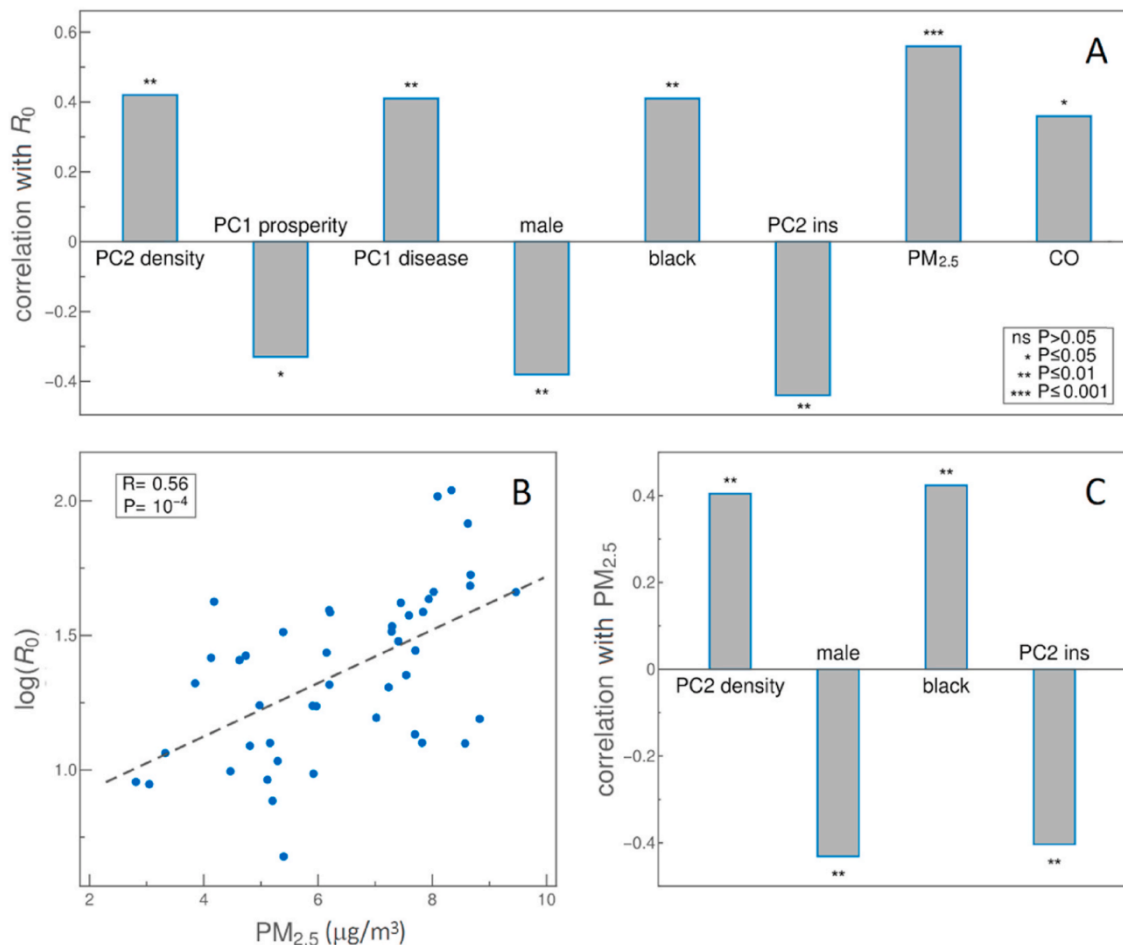
5.3.1 Inicijalna korelaciona analiza: prominencija $PM_{2.5}$

Inicijalna faza analize imala je za cilj eksploratorno mapiranje odnosa između pojedinačnih prediktorskih varijabli i procenjenih vrednosti osnovnog reproduktionog broja R_0 . Iako je jasno da korelacije parova same po sebi ne mogu uspostaviti uzročnost, niti razdvojiti međusobno zavisne efekte visoko korelisanih faktora, ovakav pristup predstavlja važan prvi korak u identifikaciji dominantnih signala u podacima i usmerava kasniju multivarijantnu analizu. Za svaku od 74 prediktorske varijable izračunat je Pirsonov koeficijent korelacije sa R_0 , uz odgovarajuću procenu statističke značajnosti. Posebna pažnja posvećena je proveriti robusnosti rezultata u odnosu na potencijalne eksremne vrednosti i nelinearne devijacije, koje su prethodno ublažene transformacijama podataka opisanim u Odeljku 5.2.5. Time je obezbeđeno da korelacione procene reflektuju strukturalne odnose u podacima, a ne artefakte distribucije.

Rezultati ove inicijalne analize ukazali su na izuzetno jasan obrazac: koncentracija finih suspendovanih čestica $PM_{2.5}$ pokazala je najjaču pojedinačnu korelaciju sa R_0 od svih analiziranih varijabli. Uočena je velika pozitivna korelacija ($R \approx 0,56$, $p \approx 10^{-4}$), što znači da su države sa višim nivoima $PM_{2.5}$ tokom inicijalne faze epidemije dosledno pokazivale veće vrednosti R_0 . Veličina ovog efekta je statistički i epidemiološki značajna: korelacija ovog intenziteta retko se pojavljuje u populacionim studijama sa velikim brojem konfundirajućih faktora, što već u ovoj fazi sugeriše da zagađenje vazduha nije marginalna, već potencijalno centralna determinanta transmisije.

Važno je naglasiti da se korelacija odnosi upravo na inicijalni eksponencijalni režim rasta, pre punog efekta javnozdravstvenih intervencija. Time se smanjuje uticaj ponašajnih i političkih faktora i omogućava da se izoluje biološko-ekološka komponenta prenosivosti virusa. Drugim rečima, uočena veza između $PM_{2.5}$ i R_0 odnosi se na inherentnu dinamiku transmisije u populaciji, a ne na kasnije efekte zaključavanja, testiranja ili zdravstvene infrastrukture.

Pored $PM_{2.5}$, statistički značajne korelacije sa R_0 ($p < 0,05$) pokazale su i sledeće varijable (Ilustracija 13): gustina stanovništva (PC2 density, pozitivna korelacija), prevalenca hroničnih bolesti (PC1 disease, pozitivna), CO (pozitivna), komponenta koja odražava pokrivenost zdravstvenim osiguranjem (PC2 ins., negativna), prosperitet države (PC1 prosperity, negativna), kao i procenat crnog stanovništva (pozitivna). Međutim, pažljivim pregledom matrice međusobnih korelacija utvrđeno je da su mnoge od ovih varijabli istovremeno značajno korelisane i sa $PM_{2.5}$, i to u istom smeru. Na primer, urbanizovanije i industrijalizovanije države imaju istovremeno veće $PM_{2.5}$ koncentracije, veću gustinu stanovništva i različite profile hroničnih bolesti. Posledica toga je da njihove značajne korelacije sa R_0 mogu biti — bar delimično — posledica korelacije sa $PM_{2.5}$, a ne nezavisnog efekta. Drugim rečima, inicijalna analiza otkriva snažan signal, ali ne razdvaja pitanje da li $PM_{2.5}$ direktno utiče na transmisiju, da li je proksi za urbanizaciju i socijalne obrasce kontakta, ili predstavlja deo šireg sindroma životne sredine i zdravstvenog opterećenja. Upravo zbog toga, sledeći korak analize zahteva prelazak sa univarijantnog na multivarijantni okvir koji može kontrolisati simultani uticaj velikog broja međusobno korelisanih faktora.



Ilustracija 13: Korelaciona analiza između koncentracije $PM_{2.5}$ i osnovnog reprodukcionog broja R_0 na nivou saveznih država SAD. Pozitivan nagib regresione linije konzistentan je kroz celokupan distribucioni raspon uzorka

Predloženi mehanizmi uključuju: vezivanje virusnih čestica za čestice $PM_{2.5}$ čime se produžava njihovo zadržavanje u vazduhu i povećava dubina prodiranja u respiratorni trakt, povećanje ekspresije ACE-2 receptora usled izloženosti česticama, te opšte slabljenje lokalnog imunog odgovora respiratornog epitela (Comunian et al., 2020; Paital & Agrawal, 2021; Qu et al., 2020). Važno je napomenuti da je u ovoj analizi korišćena kratkoročna (akutna) vrednost $PM_{2.5}$ u toku inicijalnog epidemijskog prozora — a ne dugoročni proseka — što više govori u prilog mehaničke hipoteze (vezivanje čestica) nego hipoteze o hroničnom imunosupresivnom efektu.

Svi ovi mehanizmi potencijalno povećavaju efektivnu verovatnoću transmisije pri svakom kontaktu. Stoga korelacioni nalaz nije samo statistički već i fiziološki plausibilan, što dodatno opravdava njegovo dublje ispitivanje. Zbog toga se u narednim odeljcima prelazi na regularizovane regresione modele i nelinearne ansambl metode, koje omogućavaju rigorozniju procenu stvarne uloge zagađenja vazduha u kontekstu celokupnog ekološko-demografskog sistema.

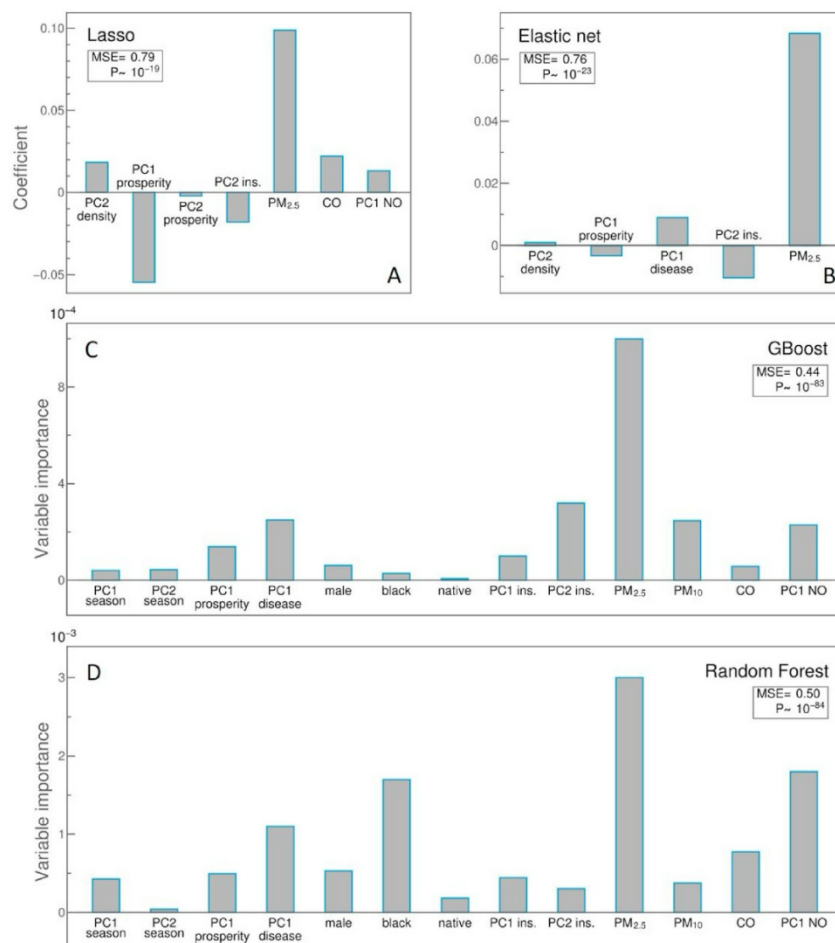
5.3.2 Selekcija obeležja putem regularizovane regresije

Nakon inicijalne korelacione analize, analiza je proširena na multivarijantni okvir primenom Lasso i Elastic Net regresija. Ove metode istovremeno vrše penalizaciju složenosti modela i automatsku selekciju varijabli skupljanjem koeficijenata neinformativnih prediktora na tačnu nulu, što omogućava identifikaciju nezavisnih doprinosa u prisustvu visoko korelisanih prediktora. Sve varijable su standardizovane pre unosa u modele, tako da apsolutne vrednosti koeficijenata direktno reflektuju relativnu važnost prediktora.

Oba modela su statistički visoko značajno nadmašila bazni konstantni model: Lasso ($MSE = 0.79$, $p \sim 10^{-19}$) i Elastic Net ($MSE = 0.76$, $p \sim 10^{-23}$). Prediktivna snaga je umerena, što je i očekivano s obzirom na ograničen broj uzoraka (46 država) i inherentnu kompleksnost sistema — cilj ovih modela bio je pre svega selekcija i rangiranje prediktora, a ne maksimizacija prediktivne tačnosti.

5.3.2.1 Dominantna uloga $PM_{2.5}$ u multivarijantnom kontekstu

Rezultati su bili izuzetno konzistentni između oba modela (Ilustracija 14). U svim optimizovanim verzijama modela, koncentracija $PM_{2.5}$ je zadržala najveći pozitivan koeficijent i bila prva varijabla koja ulazi u model tokom regularizacije. Ovo je ključan nalaz: efekat zagađenja vazduha ne nestaje nakon kontrole za desetine drugih faktora, već ostaje centralni prediktor transmisije. Drugim rečima, $PM_{2.5}$ nije samo korelisani pratilac urbanizacije ili gustine stanovništva — njegova veza sa R_0 se održava čak i kada se ti faktori eksplicitno uključe u model. Ovakva stabilnost kroz različite algoritamske okvire predstavlja snažan indikator robusnosti nalaza.



Ilustracija 14: Važnost prediktora u ansambl modelima mašinskog učenja. Prikazano je rangiranje prediktora prema doprinosu smanjenju predikcione greške u Random Forest i Gradient Boosting algoritimima. PM_{2.5} zauzima dominantnu poziciju u oba modela, sa jasnom separacijom od sekundarnih prediktora koji uključuju komponente hroničnih bolesti, okside azota i demografske faktore.

5.3.2.2 Sekundarni prediktori: višedimenzionalni profil rizika

Pored PM_{2.5}, oba modela su selektovala konzistentan skup sekundarnih prediktora (Ilustracija 14):

CO i PC1 NO (komponenta formirana od NO₂ i oksida azota) pojavljuju se kao pozitivni prediktori u oba modela, što sugeriše da efekat nije vezan isključivo za jednu hemijsku frakciju već odražava širi zagađivački ambijent urbanih i industrijskih sredina. CO je bio statistički značajan i u inicijalnoj korelacionoj analizi.

Gustina stanovništva (PC2 density) pojavljuje se kao pozitivan prediktor u oba modela, mada sa manjim koeficijentom od PM_{2.5}, što je u skladu sa teorijom mreža kontakata. Prosperitet države (PC1 prosperity) pokazuje negativan koeficijent kroz oba modela, što znači da bogatije države beleže niži R₀. Ovaj nalaz, naizgled u suprotnosti sa rezultatima prethodne studije, verovatno odražava da unutar relativno homogenog skupa razvijenih američkih država prosperitet postaje proksi za protektivno ponašanje, mogućnost rada od kuće i bolje

zdravstvene uslove — za razliku od globalnih analiza gde prosperitet korelira sa većom mobilnošću i brojem kontakata.

Prevalenca hroničnih bolesti (PC1 disease) selektovana je kao pozitivan prediktor u Elastic Net modelu. Iako hronične bolesti primarno utiču na težinu ishoda infekcije, moguće je da indirektno povećavaju transmisiju kroz produženu infektivnost ili češće zdravstvene kontakte.

Komponenta zdravstvenog osiguranja (PC2 ins.), koja efektivno odražava pokrivenost osiguranjem u hispanskoj populaciji, pokazuje negativnu korelaciju sa R_0 kroz oba modela — što je konzistentno sa nalazima inicijalne korelacione analize.

5.3.2.3 Metodološke napomene

Ključna vrednost ovog koraka nije samo u identifikaciji pojedinačnih faktora, već u potvrdi da efekat zagađenja opstaje nakon penalizacije modela. Regularizacija sistematski „kažnjava” varijable koje ne doprinose nezavisnoj informaciji. Da je $PM_{2.5}$ bio samo sporedni proksi za druge faktore, bio bi eliminisan tokom optimizacije. Njegovo zadržavanje kao glavnog prediktora predstavlja snažnu implikaciju njegove nezavisne uloge.

Takođe je važno naglasiti da se Lasso i Elastic Net razlikuju u načinu tretiranja korelisanih varijabli. Lasso teži da zadrži jednu reprezentativnu varijablu iz grupe, dok Elastic Net dozvoljava deljenje težine između njih. Činjenica da oba modela ističu isti skup ključnih prediktora dodatno povećava poverenje u stabilnost rezultata.

Iako regularizovana regresija pruža jasnu i interpretabilnu sliku, ona ostaje linearna metoda. Međutim, biološki i ekološki sistemi retko funkcionišu linearno. Efekti zagađenja, gustine i zdravstvenog statusa verovatno uključuju pragove, interakcije i saturacione efekte. Zbog toga se u sledećem odeljku analiza proširuje na nelinearne ansambl modele (Random Forest i Gradient Boost), koji mogu detektovati kompleksnije obrasce bez nametanja unapred definisane funkcionalne forme.

5.3.3 Značaj varijabli u nelinearnim ansambl modelima

Nakon regularizovanih linearnih regresija, analiza je proširena na dva nelinearna ansambl modela — Random Forest i Gradient Boosting — koji ne pretpostavljaju unapred definisanu funkcionalnu formu odnosa između prediktora i R_0 , već je uče direktno iz podataka. Primena ovih metoda imala je dvostruki cilj: nezavisnu proveru nalaza dobijenih linearnim modelima, i procenu da li deo varijabilnosti R_0 potiče iz nelinearnih odnosa koje linearni okvir ne može uhvatiti.

Pre primene ovih metoda, radi izbegavanja preprilagođavanja, sprovedena je preliminarna selekcija varijabli: zadržano je 13 varijabli koje su pokazale statistički značajnu korelaciju sa R_0 ($p < 0,1$) prema bar jednoj od tri mere — Pearsonovoj, Kendalovoj ili Spirmanovoj korelaciji.

5.3.3.1 Prediktivna tačnost

Oba nelinearna modela pokazala su značajno bolju prediktivnu tačnost u odnosu na linearne pristupe. Gradient Boosting postigao je $MSE = 0,44$, a Random Forest $MSE = 0,50$, u poređenju sa $0,79$ i $0,76$ kod Lasso i Elastic Net modela. Oba poboljšanja statistički su visoko značajna u odnosu na konstantni model ($p < 0,001$). Ovaj dobitak u prediktivnoj snazi potvrđuje da u podacima postoje nelinearne strukture koje linearni modeli ne mogu u potpunosti uhvatiti, iako treba napomenuti da je i ovde prediktivna tačnost umerena — što je očekivano s obzirom na kompleksnost epidemioloških sistema i ograničen broj uzoraka.

5.3.3.2 Dominantni prediktor: ponovna potvrda $PM_{2.5}$

U oba ansambl modela $PM_{2.5}$ je ubedljivo najvažniji prediktor R_0 , sa značajnom marginom u odnosu na sve ostale varijable. Ovo predstavlja nezavisnu potvrdu nalaza iz korelacione i linearne analize, dobijenu metodama koje su suštinski drugačije u algoritmskom pristupu i pretpostavkama. Konzistentnost $PM_{2.5}$ kao dominantnog prediktora kroz sva četiri primenjena metodološka okvira — korelaciona analiza, Lasso, Elastic Net, Random Forest i Gradient Boosting — statistički čini slučajno pojavljivanje ovog obrasca izuzetno malo verovatnim.

5.3.3.3 Sekundarni prediktori i razlike između modela

Pored primarnog efekta $PM_{2.5}$, ansambl modeli su identifikovali nekoliko dodatnih faktora, pri čemu postoje određene razlike između dva modela koje su same po sebi informativan nalaz:

- PC1 disease (prevalenca hroničnih bolesti) pojavljuje se sa umerenim značajem u oba modela, konzistentno sa Elastic Net rezultatima.
- PC1 NO (azotni oksidi) pokazuje umerenu važnost u oba modela. Naročito, Random Forest mu dodeljuje relativno visok značaj, što sugerise da doprinos azotnih oksida može biti nelinearne prirode.
- PM_{10} pojavljuje se kao umereno važan prediktor u Gradient Boosting modelu, ali nije selektovan od strane ostalih metoda. Ovo je u skladu sa biološkom hipotezom da čestice veće od $5 \mu m$ teže prodiru do ACE-2 receptora u alveolarnim ćelijama tipa II, te je njihov efekat slabiji i manje stabilan između modela.
- CO pokazuje umerenu važnost u Random Forest modelu, konzistentno sa korelacionom i linearnom analizom, ali se ne pojavljuje kao važan u Gradient Boosting modelu.
- Procenat crnog stanovništva selektovan je kao važan prediktor u Random Forest modelu, i bio je statistički značajan u inicijalnoj korelacionoj analizi, ali se ne pojavljuje kao važan u Gradient Boosting modelu. Ova varijabla je visoko korelisana sa $PM_{2.5}$, što otežava razlučivanje nezavisnog efekta — pitanje koje je u literaturi identifikovano kao inherentno teško rešivo zbog strukturne isprepletenosti rasnih, ekonomskih i ekoloških nejednakosti u SAD (Chakraborty, 2021).
- PC2 ins. (pokrivenost zdravstvenim osiguranjem) pojavljuje se u Gradient Boosting modelu sa umerenim značajem, konzistentno sa prethodnim analizama.

Važno je napomenuti da ansambl modeli tipično raspodeljuju važnost između korelisanih varijabli, umesto da biraju jednu reprezentativnu kao što to čini Lasso. Stoga se sekundarni

prediktori moraju interpretirati kao deo klastera međusobno povezanih faktora, a ne kao izolovani nezavisni efekti.

5.3.3.4 Odsustvo meteoroloških faktora

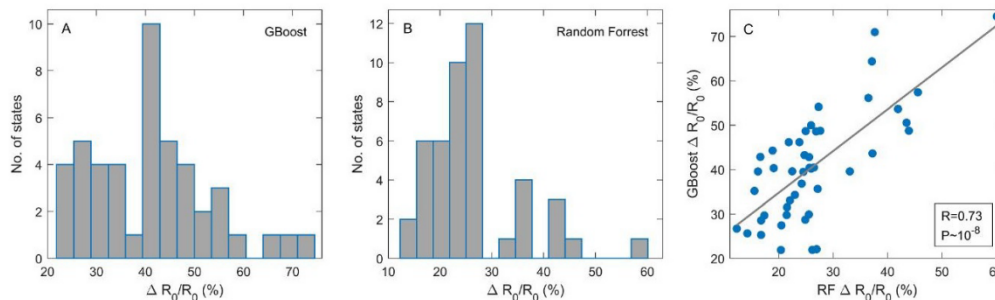
Značajno je i to što sezonske i meteorološke komponente nisu ušle u finalni skup važnih prediktora ni u jednom od primenjenih modela. Ovo sugeriše da uticaj temperature, UV zračenja i vlažnosti na transmisiju SARS-CoV-2 — u okviru varijacija koje postoje između američkih država — nije dominantan signal, što je u skladu sa nalazima pojedinih studija koje dovode u pitanje klimatsku zavisnost COVID-19 transmisije (Wang et al., 2021). Treba, međutim, imati na umu da su varijacije meteoroloških faktora između američkih država znatno manje nego na globalnom nivou, gde ti faktori pokazuju veći značaj (Salom et al., 2021).

5.3.4 Kvantifikacija uticaja zagađenja vazduha na R_0

Nakon što je $PM_{2.5}$ konzistentno identifikovana kao dominantni prediktor kroz sve primenjene metode, sledeći korak bio je prelazak sa kvalitativnog rangiranja važnosti varijabli na kvantitativnu procenu veličine efekta. Visoka prediktivna tačnost treniranih ansambl modela ($MSE = 0,44$ za Gradient Boosting i $MSE = 0,50$ za Random Forest) omogućila je generisanje kontrafaktuelnih predikcija: za svaku od 46 analiziranih država R_0 je predviđen za čitav opseg vrednosti zagađenja koji je stvarno zabeležen u svim ostalim državama, uz zadržavanje svih ostalih prediktora na stvarnim vrednostima. Na taj način izolovan je čist doprinos varijacija u zagađenju promeni R_0 , dok su ostali faktori kontrolisani.

5.3.4.1 Rezultati kvantifikacije

Relativna promena R_0 usled varijacija u zagađenju ($\Delta R_0/R_0$) procenjena je za sve analizirane države, pri čemu ΔR_0 odgovara razlici između maksimalne i minimalne predviđene vrednosti R_0 . Rezultati su prikazani u vidu histograma za oba modela (Ilustracija 15). Gradient Boosting dao je nešto veće procene sa medijanom $\Delta R_0/R_0$ od $\sim 40\%$ (vrednosti idu do $\sim 70\%$), dok je Random Forest dao konzervativnije procene sa medijanom od $\sim 25\%$. Direktno poređenje predikcija dva modela za iste države pokazuje visoku konzistentnost ($R = 0,73$, $p \sim 10^{-8}$), što potvrđuje da razlike između modela ne utiču na kvalitativni zaključak. Uzimajući oba modela zajedno, kao gruba procena uticaja zagađenja uzima se vrednost od $\sim 30\%$ promene R_0 .



Ilustracija 15: Distribucija procenjene relativne promene osnovnog reprodukcionog broja ($\Delta R_0/R_0$) po savezima SAD, dobijena kontrafaktualnom simulacijom u Random Forest i Gradient Boosting modelima. Realistične razlike u $PM_{2.5}$ koncentracijama proizvode promenu R_0 u rasponu od 25 do 40%, konzistentno između oba modela.

Važno je napomenuti da ova kvantifikacija nije ograničena samo na $PM_{2.5}$ — u ansambl modelima figurišu i ostali zagađivači (PM_{10} , CO, NO_2), ali su promene u $PM_{2.5}$ ubedljivo najvažnija komponenta ukupnog efekta zagađenja na R_0 .

5.3.4.2 Epidemiološka interpretacija veličine efekta

Da bi se razumela praktična težina procene od $\sim 30\%$, korisno je staviti je u kontekst. Novi soj SARS-CoV-2 B.1.1.7 (Alpha), koji je u vreme sprovođenja analize postajao dominantan u velikom broju zemalja, procenjen je da dovodi do povećanja R_0 za do $\sim 90\%$ (Davies et al., 2021). Efekat zagađenja od $\sim 30\%$ je, dakle, manji od efekta ovog visoko infektivnog soja, ali svakako nije zanemarljiv — predstavlja otprilike trećinu tog efekta.

Pošto se epidemijski rast u ranoj fazi ponaša eksponencijalno, relativno male promene u R_0 imaju neproporcionalno velike posledice po broj zaraženih. Na primer, u eksponencijalnom režimu koji traje ~ 10 dana, što je tipičan period inicijalnog rasta zabeležen u analiziranim državama, razlika od $\sim 30\%$ u R_0 i dva pretpostavljena ciklusa zaražavanja (tj. primarni zaraženi zaraze sekundarne pa zatim sekundarni tercijalne) može rezultovati 70% većim brojem zaraženih ($1,3 * 1,3 = 1,69$). Ovo ima direktne implikacije za vršno opterećenje zdravstvenog sistema i ukupan tok epidemijskog talasa.

5.3.4.3 Robusnost procene

Konzistentnost rezultata između Random Forest i Gradient Boosting modela, koji se suštinski razlikuju u algoritmskom pristupu, predstavlja snažnu indikaciju da nalaz nije artefakt specifične metodologije. Oba modela su nezavisno i uz visoku statističku značajnost ($p \sim 10^{-83}$ i 10^{-84}) nadmašila konstantni model, što snažno ukazuje da varijabilnost u zagađenju nosi stvarnu prediktivnu informaciju o R_0 , a ne statistički šum. Pored toga, analiza je sprovedena na skupu koji obuhvata 46 država sa različitim geografskim, demografskim i ekonomskim profilima, što povećava eksternu validnost procene.

5.4 Diskusija i zaključna razmatranja

5.4.1 Ekološko-epidemiološki okvir transmisije COVID-19

Rezultati ove studije ukazuju na konzistentan i metodološki robustan zaključak: kvalitet vazduha, a posebno koncentracija finih čestičnih materija $PM_{2.5}$, predstavlja jednu od najvažnijih ekoloških determinanti rane transmisije SARS-CoV-2. Ovaj nalaz se pojavljuje kroz više nezavisnih analitičkih pristupa — korelacione analize, regularizovane regresije i nelinearnih ansambl modela — što značajno povećava njegovu kredibilnost. U kontekstu epidemioloških studija koje uključuju veliki broj konfundirajućih faktora, ovakav nivo metodološke saglasnosti je izuzetno redak i zahteva pažljivu interpretaciju koja povezuje ekologiju, fiziologiju i dinamiku epidemija.

5.4.1.1 Zagađenje kao multiplikator transmisije: Biološki i mehanički mehanizmi

Centralni nalaz studije — da $PM_{2.5}$ konzistentno dominira kao prediktor R_0 kroz pet nezavisnih analitičkih okvira — sugerise da aerozagađenje deluje kao multiplikator transmisije. Postojeća eksperimentalna literatura predlaže nekoliko komplementarnih mehanizama:

- **Mehanička facilitacija (Aerosol-carrier hipoteza):** Virusne čestice se mogu vezati za $PM_{2.5}$, čime se produžava njihovo zadržavanje u vazduhu i povećava opseg difuzije. Čestice manje od 2,5 mikrona prodiru duboko u alveolarne regione do nivoa na kojima se nalaze ACE-2 receptori u ćelijama tipa II (Comunian et al., 2020).
- **Molekularna podložnost:** Izloženost česticama dovodi do povećane ekspresije ACE-2 receptora i enzima TMPRSS2 u respiratornim ćelijama, čime se direktno povećava verovatnoća vezivanja i ulaska virusa u ćeliju (Sagawa et al., 2021).
- **Hronična inflamacija i oštećenje barijere:** Prolongirana izloženost izaziva inflamaciju epitela i oštećenje mukociliarnog klirensa, što slabi lokalnu imunološku barijeru (Glencross et al., 2020).

Relativno mali značaj CO i NO_x u poređenju sa $PM_{2.5}$ u ovoj analizi više govori u prilog **mehaničke hipoteze** — da fine čestice fizički facilitiraju prenos kroz vazduh — nego hipoteze o generalnoj imunosupresiji. Razlika u značaju između $PM_{2.5}$ i PM_{10} je posebno instruktivna: čestice veće od 5 mikrona ne dospevaju do alveolarnih ACE-2 receptora u dovoljnoj meri da bi ostvarile isti efekat (Y. Zhu et al., 2020).

5.4.1.2 Interpretacija sekundarnih prediktora i socioekonomski kontekst

Pored dominantnog uticaja $PM_{2.5}$, analiza je identifikovala skup faktora koji opisuju kompleksan profil transmisije:

1. **Ostali zagađivači:** Azotni oksidi (NO_2 , NO_x) se pojavljuju kao pozitivni prediktori, posebno u nelinearnim modelima (Random Forest), što sugerise da doprinos azotnih oksida može biti nelinearne prirode.
2. **Hronične bolesti:** Prevalenca bolesti (gojaznost, hipertenzija, dijabetes) identifikovana je kao pozitivan prediktor osnovnog reproduktionog broja. Iako su primarno faktori rizika za težinu ishoda, ovi komorbiditeti mogu uticati na R_0 kroz

produženo trajanje infektivnog perioda ili češće kontakte unutar zdravstvenog sistema (Liu et al., 2020).

3. **Paradoks prosperiteta:** Komponenta prosperiteta (BDP, HDI, očekivano trajanje života) konzistentno pokazuje negativnu korelaciju sa R_0 . Unutar homogenog skupa američkih država, prosperitet postaje proksi za protektivno ponašanje (mogućnost rada od kuće, bolji pristup informacijama) pre nego za mobilnost, što je suprotno nalazima globalnih studija (Salom et al., 2021).
4. **Rasni sastav i zdravstveno osiguranje:** Procenat crnog stanovništva i nedostatak osiguranja koreliraju sa bržom transmisijom. Ipak, interpretacija je složena jer su ovi faktori strukturno povezani sa lokacijama visokog zagađenja (Chakraborty, 2021). Činjenica da $PM_{2.5}$ zadržava veći značaj od ovih varijabli sugerise da zagađenje ima nezavisan doprinos koji prevazilazi socioekonomske nejednakosti.

5.4.1.3 Hijerarhija pokretača transmisije: globalna naspram regionalne skale

Poređenje nalaza ove studije sa rezultatima prikazanim u Poglavlju 4 otkriva konceptualno važnu hijerarhijsku strukturu pokretača transmisije zavisnu od analitičke skale posmatranja. Ova hijerarhija nije slučajna niti stvar artefakta — ona odražava fundamentalnu osobinu kompleksnih epidemioloških sistema u kojima različiti faktori dominiraju na različitim nivoima prostorne i temporalne agregacije.

U okviru analize sprovedene u ovom poglavlju, koja je bila fokusirana na teritoriju Sjedinjenih Američkih Država, geografska i klimatska varijabilnost između analiziranih jedinica bila je relativno ograničena u poređenju sa globalnom skalom. Pored toga, analiza je bila vremenski fokusirana na inicijalni prolećni period 2020. godine, što je efektivno kontrolisalo za sezonski signal budući da su sve analizirane jedinice bile izložene istoj sezoni u istom vremenskom trenutku. U takvim uslovima, kada je sezonski signal neutralisan vremenskim i prostornim ograničenjem analize, zagađenje vazduha mereno koncentracijom finih čestičnih materija $PM_{2.5}$ iskristalisalo se kao dominantan regionalni prediktor transmisibilnosti SARS-CoV-2. Ovaj nalaz verovatno reflektuje i stvarni biološki efekat zagađenja vazduha na transmisibilnost virusa i respiratornu ranjivost populacije.

Na globalnoj skali, nasuprot tome, ogromna klimatska varijabilnost 118 zemalja koje obuhvataju tropske, suptropske, umerene i polarne klimatske zone generisala je sezonski signal toliko jak da je u prediktivnim modelima zasenio lokalizovanije efekte zagađenja vazduha i ostalih regionalnih prediktora. Kada se u isti analitički okvir uključe zemlja sa tropskom klimom i zemlja sa subarktičkom klimom, razlike u meteorološkim i ekološkim uslovima između njih toliko su velike da dominiraju nad svim ostalim izvorima varijabilnosti. U takvom kontekstu, klimatske i meteorološke promenljive — temperatura, vlažnost, UV zračenje i brzina vetra — nužno se pojavljuju kao najsnažniji prediktori jer enkoduju najveći deo ukupne varijanse u vrednostima R_0 između zemalja. Rezultati sa lokalne i globalne skale su naizgled neusaglašeni po intenzitetu efekta na R_0 i značaju pojedinih prediktivnih domena. Manja amplituda varijacije R_0 izazvana klimom u poređenju sa zagađenjem je verovatno odraz slabljenja signala na većim teritorijama celih država i heterogenijim podacima. Različiti dominantni signali potiču od manje varijacije klime u SAD sa jedne strane i loših globalnih podataka o zagađenju sa druge.

Ova skalno-zavisna hijerarhija pokretača transmisije ima važne implikacije za dizajn budućih epidemioloških studija i za interpretaciju njihovih nalaza. Studije koje se fokusiraju na relativno homogene geografske regione u kratkim vremenskim okvirima — poput analiza unutar jedne države ili kontinenta — verovatno će identifikovati lokalne socioekonomske, demografske i ekološke faktore kao dominantne prediktore, budući da je sezonski signal u takvim kontekstima neutralisan ili minimizovan. Nasuprot tome, globalne komparativne studije koje obuhvataju širok spektar klimatskih zona nužno će biti dominirane sezonskim i klimatskim signalom koji maskira lokalizovanije efekte. Ni jedan od ova dva pristupa sam po sebi nije metodološki superioran — oni odgovaraju na različita istraživačka pitanja i otkrivaju različite slojeve etiološke strukture transmisije.

Konceptualno, ova hijerarhija sugerise da sezonalnost postavlja široki, sveobuhvatni ritam transmisibilnosti na planetarnom nivou, unutar kojeg faktori poput kvaliteta vazduha, socioekonomskog statusa, demografske strukture i zdravstvenih karakteristika populacije moduliraju bazični rizik na regionalnom i nacionalnom nivou. Sezonalnost definiše gornju i donju granicu transmisijskog potencijala u datom periodu godine za datu geografsku lokaciju, dok lokalni faktori određuju gde unutar tog opsega se nalazi konkretna zemlja ili region. Razumevanje ove dvoslojne strukture pokretača transmisije ključno je kako za razvoj preciznih modela predikcije koji su validni na različitim prostornim skalama, tako i za formulisanje kontekstualno prilagođenih javnozdravstvenih strategija koje uzimaju u obzir i univerzalne sezonske determinante i lokalno specifične faktore ranjivosti populacije.

5.4.2 Metodološki doprinos i robusnost pristupa

Ova studija adresira ključne zamerke upućivane prethodnim istraživanjima (Bontempi, 2020; Villeneuve & Goldberg, 2020). Upotreba R_0 procenjenog isključivo iz inicijalne ekspanzionalne faze ublažava konfundiranje koje nastaje usled razlika u testiranju, kasnijih intervencija (*lockdown*) ili pojave novih sojeva.

Korišćenje **metodološkog pluralizma** — primena pet suštinski različitih pristupa (Pirson, Lasso, Elastic Net, Random Forest, Gradient Boost) — obezbeđuje da zaključci nisu statistički artefakt jedne metode. Primena PCA unutar tematskih blokova omogućila je upravljanje visokom dimenzionalnošću (74 varijable) bez gubitka interpretabilnosti, što predstavlja značajan iskorak u analizi kompleksnih epidemioloških sistema.

5.4.3 Ograničenja studije i merna greška

Svaka kvantitativna analiza kompleksnih sistema nosi određena ograničenja koja definišu prostor interpretacije:

- **Ekološki dizajn i agregacija:** Analiza na nivou saveznih država nosi rizik od "ekološke zablude". Odnosi na makro nivou ne moraju nužno odražavati individualne mehanizme. Iako populaciono ponderisana agregacija smanjuje problem, ona maskira mikrolokalne varijacije unutar država.

- **Prostorna i temporalna heterogenost:** Merne stanice su gušće u urbanim sredinama, što uvodi mernu grešku u ruralnim područjima. Takođe, temporalna agregacija zagađenja kroz "eksponencijalni prozor" maskira akutne vrhove zagađenja koji mogu imati neproporcionalno veliki biološki efekat (Glencross et al., 2020).
- **Unutrašnje zagađenje:** Pošto ljudi provode 80–90% vremena u zatvorenom prostoru, spoljašnje merenje je samo aproksimacija stvarnog izlaganja. Iako korelacija postoji, ona varira između tipova objekata (Glencross et al., 2020).
- **Problem uzročnosti i konfundiranja:** Strukturna isprepletenost zagađenja, rase i ekonomije u SAD čini potpuno razdvajanje ovih efekata inherentno teškim. Poseban izazov je precizno modelovanje mobilnosti, koja je ovde indirektno kontrolisana kroz komponente prosperiteta.

Važno je napomenuti da merna greška u prediktoru tipično dovodi do atenuacije efekta ka nuli, što se manifestuje manjim koeficijentima (Wooldridge, 1999). To sugeriše da su naše procene o uticaju $PM_{2.5}$ od ~30% verovatno konzervativne, a ne liberalne.

5.4.4 Pravci budućih istraživanja

Na osnovu identifikovanih ograničenja, definišu se sledeći pravci:

1. **Visoko-rezolutne analize:** Prelazak na nivo okruga (*counties*) ili gradova radi preciznije prostorne rezolucije.
2. **Longitudinalni kvazi-eksperimenti:** Korišćenje prirodnih eksperimenata (npr. nagli pad zagađenja tokom *lockdown-a*) za snažnije kauzalno zaključivanje.
3. **Direktno merenje unutrašnje izloženosti:** Korelacija seropozitivnosti sa direktnim merenjima $PM_{2.5}$ u domaćinstvima.
4. **Eksperimentalna validacija:** Laboratorijsko testiranje infektivnosti virusa vezanog za čestice i dozno-zavisnog odgovora ACE-2 ekspresije.
5. **Generalizacija:** Replikacija okvira u regionima sa ekstremno visokim zagađenjem (Azija, Afrika) radi testiranja linearnosti odnosa.

5.4.5 Zaključak: Šire implikacije za javno zdravlje

Glavni nalaz studije — da varijacija $PM_{2.5}$ tipična za SAD odgovara razlici u R_0 od ~30% — svrstava aerozagađenje u isti red veličine sa brojnim drugim faktorima kojima se posvećuje znatno veća pažnja. U eksponencijalnoj fazi, ovakva razlika može rezultovati dvostruko većim brojem zaraženih u kratkom vremenskom intervalu.

Ova studija potencijalno transformiše smanjenje zagađenja vazduha iz čisto ekološkog pitanja u **primarnu epidemiološku intervenciju**. Ako kvalitet vazduha moduliše transmisibilnost respiratornih patogena u ovoj meri, tada politike urbanog planiranja i regulacija emisija PM čestica postaju ključni alati za povećanje otpornosti populacije na buduće pandemijske pretnje. Strategija koja integriše ekološku zaštitu i javnozdravstvenu pripravnost onda nije luksuz, već racionalan odgovor na empirijski dokumentovanu neraskidivu vezu između kvaliteta životne sredine i epidemiološke dinamike.

6 Opšta diskusija

Prethodna poglavlja su predstavila dve velike empirijske istrage pokretača prenosivosti SARS-CoV-2 u aspektu faktora životne sredine i sociodemografskih faktora. Prva studija je razmatrala globalnu skalu obuhvatajući 118 zemalja, dok se druga fokusirala na Sjedinjene Američke Države, pružajući analizu visoke rezolucije unutar jedne nacije. Ovo poglavlje sintetiše nalaze iz ove dve različite ali komplementarne studije, diskutuje njihove kolektivne metodološke doprinose i istražuje njihove šire implikacije za politiku javnog zdravlja, pripravnost za pandemiju i buduća istraživanja.

6.1 Sinteza ključnih nalaza

Centralni doprinos ove teze je identifikacija dva moćna, kontekstno zavisna pokretača transmisije COVID-19 iz životne sredine: **sezonalnost na globalnom nivou i zagađenje vazduha u okolini na regionalnom nivou.**

6.1.1 Sveobuhvatni uticaj sezonalnosti na globalnoj skali

Kada je pandemija posmatrana na globalnoj skali u Poglavlju 4, pojavio se prvi jaki signal: **sezonalnost**. Globalni model, uprkos tome što je treniran samo na podacima iz prvih nekoliko nedelja pandemije, naučio je da predvidi punu godinu prenosivosti koja je sledila prepoznatljiv, hemisferski zavisani sinusoidalni obrazac. R_0 je predviđen da dostigne vrhunac tokom zimskih meseci i dosegne minimum tokom letnjih meseci i u Severnoj i Južnoj hemisferi, sa obrascima koji su bili savršeno invertovani između njih. Ovaj nalaz, statistički pojačan kroz Bajesijansku MCMC analizu, potencijalno podiže sezonalnost od potencijalne konfundirajuće varijable do dominantnog, predvidljivog pokretača globalne dinamike COVID-19, uprkos ograničenom obimu podataka o zagađenju.

6.1.2 Zagađenje vazduha kao konzistentan pokretač transmisije

Studija fokusirana na SAD predstavljena u Poglavlju 5 je pružila jasne i robusne argumente da su **fine čestične materije ($PM_{2.5}$) bile pojedinačno najvažniji prediktor osnovnog reproduktionog broja (R_0)** između država tokom inicijalne faze pandemije. Ovaj nalaz nije bila prosta korelacija; održao se kroz niz modela mašinskog učenja, od regularizovanih regresija do visokoperformantnih nelinearnih ansambla, konzistentno nadmašujući širok niz drugih potencijalnih prediktora. Kvantitativna analiza je sugerisala da varijacije u nivoima $PM_{2.5}$, unutar opsega tipično opserviranog kroz SAD, mogu promeniti virusnu prenosivost za značajnu marginu od 25-40%. Ovo uspostavlja jaku empirijsku vezu između kvaliteta vazduha i širenja SARS-CoV-2, sugerišući da slab kvalitet vazduha može značajno amplifikovati transmisiju respiratornih virusa.

6.1.3 Poređenje lokalnih (SAD) i globalnih pokretača pandemije

Na prvi pogled, primarni nalazi dve studije—zagađenje u SAD, sezonalnost na globalnoj skali—mogu izgledati kontradiktorni. Međutim, oni se bolje razumeju kao komplementarni rezultati koji ističu važnost analitičke skale. SAD, iako geografski velika država, su relativno homogene u svom opsegu klima u poređenju sa celom planetom. Fokus studije SAD na inicijalno prolećno širenje 2020. godine je efektivno "kontrolisao" za veći sezonski ciklus, omogućavajući da fini, ali ipak moćan signal regionalnih varijacija u zagađenju vazduha postane jasno vidljiv.

Na globalnoj skali, varijacija u klimi i sezonskoj vremenskoj raspodeli kroz 118 zemalja je kreirala jak signal koji je nadjačao lokalizovanije efekte zagađenja predstavljene nepotpunim podacima u prediktivnim modelima. Ovo sugeriše hijerarhijsku relaciju: sezonalnost postavlja širok, sveobuhvatni ritam prenosivosti globalno, dok faktori kao što su kvalitet vazduha i sociodemografski uslovi moduliraju bazični rizik na regionalnom ili nacionalnom nivou.

6.2 Nalazi u kontekstu evoluirajućeg pandemijskog pejzaža

Ovo poglavlje smešta nalaze ove teze unutar ogromnog i složenog korpusa naučne literature koja se pojavila od 2020. godine. Cilj je da se kritički evaluiira kako se naše razumevanje međugre između faktora životne sredine—primarno zagađenja vazduha i sezonalnosti—i transmisije i težine SARS-CoV-2 razvilo.

6.3 Metodološki doprinosi i inovacije

Pored specifičnih epidemioloških nalaza, ova teza čini nekoliko važnih metodoloških doprinosa polju.

Prvo, konzistentna upotreba osnovnog reproduktionog broja, procenjenog iz inicijalne faze eksponencijalnog rasta, predstavlja značajno poboljšanje u odnosu na mnoge savremene studije koje su se oslanjale na sirove brojeve slučajeva ili mortaliteta. Ovaj pristup efektivno ublažava pristrasnosti izazvane politikom testiranja i standardom izveštavanja, pružajući pouzdaniju osnovu za komparativnu analizu.

Drugo, ovo istraživanje služi kao svedočanstvo vrednosti **inženjeringa velikih podataka**. Pažljiva kompilacija dva sveobuhvatna, multi-domenska skupa podataka—jedan za SAD i jedan za svet—bio je značajan poduhvat koji ističe kritičnu potrebu za pristupačnim, visokokvalitetnim podacima u istraživanju javnog zdravlja. Strategije primenjene za rukovanje retkim podacima i multikolinearnošću, kao što je kreiranje "Union" seta obeležja, pružaju praktičan šablon za buduće velike ekološke studije.

Konačno, uvid u prednosti i mane svake od testiranih metoda potencira dobro poznatu teoremu u mašinskom učenju da „nema besplatnog ručka“, i da se metoda mora prilagoditi problemu i setu podataka. Naši rezultati služe kao početna tačka u budućim razmatranjima.

6.4 Pozicioniranje nalaza istraživanja u odnosu na naučnu literaturu

Veza između zagađenja vazduha, posebno $PM_{2.5}$, i COVID-19 ishoda je bila predmet intenzivnog istraživanja tokom pandemije. Naši nalazi—da je zagađenje $PM_{2.5}$ najvažniji prediktor prenosivosti u SAD—se uklapaju u rastući korpus dokaza koji povezuje kvalitet vazduha sa širenjem i težinom SARS-CoV-2. Nekoliko ključnih studija podržava ovu vezu. Wu et al. su tokom 2020. analizirali podatke iz ~3,000 američkih okruga i pronašli da je povećanje od $1 \mu g/m^3$ u dugotrajnoj $PM_{2.5}$ izloženosti bilo povezano sa 8-11% povećanjem stope mortaliteta od COVID-19. Ova veličina efekta je značajna, sugerišući da zagađenje vazduha nije mali modulator već značajan pogoršavajući faktor (Wu et al., 2020).

Slični rezultati su opservirani u Evropi. Istraživanje u Holandiji je analiziralo podatke za 355 opština i pronašlo da je povećanje od $1 \mu g/m^3$ u $PM_{2.5}$ bilo povezano sa 15 dodatnih slučajeva, 4 hospitalizacije i 3 smrti na 100,000 ljudi (Cole et al., 2020). Individualne studije su validirale ove ekološke nalaze. Jedna takva studija je zaključila da je dugotrajna $PM_{2.5}$ izloženost bila povezana sa 14-18% većim izgledima hospitalizacije među pojedincima zaraženim SARS-CoV-2 (Mendy et al., 2021).

Biološka uverljivost koja podupire ove epidemiološke asocijacije je izuzetno jaka, podržana obimnim toksikološkim i kliničkim istraživanjem. Dugotrajna izloženost zagađivačima vazduha izaziva stanje hronične, niskostepene sistemske upale i oksidativnog stresa, kompromitujući bazičnu otpornost organizma. Ovo kreira stanje povećane ranjivosti kroz nekoliko različitih, ali međusobno povezanih mehanizama: oštećenje respiratorne barijere, regulacija ACE2 receptora naviše, pojačan zapaljenjski odgovor i suprimiran imuni odgovor (Bourdrel et al., 2021; Weaver et al., 2022).

6.5 Implikacije za politiku javnog zdravlja i buduću pripravnost za pandemije

Nalazi ovog istraživanja imaju nekoliko važnih implikacija za javno zdravlje:

- 1. Kvalitet vazduha kao stub javnog zdravlja:** Jaka, kvantifikovana veza između $PM_{2.5}$ i virusne transmisije preoblikuje kvalitet vazduha kao kritičnu komponentu pripravnosti za pandemiju. Politike usmerene na smanjenje zagađenja vazduha ne treba posmatrati isključivo kroz ekološku perspektivu; one su takođe dugoročne investicije u javno zdravlje koje mogu izgraditi otpornost populacije protiv budućih respiratornih patogena.
- 2. Strateška upotreba sezonalnosti u planiranju:** Potvrda jake sezonske sile za COVID-19 pruža predvidljivi ritam koji sistemi javnog zdravlja mogu koristiti za strateško planiranje. Anticipiranje zimskih talasa omogućava proaktivnu alokaciju resursa, ciljane vakcinacione kampanje i blagovremeno javnozdravstveno slanje poruka.
- 3. Kontekstualna evaluacija intervencija:** Ovo istraživanje demonstrira da efikasnost javnozdravstvenih intervencija ne može biti naivno upoređivana između različitih regiona ili

vremenskih perioda. Zatvaranje implementirano u zemlji koja ulazi u svoju letnju sezonu niske transmisije će prirodno izgledati uspješnije od iste politike u zemlji koja ide u zimu. Svaka retrospektivna evaluacija pandemijskih odgovora mora uzeti u obzir ove osnovne ekološke i sezonske bazne linije.

6.6 Ograničenja istraživanja

Važno je priznati ograničenja ovog rada. Kao **ekološka studija**, oslanja se na agregirane podatke na nivou države i zemlje. Asocijacije opservirane na ovom makro nivou možda nisu primenjive nužno na pojedince (ekološka zablude). Dok studija demonstrira jake asocijacije, ona **ne može definitivno dokazati kauzalnost**. Složeni odnosi između zagađenja, socioekonomskog statusa, rase i zdravlja su poznato teški za potpuno razdvajanje. Štaviše, uprkos rigoroznim naporima, **kvalitet i dostupnost globalnih podataka**, posebno za zagađenje vazduha, ostaju značajno ograničenje. Konačno, ova analiza je zasnovana na ranoj fazi pandemije; naknadna pojava visoko prenosivih virusnih varijanti možda je promenila relativnu važnost ovih eksternih pokretača.

6.7 Pravci budućih istraživanja

Ova teza otvara nekoliko pravaca za buduća istraživanja. Postoji jasna potreba za studijama na **višoj prostornoj rezoluciji**, kao što je nivo grada ili susedstva, za dalje istraživanje međuigre između zagađenja, socioekonomskih faktora i transmisije. **Longitudinalne studije** koje prate pojedince tokom vremena su neophodne za približavanje uspostavljanju kauzalnosti. Štaviše, razvoj **integrisanih modela predikcije** koji eksplicitno inkorporiraju i sezonske funkcije prisiljavanja i podatke o kvalitetu vazduha u realnom vremenu bi mogao značajno poboljšati tačnost budućih pandemijskih predikcija.

7 Zaključak

7.1 Sažetak doprinosa

Ova teza je nastala sa motivom da istraži složenu mrežu faktora životne sredine i sociodemografskih faktora koji utiču na transmisiju SARS-CoV-2 tokom rane faze pandemije COVID-19. Koristeći robustan metodološki okvir utemeljen u epidemiološkom modelovanju i naprednom mašinskom učenju, ovo istraživanje je dalo nekoliko ključnih doprinosa našem razumevanju dinamike pandemije. Primarni nalazi ovog rada su dvojaki, otkrivajući hijerarhiju pokretača životne sredine zavisnu od skale:

1. Na regionalnom nivou, analiza Sjedinjenih Američkih Država je demonstrirala da je **zagađenje vazduha PM_{2.5} bilo pojedinačno najznačajniji prediktor virusne prenosivosti (R_0)**. Ovaj zaključak vođen podacima sugerise jaku, kvantitativnu vezu između kvaliteta vazduha i otpornosti javnog zdravlja, sugerisući da je ekološka politika integralni deo pripravnosti za pandemiju.
2. Na globalnom nivou, analiza 118 zemalja je otkrila da je **sezonalnost veoma značajan, predvidljiv i hemisferski zavisan pokretač transmisije**. Podaci ukazuju da ovaj sveobuhvatni ekološki ritam diktira vremenski oseku i plimu pandemije na globalnoj skali, što predstavlja kritičan uvid za predikciju i za kontekstualnu evaluaciju javnozdravstvenih intervencija.

Metodološki, ova teza je unapredila polje primenom metrike svesne dinamike (R_0) da bi osigurala standardizovanu osnovu za poređenje, i preduzimanjem značajnog napora inženjeringa podataka da bi kreirala dva sveobuhvatna, multi-domenska skupa podataka.

7.2 Završne napomene

Putovanje ovog istraživanja, preduzeto usred neizvesnosti globalne zdravstvene krize, je naglasilo duboku međupovezanost naše životne sredine, naših društava i našeg zdravlja. Nalazi predstavljeni u ovoj disertaciji ističu da širenje virusa nije samo biološki događaj, već složen fenomen oblikovan svetom koji nastanjujemo.

I pored prividne organizacije celog sveta u borbi protiv pandemije od kojih je važan aspekt bilo prikupljanje podataka, naišli smo na mnoge prepreke vezane za dostupnost i kvalitet podataka. Ovde se ne misli samo na COVID-19-specifične podatke, već i na opšte prediktore koji se godinama akumuliraju i ažuriraju a i dalje bez zadovoljavajućih rezultata.

U krajnjoj liniji, ovaj rad služi kao poziv za integrisaniji, interdisciplinarni pristup rešavanju globalnih zdravstvenih izazova. Jasni signali i zagađenja i sezonalnosti u podacima su moćan podsetnik da ne možemo razdvojiti ljudsko zdravlje od planetarnog zdravlja. Kako se krećemo kroz eru rastuće globalne povezanosti i ekoloških promena, uvidi stečeni iz

razumevanja ovih složenih interakcija će biti kritičniji nego ikad. Nada autora je da će istraživanje predstavljeno u ovoj tezi doprineti izgradnji otpornije i pripremljenije globalne zajednice za izazove koji leže pred nama.

8 Bibliografija

Aizerman, M., Braverman, E., & Rozonoer, L. (2019). *Theoretical foundation of potential functions method in pattern recognition*.

<https://www.semanticscholar.org/paper/Theoretical-foundation-of-potential-functions-in-Aizerman-Braverman/d94b70f294d2aa58d1d95d0b9e2a8a614ce6e92b>

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>

Almotairi, K. H., Hussein, A. M., Abualigah, L., Abujayyab, S. K. M., Mahmoud, E. H., Ghanem, B. O., & Gandomi, A. H. (2023). Impact of Artificial Intelligence on COVID-19 Pandemic: A Survey of Image Processing, Tracking of Disease, Prediction of Outcomes, and Computational Medicine. *Big Data and Cognitive Computing*, 7(1), 11. <https://doi.org/10.3390/bdcc7010011>

Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8), 879–891. <https://doi.org/10.1002/gepi.20543>

Ayers, L., & Cordell, H. J. (2010). SNP-based elastic net for identifying and ranking genetic predictors of complex traits. *Bioinformatics*, 26(15), 1886–1892.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, 32(3), 259–280. <https://doi.org/10.1257/jep.32.3.259>

Billionnet, C., Sherrill, D., Annesi-Maesano, I., & GERIE study. (2012). Estimating the health effects of exposure to multi-pollutant mixture. *Annals of Epidemiology*, 22(2), 126–141. <https://doi.org/10.1016/j.annepidem.2011.11.004>

Bontempi, E. (2020). Commercial exchanges instead of air pollution as possible origin of COVID-19 initial diffusion phase in Italy: More efforts are necessary to address interdisciplinary research. *Environmental Research*, 188, 109775. <https://doi.org/10.1016/j.envres.2020.109775>

Bontempi, E., Vergalli, S., & Squazzoni, F. (2020). Understanding COVID-19 diffusion requires an interdisciplinary, multi-dimensional approach. *Environmental Research*, 188, 109814. <https://doi.org/10.1016/j.envres.2020.109814>

Bourdrel, T., Annesi-Maesano, I., Alahmad, B., Maesano, C., & Bind, M.-A. (2021). The impact of outdoor air pollution on COVID-19: A review of evidence from in vitro, animal, and human studies. *European Respiratory Review*, 30(159).

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>

Byun, W. S., Heo, S. W., Jo, G., Kim, J. W., Kim, S., Lee, S., Park, H. E., & Baek, J.-H. (2021). Is coronavirus disease (COVID-19) seasonal? A critical analysis of empirical and epidemiological studies at global and local scales. *Environmental Research*, 196, 110972.
<https://doi.org/10.1016/j.envres.2021.110972>

CDC. (2021, May 7). *Scientific brief: SARS-CoV-2 transmission*. <https://stacks.cdc.gov>

Census Reporter. (2020). *Census Reporter*. <https://censusreporter.org/>

Chakraborty, J. (2021). Convergence of COVID-19 and chronic air pollution risks: Racial/ethnic and socioeconomic inequities in the U.S. *Environmental Research*, 193, 110586. <https://doi.org/10.1016/j.envres.2020.110586>

Chen, Q., Zheng, X., Shi, H., Zhou, Q., Hu, H., Sun, M., Xu, Y., & Zhang, X. (2024). Prediction of influenza outbreaks in Fuzhou, China: Comparative analysis of forecasting models. *BMC Public Health*, 24(1), 1399. <https://doi.org/10.1186/s12889-024-18583-x>

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, 785–794. <https://doi.org/10.1145/2939672.2939785>

Cho, S., Kim, K., Kim, Y. J., Lee, J.-K., Cho, Y. S., Lee, J.-Y., Han, B.-G., Kim, H., Ott, J., & Park, T. (2010). Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Annals of Human Genetics*, 74(5), 416–428.
<https://doi.org/10.1111/j.1469-1809.2010.00597.x>

Cole, M. A., Ozgen, C., & Strobl, E. (2020). Air Pollution Exposure and Covid-19 in Dutch Municipalities. *Environmental & Resource Economics*, 76(4), 581–610.
<https://doi.org/10.1007/s10640-020-00491-4>

Comunian, S., Dongo, D., Milani, C., & Palestini, P. (2020). Air pollution and Covid-19: The role of particulate matter in the spread and increase of Covid-19's morbidity and mortality. *International Journal of Environmental Research and Public Health*, 17(12), 4487.
<https://doi.org/10.3390/ijerph17124487>

Contini, D., & Costabile, F. (2020). Does Air Pollution Influence COVID-19 Outbreaks? *Atmosphere*, 11(4), 377. <https://doi.org/10.3390/atmos11040377>

Cook, R. D. (1982). Cook, R. D., S. Weisberg: Residuals and influence in regression. Chapman and Hall, New York — London 1982. VIII, 229 pp., £ 12,-. *Biometrical Journal*, 27(1), 80–80.
<https://doi.org/10.1002/bimj.4710270110>

Copat, C., Cristaldi, A., Fiore, M., Grasso, A., Zuccarello, P., Signorelli, S. S., Conti, G. O., & Ferrante, M. (2020). The role of air pollution (PM and NO₂) in COVID-19 spread and lethality: A systematic review. *Environmental Research*, 191, 110129. <https://doi.org/10.1016/j.envres.2020.110129>

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. <https://doi.org/10.1007/BF00994018>

Crippa, M., Solazzo, E., Huang, G., Guizzardi, D., Koffi, E., Muntean, M., Schieberle, C., Friedrich, R., & Janssens-Maenhout, G. (2020). High resolution temporal profiles in the Emissions Database for Global Atmospheric Research. *Scientific Data*, 7(1), 121. <https://doi.org/10.1038/s41597-020-0462-2>

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., Wenseleers, T., Gimma, A., Waite, W., Wong, K. L. M., van Zandvoort, K., Silverman, J. D., CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, Diaz-Ordaz, K., ... Edmunds, W. J. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, 372(6538), eabg3055. <https://doi.org/10.1126/science.abg3055>

Djordjevic, M., Salom, I., Markovic, S., Rodic, A., Milicevic, O., & Djordjevic, M. (2021). Inferring the Main Drivers of SARS-CoV-2 Global Transmissibility by Feature Selection Methods. *GeoHealth*, 5(9), e2021GH000432. <https://doi.org/10.1029/2021GH000432>

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5), 533–534. [https://doi.org/10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1)

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons.

Ekundayo, F. (2024). Using machine learning to predict disease outbreaks and enhance public health surveillance. *World Journal of Advanced Research and Reviews*, 24(3), 794–811. <https://doi.org/10.30574/wjarr.2024.24.3.3732>

European Commission Global Human Settlement. (2020). *Urban Centre Database UCDB R2019A*. https://ghsl.jrc.ec.europa.eu/ghs_stat_ucdb2015mt_r2019a.php

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/b16018>

Glencross, D. A., Ho, T.-R., Camiña, N., Hawrylowicz, C. M., & Pfeffer, P. E. (2020). Air pollution and its effects on the immune system. *Free Radical Biology and Medicine, Air Pollution: Consequences for Cellular Redox Signaling, Antioxidant Defenses and Disease*, 151, 56–68. <https://doi.org/10.1016/j.freeradbiomed.2020.01.179>

Gohari, M. R., Doggett, A., Patte, K. A., Ferro, M. A., Dubin, J. A., Hilario, C., & Leatherdale, S. T. (2024). Using random forest to identify correlates of depression symptoms among adolescents. *Social Psychiatry and Psychiatric Epidemiology*, 59(11), 2063–2071. <https://doi.org/10.1007/s00127-024-02695-1>

Green, E. D., Guyer, M. S., & National Human Genome Research Institute. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333), 204–213. <https://doi.org/10.1038/nature09764>

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information*, 15(4), 235. <https://doi.org/10.3390/info15040235>

Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J., Zhang, Q., Luo, G., Li, Z., He, J., Zhang, Y., & Ma, W. (2017). Developing a dengue forecast model using machine learning: A case study in China. *PLoS Neglected Tropical Diseases*, 11(10), e0005973. <https://doi.org/10.1371/journal.pntd.0005973>

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., & Tatlow, H. (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, 5(4), 529–538. <https://doi.org/10.1038/s41562-021-01079-8>

Hamidi, S., Sabouri, S., & Ewing, R. (2020). Does Density Aggravate the COVID-19 Pandemic?: Early Findings and Lessons for Planners. *Journal of the American Planning Association*, 86(4), 495–509. <https://doi.org/10.1080/01944363.2020.1777891>

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Springer Series in Statistics*. <https://doi.org/10.1007/978-0-387-84858-7>

Hayashi, F. (2011). *Econometrics*. Princeton University Press.

He, X., Lau, E. H. Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., ... Leung, G. M. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26(5), 672–675. <https://doi.org/10.1038/s41591-020-0869-5>

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C., & Pöhlmann, S.

(2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, 181(2), 271-280.e8.
<https://doi.org/10.1016/j.cell.2020.02.052>

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1), 73–101. <https://doi.org/10.1214/aoms/1177703732>

Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The **forecast** Package for R. *Journal of Statistical Software*, 27(3). <https://doi.org/10.18637/jss.v027.i03>

Jo, Y., Hong, A., & Sung, H. (2021). Density or Connectivity: What Are the Main Causes of the Spatial Proliferation of COVID-19 in Korea? *International Journal of Environmental Research and Public Health*, 18(10), 5084. <https://doi.org/10.3390/ijerph18105084>

Jolliffe, I. T. (Ed.). (2002). Principal Components in Regression Analysis. In *Principal Component Analysis* (pp. 167–198). Springer. https://doi.org/10.1007/0-387-22440-8_8

Keeling, M. J., & Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press. <https://doi.org/10.2307/j.ctvcm4gk0>

Keeling, M. J., & Rohani, P. (2011). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

Kendall, M. G. (1938). A New Measure Of Rank Correlation. *Biometrika*, 30(1–2), 81–93. <https://doi.org/10.1093/biomet/30.1-2.81>

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772), 700–721.
<https://doi.org/10.1098/rspa.1927.0118>

Keshavamurthy, R., Pazdernik, K. T., Ham, C., Dixon, S., Erwin, S., & Charles, L. E. (2025). Meeting Global Health Needs via Infectious Disease Forecasting: Development of a Reliable Data-Driven Framework. *JMIR Public Health and Surveillance*, 11(1), e59971.
<https://doi.org/10.2196/59971>

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17. <https://doi.org/10.1016/j.csbj.2014.11.005>

Laden, F., Neas, L. M., Dockery, D. W., & Schwartz, J. (2000). Association of fine particulate matter from different sources with daily mortality in six U.S. cities. *Environmental Health Perspectives*, 108(10), 941–947. <https://doi.org/10.1289/ehp.00108941>

- Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. *Annals of Internal Medicine*, M20-0504. <https://doi.org/10.7326/M20-0504>
- Levin, A. T., Hanage, W. P., Owusu-Boaitey, N., Cochran, K. B., Walsh, S. P., & Meyerowitz-Katz, G. (2020). Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *European Journal of Epidemiology*, 35(12), 1123–1138. <https://doi.org/10.1007/s10654-020-00698-1>
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., ... Feng, Z. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine*, 382(13), 1199–1207. <https://doi.org/10.1056/NEJMoa2001316>
- Liu, Y., Yang, Y., Zhang, C., Huang, F., Wang, F., Yuan, J., Wang, Z., Li, J., Li, J., Feng, C., Zhang, Z., Wang, L., Peng, L., Chen, L., Qin, Y., Zhao, D., Tan, S., Yin, L., Xu, J., ... Liu, L. (2020). Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury. *Science China Life Sciences*, 63(3), 364–374. <https://doi.org/10.1007/s11427-020-1643-8>
- Long, H., Chen, Y., Feng, J., Chen, J., Zhang, X., Han, W., Kang, M., & Du, X. (2025). Annual global dengue dynamics are related to multi-source factors revealed by a machine learning prediction analysis. *PLOS Neglected Tropical Diseases*, 19(6), e0013232. <https://doi.org/10.1371/journal.pntd.0013232>
- Maier, B. F., & Brockmann, D. (2020). Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, 368(6492), 742–746. <https://doi.org/10.1126/science.abb4557>
- Martineau, A. R., Jolliffe, D. A., Hooper, R. L., Greenberg, L., Aloia, J. F., Bergman, P., Dubnov-Raz, G., Esposito, S., Ganmaa, D., Ginde, A. A., Goodall, E. C., Grant, C. C., Griffiths, C. J., Janssens, W., Laaksi, I., Manaseki-Holland, S., Mauger, D., Murdoch, D. R., Neale, R., ... Camargo, C. A. (2017). *Vitamin D supplementation to prevent acute respiratory tract infections: Systematic review and meta-analysis of individual participant data*. <https://doi.org/10.1136/bmj.i6583>
- Mehta, P., McAuley, D. F., Brown, M., Sanchez, E., Tattersall, R. S., & Manson, J. J. (2020). COVID-19: Consider cytokine storm syndromes and immunosuppression. *The Lancet*, 395(10229), 1033–1034. [https://doi.org/10.1016/S0140-6736\(20\)30628-0](https://doi.org/10.1016/S0140-6736(20)30628-0)

Mendy, A., Wu, X., Keller, J. L., Fassler, C. S., Apewokin, S., Mersha, T. B., Xie, C., & Pinney, S. M. (2021). Air pollution and the pandemic: Long-term PM_{2.5} exposure and disease severity in COVID-19 patients. *Respirology (Carlton, Vic.)*, 26(12), 1181–1187. <https://doi.org/10.1111/resp.14140>

Milicevic, O., Salom, I., Rodic, A., Markovic, S., Tumbas, M., Zigic, D., Djordjevic, M., & Djordjevic, M. (2021). PM_{2.5} as a major predictor of COVID-19 basic reproduction number in the USA. *Environmental Research*, 201, 111526. <https://doi.org/10.1016/j.envres.2021.111526>

NASA Langley Research Center. (2020). *The Prediction of Worldwide Energy Resources (POWER) Project*. <https://power.larc.nasa.gov/>

Nishiura, H., Linton, N. M., & Akhmetzhanov, A. R. (2020). Serial interval of novel coronavirus (COVID-19) infections. *International Journal of Infectious Diseases*, 93, 284–286. <https://doi.org/10.1016/j.ijid.2020.02.060>

O'Driscoll, M., Ribeiro Dos Santos, G., Wang, L., Cummings, D. A. T., Azman, A. S., Paireau, J., Fontanet, A., Cauchemez, S., & Salje, H. (2021). Age-specific mortality and immunity patterns of SARS-CoV-2. *Nature*, 590(7844), 140–145. <https://doi.org/10.1038/s41586-020-2918-0>

OpenUV. (2020). *Global Real-Time UV Index REST JSON API*. OpenUV. <https://www.openuv.io/>

Paital, B., & Agrawal, P. K. (2021). Air pollution by NO₂ and PM_{2.5} explains COVID-19 infection severity by overexpression of angiotensin-converting enzyme 2 in respiratory cells: A review. *Environmental Chemistry Letters*, 19, 25–42. <https://doi.org/10.1007/s10311-020-01091-w>

Pearson, K. (1896). VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, (187), 253–318. <https://doi.org/10.1098/rsta.1896.0007>

Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., Flasche, S., Clifford, S., Pearson, C. A. B., Munday, J. D., Abbott, S., Gibbs, H., Rosello, A., Quilty, B. J., Jombart, T., Sun, F., Diamond, C., Gimma, A., Zandvoort, K. van, ... Klepac, P. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. *The Lancet Public Health*, 5(5), e261–e270. [https://doi.org/10.1016/S2468-2667\(20\)30073-6](https://doi.org/10.1016/S2468-2667(20)30073-6)

Qu, G., Li, X., Hu, L., & Jiang, G. (2020). An Imperative Need for Research on the Role of Environmental Factors in Transmission of Novel Coronavirus (COVID-19). *Environmental Science & Technology*, 54(7), 3730–3732. <https://doi.org/10.1021/acs.est.0c01102>

Remes, O., Mendes, J. F., & Templeton, P. (2021). Biological, Psychological, and Social Determinants of Depression: A Review of Recent Literature. *Brain Sciences*, 11(12), 1633. <https://doi.org/10.3390/brainsci11121633>

Roberts, S., & Martin, M. A. (2006). Using supervised principal components analysis to assess multiple pollutant effects. *Environmental Health Perspectives*, 114(12), 1877–1882. <https://doi.org/10.1289/ehp.9226>

Rousseeuw, P. J., & Leroy, A. M. (1987, October 19). *Robust Regression and Outlier Detection*. *Wiley Series in Probability and Statistics*. <https://doi.org/10.1002/0471725382>

Sagawa, T., Tsujikawa, T., Honda, A., Miyasaka, N., Tanaka, M., Kida, T., Hasegawa, K., Okuda, T., Kawahito, Y., & Takano, H. (2021). Exposure to particulate matter upregulates ACE2 and TMPRSS2 expression in the murine lung. *Environmental Research*, 195, 110722. <https://doi.org/10.1016/j.envres.2021.110722>

Salom, I., Rodic, A., Milicevic, O., Zigic, D., Djordjevic, M., & Djordjevic, M. (2021). Effects of Demographic and Weather Parameters on COVID-19 Basic Reproduction Number. *Frontiers in Ecology and Evolution*, 8. <https://doi.org/10.3389/fevo.2020.617841>

Santesmasses, D., Castro, J. P., Zenin, A. A., Shindyapina, A. V., Gerashchenko, M. V., Zhang, B., Kerepesi, C., Yim, S. H., Fedichev, P. O., & Gladyshev, V. N. (2020). COVID-19 is an emergent disease of aging. *Aging Cell*, 19(10), e13230. <https://doi.org/10.1111/accel.13230>

Sarkodie, S. A., & Owusu, P. A. (2020). Impact of meteorological factors on COVID-19 pandemic: Evidence from top 20 countries with confirmed cases. *Environmental Research*, 191, 110101. <https://doi.org/10.1016/j.envres.2020.110101>

Schuit, M., Gardner, S., Wood, S., Bower, K., Williams, G., Freeburger, D., & Dabisch, P. (2020). The Influence of Simulated Sunlight on the Inactivation of Influenza Virus in Aerosols. *The Journal of Infectious Diseases*, 221(3), 372–378. <https://doi.org/10.1093/infdis/jiz582>

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4), 591–611. <https://doi.org/10.1093/biomet/52.3-4.591>

Smits, J., & Permanyer, I. (2019). The Subnational Human Development Database. *Scientific Data*, 6(1), 190038. <https://doi.org/10.1038/sdata.2019.38>

Social Science Research Council. (2020). *Measure of America*.
<https://measureofamerica.org/>

Spearman, C. (1904). The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5), 1137–1150.
<https://doi.org/10.1093/ije/dyq191>

Stieb, D. M., Evans, G. J., To, T. M., Brook, J. R., & Burnett, R. T. (2020). An ecological analysis of long-term exposure to PM_{2.5} and incidence of COVID-19 in Canadian health regions. *Environmental Research*, 191, 110052. <https://doi.org/10.1016/j.envres.2020.110052>

Suarez-Lledo, V., & Alvarez-Galvez, J. (2019, July 30). *A Random Forest Approach to Study Social Determinants of Depression: Turning the Black Box into a White Box in Social Sciences*.

Thurston, G. D., & Ito, K. (2001). Epidemiological studies of acute ozone exposures and mortality. *Journal of Exposure Analysis and Environmental Epidemiology*, 11(4), 286–294.
<https://doi.org/10.1038/sj.jea.7500169>

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
<https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

Uddin, S., & Lu, H. (2024). Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. *PLOS ONE*, 19(4), e0301541. <https://doi.org/10.1371/journal.pone.0301541>

United Health Foundation. (2020). *America's Health Rankings*.
<https://www.americashealthrankings.org/>

U.S. Census Bureau. (2020). *Explore Census Data*. <https://data.census.gov/>

US EPA, O. (2013, August 1). *Air Quality System (AQS)* [Data and Tools].
<https://www.epa.gov/aqs>

Venables, W., & Ripley, B. (1997). *Modern Applied Statistics with S-Plus Second edition*.
<https://www.semanticscholar.org/paper/Modern-Applied-Statistics-with-S-Plus-Second-Venables-Ripley/c38458d4d51b1c14ac5867ddcd42746422b53edf>

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G. T., Fu, H., Dighe, A., Griffin, J. T., Baguelin, M., Bhatia, S., Boonyasiri, A., Cori, A., Cucunubá, Z., FitzJohn, R., Gaythorpe, K., ... Ferguson, N. M. (2020). Estimates of the severity of coronavirus disease 2019: A model-based analysis. *The Lancet Infectious Diseases*, 20(6), 669–677. [https://doi.org/10.1016/S1473-3099\(20\)30243-7](https://doi.org/10.1016/S1473-3099(20)30243-7)

Villanueva-Miranda, I., Xiao, G., & Xie, Y. (2025). Artificial intelligence in early warning systems for infectious disease surveillance: A systematic review. *Frontiers in Public Health*, 13, 1609615. <https://doi.org/10.3389/fpubh.2025.1609615>

Villeneuve, P. J., & Goldberg, M. S. (2020). Methodological Considerations for Epidemiological Studies of Air Pollution and the SARS and COVID-19 Coronavirus Outbreaks. *Environmental Health Perspectives*, 128(9), 095001. <https://doi.org/10.1289/EHP7411>

Wang, J., Tang, K., Feng, K., Lin, X., Lv, W., Chen, K., & Wang, F. (2021). *Impact of temperature and relative humidity on the transmission of COVID-19: A modelling study in China and the United States*. <https://doi.org/10.1136/bmjopen-2020-043863>

Weaver, A. K., Head, J. R., Gould, C. F., Carlton, E. J., & Remais, J. V. (2022). Environmental Factors Influencing COVID-19 Incidence and Severity. *Annual Review of Public Health*, 43, 271–291. <https://doi.org/10.1146/annurev-publhealth-052120-101420>

Wiemken, T. L., & Kelley, R. R. (2020). Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health*, 41(Volume 41, 2020), 21–36. <https://doi.org/10.1146/annurev-publhealth-040119-094437>

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P., Cockburn, J., McDonald, H. I., MacKenna, B., Tomlinson, L., Douglas, I. J., Rentsch, C. T., Mathur, R., Wong, A. Y. S., Grieve, R., ... Goldacre, B. (2020). Factors associated with COVID-19-related death using OpenSAFELY. *Nature*, 584(7821), 430–436. <https://doi.org/10.1038/s41586-020-2521-4>

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. <https://doi.org/10.1109/4235.585893>

Wooldridge, J. (1999, August 6). *Introductory Econometrics: A Modern Approach*. <https://www.semanticscholar.org/paper/Introductory-Econometrics%3A-A-Modern-Approach-Wooldridge/d841c77e5f142dd73e49039381372755bdd78342>

World Health Organization. (2020). *WHO Director-General's opening remarks at the media briefing on COVID-19—11 March 2020*. <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>

Worldometer. (2020, June 27). *Worldometer: COVID-19 coronavirus pandemic—Google Scholar*. https://scholar.google.com/scholar_lookup?title=COVID-19%20Coronavirus%20Pandemic&publication_year=2020&author=Worldometer

Wu, X., Nethery, R. C., Sabath, M. B., Braun, D., & Dominici, F. (2020). Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances*, 6(45), eabd4049. <https://doi.org/10.1126/sciadv.abd4049>

Yao, Y., Pan, J., Liu, Z., Meng, X., Wang, W., Kan, H., & Wang, W. (2021). Ambient nitrogen dioxide pollution and spreadability of COVID-19 in Chinese cities. *Ecotoxicology and Environmental Safety*, 208, 111421. <https://doi.org/10.1016/j.ecoenv.2020.111421>

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, 382(8), 727–733. <https://doi.org/10.1056/NEJMoa2001017>

Zhu, Y., Xie, J., Huang, F., & Cao, L. (2020). Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China. *Science of The Total Environment*, 727, 138704. <https://doi.org/10.1016/j.scitotenv.2020.138704>

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>

9 Apendiks

Tabela 3: Pregled svih varijabli konačnog skupa podataka podeljenih u pet grupa. Varijable vezane za izračunavanje osnovnog reprodukcionog broja nisu korišćene za predikciju, kao ni duplikati meteoroloških podataka iz WAQI baze, pa je ukupan broj razmatranih prediktora 96.

#	Naziv varijable	Opis	Izvor
1. Meteorološki podaci (NASA POWER)			
1	T2M	Srednja temperatura vazduha na 2 m (°C)	NASA POWER
2	T2M_MAX	Maksimalna temperatura na 2 m (°C)	NASA POWER
3	T2M_MIN	Minimalna temperatura na 2 m (°C)	NASA POWER
4	T10M	Srednja temperatura vazduha na 10 m (°C)	NASA POWER
5	T10M_MAX	Maksimalna temperatura na 10 m (°C)	NASA POWER
6	T10M_MIN	Minimalna temperatura na 10 m (°C)	NASA POWER
7	TS	Temperatura površine zemljišta — skin temperature (°C)	NASA POWER
8	T2MWET	Temperatura vlažnog termometra na 2 m (°C)	NASA POWER
9	RH2M	Relativna vlažnost vazduha na 2 m (%)	NASA POWER
10	QV2M	Specifična vlažnost na 2 m (g/kg)	NASA POWER
11	T2MDEW	Temperatura tačke rose na 2 m (°C)	NASA POWER
12	WS2M	Brzina vetra na 2 m (m/s)	NASA POWER
13	WS10M	Brzina vetra na 10 m (m/s)	NASA POWER
14	PRECTOT	Ukupne padavine (mm/dan)	NASA POWER
15	TQV	Ukupna količina vodene pare u stubcu atmosfere (kg/m²)	NASA POWER
16	ALLSKY_SFC_SW_DWN	Površinsko sunčevo kratkovalno zračenje — sve nebo (W/m²)	NASA POWER
17	CLRSKY_SFC_SW_DWN	Površinsko sunčevo kratkovalno zračenje — vedro nebo (W/m²)	NASA POWER
18	ALLSKY_SFC_LW_DWN	Površinsko dugovalasno zračenje — sve nebo (W/m²)	NASA POWER
19	OpenUV_max	Maksimalni dnevni UV indeks (ponderisano populacijom)	OpenUV API
20	temperature	WAQI temperatura (direktno merenje u urbanim centrima, °C)	WAQI
2. Sociodemografski podaci (GHS-UCDB, World Bank, Facebook SCI)			
1	BUCAP15	Površina izgrađenog gradskog prostora po stanovniku, 2015 (m²/st.)	GHS-UCDB R2019A (JRC)
2	GDP15_SM	Bruto domaći proizvod na nivou gradova, 2015 (USD)	GHS-UCDB R2019A (JRC)
3	density	Gustina stanovništva (st./km²)	GHS-UCDB / World Bank
4	urban.population.2018.	Procenat urbanog stanovništva, 2018 (%)	World Bank

#	Naziv varijable	Opis	Izvor
5	Median.age..CIA.	Medijana starosti populacije (godine)	CIA World Factbook
6	over.65....	Procenat populacije starije od 65 godina (%)	World Bank
7	PPP.per.capita	Paritet kupovne moći po stanovniku (intl. USD)	World Bank
8	HDI...OWID.	Indeks humanog razvoja	OWID / UNDP
9	life.expectancy.born.1990.	Očekivano trajanje života (osobe rođene 1990.)	World Bank
10	infant.mort..2018.	Stopa smrtnosti odojčadi na 1.000 živorođenih, 2018	World Bank
11	Under.5.mort..2018.	Smrtnost dece do 5 godina na 1.000 živorođenih, 2018	World Bank
12	prc.rodj.u.inostranstvu.WB.2015.	Procenat stanovnika rođenih u inostranstvu, 2015 (%)	World Bank
13	immigrants.emigrants.2017.	Neto migracija (imigranti – emigranti), 2017	World Bank
14	n.nastanjenih.izbeglica.2018.	Broj nastanjenih izbeglica, 2018	UNHCR
15	Smokers.2016	Prevalenca pušenja, oba pola, modelovan iz trenda 2016 (%)	WHO GHO
16	Population_2015	Ukupan broj stanovnika, 2015	World Bank

3. Zdravstveni indikatori (WHO Global Health Observatory)

1	Raised.BP	Prevalenca povišenog krvnog pritiska, odrasli (%)	WHO GHO (BP_04)
2	CholesterolM	Srednji nivo holesterola — muškarci (mmol/L)	WHO GHO (CHOL_03)
3	CholesterolF	Srednji nivo holesterola — žene (mmol/L)	WHO GHO (CHOL_03)
4	CHEpGDP	Tekući izdaci za zdravstvo kao % BDP-a	WHO GHED
5	doc10K	Broj lekara na 10.000 stanovnika	WHO HWF
6	pharm10K	Broj farmaceuta na 10.000 stanovnika	WHO HWF
7	tobaccoMFS	Prevalenca pušenja iz ankete, oba pola, standardizovana (%)	WHO GHO
8	suicideMF2016	Stopa suicida na 100.000 — oba pola, 2016	WHO GHO
9	obesity_prev	Prevalenca gojaznosti kod odraslih (%)	WHO GHO (NCD_BMI_30A)
10	meanBMI	Srednji indeks telesne mase (BMI)	WHO GHO (NCD_BMI_MEAN)
11	fast_glucM	Srednji nivo glukoze našte — muškarci (mmol/L)	WHO GHO (NCD_GLUC_01)
12	fast_glucF	Srednji nivo glukoze našte — žene (mmol/L)	WHO GHO (NCD_GLUC_01)
13	inactive	Prevalenca fizičke neaktivnosti, odrasli (%)	WHO GHO (NCD_PAC/PAA)
14	prob_death	Verovatnoća prerane smrti od NCD (30–70 god.), %	WHO GHO (NCDMORT3070)

#	Naziv varijable	Opis	Izvor
15	mort_WASH	Smrtnost pripisiva lošim WASH uslovima (na 100.000)	WHO GHO (SDGWSHBOD)
16	total_alc	Ukupna potrošnja alkohola po glavi (litara čistog alkohola)	WHO GHO (WHOSIS_000011)
17	BCG1	Pokrivenost BCG vakcinacijom (%)	WHO GHO (WHS4_543)
18	hosp_beds10K	Broj bolničkih kreveta na 10.000 stanovnika	WHO GHO (WHS6_102)
19	basic_washU	Pristup osnovnim uslovima higijene, urbana populacija (%)	WHO GHO (WSH_HYGIENE_BASIC)
20	open_defT2017	Prevalenca otvorene defekacije — ukupno, 2017 (%)	WHO GHO
21	bas_sanU2017	Pristup osnovnoj sanitaciji, urbana populacija, 2017 (%)	WHO GHO
22	bas_watU2017	Pristup bezbednoj vodi, urbana populacija, 2017 (%)	WHO GHO
23	CVD_canc_diab_CRD..2016.	Mortalitet od KVB, raka, dijabetesa i HBP, 2016 (na 100.000)	WHO GHO
24	Share.adults.obese.OWID.	Udeo gojaznih odraslih (OWID, %)	OWID
25	avg.BMI...OWID.	Srednji BMI (OWID)	OWID
4. Podaci o zagađenju vazduha (WAQI, EDGAR, IQAIR)			
1	pollution.2.5.2017.	Koncentracija PM _{2.5} , 2017 (µg/m³)	WAQI / satelitski podaci
2	pollution.2.5.2013.2017.avg	Prosečna PM _{2.5} koncentracija 2013–2017 (µg/m³)	WAQI / satelitski podaci
3	PM2.5_2019	PM _{2.5} koncentracija, 2019 (µg/m³)	IQAir World Air Quality Report
4	pm25	PM _{2.5} (WAQI direktno merenje, µg/m³)	WAQI
5	pm10	PM ₁₀ (WAQI direktno merenje, µg/m³)	WAQI
6	pm1	PM ₁ (WAQI direktno merenje, µg/m³)	WAQI
7	no2	NO ₂ (WAQI, µg/m³)	WAQI
8	o3	O ₃ — ozon (WAQI, µg/m³)	WAQI
9	so2	SO ₂ (WAQI, µg/m³)	WAQI
10	co	CO — ugljen-monoksid (WAQI, µg/m³)	WAQI
11	aqi	Indeks kvaliteta vazduha — AQI (WAQI)	WAQI
12	pol	Dominantni zagađivač (kategorijska promenljiva)	WAQI
13	neph	Nefelometrija — optički pokazatelj zamućenosti vazduha	WAQI
14	mepaqi	MEP AQI indeks (kineska skala)	WAQI
15	uvi	UV indeks (WAQI)	WAQI
16	total_NOx_2015	Ukupne emisije NO _x , 2015 (kt/god.)	EDGAR v5.0
17	NOx_pc	Emisije NO _x po glavi stanovnika, 2015	EDGAR v5.0
18	total_BC_2015	Ukupne emisije crnog ugljenika (BC), 2015 (kt/god.)	EDGAR v5.0

#	Naziv varijable	Opis	Izvor
19	BC_pc	Emisije BC po glavi stanovnika, 2015	EDGAR v5.0
20	total_CO_2015	Ukupne emisije CO, 2015 (kt/god.)	EDGAR v5.0
21	CO_pc	Emisije CO po glavi stanovnika, 2015	EDGAR v5.0
22	total_NH3_2015	Ukupne emisije NH ₃ , 2015 (kt/god.)	EDGAR v5.0
23	NH3_pc	Emisije NH ₃ po glavi stanovnika, 2015	EDGAR v5.0
24	total_NMVOC_2015	Ukupne emisije NMVOC, 2015 (kt/god.)	EDGAR v5.0
25	NMVOC_pc	Emisije NMVOC po glavi stanovnika, 2015	EDGAR v5.0
26	total_OC_2015	Ukupne emisije organskog ugljenika (OC), 2015 (kt)	EDGAR v5.0
27	OC_pc	Emisije OC po glavi stanovnika, 2015	EDGAR v5.0
28	total_PM10_2015	Ukupne emisije PM ₁₀ , 2015 (kt/god.)	EDGAR v5.0
29	PM10_pc	Emisije PM ₁₀ po glavi stanovnika, 2015	EDGAR v5.0
30	total_SO2_2015	Ukupne emisije SO ₂ , 2015 (kt/god.)	EDGAR v5.0
31	SO2_pc	Emisije SO ₂ po glavi stanovnika, 2015	EDGAR v5.0
32	total_PM2.5_2015	Ukupne emisije PM _{2.5} , 2015 (kt/god.)	EDGAR v5.0
33	PM2.5_pc	Emisije PM _{2.5} po glavi stanovnika, 2015	EDGAR v5.0
34	dew	Temperatura rose (WAQI, °C)	WAQI
35	wind.speed	Brzina vetra — WAQI (m/s)	WAQI
36	wind.gust	Udari vetra — WAQI (m/s)	WAQI
37	humidity	Relativna vlažnost — WAQI (%)	WAQI
38	pressure	Atmosferski pritisak — WAQI (hPa)	WAQI
39	wind.speed.1	Brzina vetra (alternativni izvor, m/s)	WAQI
40	wind.gust.1	Udari vetra (alternativni izvor, m/s)	WAQI
41	precipitation	Padavine — WAQI (mm)	WAQI
42	wd	Smer vetra — WAQI (°)	WAQI
5. Ostale varijable (R₀ procena, Mere centralnosti)			
1	R0	Procenjeni osnovni reprodukcionni broj (zavisna varijabla)	Procenjeno iz epidemioloških podataka
2	c	Eksponecijalna stopa rasta λ^+ (dan ⁻¹)	Procenjeno iz epidemioloških podataka
3	start	Datum početka eksponecijalne faze za datu zemlju	Worldometer / Our World in Data
4	end	Datum kraja eksponecijalne faze za datu zemlju	Worldometer / Our World in Data
5	Country.Code	ISO 3166-1 alpha-3 kod zemlje	ISO 3166
6	Flow.Betweenness.Centrality	Centralnost protoka posredništva u mreži SCI	Facebook SCI (Bailey et al., 2018)
7	Weighted.Vertex.Degree	Ponderisani stepen čvora u mreži SCI	Facebook SCI (Bailey et al., 2018)
8	Weight	Ponder za ponderisanu regresiju (interna promenljiva)	Interno

Tabela 4: Odabrani prediktori za svaki od testiranih setova za 118 država

Set	Linearna formula
SAIC	$R0 \sim \text{so2} + \text{Weight} + \text{NMVOC_pc} + \text{o3} + \text{PPP.per.capita} + \text{fast_glucF} + \text{obesity_prev}$
Lasso1	$R0 \sim \text{T2M_MIN} + \text{so2} + \text{CholesterolF} + \text{total_alc}$
Lasso2	$R0 \sim \text{T2MWET} + \text{ALLSKY_SFC_SW_DWN} + \text{so2} + \text{CholesterolF} + \text{total_alc}$
RF1	$R0 \sim \text{total_alc} + \text{OpenUV_max} + \text{wind.speed} + \text{GDP15_SM} + \text{obesity_prev} + \text{total_NH3_2015} + \text{urban.population.2018.} + \text{total_SO2_2015} + \text{no2} + \text{CLRSKY_SFC_SW_DWN} + \text{ALLSKY_SFC_SW_DWN} + \text{CholesterolF} + \text{so2}$
RF2	$R0 \sim \text{total_alc} + \text{OpenUV_max} + \text{wind.speed} + \text{GDP15_SM} + \text{obesity_prev} + \text{total_NH3_2015} + \text{total_SO2_2015} + \text{Under.5.mort..2018.} + \text{no2} + \text{CholesterolF}$
Union	$R0 \sim \text{total_alc} + \text{OpenUV_max} + \text{wind.speed} + \text{GDP15_SM} + \text{obesity_prev} + \text{total_NH3_2015} + \text{urban.population.2018.} + \text{total_SO2_2015} + \text{no2} + \text{CLRSKY_SFC_SW_DWN} + \text{ALLSKY_SFC_SW_DWN} + \text{CholesterolF} + \text{so2} + \text{Under.5.mort..2018.} + \text{T2MWET} + \text{T2M_MIN} + \text{Weight} + \text{NMVOC_pc} + \text{o3} + \text{PPP.per.capita} + \text{fast_glucF}$

Tabela 5: Pregled varijabli — analiza saveznih država SAD (70 varijabli u 4 kategorije + 4 izvedene). Vrednosti su populaciono ponderisane i vremenski pomerene za 5 dana unazad u odnosu na epidemiološke podatke. U analizi su korišćene i 4 izvedene varijable: dPol2017 (razlika $PM_{2.5}$ između 2017. i proseka 2013–2017), kao i PC komponente temperature, vlažnosti i radijacije dobijene PCA transformacijom (Tabela 2).

#	Naziv varijable (original)	Opis	Izvor
1. Zagađenje vazduha — 10 varijabli (US EPA Air Quality System)			
1	PM2.5	Fine čestične materije $PM_{2.5}$ — populaciono ponderisana srednja vrednost ($\mu\text{g}/\text{m}^3$)	US EPA AQS
2	PM10	Čestične materije PM_{10} ($\mu\text{g}/\text{m}^3$)	US EPA AQS
3	CO	Ugljen-monoksid CO (ppm)	US EPA AQS
4	NO2	Azot-dioksid NO_2 (ppb)	US EPA AQS
5	Ozone	Ozon O_3 (ppb)	US EPA AQS
6	SO2	Sumpor-dioksid SO_2 (ppb)	US EPA AQS
7	VOC	Isparljiva organska jedinjenja — VOC (ppb)	US EPA AQS
8	Lead	Olovo Pb ($\mu\text{g}/\text{m}^3$)	US EPA AQS
9	HAPS	Opasni zagađivači vazduha — HAPs ($\mu\text{g}/\text{m}^3$)	US EPA AQS
10	NONOxNOy	Oksidi azota — NO_x/NO_y frakcija (ppb)	US EPA AQS
2. Meteorološki podaci — 20 varijabli (NASA POWER + OpenUV + WAQI)			

#	Naziv varijable (original)	Opis	Izvor
1	T2M	Srednja temperatura vazduha na 2 m (°C)	NASA POWER
2	T2M_MAX	Maksimalna temperatura na 2 m (°C)	NASA POWER
3	T2M_MIN	Minimalna temperatura na 2 m (°C)	NASA POWER
4	T10M	Srednja temperatura na 10 m (°C)	NASA POWER
5	T10M_MAX	Maksimalna temperatura na 10 m (°C)	NASA POWER
6	T10M_MIN	Minimalna temperatura na 10 m (°C)	NASA POWER
7	TS	Temperatura površine zemljišta — skin temperature (°C)	NASA POWER
8	T2MWET	Temperatura vlažnog termometra na 2 m (°C)	NASA POWER
9	RH2M	Relativna vlažnost na 2 m (%) — ulazi samostalno, bez PCA	NASA POWER
10	QV2M	Specifična vlažnost na 2 m (g/kg)	NASA POWER
11	T2MDEW	Temperatura tačke rose na 2 m (°C)	NASA POWER
12	WS2M	Brzina vetra na 2 m (m/s)	NASA POWER
13	WS10M	Brzina vetra na 10 m (m/s)	NASA POWER
14	PRECTOT	Ukupne padavine (mm/dan)	NASA POWER
15	TQV	Ukupna vodena para u stubcu atmosfere (kg/m²)	NASA POWER
16	ALLSKY_SFC_SW_DWN	Sunčevo kratkovalno zračenje — sve nebo (W/m²)	NASA POWER
17	CLRSKY_SFC_SW_DWN	Sunčevo kratkovalno zračenje — vedro nebo (W/m²)	NASA POWER
18	ALLSKY_SFC_LW_DWN	Dugovalasno površinsko zračenje — sve nebo (W/m²)	NASA POWER
19	OpenUV_max	Maksimalni dnevni UV indeks (ponderisano populacijom gradova)	OpenUV API
20	pressure	Atmosferski pritisak (hPa)	WAQI

3. Sociodemografski i ekonomski podaci — 27 varijabli

1	total_population	Ukupan broj stanovnika	US Census Bureau
2	total_population_men	Ukupno muško stanovništvo	US Census Bureau
3	total_population_women	Ukupno žensko stanovništvo	US Census Bureau
4	youth_population(%under18)	Procenat populacije mlađe od 18 godina (%)	US Census Bureau
5	Cesus 2010 median age	Medijana starosti, popis 2010 (godine)	US Census Bureau
6	total_population_white	Broj belog stanovništva	US Census Bureau
7	total_population_balck	Broj crnog stanovništva	US Census Bureau
8	total_population_native	Broj starosedelačke populacije	US Census Bureau
9	total_population_asian	Broj azijskog stanovništva	US Census Bureau
10	total_population_latino	Broj latinoameričkog stanovništva	US Census Bureau
11	foreign_born_population%	Procenat stanovnika rođenih u inostranstvu (%)	Census Reporter
12	Population density p/km**2	Gustina stanovništva (st./km²)	US Census Bureau
13	Urban Population (%)	Procenat urbanog stanovništva (%)	US Census Bureau

#	Naziv varijable (original)	Opis	Izvor
14	Life Expectancy at Birth	Očekivano trajanje života pri rođenju (godine)	Measure of America
15	Infant Mortality Rate	Smrtnost odojčadi na 1.000 živorođenih	Measure of America
16	Child Mortality	Smrtnost dece 1–4 god. na 1.000 st.	Measure of America
17	State Per Capita GDP (\$)	BDP po glavi stanovnika (USD)	Measure of America / BEA
18	HDI_0-1_2010	Subnacionalni Indeks humanog razvoja, 2010 (0–1)	Global Data Lab
19	BUCAP15	Izgrađena urbana površina po stanovniku, 2015 (m²/st.)	GHS-UCDB (JRC)
20	no_health_insurance (%of_children_under_18)	Deca bez zdravstvenog osiguranja (%)	Measure of America
21	no_health_insurance (%Of_adults_ages_18-64)	Odrasli 18–64 bez zdravstvenog osiguranja (%)	Measure of America
22	no_health_insurance_total_population%	Ukupna populacija bez zdravstvenog osiguranja (%)	Measure of America
23	no_health_insurance_black%	Crno stanovništvo bez zdravstvenog osiguranja (%)	Measure of America
24	no_health_insurance_native%	Starosedelačka pop. bez zdravstvenog osiguranja (%)	Measure of America
25	no_health_insurance_asian%	Azijsko stanovništvo bez zdravstvenog osiguranja (%)	Measure of America
26	no_health_insurance_latino%	Latinoameričko st. bez zdravstvenog osiguranja (%)	Measure of America
27	no_health_insurance_white%	Belo stanovništvo bez zdravstvenog osiguranja (%)	Measure of America
4. Zdravstveni i medicinski podaci — 13 varijabli (America's Health Rankings / CDC)			
1	AdultsAlcoholConsumption%	Prevalenca prekomerne konzumacije alkohola — binge drinking (%)	America's Health Rankings (CDC)
2	Obesity (% age 20 and older)	Prevalenca gojaznosti, odrasli 20+ (%)	America's Health Rankings (CDC)
3	adults_with_hypertension%	Prevalenca hipertenzije kod odraslih (%)	America's Health Rankings (CDC)
4	population_with_high_cholesterol%	Prevalenca visokog holesterola (%)	America's Health Rankings (CDC)
5	population_smoking%	Prevalenca pušenja (%)	America's Health Rankings (CDC)
6	population_with_CVD%	Prevalenca kardiovaskularnih bolesti (%)	America's Health Rankings (CDC)
7	population_with_diabetes%	Prevalenca dijabetesa (%)	America's Health Rankings (CDC)
8	population_with_cancer%	Prevalenca raka (%)	America's Health Rankings (CDC)
9	population_chronic_kidney_disease%	Prevalenca hronične bubrežne bolesti (%)	America's Health Rankings (CDC)
10	population_COBP%	Prevalenca hronične opstruktivne plućne bolesti — HOBP (%)	America's Health Rankings (CDC)

#	Naziv varijable (original)	Opis	Izvor
11	population_multiple_chronic_conditions%	Procenat sa dve ili više hroničnih bolesti (%)	America's Health Rankings (CDC)
12	population_physiscal_inactivity%	Prevalenca fizičke neaktivnosti (%)	America's Health Rankings (CDC)
13	Age 65+ CVD deaths	Smrtnost od KVB, populacija 65+, na 100.000	America's Health Rankings (CDC)

Biografija autora

Ognjen Milićević je rođen 18. maja 1985. godine. Medicinski fakultet Univerziteta u Beogradu završio je 2010. godine sa prosečnom ocenom 9,46. Osnovne studije na Elektrotehničkom fakultetu Univerziteta u Beogradu (Odsek za signale i sisteme) završio je 2014. godine sa prosečnom ocenom 8,38, na kojem je potom pohađao i master akademske studije do 2016. godine. Interdisciplinarne doktorske studije iz oblasti Biofizike na Univerzitetu u Beogradu upisao je u oktobru 2016. godine.

Svoju profesionalnu karijeru uspešno gradi u domenu medicine, bioinformatike i nauke o podacima. Od jula 2018. godine angažovan je na Medicinskom fakultetu Univerziteta u Beogradu na poziciji asistenta za oblast statistike i informatike. Pored akademskog angažovanja, od aprila 2021. godine radi kao vođa tima za mašinsko učenje i ekspert za medicinski domen u kompaniji HTEC Group. Prethodno profesionalno iskustvo sticao je radeći na pozicijama bioinformatičkog analitičara, konsultanta za podatke i produkt menadžera u organizacijama kao što su Seven Bridges Genomics, Genieus Genomics, Formdwell i na međunarodnom PULSE projektu.

Kao autor i koautor do sada je objavio više od 40 naučnih radova (od čega dva kao prvi autor) i redovno učestvuje na domaćim i međunarodnim naučnim konferencijama. Posедуje višegodišnje iskustvo u primeni programskih jezika Python, R i MATLAB u naučne i komercijalne svrhe. Tečno govori engleski jezik.

Izjava o autorstvu

Ime i prezime autora Ognjen Milićević

Broj indeksa 56/2016

Izjavljujem

da je doktorska disertacija pod naslovom

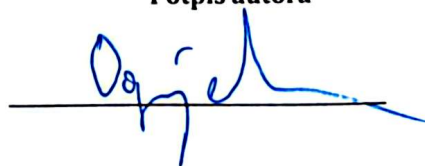
Računarsko modelovanje i mogućnosti predikcije reproduktionog broja COVID-19
pomoću sredinskih i sociodemografskih parametara

- rezultat sopstvenog istraživačkog rada;
- da disertacija u celini ni u delovima nije bila predložena za sticanje druge diplome prema studijskim programima drugih visokoškolskih ustanova;
- da su rezultati korektno navedeni i
- da nisam kršio/la autorska prava i koristio/la intelektualnu svojinu drugih lica.

U Beogradu,

25.5.2026.

Potpis autora



Izjava o istovetnosti štampane i elektronske verzije doktorskog rada

Ime i prezime autora Ognjen Milićević

Broj indeksa 56/2016

Studijski program Biofizika

Naslov rada Računarsko modelovanje i mogućnosti predikcije reproduktionog broja COVID-19 pomoću sredinskih i sociodemografskih parametara

Mentor Marko Đorđević; Igor Salom

Izjavljujem da je štampana verzija mog doktorskog rada istovetna elektronskoj verziji koju sam predao/la radi pohranjivanja u **Digitalnom repozitorijumu Univerziteta u Beogradu**.

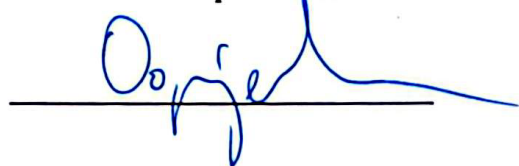
Dozvoljavam da se objave moji lični podaci vezani za dobijanje akademskog naziva doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada.

Ovi lični podaci mogu se objaviti na mrežnim stranicama digitalne biblioteke, u elektronskom katalogu i u publikacijama Univerziteta u Beogradu.

U Beogradu,

25.5.2026.

Potpis autora



Izjava o korišćenju

Ovlašćujem Univerzitetsku biblioteku „Svetozar Marković“ da u Digitalni repozitorijum Univerziteta u Beogradu unese moju doktorsku disertaciju pod naslovom:

Računarsko modelovanje i mogućnosti predikcije reproduktionog broja COVID-19 pomoću sredinskih i sociodemografskih parametara

koja je moje autorsko delo.

Disertaciju sa svim prilogima predao/la sam u elektronskom formatu pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju pohranjenu u Digitalnom repozitorijumu Univerziteta u Beogradu i dostupnu u otvorenom pristupu mogu da koriste svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons) za koju sam se odlučio/la.

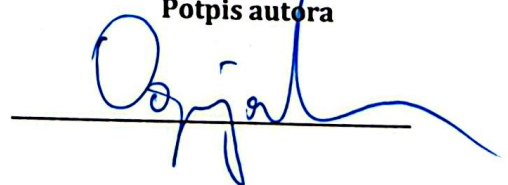
1. Autorstvo (CC BY)
2. Autorstvo nekomercijalno (CC BY-NC)
3. Autorstvo nekomercijalno bez prerada (CC BY-NC-ND)
4. Autorstvo nekomercijalno deliti pod istim uslovima (CC BY-NC-SA)
5. Autorstvo bez prerada (CC BY-ND)
6. Autorstvo deliti pod istim uslovima (CC BY-SA)

(Molimo da zaokružite samo jednu od šest ponuđenih licenci. Kratak opis licenci je sastavni deo ove izjave).

U Beogradu,

25.5.2026.

Potpis autora



1. **Autorstvo.** Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence, čak i u komercijalne svrhe. Ovo je najslobodnija od svih licenci.

2. **Autorstvo - nekomercijalno.** Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca ne dozvoljava komercijalnu upotrebu dela.

3. **Autorstvo - nekomercijalno - bez prerada.** Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, bez promena, preoblikovanja ili upotrebe dela u svom delu, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca ne dozvoljava komercijalnu upotrebu dela. U odnosu na sve ostale licence, ovom licencom se ograničava najveći obim prava korišćenja dela.

4. **Autorstvo - nekomercijalno - deliti pod istim uslovima.** Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence i ako se prerada distribuira pod istom ili sličnom licencom. Ova licenca ne dozvoljava komercijalnu upotrebu dela i prerada.

5. **Autorstvo - bez prerada.** Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, bez promena, preoblikovanja ili upotrebe dela u svom delu, ako se navede ime autora na način određen od strane autora ili davaoca licence. Ova licenca dozvoljava komercijalnu upotrebu dela.

6. **Autorstvo - deliti pod istim uslovima.** Dozvoljavate umnožavanje, distribuciju i javno saopštavanje dela, i prerade, ako se navede ime autora na način određen od strane autora ili davaoca licence i ako se prerada distribuira pod istom ili sličnom licencom. Ova licenca dozvoljava komercijalnu upotrebu dela i prerada. Slična je softverskim licencama, odnosno licencama otvorenog koda.