



УНИВЕРЗИТЕТ У БЕОГРАДУ
ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Modelovanje i predikcija koncentracije polena u Srbiji korišćenjem geostatistike i mašinskog učenja

DIPLOMSKI RAD

Student

Luka Milutinović
2021/0103

Mentor

prof. dr Predrag Tadić
vanredni profesor

BEOGRAD, 2025.

Sadržaj

1 Uvod	1
1.1 Cilj rada	1
1.2 Struktura rada	1
2 Teorijski osnov	2
2.1 Polen i alergije	2
2.2 Vremenske serije	2
2.2.1 Primeri primene	3
2.2.2 Matematičko definisanje i komponente vremenskih serija	3
2.2.3 Stacionarnost i testovi stacionarnosti (ADF i KPSS)	4
2.2.4 Box-Cox transformacija	7
2.2.5 Beli šum	7
2.2.6 Autokorelacija (ACF) i parcijalna autokorelacija (PACF)	8
2.3 SARIMAX	9
2.3.1 AR model	9
2.3.2 MA model	10
2.3.3 ARMA modeli	11
2.3.4 ARIMA model	11
2.3.5 SARIMA i SARIMAX modeli	12
2.3.6 Kriterijumi za izbor modela	13
2.4 Prophet	14
2.5 Random Forest	15
2.6 Geostatistika i <i>kriging</i>	16
2.6.1 Prostorno-vremenski <i>kriging</i>	18
3 Podaci i metode	20
3.1 Opis podataka	20
3.2 Veza između meteoroloških parametara i koncentracije polena ambrozije	22
3.3 Preprocesiranje i čišćenje podataka	24
3.4 Imputacija	27
3.5 Modelovanje vremenske serije i podešavanje parametara SARIMAX modela	31
3.6 Modelovanje vremenske serije korišćenjem Prophet modela	34
3.7 Modelovanje koncentracija polena korišćenjem Random Forest modela	35
3.8 Metrike evaluacije modela	36
4 Rezultati	38
4.1 Imputacija	38
4.1.1 Metodologija evaluacije	38
4.1.2 Rezultati po alergenima	38
4.1.3 Diskusija	40
4.2 Predikcija	41
4.2.1 Metodologija evaluacije	41
4.2.2 SARIMAX	41
4.2.3 Prophet model	45
4.2.4 Random Forest	49
4.3 Uporedna analiza modela	54

5 Zaključak

56

Literatura

57

1 Uvod

U poslednjih nekoliko decenija, sve veća pažnja posvećuje se praćenju koncentracije polena u vazduhu zbog njegovog značajnog uticaja na zdravlje ljudi i svakodnevni život. Polen je jedan od glavnih alergena u spoljašnjoj sredini, pa osobe koje pate od polenskih alergija često osećaju pogoršanje simptoma upravo u vreme cvetanja biljaka. Tačna i pravovremena prognoza koncentracije polena može da bude ključna za blagovremeno započinjanje terapije, što pomaže u smanjenju jačine i trajanja neprijatnih simptoma.

Zbog urbanizacije, zagađenja vazduha i klimatskih promena, broj osoba koje pate od polenskih alergija u stalnom je porastu širom sveta, pa i u Srbiji. Prema dostupnim podacima, procenjuje se da između 30% i 40% stanovništva Srbije ima neku formu alergije na polen [1, 2].

Klinička istraživanja su pokazala da osobe koje započnu terapiju protiv polenske alergije nedelju dana pre početka sezone imaju značajno blaže simptome u poređenju sa onima koji terapiju započinju tek kada simptomi nastupe, što naglašava značaj lokalno precizne prognoze polena za planiranje terapije i smanjenje zdravstvenih problema [3, 4].

Predviđanje koncentracije polena predstavlja složen zadatak, jer zavisi od sezonskih obrazaca, meeteoroloških faktora (temperatura, vetar, padavine, vlažnost), kao i od učestalog problema nedostajućih podataka. Proces merenja zahteva kontinuirano praćenje i laboratorijsku analizu, što se zbog tehničkih i finansijskih ograničenja često ne sprovodi u kontinuitetu, pa podaci ostaju nepotpuni ili nedostupni u određenim vremenskim i prostornim intervalima [5, 6].

Za rešavanje problema nedostajućih vrednosti primjenjen je prostorno-vremenski *kriging* - geostistički model mašinskog učenja za interpolaciju podataka. Nakon toga, za predikciju koncentracije polena korišćeni su modeli vremenskih serija i mašinskog učenja: **SARIMAX**, **Prophet** i **Random Forest** regresija.

1.1 Cilj rada

Cilj ovog rada je da se razvije i implementira sistem za predikciju koncentracije polena korišćenjem modela vremenskih serija i mašinskog učenja, uz prethodnu imputaciju nedostajućih vrednosti pomoću geostatističkih modela zasnovanih na principima mašinskog učenja, kako bi se unapredila tačnost i pouzdanost prognoze, a time i omogućila pravovremena reakcija osoba sa alergijama, unapređenje pčelarske proizvodnje i doprinosi uspešnijem oprašivanju biljaka u poljoprivredi.

1.2 Struktura rada

Prvo poglavlje predstavlja uvod i motivaciju za istraživanje. Drugo poglavlje daje teorijski osnov o polenu, modelima vremenskih serija, mašinskom učenju i geostatistici. Treće poglavlje opisuje podatke i metode, uključujući opis podataka, pretprocesiranje, imputaciju i treniranje modela. Četvrto poglavlje prikazuje rezultate eksperimenta i diskusiju o dobijenim rezultatima. Peto poglavlje donosi zaključke i predloge za budući rad.

2 Teorijski osnov

2.1 Polen i alergije

Polen su mikroskopske čestice koje biljke proizvode tokom reprodukcije, a njihov osnovni biološki zadatak je oprašivanje [7]. U vazduhu se najčešće nalaze poleni drveća, trava i korova, posebno oni koje vetr lako prenosi na velike udaljenosti [6]. Veličina polenovih zrna varira od 10 do 100 mikrometara, što im omogućava da se lako šire i dugo zadrže u atmosferi [8].

Svaka biljna vrsta proizvodi svoj karakteristični polen, a dominantni tipovi polena u vazduhu menjaju se tokom godine. U proleće preovlađuju poleni drveća poput breze, leske i jove, u kasno proleće i početkom leta dominiraju poleni trava, dok krajem leta i početkom jeseni dominira polen korova poput ambrozije [4]. Najčešće alergijske reakcije izazivaju poleni trava, ambrozije i drveća, a posebno je problematična ambrozija zbog dugog perioda cvetanja i visoke koncentracije polena, zbog čega predstavlja jedan od najzastupljenijih i najjačih alergena u regionu [3].

Alergijska reakcija nastaje kada imuni sistem prepozna polen kao stranu i potencijalno opasnu supstancu, iako on sam po sebi nije štetan [1]. Tipični simptomi uključuju svrab očiju i nosa, kijanje, curenje nosa, suzenje očiju, dok kod osoba sa astmom može doći do otežanog disanja i napada gušenja. Ovi simptomi često dovode do problema sa spavanjem, osećaja hroničnog umora i smanjene koncentracije, što značajno utiče na svakodnevnu produktivnost [5]. Istraživanja pokazuju da alergijski rinitis može dovesti do smanjenja radne efikasnosti i do 40%, posebno u periodima visoke koncentracije polena [6]. Ipak, polen ima i korisne aspekte. Za pčelarstvo je od presudne važnosti jer predstavlja glavni izvor proteina za pčele, neophodan za razvoj legla i jačanje pčelinjeg društva, čime se direktno utiče na proizvodnju meda i polena kao dijetetskog suplementa [7]. U poljoprivredi obezbeđuje oprašivanje, ključno za stabilne i visoke prinose mnogih kultura, dok se u ekologiji i palinologiji koristi za proučavanje promena vegetacije i klime tokom istorije, budući da polen ostaje sačuvan u sedimentima hiljadama godina [8].

Koncentracija polena u vazduhu zavisi od vrste biljaka, geografskog područja, kao i od meteoroških faktora poput temperature, padavina, vlažnosti i brzine vetra [9]. Ovi faktori zajedno određuju vreme i intenzitet oslobođanja polena, zbog čega predikcija koncentracije polena zahteva modele koji uzimaju u obzir sezonske obrasce i promene u okruženju [6]. Razumevanje i predviđanje dinamike polena u vazduhu od velikog su značaja za javno zdravlje, pčelarstvo, poljoprivredu i ekološka istraživanja [5].

2.2 Vremenske serije

Predviđanje budućih događaja oduvek je fasciniralo čovečanstvo. Još od najranijih vremena ljudi su posmatrali kretanje zvezda, Mesečeve mene i promene godišnjih doba kako bi znali kada da seju, žanju ili migriraju. Danas se predviđanje temelji na matematičkim i statističkim metodama, ali osnovna ideja ostaje nepromenjena – težnja ka sagledavanju i razumevanju budućih pojava i procesa. Na primer, stari Egipćani su pre više od 4000 godina beležili nivo reke Nil kako bi procenili plodnost zemljišta i planirali raspodelu resursa. Time su, iako nesvesno, primenjivali principe vremenskih serija, iako formalni statistički okvir kakav se danas poznaje tada još nije postojao [10].

Kroz istoriju, međutim, predviđanje nije uvek bilo prihvaćeno. U starom Rimu, car Konstantin II je 357. godine nove ere izdao dekret kojim je zabranio svako proricanje budućnosti i konsultovanje matematičara ili astrologa, smatrujući da takve aktivnosti mogu ugroziti vlast i izazvati nemire. Slično tome, u Engleskoj je 1824. godine donet zakon prema kojem su svi koji su tvrdili da mogu da predviđaju budućnost proglašavani za prevarante i osuđivani na zatvor i prinudni rad. Iako su ove zabrane bile usmerene na astrologiju i gatanje, one pokazuju da je pokušaj predikcije kroz istoriju imao i snažnu društvenu i političku težinu [11].

Prvi značajni matematički radovi o vremenskim serijama pojavili su se početkom 20. veka, kada su Yule i Slutsky postavili temelje teorije linearnih modela, uvodeći ključne pojmove poput autoregresivnih procesa i analize periodičnih komponenti ekonomskih podataka [12, 13]. Tokom 1950-ih i 1960-ih, Robert G. Brown je popularizovao metodu eksponencijalnog zaglađivanja, omogućivši jednostavna i efikasna predviđanja u industrijskoj praksi. Dalje, 1970. godine, Box i Jenkins razvili su **ARIMA** metodologiju, čime su otvorili put za praktičnu i sistematsku primenu analiza vremenskih serija u ekonomiji, inženjerstvu i brojnim drugim oblastima [14].

Danas se vremenske serije definišu kao niz uzastopnih posmatranja neke promenljive Y_t prikupljenih kroz vreme, gde je redosled posmatranja od suštinskog značaja.

2.2.1 Primeri primene

- U meteorologiji, vremenske serije se koriste za prognozu temperature, padavina i veta, što direktno utiče na planiranje u poljoprivredi.
- U finansijama, primenjuju se za analizu i predikciju cena akcija, deviznih kurseva i tržišnih indeksa.
- U energetici, omogućavaju planiranje proizvodnje i balansiranje potrošnje električne energije.
- U biomedicini, koriste se za analizu srčanog ritma, EEG i drugih fizioloških signala radi dijagnostike i praćenja stanja pacijentata.

2.2.2 Matematičko definisanje i komponente vremenskih serija

Vremenska serija predstavlja realizaciju stohastičkog procesa čije se vrednosti posmatraju u različitim vremenskim tačkama [12, 14]. Formalno, vremenska serija je skup slučajnih promenljivih:

$$\{Y_t\}, \quad t \in T$$

gde je T diskretan skup vremena, a Y_t realizacija promenljive u trenutku t .

U praktičnoj analizi vremenskih serija, pretpostavlja se da jedna realizacija odražava svojstva celog stohastičkog procesa [13].

Svaka vremenska serija može se razložiti na nekoliko komponenti koje zajedno definišu njen ponasanjanje. Te komponente su:

- **Trend (T_t)**: dugoročna promena u nivou serije kroz vreme. Može biti rastući, opadajući ili stabilan.
- **Sezonalnost (S_t)**: obrazac koji se ponavlja u pravilnim vremenskim intervalima, npr. godišnja sezonalnost koncentracije polena [6].
- **Cikličnost (C_t)**: oscilacije koje se javljaju u nepravilnim i dužim vremenskim intervalima, često povezane sa ekonomskim ciklusima ili klimatskim promenama.
- **Slučajna komponenta (R_t)**: reziduali koji se ne mogu objasniti trendom, sezonalnošću ili cikličnošću; predstavljaju šum u podacima.

Postoje dva osnovna modela dekompozicije vremenskih serija [12, 14]:

- **Aditivni model:**

$$Y_t = T_t + S_t + C_t + R_t$$

Koristi se kada su amplitudne sezonalnosti i fluktuacija nezavisne od nivoa serije.

- **Multiplikativni model:**

$$Y_t = T_t \times S_t \times C_t \times R_t$$

Koristi se kada amplitudne sezonalnosti i šuma rastu proporcionalno nivou serije.

2.2.3 Stacionarnost i testovi stacionarnosti (ADF i KPSS)

Za većinu metoda analize vremenskih serija, posebno za **ARIMA** modele, stacionarnost je osnova pretpostavka za statistički validno i pouzdano modeliranje [14, 15, 16]. U praktičnom smislu, stacionarna serija ima:

- konstantnu srednju vrednost
- konstantnu varijansu kroz vreme
- kovarijansu koja zavisi samo od vremenskog pomaka, a ne od samog vremena.

Formalno, vremenska serija $\{Y_t\}$ je slabo stacionarna ako:

$$\begin{aligned} E[Y_t] &= \mu \\ \text{Var}(Y_t) &= \sigma^2 \\ \text{Cov}(Y_t, Y_{t+h}) &= \gamma(h) \end{aligned}$$

gde su μ i σ^2 konstante, a $\gamma(h)$ zavisi samo od vremenskog pomaka h , a ne od samog vremena t [14].

Serija je strogo stacionarna ako zajednička raspodela bilo kog vektora

$$F_X(Y_{t_1}, Y_{t_2}, \dots, Y_{t_k})$$

ostaje nepromenjena za svaki vremenski pomak h , odnosno:

$$F_X(Y_{t_1}, Y_{t_2}, \dots, Y_{t_k}) = F_X(Y_{t_1+h}, Y_{t_2+h}, \dots, Y_{t_k+h})$$

za svako k i h [14].

U praksi, uslov striktne stacionarnosti je veoma jak i teško proverljiv, te se za većinu modela koristi slaba stacionarnost, koja zahteva konstantnu sredinu, konstantnu varijansu i kovarijansu koja zavisi samo od vremenskog pomaka [15, 16].

Augmented Dickey-Fuller (ADF) test Jedan od najčešće korišćenih testova za proveru stacionarnosti vremenskih serija je **Augmented Dickey-Fuller (ADF) test** [15, 14]. Ovaj test se primarno koristi za ispitivanje prisustva **jediničnog korena (unit root)** u vremenskoj seriji.

Jedinični koren. Vremenska serija poseduje jedinični koren ako sadrži komponentu slučajnog hoda, što implicira nestacionarnost. Na primer, AR(1) proces definisan kao:

$$Y_t = \phi Y_{t-1} + \varepsilon_t,$$

gde je ε_t beli šum, biće stacionaran ako $|\phi| < 1$. Kada $\phi = 1$, model prelazi u formu slučajnog hoda:

$$Y_t = Y_{t-1} + \varepsilon_t,$$

kod koje varijansa raste linearno kroz vreme, što znači da proces nije stacionaran.

Izvodenje regresione forme ADF testa.

Osnovni *Dickey-Fuller* test polazi od AR(1) modela:

$$Y_t = \phi Y_{t-1} + \varepsilon_t.$$

Oduzimanjem Y_{t-1} sa obe strane, dobija se:

$$Y_t - Y_{t-1} = \phi Y_{t-1} - Y_{t-1} + \varepsilon_t,$$

$$\Delta Y_t = (\phi - 1)Y_{t-1} + \varepsilon_t.$$

Neka je $\gamma = \phi - 1$, dobija se osnovni oblik regresione jednačine:

$$\Delta Y_t = \gamma Y_{t-1} + \varepsilon_t.$$

Testiranje hipoteze $\phi = 1$ ekvivalentno je testiranju $\gamma = 0$.

Augmented Dickey-Fuller (ADF) regresioni model.

ADF test proširuje osnovni *Dickey-Fuller* test uključivanjem:

- konstantnog člana (α),
- determinističkog trenda (βt),
- i p zaostalih diferenciranih vrednosti ΔY_{t-i} radi eliminacije autokorelacije u rezidualima.

Konačna regresiona forma ADF testa glasi:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^p \delta_i \Delta Y_{t-i} + \varepsilon_t.$$

Hipoteze ADF testa:

- H_0 : serija ima jedinični koren (nije stacionarna), tj. $\gamma = 0$,
- H_1 : serija je stacionarna, tj. $\gamma \neq 0$.

Odluka se donosi na osnovu procene parametra γ . Ako je γ značajno manji od nule, odbacuje se nulta hipoteza i zaključuje da je serija stacionarna. Kritične vrednosti za ADF test razlikuju se od standardnih vrednosti t -raspodele i određene su tablično.

KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test KPSS test je statistički test koji se koristi za proveru stacionarnosti vremenskih serija [16]. Za razliku od ADF testa, KPSS test ima **suprotne hipoteze**.

Hipoteze KPSS testa:

- H_0 : vremenska serija je stacionarna (oko konstante ili determinističkog trenda),
- H_1 : vremenska serija nije stacionarna (ima jedinični koren).

Model KPSS testa. KPSS test polazi od modela vremenske serije zapisanog kao:

$$Y_t = r_t + \beta t + \varepsilon_t,$$

gde je:

- r_t – komponenta slučajnog hoda,
- βt – deterministički trend,
- ε_t – stacionarna greška (beli šum).

Komponenta slučajnog hoda definiše se kao:

$$r_t = r_{t-1} + u_t,$$

gde je u_t beli šum. Ako je varijansa u_t jednaka nuli, tada je r_t konstantan i Y_t je trend-stacionaran [16].

Interpretacija rezultata.

KPSS test ispituje prisustvo komponente slučajnog hoda u vremenskoj seriji:

- Ako komponenta slučajnog hoda *ne postoji* ($\text{Var}(u_t) = 0$), serija je stacionarna oko konstante ili trenda.
- Ako komponenta slučajnog hoda *postoji* ($\text{Var}(u_t) > 0$), serija je nestacionarna jer varijansa raste tokom vremena.

Test statistika.

KPSS test statistika se računa kao:

$$\text{KPSS} = \frac{1}{T^2} \frac{\sum_{t=1}^T S_t^2}{\hat{\sigma}^2}$$

Reziduali $e_t = Y_t - \hat{Y}_t = Y_t - (\hat{r}_t + \hat{\beta}t)$ predstavljaju empirijske procene stacionarne komponente ε_t iz modela $Y_t = r_t + \beta t + \varepsilon_t$ i koriste se za izračunavanje kumulativne sume S_t .

gde je:

- S_t – kumulativna suma reziduala, $S_t = \sum_{i=1}^t e_i$,
- $\hat{\sigma}^2$ – procena dugoročne varijanse reziduala,
- T – veličina uzorka.

Ako je vrednost test statistike veća od kritične vrednosti, odbacuje se nulta hipoteza i zaključuje da serija nije stacionarna [16].

Kombinacija sa ADF testom.

U praksi se preporučuje upotreba KPSS testa zajedno sa ADF testom, jer imaju suprotne hipoteze:

- Ako ADF odbacuje H_0 (serija je stacionarna) i KPSS ne odbacuje H_0 (serija je stacionarna), zaključuje se stacionarnost.
- Ako ADF ne odbacuje H_0 (serija nije stacionarna) i KPSS odbacuje H_0 (serija nije stacionarna), zaključuje se nestacionarnost.
- Ako su rezultati kontradiktorni, potrebna je dodatna analiza (diferenciranje, transformacije, vizuelna analiza) [15, 16].

2.2.4 Box-Cox transformacija

Box-Cox transformacija predstavlja metodu koja primenjuje različite funkcionalne oblike na podatke s ciljem stabilizacije varijanse i približavanja raspodele normalnoj [17].

Transformacija se definiše na sledeći način:

$$Y_t^{(\lambda)} = \begin{cases} \frac{Y_t^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(Y_t), & \lambda = 0 \end{cases}$$

gde je Y_t posmatrana vrednost vremenske serije.

Parametar λ određuje oblik transformacione funkcije i bira se tako da maksimizuje verodostojnost modela, omogućavajući da transformisani podaci što bolje prate normalnu raspodelu.

Box-Cox transformacija ima dva osnovna cilja:

1. Stabilizacija varijanse.

Kod mnogih vremenskih serija varijansa zavisi od nivoa serije (heteroskedastičnost). Primena ove transformacije omogućava da varijansa bude približno konstantna bez obzira na nivo, što je ključno za korišćenje linearnih modela poput **ARIMA** i **SARIMAX**.

2. Približavanje normalnosti raspodele.

Brojne statističke metode prepostavljaju normalno distribuirane greške. Box-Cox transformacija omogućava da distribucija podataka bude bliža normalnoj, čime se poboljšava tačnost procene parametara i predikcija modela.

2.2.5 Beli šum

Beli šum predstavlja osnovni koncept u teoriji vremenskih serija i stohastičkih procesa [14, 18]. Formalno, beli šum je sekvenca nezavisnih i identično distribuiranih slučajnih promenljivih sa konstantnim očekivanjem i varijansom.

Matematički, niz ε_t predstavlja beli šum ako:

$$\begin{aligned} E[\varepsilon_t] &= 0, \\ \text{Var}(\varepsilon_t) &= \sigma^2, \\ \text{Cov}(\varepsilon_t, \varepsilon_{t+\tau}) &= 0, \quad \text{za svaki } \tau \neq 0. \end{aligned}$$

Karakteristike belog šuma:

- Vrednosti su statistički među sobom nezavisne.
- Vrednosti imaju istu varijansu.
- Nema serijske autokorelaciјe – koeficijent autokorelaciјe jednak je nuli za sva vremenska kašnjenja osim za $\tau = 0$.
- Spektralna gustina je konstantna, što znači da su sve frekvencije podjednako zastupljene, analogno „belom svetlu“ koje sadrži sve talasne dužine.

Beli šum ima ključnu ulogu u modeliranju vremenskih serija jer predstavlja osnovnu prepostavku za reziduale – nakon uklanjanja trenda, sezonalnosti i cikličnih komponenti, reziduali bi trebalo da se ponašaju kao beli šum. Analiza njihove autokorelacijske funkcije koristi se za proveru adekvatnosti modela; ako reziduali nisu beli šum, to ukazuje da model nije obuhvatio sve obrasce u podacima i potrebno ga je unaprediti [14, 18].

2.2.6 Autokorelacija (ACF) i parcijalna autokorelacija (PACF)

Autokorelacija predstavlja meru povezanosti između posmatranja vremenske serije koja su međusobno udaljena za k vremenskih jedinica (*lag*) [14, 18]. Formalno, autokorelacija reda k definiše se kao:

$$\rho_k = \frac{\text{Cov}(Y_t, Y_{t-k})}{\text{Var}(Y_t)}.$$

Autokorelaciona funkcija (ACF) prikazuje autokorelacije za sve vrednosti kašnjenja k , pri čemu za slabostacionarne procese važi:

- $\rho_0 = 1$,
- $|\rho_k| < 1$, za $k > 0$,
- $\lim_{k \rightarrow \infty} \rho_k = 0$.

Intervali poverenja za ACF

Za stacionarnu seriju bez autokorelacijske funkcije (beli šum), standardna greška (SE) procene autokorelacijske funkcije reda k aproksimira se kao:

$$\text{SE}(\rho_k) \approx \frac{1}{\sqrt{N}}.$$

gde je N broj posmatranja. Za 95% interval poverenja koristi se aproksimacija:

$$\pm \frac{1.96}{\sqrt{N}}.$$

Preciznija formula (Box-Jenkins) uzima u obzir prethodne autokorelacijske funkcije:

$$\text{SE}(\rho_k) = \sqrt{\frac{1 + 2 \sum_{i=1}^{k-1} \rho_i^2}{N}}.$$

Parcijalna autokorelacija (PACF)

Parcijalna autokorelacija meri povezanost između Y_t i Y_{t-k} nakon što se eliminiše uticaj svih posrednih vrednosti između njih [18, 19]. PACF reda k definiše se kao poslednji koeficijent ϕ_{kk} u AR(k) modelu:

$$Y_t = \phi_{k1}Y_{t-1} + \phi_{k2}Y_{t-2} + \cdots + \phi_{kk}Y_{t-k} + \varepsilon_t.$$

Intuitivno:

- ϕ_{kk} predstavlja direktnu (parcijalnu) korelaciju između Y_t i Y_{t-k} uz eliminaciju uticaja svih kašnjenja između.

- Prethodni koeficijenti $\phi_{k1}, \dots, \phi_{k(k-1)}$ uklanjaju indirektne veze preko bližih vremenskih kašnjenja.

Računanje PACF

PACF se računa pomoću:

1. Yule-Walker jednačina (rekurzivno) [18],
2. Regresionih metoda, gde se za svako vremensko kašnjenje k određuje AR(k) model, a PACF reda k je poslednji koeficijent ϕ_{kk} .

Intervali poverenja za PACF

Pod pretpostavkom belog šuma, standardna greška PACF približno je jednaka standardnoj grešci ACF:

$$SE \approx \frac{1}{\sqrt{N}}.$$

te se za 95% interval poverenja koristi:

$$\pm \frac{1.96}{\sqrt{N}}.$$

Osobine PACF

Za slabo stacionarne procese:

- Kod AR(p) procesa, PACF ima nenulte vrednosti do reda p i naglo seče na nulu za $k > p$.
- Kod MA(q) procesa, PACF eksponencijalno ili oscilatorno opada bez oštrog prekida.
- Kod belog šuma, PACF je približno jednaka nuli za sve $k > 0$.

2.3 SARIMAX

2.3.1 AR model

Autoregresioni (AR) modeli predstavljaju osnovnu klasu linearnih vremenskih modela, gde se trenutna vrednost posmatranog procesa modeluje kao linearna kombinacija prethodnih vrednosti i belog šuma [14, 18].

Formalna definicija:

Autoregresioni model reda p , označen kao AR(p), definisan je jednačinom:

$$X_t = c + \sum_{i=1}^p \phi_i X_{t-i} + \varepsilon_t \tag{1}$$

gde je:

- X_t — trenutna vrednost procesa u vremenskom trenutku t ,
- c — konstanta (često $c = 0$ ako je serija centrirana, tj. $E[X_t] = 0$),
- ϕ_i — koeficijenti AR modela,

- p — red modela,
- ε_t — beli šum, $\varepsilon_t \sim N(0, \sigma^2)$.

ACF i PACF karakteristike:

- **ACF** kod AR(p) modela eksponencijalno ili oscilatorno opada [14].
- **PACF** pokazuje značajne vrednosti do reda p , dok su svi vremenski pomaci većeg reda unutar intervala poverenja [18, 19].

Praktično određivanje reda modela:

Red AR modela (p) određuje se analizom PACF grafa. Parametri modela ϕ_i procenjuju se korišćenjem Yule-Walker jednačina, metode najmanjih kvadrata ili maksimalne verodostojnosti.

2.3.2 MA model

Model pomerajućeg proseka (MA) predstavlja drugu osnovnu klasu linearnih vremenskih modela, gde se trenutna vrednost procesa izražava kao linearna kombinacija trenutne i prethodnih vrednosti belog šuma [14, 18].

Formalna definicija:

Model pomerajućeg proseka reda q , MA(q), definisan je jednačinom:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i} \quad (2)$$

gde je:

- X_t — trenutna vrednost procesa,
- μ — srednja vrednost procesa,
- θ_i — koeficijenti MA modela,
- q — red modela,
- ε_t — beli šum $\sim N(0, \sigma^2)$.

ACF i PACF karakteristike:

- **ACF** kod MA(q) modela pokazuje značajne vrednosti za vremenska kašnjenja do reda q , a zatim naglo opada [14].
- **PACF** kod MA modela ne pokazuje karakterističan obrazac i stoga se primarno koristi ACF za određivanje reda modela [18].

Praktično određivanje reda modela:

Red MA modela (q) određuje se analizom ACF grafa. Parametri modela θ_i procenjuju se metodama najmanjih kvadrata ili maksimalne verodostojnosti.

2.3.3 ARMA modeli

Modeli autoregresije i pomerajućeg proseka (*Autoregressive Moving Average*, ARMA) predstavljaju kombinaciju AR i MA modela i koriste se za modelovanje stacionarnih vremenskih serija koje pokazuju i autoregresivne i pokretne prosečne komponente [14, 18].

Formalna definicija:

ARMA model reda (p, q) , označen kao $\text{ARMA}(p, q)$, definisan je jednačinom:

$$X_t = c + \sum_{i=1}^p \phi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \quad (3)$$

gde je:

- X_t – trenutna vrednost procesa u vremenskom trenutku t ,
- c – konstanta (često $c = 0$ ako je serija centrirana),
- ϕ_i – koeficijenti AR dela modela,
- θ_j – koeficijenti MA dela modela,
- p – red AR dela modela,
- q – red MA dela modela,
- ε_t – beli šum $\sim N(0, \sigma^2)$.

Intuitivno objašnjenje:

ARMA modeli kombinuju sposobnost AR modela da koristi prethodne vrednosti serije i MA modela da uključi uticaj prethodnih nasumičnih šumova. Time omogućavaju modelovanje kompleksnih obrazaca u stacionarnim vremenskim serijama.

ACF i PACF karakteristike:

Kod ARMA modela oba grafa, ACF i PACF, mogu pokazivati eksponencijalno ili oscilatorno opadanje, bez jasnog naglog „odsecanja“ kao kod čistih AR ili MA modela. Analiza ACF i PACF koristi se kao smernica za inicijalnu procenu strukture modela, dok konačan izbor p i q potvrđuju kriterijumi poput AIC ili BIC [14, 18].

Praktično određivanje reda modela:

Pri određivanju redova p i q , koristi se heuristika: AR red posmatra se preko PACF, MA red preko ACF, a parametri se procenjuju metodom maksimalne verodostojnosti.

2.3.4 ARIMA model

Iako ARMA model uspešno modeluje stacionarne serije, često se sreću serije sa trendovima ili sezonskim obrascima. ARIMA modeli rešavaju problem **trendova** primenom diferenciranja [14, 19].

Operator vremenskog kašnjenja:

Uvodi se operator vremenskog kašnjenja L , definisan kao $L^k X_t = X_{t-k}$. Na primer, $LX_t = X_{t-1}$.

Diferenciranje:

Nestacionarna serija transformiše se u stacionarnu primenom diferenciranja:

$$\nabla X_t = X_t - X_{t-1} = (1 - L)X_t \quad (4)$$

Za d -to diferenciranje primenjuje se $(1 - L)^d X_t$.

Formalna definicija:

ARIMA(p, d, q) model definisan je kao ARMA model primenjen na d -tu diferenciranu seriju:

$$(1 - L)^d X_t = c + \sum_{i=1}^p \phi_i (1 - L)^d X_{t-i} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}. \quad (5)$$

gde je $(1 - L)^d X_t$ d -ta diferencirana serija, d broj diferenciranja, a ostale označke kao kod ARMA modela.

Napomena o diferenciranju:

Prekomerno diferenciranje može ukloniti dugoročne obrasce i uvoditi neželjene oscilacije, stoga je važno diferencirati samo po potrebi.

Praktično određivanje reda modela:

1. Analiza originalne serije radi utvrđivanja potrebe za diferenciranjem.
2. Diferenciranje dok serija ne postane stacionarna (d).
3. Određivanje redova p i q analizujući diferenciranu seriju.
4. Procena parametara metodom maksimalne verodostojnosti i validacija modela AIC/BIC.

2.3.5 SARIMA i SARIMAX modeli

Za vremenske serije koje pored trendova pokazuju i sezonske obrasce, koristi se proširenje ARIMA modela poznato kao **SARIMA (Seasonal ARIMA)** model. Ovaj model omogućava simultano modelovanje nesezonskih i sezonskih komponenti vremenske serije [14, 18].

Formalna definicija:

SARIMA model se označava kao ARIMA(p, d, q)(P, D, Q) $_s$, gde:

- p, d, q – redovi nesezonskih AR, diferenciranja i MA komponenti,
- P, D, Q – redovi sezonskih AR, diferenciranja i MA komponenti,
- s – sezonski period (npr. $s = 12$ za mesečne podatke sa godišnjom sezonskom komponentom).

Matematički, model uključuje i nesezonske i sezonske polinome u operatoru vremenskog kašnjenja L :

$$\Phi_P(L^s)\phi_p(L)(1 - L)^d(1 - L^s)^D X_t = \Theta_Q(L^s)\theta_q(L)\varepsilon_t \quad (6)$$

gde su:

- $\phi_p(L)$ i $\theta_q(L)$ – nesezonski AR i MA polinomi,
- $\Phi_P(L^s)$ i $\Theta_Q(L^s)$ – sezonski AR i MA polinomi.

Praktično određivanje parametara:

Prilikom izbora sezonskih parametara P i Q , posmatraju se ACF i PACF na sezonskim vremenskim kašnjenjima:

- **PACF** na sezonskim vremenskim kašnjenjima koristi se za određivanje reda sezonske AR (P) komponente.
- **ACF** na sezonskim vremenskim kašnjenjima koristi se za određivanje reda sezonske MA (Q) komponente.

Parametar sezonskog diferenciranja D bira se ukoliko postoji sezonski trend koji je potrebno eliminisati da bi se postigla stacionarnost.

SARIMA vs SARIMAX:

SARIMA model omogućava modelovanje samo na osnovu internih (endogenih) komponenti serije, dok **SARIMAX (Seasonal ARIMA with eXogenous regressors)** predstavlja proširenje koje uključuje i egzogene regresorske promenljive (npr. temperaturu, vlažnost, brzinu veta) kao dodatne ulaze u model. Na taj način SARIMAX model omogućava precizniju predikciju kada su poznate spoljašnje varijable koje utiču na posmatrani proces.

Formalna definicija SARIMAX modela:

$$\Phi_P(L^s)\phi_p(L)(1-L)^d(1-L^s)^D X_t = \Theta_Q(L^s)\theta_q(L)\varepsilon_t + \beta^T Z_t \quad (7)$$

gde je:

- Z_t - vektor egzogenih regresorskih promenljivih u trenutku t ,
- β - vektor koeficijenata regresora.

SARIMAX model kombinuje sezonske i nesezonske ARIMA komponente sa efektima eksternih varijabli, čime omogućava modelovanje i predikciju kompleksnih vremenskih serija pod uticajem poznatih spoljnih faktora [14, 18].

2.3.6 Kriterijumi za izbor modela

Za poređenje i odabir najboljeg modela koriste se kriterijumi poput **AIC**, **AICc** i **BIC**. Ovi kriterijumi omogućavaju izbor modela koji postiže dobru ravnotežu između preciznosti u prilagođavanju podacima i kompleksnosti modela, čime se smanjuje rizik od preobučavanja [20].

- **AIC (Akaike Information Criterion)** je definisan kao:

$$AIC = -2 \ln(L) + 2k, \quad (8)$$

gde je L maksimalna verodostojnost modela, a k broj procenjenih parametara. AIC meri kvalitet modela balansirajući između njegove složenosti (broja parametara) i njegove prilagođenosti posmatranim podacima (maksimalna verodostojnost). Model sa nižom vrednošću AIC smatra se boljim.

- **AICc** predstavlja korigovanu verziju AIC za male uzorke i definisan je kao:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}, \quad (9)$$

gde je n broj opservacija. Kada je n mali u odnosu na k , AICc daje pouzdaniju procenu jer dodatno penalizuje kompleksnost modela.

- **BIC (Bayesian Information Criterion)** je definisan kao:

$$BIC = -2 \ln(L) + k \ln(n), \quad (10)$$

gde je n broj opservacija. BIC uvodi jaču penalizaciju za kompleksnost modela u poređenju sa AIC, te preferira jednostavnije modele ukoliko razlika u prilagođenosti nije značajna.

2.4 Prophet

Prophet je model razvijen od strane *Facebook-a (Meta)* za predikciju vremenskih serija koje imaju izražene trendove, sezonske obrasce i efekte praznika. Namjenjen je brzim i robusnim predikcijama, čak i kada podaci sadrže *outlier-e* ili nedostajuće vrednosti, a posebno je praktičan za primenu od strane analitičara koji nemaju duboko znanje iz vremenskog modelovanja [21].

Formalna definicija:

Prophet modelira vremensku seriju kao kombinaciju trend komponente, sezonskih obrazaca i prazničnih efekata, uz prisustvo nasumične greške.

- **Aditivni model:**

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \quad (11)$$

- **Multiplikativni model:**

$$y(t) = g(t) \times (1 + s(t)) + h(t) + \varepsilon_t \quad (12)$$

Gde su:

- $g(t)$ – trend komponenta, linearna ili logistička,
- $s(t)$ – sezonska komponenta modelovana Furijeovim redovima,
- $h(t)$ – praznični (*holidays*) efekti,
- ε_t – nasumična greška (beli šum).

Trend komponenta:

Kod segmentnog linearног trenda:

$$g(t) = (k + a(t)^T \delta)(t - t_0) + (m + a(t)^T \gamma) \quad (13)$$

- k – osnovna brzina promene trenda,
- m – početna vrednost funkcije,
- δ – promene u nagibu trenda u tačkama promene (*changepoints*),
- $a(t)$ – indikator funkcija za *changepoints*,
- γ – korekcija za promene nagiba, da funkcija ostane kontinualna.

Sezonska komponenta:

Sezonski obrasci modeluju se Furijeovim redovima:

$$s(t) = \sum_{n=1}^N \left[a_n \cos\left(\frac{2\pi n t}{T}\right) + b_n \sin\left(\frac{2\pi n t}{T}\right) \right] \quad (14)$$

- T – period (npr. 365.25 za godišnju sezonskost),
- N – red Furijevog reda,
- a_n, b_n – koeficijenti Furijeovih redova.

Glavni parametri Prophet modela:

- `growth` – tip trenda: "linear" ili "logistic",
- `changepoint_prior_scale` – kontrola fleksibilnosti modela u tačkama promene,
- `seasonality_mode` – aditivna ili multiplikativna sezonska komponenta,
- `seasonality_prior_scale` – fleksibilnost sezonskih obrazaca,
- `holidays_prior_scale` – uticaj prazničnih efekata,
- `changepoints` – korisnički definisana lista tačaka promene,
- `n_changepoints` – broj automatski detektovanih changepoints,
- `fourier_order` – red Furijevog reda za sezonske komponente.

2.5 Random Forest

Random Forest je popularan ansambl algoritam mašinskog učenja zasnovan na skupu stabala odlučivanja. Razvijen je sa ciljem poboljšanja preciznosti predikcije i robusnosti modela smanjenjem varijanse i problema preobučavanja koji su česti kod pojedinačnih stabala [22].

Ansambl (*ensemble*) metode:

Ansambl metode kombinuju više osnovnih modela kako bi proizvele jači model. Ideja je da kombinacija više slabijih modela (npr. stabala odlučivanja) rezultuje boljim performansama nego bilo koji od pojedinačnih modela. Random Forest koristi ***bagging*** (*Bootstrap Aggregation*), gde se stabla treniraju na različitim podskupovima trening skupa dobijenim bootstrap uzorkovanjem, a zatim kombinuju prosečno (za regresiju) ili glasanjem (za klasifikaciju).

Stablo odlučivanja (*decision trees*):

Stablo odlučivanja je struktura koja donosi odluke tako što u svakom čvoru postavlja uslov zasnovan na nekoj ulaznoj karakteristici. Podaci se razdvajaju rekurzivno dok se ne dođe do listova stabla, koji predstavljaju konačnu predikciju. Za svaki čvor bira se karakteristika i granica podele koja najbolje razdvaja podatke prema nekoj meri homogenosti.

Merenje homogenosti (*purity*):

- **Gini nečistoća:**

$$\text{Gini} = 1 - \sum_{i=1}^K p_i^2, \quad (15)$$

gde je p_i verovatnoća da uzorak pripada klasi i , a K broj klasa.

- **Entropija:**

$$\text{Entropija} = - \sum_{i=1}^K p_i \log_2 p_i, \quad (16)$$

gde je p_i verovatnoća da uzorak pripada klasi i .

- **Srednja kvadratna greška (MSE):**

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2, \quad (17)$$

gde je N broj uzoraka u čvoru, y_i stvarna vrednost ciljne promenljive, a \bar{y} srednja vrednost ciljne promenljive u čvoru.

Varijansa i ansambl efekat:

Jedno stablo odlučivanja ima visoku varijansu, što može dovesti do preobučavanja i nestabilnih predikcija. Random Forest smanjuje varijansu tako što kombinuje više stabala, povećavajući robusnost modela. Međutim, zbog delimične korelacije stabala, posle određenog broja stabala dolazi do efekta **zasićenja**, gde dodavanje novih stabala ne poboljšava performanse, ali se povećava vreme treniranja.

Parametri Random Forest modela:

- `n_estimators` – broj stabala odlučivanja u šumi.
- `max_depth` – maksimalna dubina stabala.
- `min_samples_split` – minimalan broj uzoraka potreban za podelu čvora.
- `min_samples_leaf` – minimalan broj uzoraka u listu stabla.
- `max_features` – broj prediktora-a razmatranih pri svakoj podeli čvora, radi smanjenja korelacije između stabala.

Važnost prediktora (*Feature importance*):

Random Forest omogućava procenu važnosti prediktora, koja meri koliko je svaki prediktor doprineo smanjenju nečistoće kroz sva stabla i njihove čvorove. Veća vrednost znači da je prediktor češće korišćen za podelu podataka i imao veći uticaj na predikciju.

2.6 Geostatistika i kriging

Geostatistika:

Geostatistika predstavlja skup metoda za analizu, modelovanje i predikciju prostornih podataka [23, 24]. Osnovna prepostavka geostatistike je da vrednosti posmatrane promenljive zavise ne samo od globalnih trendova, već i od lokacije u prostoru (i/ili vremenu) zbog prostorne autokorelaciјe. Drugim rečima, bliže tačke imaju sličnije vrednosti nego udaljene, što omogućava interpolaciju i predikciju na nepoznatim lokacijama.

Variogram:

Osnovni alat geostatistike je variogram, funkcija koja opisuje prostornu zavisnost podataka [25]. Variogram meri koliko se vrednosti promenljive razlikuju u zavisnosti od udaljenosti između tačaka.

Formalno, empirijski variogram se definiše kao:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (18)$$

gde je:

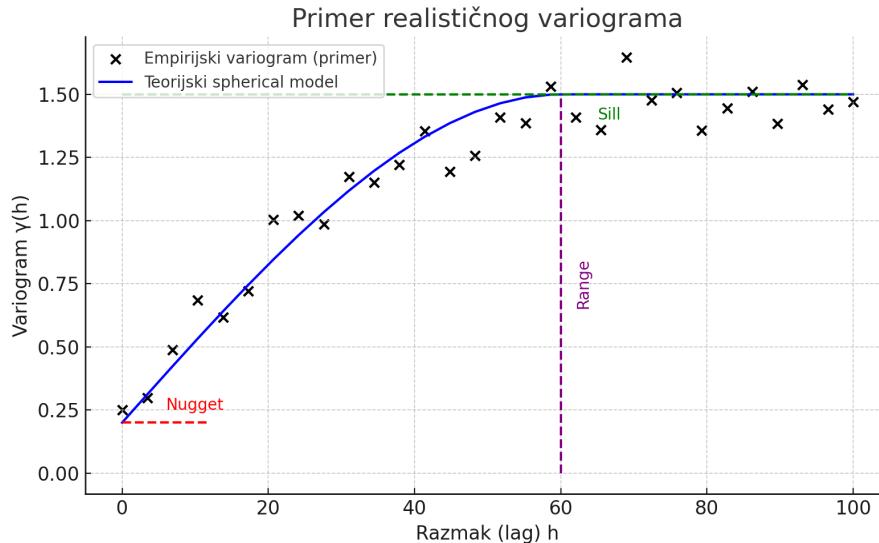
- h – prostorni razmak,
- $N(h)$ – broj parova tačaka razdvojenih udaljenošću h ,
- $Z(x_i)$ – vrednost posmatrane promenljive u tački x_i .

Elementi variograma:

Teorijski variogram model se definiše pomoću tri glavna parametra:

- **Nugget (c_0):** vrednost variograma za $h = 0$, predstavlja mikroskopske varijacije ili mernu grešku.

- **Sill** ($c_0 + c$): asimptotska vrednost variograma, dostiže se pri velikim udaljenostima i predstavlja ukupnu varijansu.
- **Range** (a): udaljenost na kojoj variogram dostiže sill. Posle te udaljenosti vrednosti više nisu korelisane.



Slika 1: Primer teorijskog variograma sa označenim parametrima *nugget*, *sill* i *range*.

Najčešće korišćeni teorijski modeli variograma su **gausovski**, **sferni** i **eksponencijalni** modeli.

Gausovski model karakteriše vrlo gladak porast na malim udaljenostima, sferni model brzo dostiže *sill*, dok eksponencijalni model opisuje situacije u kojima zavisnost opada postepeno i nikada potpuno ne nestaje, već teži *sill*-u asimptotski.

U ovom radu korišćen je **eksponencijalni model variograma**, jer se pokazao kao najprikladniji za modelovanje prostornih i vremenskih zavisnosti koncentracije polena, koje imaju postepeno opadajući obrazac korelacije bez nagle saturacije na određenom rastojanju.

Matematički, eksponencijalni model se definiše kao:

$$\gamma(h) = c_0 + c \left[1 - \exp \left(-\frac{h}{a} \right) \right] \quad (19)$$

gde je:

- c_0 - nugget efekat,
- c - partial sill (razlika između sill i nugget),
- a - parametar rasprostiranja koji određuje brzinu rasta variograma (pravi range približno odgovara $3a$).

Kriging:

Kriging predstavlja metod geostatističke interpolacije koji procenjuje vrednost u nepoznatoj tački kao linearu kombinaciju poznatih vrednosti u okolini, pri čemu se težinski koeficijenti određuju tako da se dobije nepomereni (*unbiased*) estimator minimalne varijanse, poznat i kao **BLUE** (*Best Linear Unbiased Estimator*) [23, 24].

Model je koncipiran tako da daje veću težinu tačkama koje su bliže tački procene, ali kada su dve tačke međusobno veoma blizu, njihova ukupna težina se deli između njih, čime estimator koristi što veći broj različitih tačaka za dobijanje optimalne procene.

Za primenu *kriging-a* pretpostavlja se:

- **Prostorna stacionarnost:** proces ima konstantnu srednju vrednost i variogram zavisi samo od razmaka između tačaka, a ne od apsolutnih koordinata.
- **Poznavanje variograma:** poznat ili procenjen model variograma koji tačno opisuje prostornu zavisnost posmatranih podataka.
- **Nepostojanje značajnog globalnog trenda:** u slučaju postojanja trenda, potrebno ga je ukloniti.

Procena vrednosti u tački x_0 data je:

$$\hat{Z}(x_0) = \sum_{i=1}^N \lambda_i Z(x_i) \quad (20)$$

gde su:

- $Z(x_i)$ – poznate vrednosti na lokacijama x_i ,
- λ_i – težinski koeficijenti.

Određivanje koeficijenata:

Koeficijenti λ_i dobijaju se rešavanjem sistema jednačina, koji se bazira na teorijskom variogramu i uslovu nepomerene estimacije. Za običan *kriging* sistem jednačina glasi:

$$\begin{bmatrix} \gamma(x_1, x_1) & \cdots & \gamma(x_1, x_N) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_N, x_1) & \cdots & \gamma(x_N, x_N) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_N \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(x_1, x_0) \\ \vdots \\ \gamma(x_N, x_0) \\ 1 \end{bmatrix} \quad (21)$$

gde je μ Lagranžov multiplikator.

2.6.1 Prostorno-vremenski *kriging*

Opis:

Klasični *kriging* modelira zavisnost samo u prostoru. Međutim, u slučaju podataka koji zavise i od vremena, koristi se **prostorno-vremenski *kriging***, koji istovremeno modeluje prostornu i vremensku autokorelaciju podataka [26].

Definicija variograma:

Kod prostorno-vremenskog *kriging-a*, variogram zavisi i od prostorne udaljenosti h_s i vremenske udaljenosti h_t , te se definiše kao:

$$\gamma(h_s, h_t) = \frac{1}{2N(h_s, h_t)} \sum_{i=1}^{N(h_s, h_t)} [Z(x_i, t_i) - Z(x_i + h_s, t_i + h_t)]^2 \quad (22)$$

gde je:

- h_s – prostorni razmak između tačaka,
- h_t – vremenski razmak između tačaka,

- $N(h_s, h_t)$ – broj parova tačaka razdvojenih prostornim razmakom h_s i vremenskim razmakom h_t .

Modeli prostorno-vremenskog variograma:

Za modelovanje se koriste:

- **Separabilni modeli:** variogram se modeluje kao proizvod ili zbir prostornog i vremenskog variograma.
- **Neseparabilni modeli:** koriste jedinstvene funkcije koje simultano opisuju zavisnost u prostoru i vremenu bez mogućnosti razdvajanja na prostornu i vremensku komponentu.

3 Podaci i metode

3.1 Opis podataka

Podaci o polenu:

Podaci o koncentracijama polena korišćeni u ovom radu preuzeti su sa Portala otvorenih podataka Republike Srbije, iz baze "Polen – objedinjeni podaci od 2016. godine" [27].

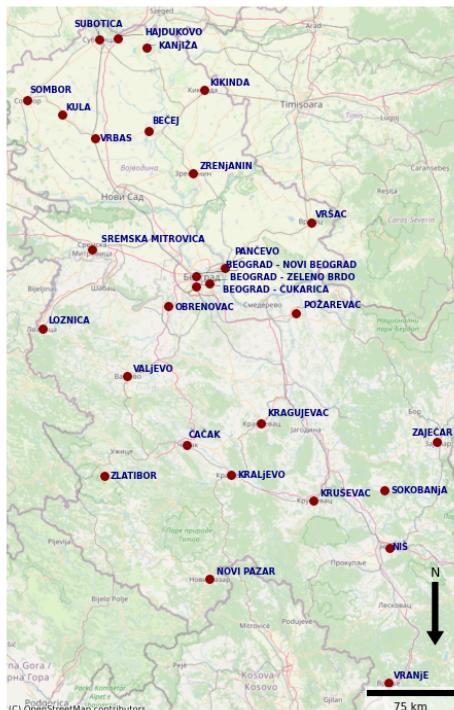
Skup podataka obuhvata dnevne koncentracije polena različitih biljnih vrsta izmerene na više mernih stanica širom Srbije u periodu od 2016. godine do danas. Svaki zapis sadrži sledeće atribute:

- Datum merenja
- Lokacija merenja (naziv stanice)
- Geografske koordinate stanice (širina, dužina)
- Vrsta polena
- Koncentracija polena (zrna/m^3)
- Donja vrednost – prag koncentracije ispod kojeg nema biološkog efekta (za ambroziju 30 zrna/m^3 , za sve ostale alergene 60 zrna/m^3) [28]
- Gornja vrednost – koncentracija iznad koje se klasificuje kao „vrlo visoka“ (100 zrna/m^3 za sve alergene) [29]

U daljoj analizi, prioritet će biti dat alergenima koji imaju **izraženu alergenost i visok značaj za zdravlje stanovništva**, te će fokus biti najviše na **ambroziji i jovi**. Od gradova, za detaljnju analizu odabrani su **Požarevac, Kragujevac i Pančevo**.

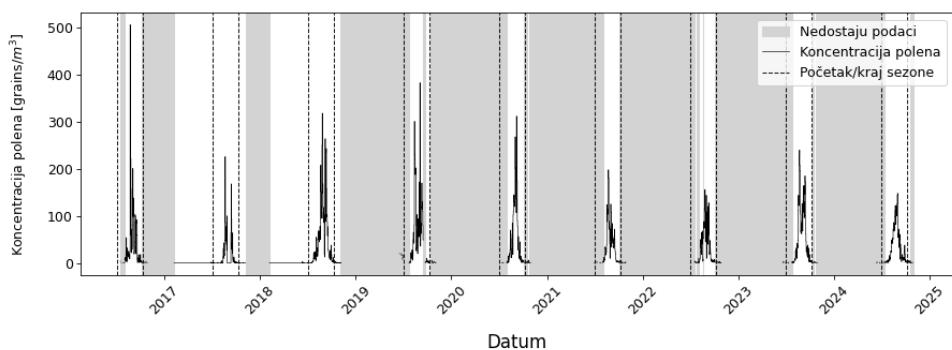
Kartografski prikaz svih lokacija na kojima se vrši merenje koncentracije polena dat je na slici 2. Na osnovu ovih lokacija, u nastavku su prikazane vremenske serije koncentracija odabralih alergena u gradovima Kragujevac i Pančevo, što je prikazano na slici 3.

Lokacije praćenja polena u Srbiji

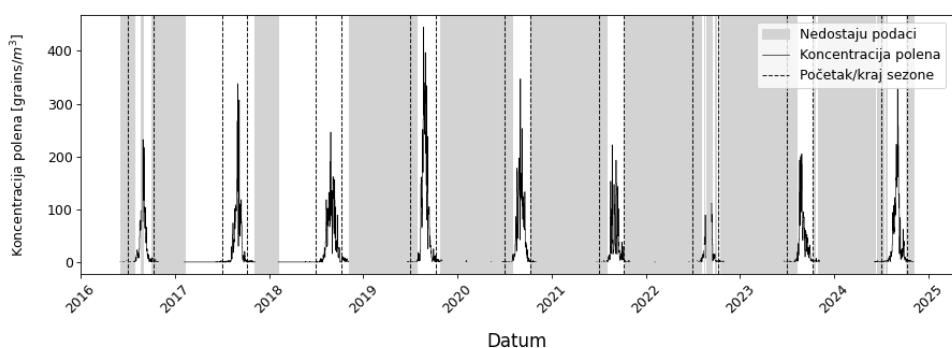


Slika 2: Lokacije mernih stanica za praćenje koncentracije polena u Srbiji.

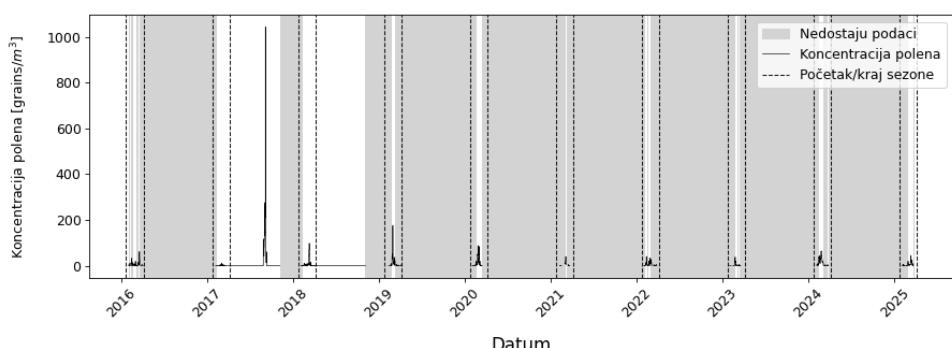
Dnevna koncentracija polena - AMBROZIJA (Kragujevac)



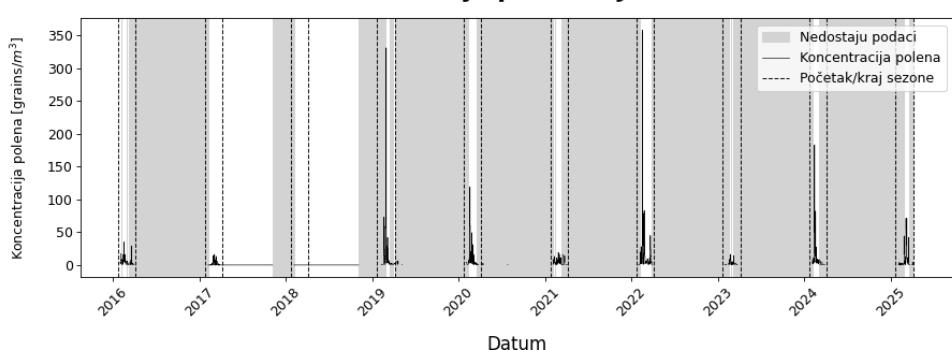
Dnevna koncentracija polena - AMBROZIJA (Pančevo)

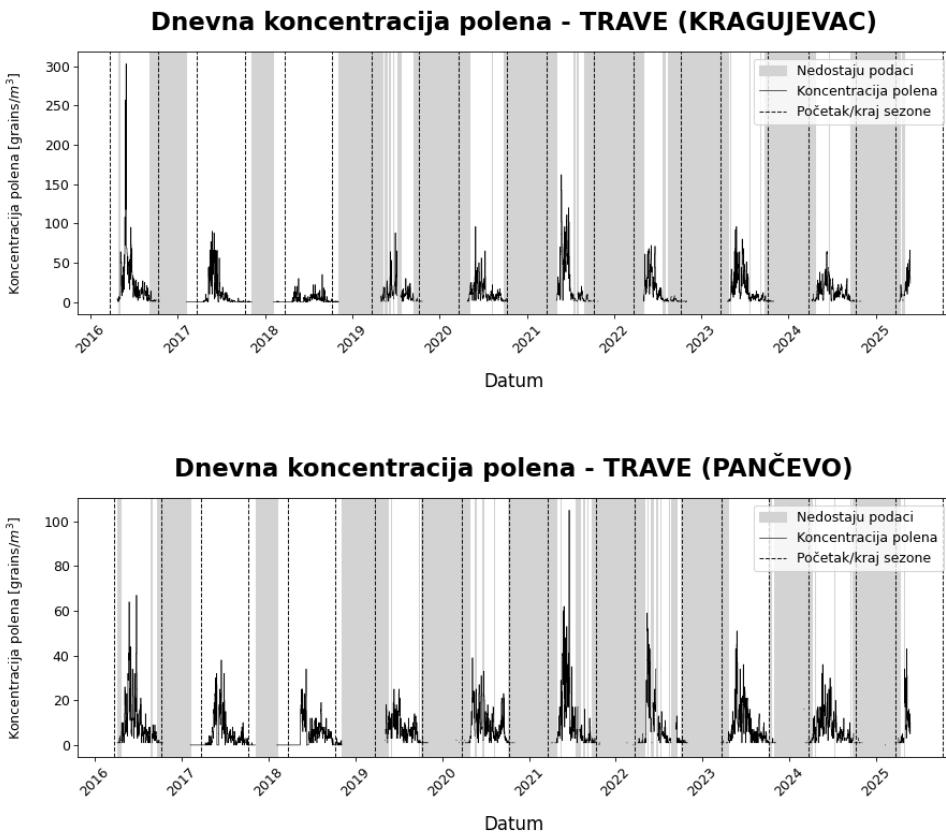


Dnevna koncentracija polena - JOVA (Kragujevac)



Dnevna koncentracija polena - JOVA (Pančevo)





Slika 3: Vremenske serije dnevnih koncentracija polena odabralih alergena u Kragujevcu i Pančevu.

Meteorološki podaci:

Meteorološki podaci preuzeti su sa *Copernicus Climate Data Store*, iz baze ERA5 "Reanalysis - ERA5 Single Levels" [30].

Podaci obuhvataju meteorološke varijable sa vremenskom rezolucijom od jednog sata i prostornom rezolucijom od 31 km, uključujući:

- Temperatura vazduha na 2 m ($^{\circ}\text{C}$)
- Relativna vlažnost vazduha (%)
- Brzina veta (m/s)
- Padavine (mm)
- Pravac duvanja veta (rad)

Za potrebe analize, podaci su agregirani u dnevne prosečne vrednosti i prostorno interpolirani na lokacije mernih stanica polena.

3.2 Veza između meteoroloških parametara i koncentracije polena ambrozije

Analiza zavisnosti između koncentracije polena alergena i meteoroloških parametara sprovedena je za tri lokacije: Kragujevac, Požarevac i Pančevo. Rezultati ukazuju na sledeće obrasce:

Temperatura kod ambrozije i jove uglavnom pokazuje **slabu pozitivnu korelaciju** sa koncentracijom polena, sa vrednostima koeficijenta oko $r \approx 0.1$. Kod trava, korelacija sa temperaturom

je uglavnom nešto niža, ali u poređenju sa ostalim meteorološkim parametrima i dalje predstavlja značajan prediktor koncentracije polena.

Vlažnost vazduha kod ambrozije i jove pokazuje **blago negativnu korelaciju**, što je u skladu sa biološkim mehanizmom otežanog širenja polena pri višoj vlažnosti vazduha. Kod trava, uticaj vlažnosti je manji, ali i dalje postoji tendencija ka negativnom efektu.

Padavine, brzina vetra i pravac vetra uglavnom pokazuju **vrlo malu ili zanemarljivu korelaciju** sa koncentracijom polena za sve alergene, sa vrednostima Pirsonovog koeficijenta bliskim nuli.

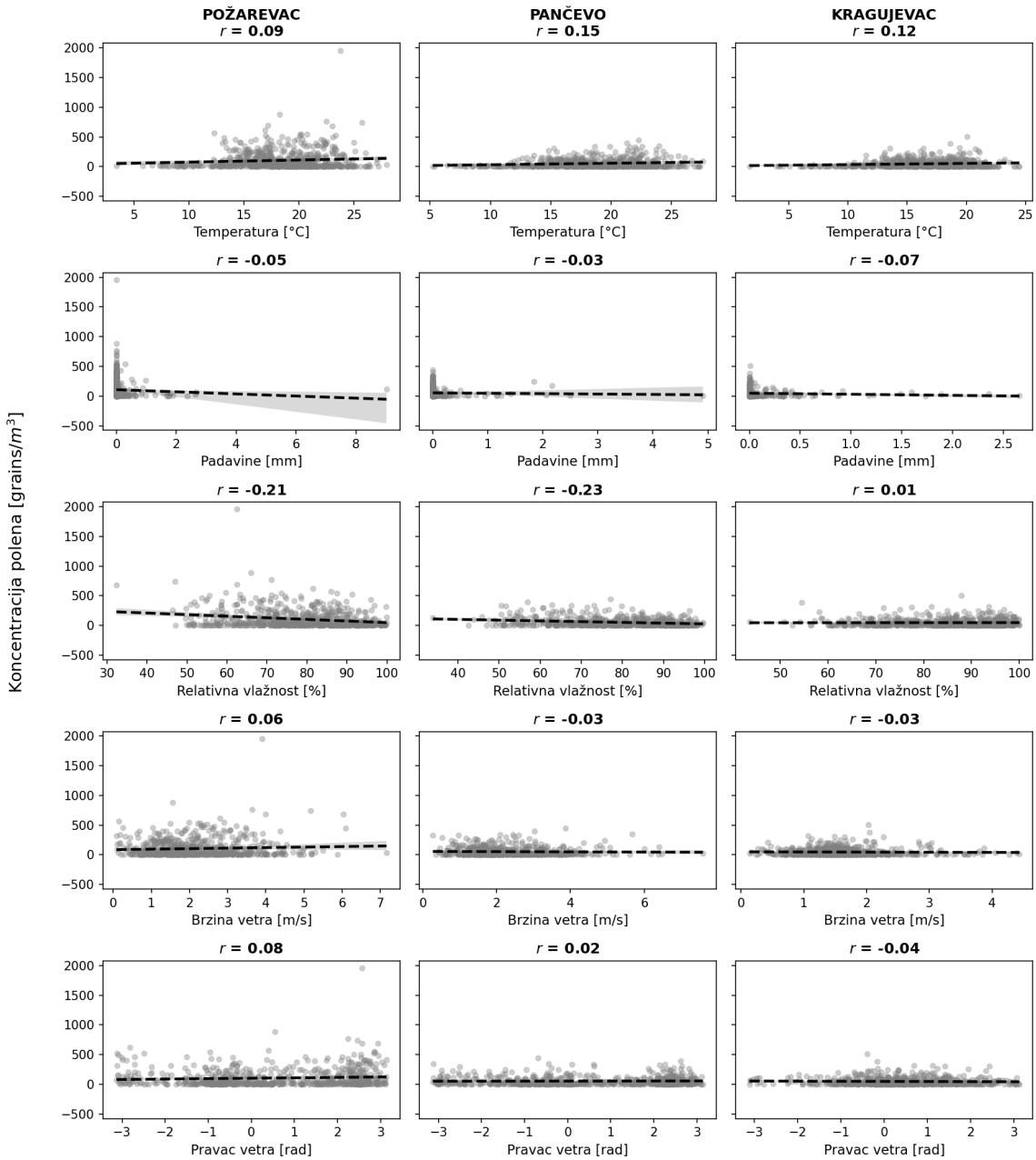
Ovi rezultati su u skladu sa prethodnim istraživanjima u regionu centralne i jugoistočne Evrope [31, 32, 33], koja ukazuju da temperatura podstiče otpuštanje i emisiju polena, dok viša vlažnost vazduha može otežati širenje polena u vazduhu. Padavine dovode do spiranja polena iz atmosfere, dok brzina i pravac vetra imaju složen nelinearan uticaj, zavisani od pravca transporta i regionalne orografske, koji linearna korelacija ne može u potpunosti da opiše.

S obzirom na to da za padavine, brzinu i pravac vetra **gotovo da ne postoji linearna veza sa koncentracijom polena**, ovi parametri **neće biti u razmatranje prilikom projektovanja linearnih modela**, kao što su *kriging*, SARIMAX i Prophet. Međutim, zbog mogućih nelinearnih interakcija, ovi parametri mogu biti **uključeni u Random Forest modele**, koji su robusniji na ovakve tipove zavisnosti.

Alergen	Parametar	POŽAREVAC	PANČEVO	KRAGUJEVAC
AMBROZIJA	Temperatura	0.089	0.146	0.118
	Padavine	-0.050	-0.027	-0.066
	Vlažnost vazduha	-0.212	-0.234	-0.007
	Brzina vetra	-0.062	-0.028	-0.026
	Pravac duvanja vetra	0.076	-0.017	-0.037
JOVA	Temperatura	0.115	0.060	0.170
	Padavine	-0.099	-0.059	-0.068
	Vlažnost vazduha	-0.095	-0.082	-0.180
	Brzina vetra	0.030	0.022	0.076
	Pravac duvanja vetra	0.101	-0.006	0.109
TRAVE	Temperatura	0.069	0.094	0.013
	Padavine	-0.077	-0.063	-0.058
	Vlažnost vazduha	-0.011	-0.005	0.084
	Brzina vetra	0.007	0.010	0.010
	Pravac duvanja vetra	0.108	0.025	0.016

Na slici 4 prikazana je zavisnost koncentracije polena ambrozije od meteoroloških parametara za sve analizirane lokacije i alergene, zajedno sa regresionim linijama i vrednostima Pirsonovog koeficijenta korelacije za svaku promenljivu.

AMBROZIJA: Zavisnost koncentracije polena od meteoroloških parametara po gradovima



Slika 4: Zavisnost koncentracije polena ambrozije od meteoroloških parametara (temperatura, vlažnost vazduha, padavine, brzina i pravac vetra) na analiziranim lokacijama.

3.3 Preprocesiranje i čišćenje podataka

Kako bi analiza bila precizna i pouzdana, podaci o koncentracijama polena najpre su detaljno preprocesirani i očišćeni. Prvi korak u ovom procesu bilo je filtriranje podataka kako bi se u analizi posmatrali samo oni periodi godine koji odgovaraju sezoni cvetanja alergena. Ovakav pristup ima dvostruko opravdanje: s jedne strane, van sezone često nedostaju merenja jer koncentracije polena nisu prisutne u vazduhu, dok s druge strane, upravo je tokom sezone najvažnije imati tačne informacije

i predikcije koncentracija zbog uticaja na zdravlje stanovništva [28, 29].

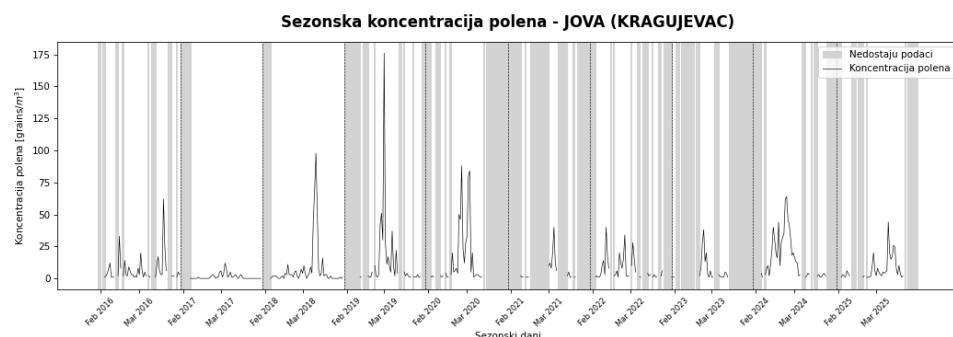
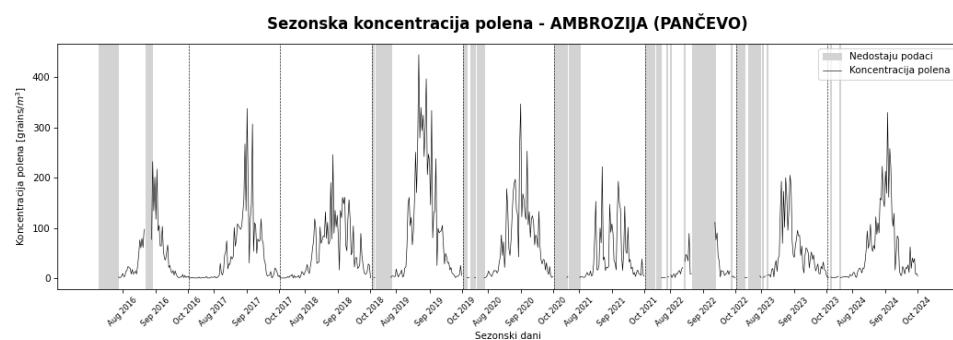
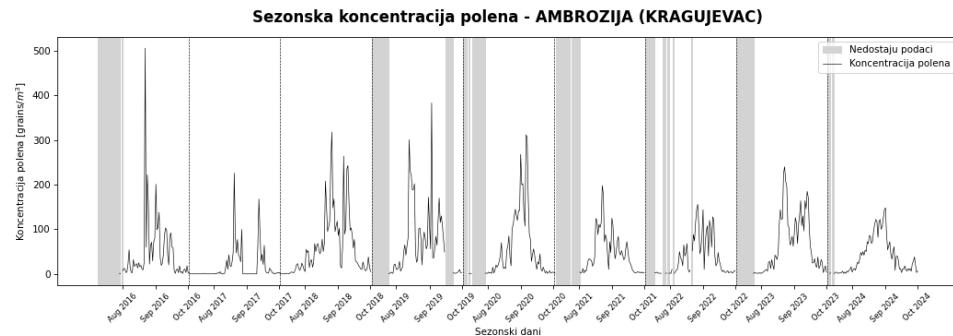
Na osnovu informacija sa Portala otvorenih podataka Republike Srbije sezonski periodi za odabrane alergene su sledeći:

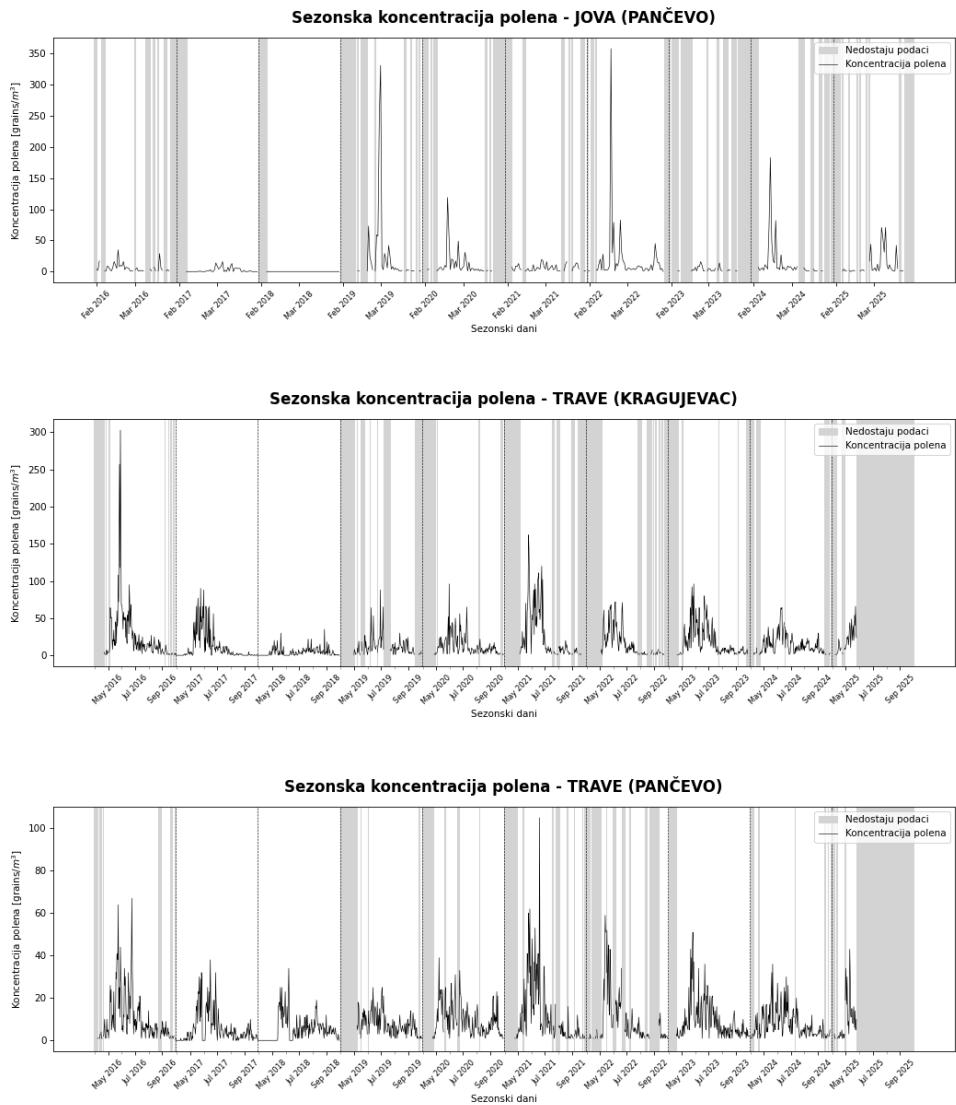
- **Ambrozija:** sezona traje od sredine jula do septembra.
- **Jova:** sezona traje od februara do sredine marta.
- **Trave:** sezona traje od aprila do septembra.

Kako bi predikcija bila robusnija i pokrivala potencijalne rane početke i kasne završetke sezona, ovi intervali su u radu blago prošireni i definisani na sledeći način:

- **Ambrozija:** od 9. jul do 1. oktobar
- **Jova:** od 30. januara do 31. mart
- **Trave:** od 30. marta do 1. oktobra

Na slici 5 prikazani su grafici koncentracija za odabrane alergene u sezonskim periodima za gradove Kragujevac i Pančevo.



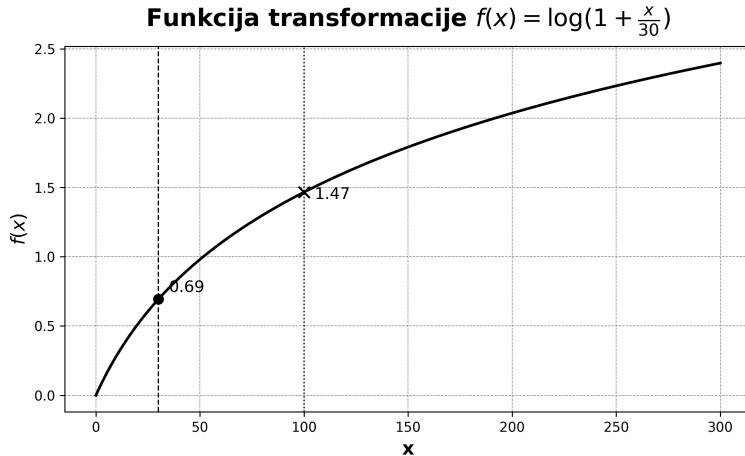


Slika 5: Sezonske koncentracije polena ambrozije, jove i trave u gradovima Kragujevac i Pančevo.

Transformacija podataka.

U cilju smanjenja varijanse i stabilizacije disperzije podataka, u analizi će biti primenjene sledeće transformacije:

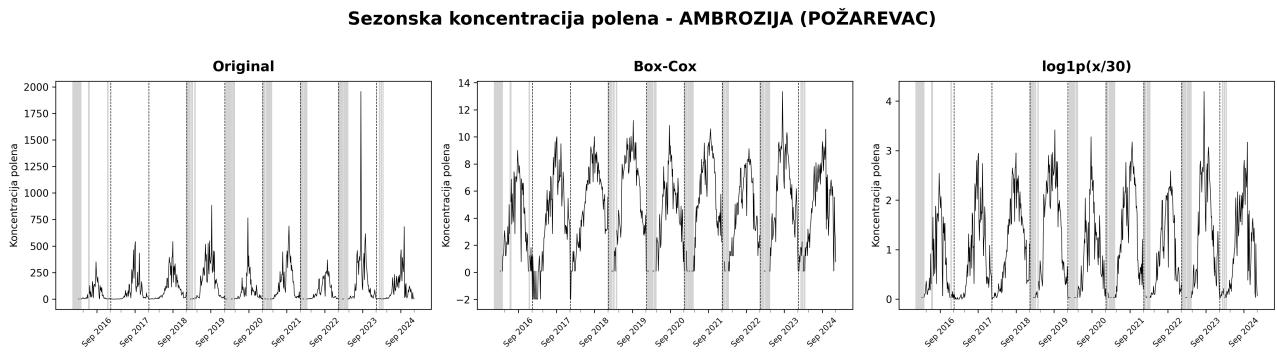
- **Box-Cox transformacija:** standardna metoda za normalizaciju podataka i smanjenje varijabilnosti varijanse [17].
- **Logaritamska transformacija:** transformacija oblika $\log(1 + \frac{x}{30})$, gde je x koncentracija polena. Ova transformacija je izabrana kako bi se vrednosti koncentracija oko 30 zrna/m^3 najviše razmakle, s obzirom na to da ta vrednost predstavlja prag kada kod ljudi dolazi do izraženijih alergijskih simptoma [28].



Slika 6: Primena logaritamske transformacije na koncentracije polena ambrozije.

Primena ovih transformacija omogućava da modeli predikcije bolje uoče strukturu podataka i umanjuje uticaj ekstremnih vrednosti, čime se povećava tačnost i stabilnost modela u daljim analizama.

Slika 7 prikazuje kako izgleda signal koncentracije polena ambrozije nakon primene različitih transformacija. Na originalnom signalu uočavaju se izraženi pikovi koncentracije tokom sezona cvetanja, sa velikim razlikama između godina. Nakon primene Box-Cox transformacije, varijansa signala je značajno smanjena, a raspodela približena normalnoj, što je korisno za primenu linearnih modela. Transformacija $\log(1 + \frac{x}{30})$ posebno ističe vrednosti oko 30 zrna/m^3 , praga iznad kog koncentracija ambrozije najčešće izaziva alergijske simptome kod osetljivih osoba.



Slika 7: Primer uticaja Box-Cox i logaritamske transformacije na signal koncentracije polena ambrozije u Požarevcu.

3.4 Imputacija

U cilju popunjavanja nedostajućih vrednosti koncentracija polena, u ovom radu primenjena je metoda *kriging* interpolacije [23, 34]. Proces imputacije sproveden je kroz sledeće korake:

1. Transformacija podataka.

Kriging model je primenjen nakon odgovarajućih transformacija podataka kako bi se stabilizovala varijansa i poboljšala prostorno-vremenska interpolacija koncentracija polena. Transformacije su sprovedene na dva različita načina:

- **Prvi pristup** podrazumeva je direktnu primenu Box–Cox ili $\log(1 + \frac{x}{30})$ transformacije na sirove podatke. U ovom slučaju, za Box–Cox transformaciju određivana je posebna vrednost parametra λ za svaku lokaciju pojedinačno, čime je omogućeno optimalno prilagođavanje distribuciji podataka svake merne stanice.
- **Drugi pristup** uključiva je prethodnu standardizaciju podataka za svaku lokaciju, tako što su koncentracije podeljene sa standardnom devijacijom odgovarajuće vremenske serije. Nakon standardizacije primenjivane su iste transformacije kao u prvom pristupu, pri čemu je u slučaju Box–Cox transformacije korišćena ista vrednost parametra λ za sve lokacije, čime je obezbeđena veća konzistentnost između različitih mernih stanica.“.

2. Procena trenda i sezonalnosti.

Nakon transformacije podataka, sprovedeno je modelovanje determinističkih komponenti koncentracija polena, koje obuhvataju trend i sezonalnost. Trend je modelovan primenom linearne regresije, gde su u model uključeni konstanta i linearna promena u vremenu, uz mogućnost dodavanja egzogenih promenljivih, poput temperature i vlažnosti vazduha.

Sezonska komponenta modelovana je pomoću Furijeovih redova opšte forme [35]:

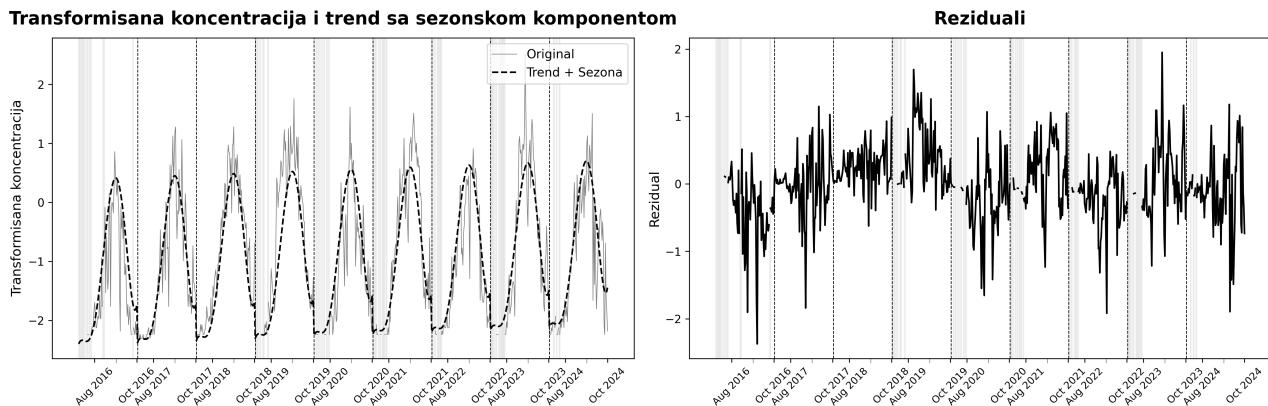
$$S_t = \sum_{k=1}^K \alpha_k \sin\left(\frac{2\pi kt}{T}\right) + \beta_k \cos\left(\frac{2\pi kt}{T}\right),$$

gde je T period sezonalnosti (u ovom slučaju $T = 365$, što odgovara godišnjem periodu), t trenutni dan u godini, α_k i β_k parametri sinusnih i kosinusnih komponenti, a K broj Furijeovih parova uključenih u model.

U radu je korišćen broj redova $K = 2$, kako bi se izbeglo preobučavanje modela i sačuvala glatkoća sezonske komponente.

Ovako definisan model omogućava izdvajanje trenda i sezonske strukture iz podataka, čime se preostala komponenta signala (reziduali) može smatrati stacionarnom i pogodnom za dalju analizu.

Sezonska dekompozicija - AMBROZIJA, lokacija: POŽAREVAC



Slika 8: Sezonska dekompozicija koncentracije polena ambrozije u Požarevcu: trend, sezonalnost i reziduali.

3. Modelovanje variograma.

Nakon izdvajanja trenda i sezonske komponente, sprovedeno je modelovanje variograma u cilju opisivanja strukture prostorno-vremenske zavisnosti koncentracija polena [23, 34]. U radu je pretpostavljeno da je variogram separabilan, odnosno da se može modelovati kao kombinacija prostornog i vremenskog variograma.

Empirijski variogram je izračunat prema standardnoj formuli:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(u_i) - z(u_i + h)]^2,$$

gde je $N(h)$ broj parova tačaka razdvojenih rastojanjem h , a $z(u_i)$ koncentracija polena u tački u_i .

Za oba variograma (vremenski i prostorni) korišćen je **eksponencijalni model** [34], budući da prirodne pojave često najbolje odgovaraju eksponencijalnoj strukturi zavisnosti. Opšta forma eksponencijalnog variograma definisana je izrazom:

$$\gamma(h) = c_0 + c \left(1 - e^{-\frac{h}{a}}\right),$$

gde je c_0 *nugget* efekt, c *sill*, a *range* parametar, a h rastojanje (prostorno ili vremensko).

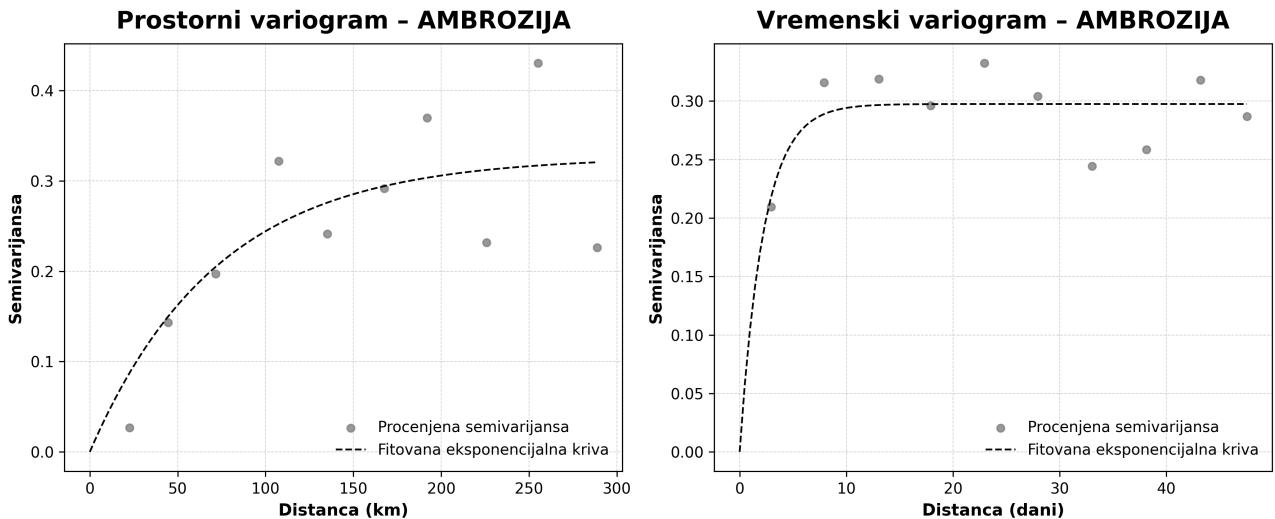
Vremenski variogram procenjen je korišćenjem parova posmatranja iz iste lokacije u različitim vremenskim periodima. Diskretizacija je sprovedena sa korakom od **5 dana**, čime su vremenska kašnjenja grupisana u klase radi stabilnije procene variograma i smanjenja varijabilnosti.

Prostorni variogram procenjen je korišćenjem parova posmatranja za isti dan na različitim lokacijama. Diskretizacija je u ovom slučaju sprovedena sa korakom od **40 km**. Budući da su mnoge tačke bile prostorno veoma blizu, grupisane su u iste klase rastojanja kako bi se izbegla dominacija velikog broja parova sa malim rastojanjima i omogućilo stabilnije modelovanje.

Na osnovu dobijenih prostornog i vremenskog variograma, zajednički (prostorno-vremenski) variogram određen je prema sledećoj formuli:

$$\gamma(u, v) = \gamma(u, 0) + \gamma(0, v) - k\gamma(u, 0)\gamma(0, v),$$

gde $\gamma(u, 0)$ predstavlja prostorni variogram, $\gamma(0, v)$ vremenski variogram, dok je k parametar koji se određuje optimizacijom kao onaj koji najbolje zadovoljava uslove separabilnosti i omogućava najtačnije modelovanje prostorno-vremenskog variograma.



Slika 9: Empirijski i modelovani prostorni i vremenski variogrami koncentracije polena ambrozije.

4. Interpolacija nedostajućih podataka.

Nakon modelovanja prostornog i vremenskog variograma, sprovedena je prostorno-vremenska *kriging* interpolacija u cilju procene nedostajućih vrednosti koncentracija polena.

Da bi se vremenska komponenta mogla upoređivati i kombinovati sa prostornim rastojanjima u prostorno-vremenskom modelovanju, izvršeno je skaliranje vremenskih razlika faktorom a_s/a_t , gde je a_s prostorni *range*, a a_t vremenski *range* parametar variograma. Vrednosti ovog faktora su **ograničene na interval 10–50**, kako bi se izbegle ekstremne vrednosti i obezbedilo da u modelu **jedan dan odgovara prostornom rastojanju između približno 10 i 50 km**. Ovakvo ograničenje je uvedeno kako bi se sprečile nerealistično velike ili male vrednosti skaliranja, koje bi narušile balans između prostorne i vremenske komponente u modelu. Konkretno, prevelike vrednosti bi dovele do toga da se promene u vremenu posmatraju kao zanemarljive u poređenju sa prostorom, dok bi premale vrednosti učinile da vremenska komponenta potpuno nadjača prostornu strukturu podataka.

Za procenu reziduala koncentracije u nekoj tački, korišćene su **najbliže poznate vrednosti** po prostorno-vremenskoj distanci, pri čemu je maksimalno rastojanje za uključivanje u predikciju bilo ograničeno na **200 km**. Ukoliko za neku tačku nije bilo dovoljno poznatih posmatranja koja ispunjavaju ovaj uslov, prepostavljano je da je rezidual jednak nuli, te je **konačna procena koncentracije** za tu tačku dobijena kao zbir prethodno procenjenog trenda i sezonske komponente.

Na kraju ovog dela, primenjeno je *winsorizing* (ograničavanje ekstremnih vrednosti) reziduala [36], kako bi se smanjio uticaj izuzetno visokih vrednosti koje bi mogle narušiti stabilnost modela i dovesti do nerealnih predikcija koncentracije polena.

Konkretno, za svaku lokaciju izračunati su prvi ($Q1$) i treći kvartil ($Q3$) reziduala, kao i interkvartilni raspon ($IQR = Q3 - Q1$). Na osnovu toga je određena gornja granica, definisana kao $Q3 + IQR$. Sve vrednosti reziduala koje su prelazile ovu granicu zamenjene su upravo tom graničnom vrednošću.

5. Rekonstrukcija konačnih vrednosti.

Poslednji korak u procesu imputacije je dobijanje konačnih procena koncentracija polena. To je postignuto jednostavnim sabiranjem prethodno procenjenih komponenti – trenda i sezonalnosti, zajedno sa imputiranim rezidualima dobijenim *kriging* interpolacijom [23].

Nakon toga, na dobijene vrednosti primenjena je **inverzna transformacija**, odnosno obrnuti postupak transformacija opisanih u prvom koraku (Box-Cox ili $\log(1 + \frac{x}{30})$ transformacija), kako bi se rezultati vratili u originalnu skalu koncentracija polena (zrna/m³).

6. Evaluacija modela.

Nakon rekonstrukcije konačnih vrednosti, sprovedena je evaluacija performansi *kriging* modela kako bi se procenila njegova tačnost i generalizacija.

U radu je primenjena **5-fold kros-validacija** metoda [35]. Za svaki alergen, kompletan skup podataka podeljen je na pet podskupova (*fold*-ova). U svakoj iteraciji, model je treniran na četiri podskupa, dok je preostali podskup korišćen za evaluaciju. Ovaj postupak ponovljen je pet puta, tako da je svaki podskup jednom korišćen kao test skup.

3.5 Modelovanje vremenske serije i podešavanje parametara SARIMAX modela

Za modelovanje i predikciju koncentracija polena korišćen je SARIMAX model, koji omogućava istovremeno opisivanje sezonskih obrazaca u podacima i uključivanje dodatnih meteoroloških faktora kao prediktora, poput temperature i vlažnosti vazduha [14, 18, 35].

Modelovanje je sprovedeno kroz sledeće korake:

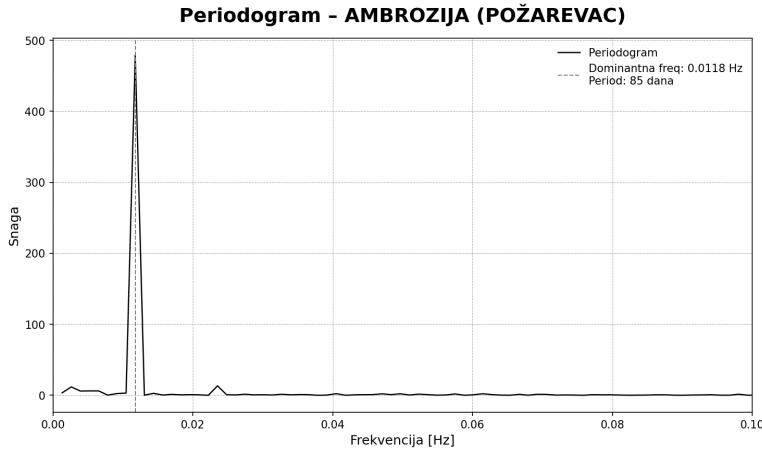
1. Priprema podataka.

Kao ulaz u SARIMAX model korišćeni su prethodno imputirani podaci, na koje je primenjena jedna od transformacija opisanih u ranijim koracima (Box-Cox ili $\log(1 + \frac{x}{30})$ transformacija). Ove transformacije su korišćene u cilju smanjenja varijanse i postizanja stacionarnosti podataka, što je preduslov za stabilno i precizno modelovanje vremenskih serija.

Serija je vizuelno analizirana radi procene potrebe za diferencijacijom. S obzirom na to da su koncentracije polena posmatrane u periodu od deset godina, tokom kojeg se ne očekuje postojanje značajnog linearne ili sezonskog trenda, odlučeno je da se parametri diferenciranja fiksiraju na $d = 0$ i $D = 0$. Pored vizuelne analize, stacionarnost serije je dodatno testirana uz pomoć statističkih testova ADF i KPSS, koji su opisani u odeljku 2.2.3.

2. Određivanje perioda sezonalnosti.

Budući da su iz analize izostavljeni periodi van sezona cvetanja, standardna godišnja sezonalnost (npr. 365 dana) nije mogla biti direktno korišćena u modelu. Jedno od potencijalnih rešenja bilo bi da se za datume van sezone unesu vrednosti nula, čime bi model formalno mogao da koristi godišnji period. Međutim, time bi SARIMAX model bio pristrasan ka predviđanju nižih koncentracija, smanjila bi se preciznost procene sezonskih amplituda, a kvalitet predikcija u sezonskim periodima bio bi značajno degradiran.



Slika 10: Periodogram koncentracija polena ambrozije u Požarevcu, korišćen za određivanje dominantnog sezonskog perioda.

Zbog toga je za određivanje sezonskog perioda korišćen **maksimalni pik periodograma** vremenske serije [12, 13]. Ovim pristupom, dominantna frekvencija u podacima određena je na osnovu realnih obrazaca, a ne unapred nametnutih sezonskih prepostavki, što omogućava modelu da preciznije uhvati specifične sezonske fluktuacije koncentracija polena.

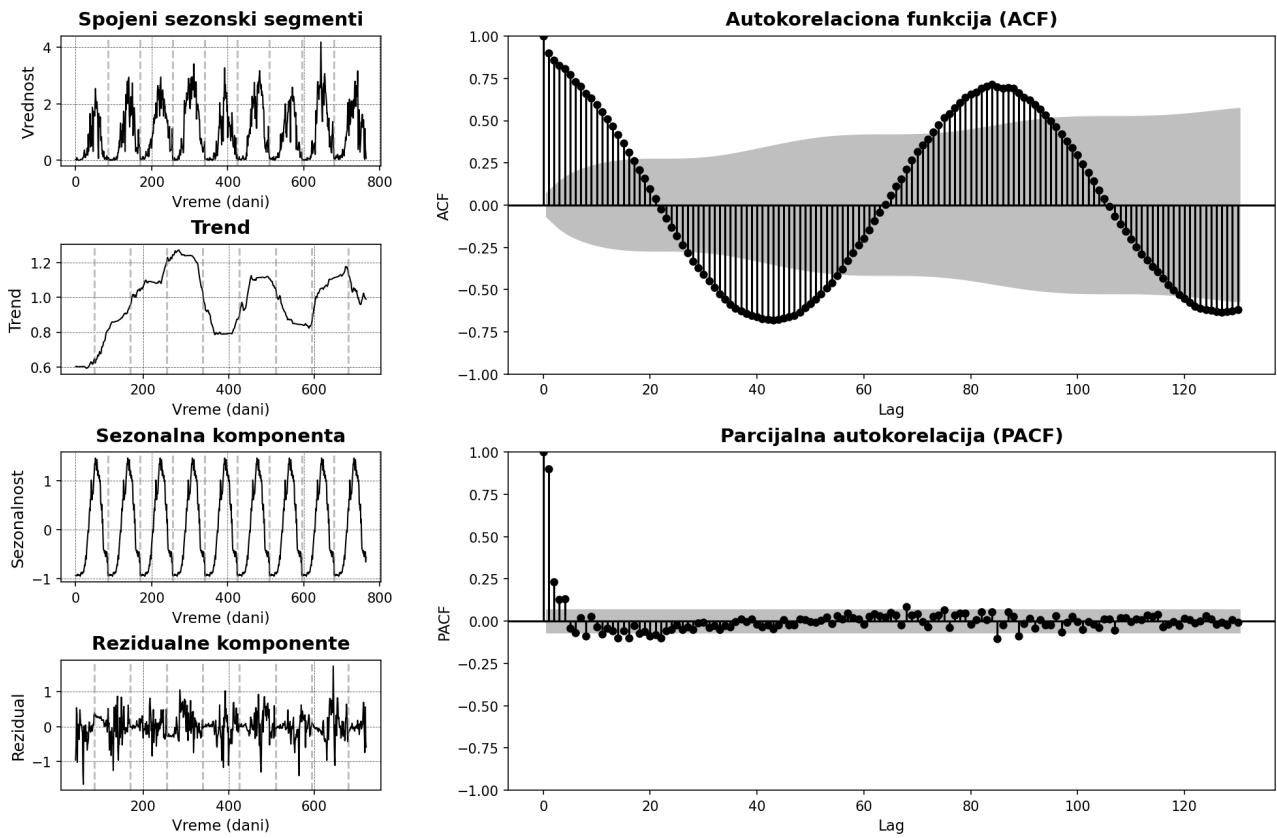
3. Definisanje strukture modela.

SARIMAX model karakterišu nesezonski parametri (p, d, q), sezonski parametri (P, D, Q, s), kao i egzogene promenljive [14, 18]. U ovom radu, vrednosti d i D su fiksirane na 0, dok su ostali parametri određivani kombinacijom sledećih metoda:

- Vizuelne analize ACF i PACF grafova za inicijalnu procenu potencijalnih vrednosti p, q, P i Q .
- *Grid search* procedure, uz minimizaciju AICc kriterijuma, radi pronalaženja optimalne kombinacije parametara i izbegavanja preobučavanja modela [20].

Kao egzogene promenljive u model su uključivani **Furijeovi redovi reda 3**, kako bi se modelovala sezonalnost bez rizika od preobučavanja, kao i potencijalno meteorološki podaci (temperatura, vlažnost vazduha) za poboljšanje prediktivne sposobnosti modela.

Spojeni sezonski signali - AMBROZIJA (POŽAREVAC)



Slika 11: STL dekompozicija vremenske serije koncentracije polena ambrozije: prikaz originalnih podataka, izdvojenog trenda, sezonske komponente i reziduala, uz ACF i PACF signala korišćenih za inicijalnu procenu parametara SARIMAX modela.

4. Trening modela i *rolling forecast* evaluacija.

Model je treniran na podacima iz perioda do kraja 2023. godine. Na osnovu ovih podataka određeni su optimalni modela (p , q , P i Q), uz primenu kriterijuma minimizacije AICc.

Evaluacija modela sprovedena je za 2024. godinu korišćenjem *rolling forecast* pristupa [37]. Ovaj pristup podrazumeva da se, nakon svakog novog posmatranja u tekućoj godini, model ponovo prilagođava i koristi za predikciju koncentracija polena za narednih nekoliko dana unapred. Na taj način simuliran je realni operativni scenario, gde se modeli svakodnevno ažuriraju najnovijim merenjima i koriste za kratkoročne prognoze u cilju pravovremenog informisanja javnosti i zdravstvenih sistema.

5. Rekonstrukcija predikcija.

Konačne predikcije dobijene su primenom inverzne transformacije (Box-Cox ili $\log(1 + \frac{x}{30})$), u zavisnosti od prethodno korišćene transformacije) na prediktovane vrednosti modela, čime su rezultati vraćeni u originalnu skalu koncentracija polena (zrna/m^3).

6. Metrike evaluacije.

Evaluacija tačnosti SARIMAX modela sprovedena je korišćenjem metrika definisanih u posebnom odeljku 3.8.

3.6 Modelovanje vremenske serije korišćenjem Prophet modela

U cilju predikcije koncentracija polena, pored SARIMAX modela, primjenjen je i **Prophet** model [21], razvijen od strane *Facebook Research* tima.

Proces modelovanja Prophet-om sastojao se od sledećih koraka:

1. Transformacija podataka.

Za razliku od nekih drugih modela vremenskih serija, Prophet može efikasno raditi i bez pretvodnih transformacija podataka. U ovom radu, model je treniran na podacima unutar sezonskih perioda, koristeći prethodno imputirane vrednosti koncentracija polena. Takođe, kao i u prethodnim analizama, primenjivane su i transformacije (Box-Cox ili $\log(1 + \frac{x}{30})$), kako bi se ispitalo da li transformacije dodatno poboljšavaju performanse modela.

2. Definisanje strukture modela.

Struktura Prophet modela podešavana je optimizacijom sledećih hiperparametara:

- **changepoint_prior_scale**: [0.1, 0.2, 0.5] – kontroliše fleksibilnost modela u detektovanju promena trenda. Niže vrednosti daju glatkiji trend, dok veće omogućavaju praćenje naglih promena u seriji, što je važno za alergene sa izraženim varijacijama.
- **seasonality_prior_scale**: [1.0, 2.0, 5.0] – određuje koliko sezonalnost utiče na predikciju. Veće vrednosti omogućavaju veću slobodu sezonskim komponentama, što poboljšava prilagođavanje modela kod alergena sa izraženim sezonskim obrascima.
- **seasonality_mode**: ['additive', 'multiplicative'] – definiše da li je sezonalnost aditivna (konstantna amplituda) ili multiplikativna (amplituda proporcionalna nivou serije).
- **changepoint_range**: [0.6, 0.8, 0.95] – definiše proporciju podataka unutar koje model detektuje tačke promene trenda. Veće vrednosti omogućavaju detekciju promena i u kasnijim delovima serije, što je korisno za alergene sa dugom sezonom cvetanja.

U model je dodat i **godišnji sezonski efekat**, aproksimiran korišćenjem Furijeovih redova [35], što omogućava preciznije hvatanje sezonskih obrazaca koncentracije polena tokom godine. Takođe, kao egzogene promenljive uključeni su i meteorološki podaci (npr. temperatura i vlažnost vazduha), kako bi se unapredila prediktivna sposobnost modela.

Optimalni hiperparametri birani su korišćenjem *grid search* procedure na podacima do kraja 2023. godine, dok je evaluacija modela sprovedena na podacima iz 2024. godine.

3. Treniranje modela i *rolling forecast* evaluacija.

Model je treniran i evaluiran primenom ***rolling forecast*** pristupa, kojim se simulira realni operativni scenario predikcije [37]. Nakon svakog novog posmatranja u tekućoj godini (2024), model se ponovo prilagođava i koristi za predikciju koncentracija polena za narednih nekoliko dana unapred.

Trening je sproveden na dva načina:

- Prvi pristup uključivao je korišćenje svih dostupnih poznatih vrednosti u trening skupu, kako bi model naučio dugoročne obrasce i sezonske fluktuacije koncentracija polena.
- Drugi pristup ograničavao je trening skup samo na posmatranja unutar **±30 dana** oko datuma predikcije svake godine. Ovakav pristup omogućava modelu da se fokusira na lokalne sezonske obrasce specifične za posmatrani period godine, s obzirom na to da koncentracije polena u različitim mesecima imaju različite obrasce i amplitudu.

4. Metrike evaluacije.

Metrike korišćene za evaluaciju performansi Prophet modela detaljno su opisane u odeljku 3.8, gde su predstavljeni RMSLE, MAE, RMSE, kao i dodatne evaluacije na osnovu klasifikacije koncentracija i predikcije datuma početka sezone.

3.7 Modelovanje koncentracija polena korišćenjem Random Forest modela

Za predikciju koncentracija polena primjenjen je i Random Forest (RF) model [22]. Sam proces modelovanja se sastoji iz sledećih koraka:

1. Transformacija podataka.

Za razliku od linearnih modela vremenskih serija, Random Forest može efikasno raditi i bez prethodnih transformacija podataka. U ovom radu, model je treniran na podacima unutar sezonskih perioda, koristeći prethodno imputirane vrednosti koncentracija polena. Takođe, kao i u prethodnim analizama, primenjivane su i transformacije (Box-Cox ili $\log(1 + \frac{x}{30})$), kako bi se ispitalo da li transformacije dodatno poboljšavaju performanse modela [17].

2. Uključivanje dodatnih promenljivih.

Budući da Random Forest ne prepostavlja linearost odnosa između prediktora i ciljne promenljive, u model su uključene i druge meteorološke promenljive, kao što su brzina vetra, pravac vetra, padavine, temperatura i vlažnost vazduha, kako bi se obuhvatile sve potencijalno relevantne informacije za predikciju koncentracija polena.

3. Podešavanje hiperparametara modela.

Optimalni parametri modela određivani su korišćenjem *grid search* procedure, pri čemu su testirane sledeće vrednosti:

- **n_estimators:** [250] – broj stabala u šumi; veći broj smanjuje varijansu modela, ali povećava vreme treniranja.
- **max_depth:** [5, 10, *None*] – maksimalna dubina svakog stabla; manja dubina smanjuje mogućnost preobučavanja.
- **max_features:** [' \log_2 ', ' $\sqrt{}$ '] – broj prediktora razmatranih pri svakom *split*-u; koristi se \log_2 ili kvadratni koren broja prediktora radi smanjenja korelacije među stablima.
- **min_samples_split:** [2, 5] – minimalni broj uzoraka potreban za podelu čvora; veće vrednosti smanjuju kompleksnost stabla.
- **max_lag:** [3, 5] – maksimalni broj vremenskih kašnjenja uključenih kao prediktora u model.

4. Kreiranje karakteristika (*Feature engineering*).

Kao ulazni prediktori u model su uključeni:

- vrednosti koncentracije polena sa zaostatkom do *max_lag* dana,
- prosečna koncentracija u prethodnih 7 dana,
- prosečna koncentracija u istom periodu prethodne godine (± 3 dana),
- koncentracija na isti dan prethodne godine,
- broj dana od početka sezone te godine i godina,
- Furijeovim redovi reda 3, sa periodom određivanim na isti način kao kod SARIMAX modela, radi aproksimacije sezonskih obrazaca [12, 13].

5. Treniranje i evaluacija modela.

Optimalni parametri modela birani su na podacima do kraja 2023. godine, dok je evaluacija sprovedena na podacima iz 2024. godine, korišćenjem *rolling forecast* pristupa [37]. Ovaj pristup podrazumeva da se nakon svakog novog posmatranja u tekućoj godini model ažurira i koristi za predikciju koncentracija polena za narednih nekoliko dana unapred, čime se simulira realni operativni scenario.

Budući da pri *rolling forecast* predikciji nisu uvek dostupne sve vrednosti za prediktore zasnovane na vremenskom kašnjenju (npr. pri predikciji za dva dana unapred, vrednost prethodnog dana nije poznata u trenutku predikcije), ovaj problem je rešen korišćenjem prethodno prediktovanih vrednosti modela kao ulaza za naredne korake predikcije.

6. Metrike evaluacije.

Metrike korišćene za evaluaciju performansi Random Forest modela detaljno su opisane u odeljku 3.8, gde su predstavljeni RMSLE, MAE, RMSE, kao i dodatne evaluacije na osnovu klasifikacije koncentracija i predikcije datuma početka sezone.

3.8 Metrike evaluacije modela

Za proveru ispravnosti modela, pre same evaluacije performansi, ispitivano je da li reziduali predstavljaju beli šum primenom Ljung–Box testa [38]. Rezultati su pokazali da su reziduali nekorelisani kod svih primenjenih modela, što potvrđuje njihovu konzistentnost, iako ova analiza nije dalje detaljno razmatrana u okviru ovog rada.

Za evaluaciju performansi modela vremenskih serija korišćene su standardne metrike koje omogućavaju procenu odstupanja predikcija od stvarnih vrednosti, kao i dodatne metrike vezane za sezonalnost i klasifikaciju koncentracija polena. U ovom radu fokus je stavljen na sledeće metrike:

- **Root Mean Squared Logarithmic Error (RMSLE)**

RMSLE naglašava proporcionalne greške i posebno je pogodan za podatke sa velikim varijacijama i ekstremnim vrednostima, kakvi su podaci o koncentraciji polena. Definiše se kao:

$$RMSLE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(1 + \frac{\hat{y}_i}{30}) - \log(1 + \frac{y_i}{30}))^2},$$

gde su \hat{y}_i prediktovane vrednosti, a y_i posmatrane vrednosti koncentracije polena. Ova metrika omogućava da se greške pri malim vrednostima tretiraju proporcionalno značajno, dok se uticaj velikih ekstremnih odstupanja ublažava [18, 35].

- **Mean Absolute Error (MAE)**

MAE predstavlja prosečno apsolutno odstupanje predikcija od stvarnih vrednosti i često se koristi zbog svoje intuitivnosti i robusnosti prema ekstremnim vrednostima.

- **Root Mean Squared Error (RMSE)**

RMSE kvadrira odstupanja, pa veću težinu daje većim greškama. Koristi se kao dopuna MAE kako bi se ocenilo koliko su predikcije podložne velikim odstupanjima.

- **Klasifikacija koncentracija**

Pored standardnih regresionih metrika, izvršena je i klasifikacija koncentracija polena. Definišane su tri kategorije:

- **Niska** — vrednosti manje od donje granične vrednosti.

- **Srednja** — vrednosti između donje i gornje granične vrednosti.
- **Visoka** — vrednosti veće od gornje granične vrednosti.

Ova klasifikacija omogućava praktičnu evaluaciju modela u smislu korisnosti za zdravstveni sistem i javnost, jer kategorizovane vrednosti imaju direktni uticaj na zdravstvene preporuke.

- **Predikcija datuma početka sezone visoke koncentracije polena**

Ova metrika procenjuje tačnost modela u predviđanju datuma početka visoke koncentracije polena.

- **Stvarna vrednost** početka sezone visoke koncentracije definisana je kao prvi dan kada vremenska serija pokazuje koncentraciju veću od donje granice u **tri uzastopna dana**.
- **Predikcija** je vršena tako što je svakog dana model radio prognozu za 7., 8. i 9. dan unapred. Prvi trenutak kada su sve tri prognozirane vrednosti bile iznad donje granice smatran je početkom sezone, pri čemu je kao zvanični datum uzet sedmi dan iz tog intervala.

Ova metrika omogućava kvantifikaciju sposobnosti modela da pravovremeno identificuje početak sezone, što je od posebne važnosti za preventivne mere.

4 Rezultati

4.1 Imputacija

Za sve alergene sprovedena je detaljna evaluacija različitih metoda za imputaciju nedostajućih vrednosti koncentracija polena, međutim, u ovom radu biće prikazani rezultati samo za tri alergena: ambroziju, jovu i trave. Kao metrika greške korišćen je **RMSLE**.

4.1.1 Metodologija evaluacije

Za evaluaciju imputacionih modela korišćena je **5-fold kros-validacija** procedura [37]. Svi dostupni podaci su podeljeni na pet podskupova (*fold-ova*). Trening modela vršen je na četiri folda, dok je performansa evaluirana na preostalom foldu. Ovaj proces je ponovljen pet puta, tako da je svaki podskup korišćen kao test set jednom, a rezultati su agregirani za svaki alergen posebno.

Poređenje metoda imputacije:

- **Prostorno-vremenski Kriging (ST Kriging)** – klasičan prostorno-vremenski *kriging* [23, 25] nakon transformacije podataka Box-Cox ili $\log(1 + \frac{x}{30})$.
- **Standardizovani ST Kriging** – podaci standardizovani po lokaciji i alergenu, zatim transformi-sani (Box-Cox ili $\log(1 + \frac{x}{30})$).
- **Standardizovani ST Kriging sa egzogenim promenljivama** – dodat je uticaj temperature i vlažnosti vazduha.
- **Naive Temporal Nearest** – imputacija nedostajuće vrednosti korišćenjem najbliže poznate vremenske vrednosti.
- **IDW (Inverse Distance Weighting)** – metoda imputacije koja koristi prostornu interpolaciju inverznim ponderisanjem po udaljenosti unutar istog dana [39], a u slučaju nedostatka suseda koristi vrednost sa najbližeg datuma iste lokacije.

4.1.2 Rezultati po alergenima

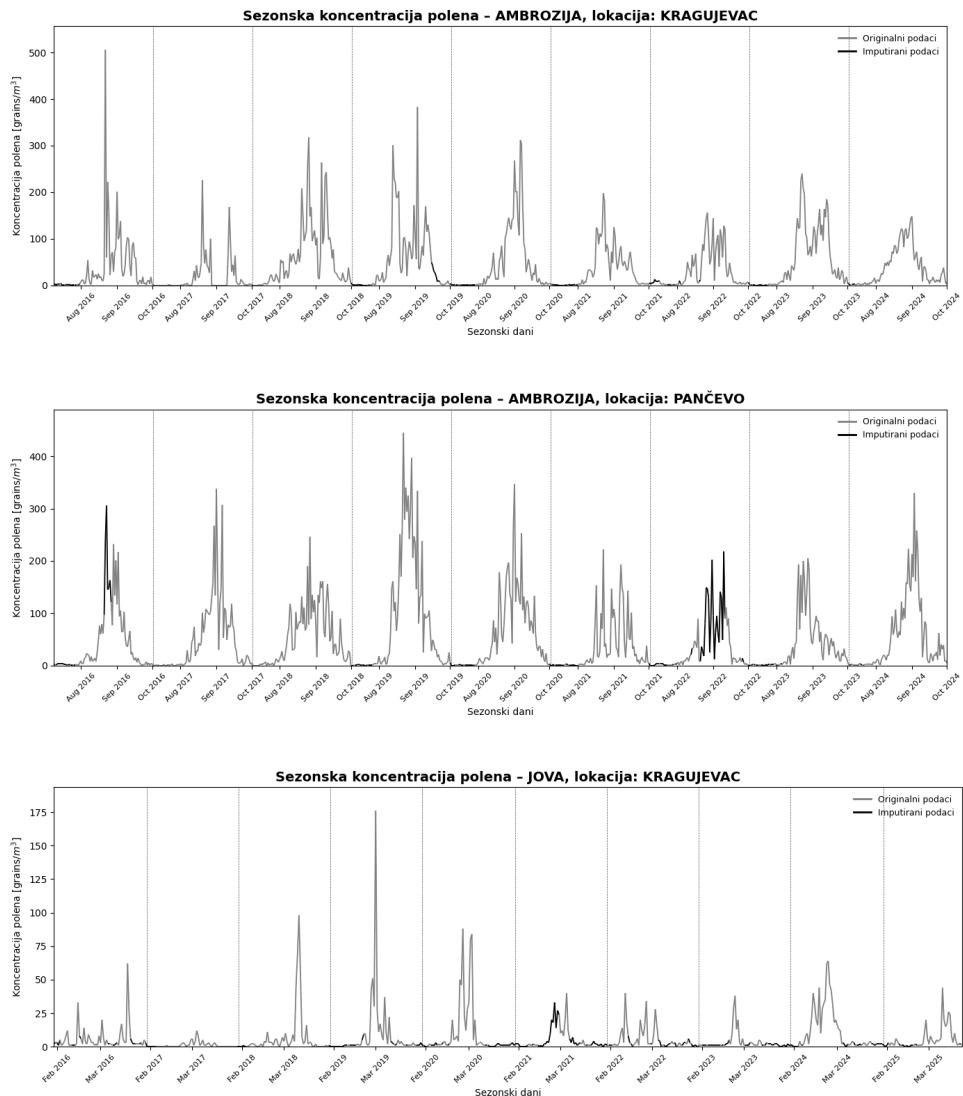
Ambrozija. Za ambroziju, najniži RMSLE postignut je standardizovanim *ST Kriging* modelom sa Box-Cox transformacijom, iznoseći **0.260** ($\sigma = 0.011$). Slične rezultate dali su i standardizovani modeli sa Box-Cox transformacijom i dodatim egzogenim promenljivama. Naivna metoda imala je RMSLE od 0.309 ($\sigma = 0.007$), dok je IDW metoda imala značajno višu grešku od 1.085 ($\sigma = 0.009$).

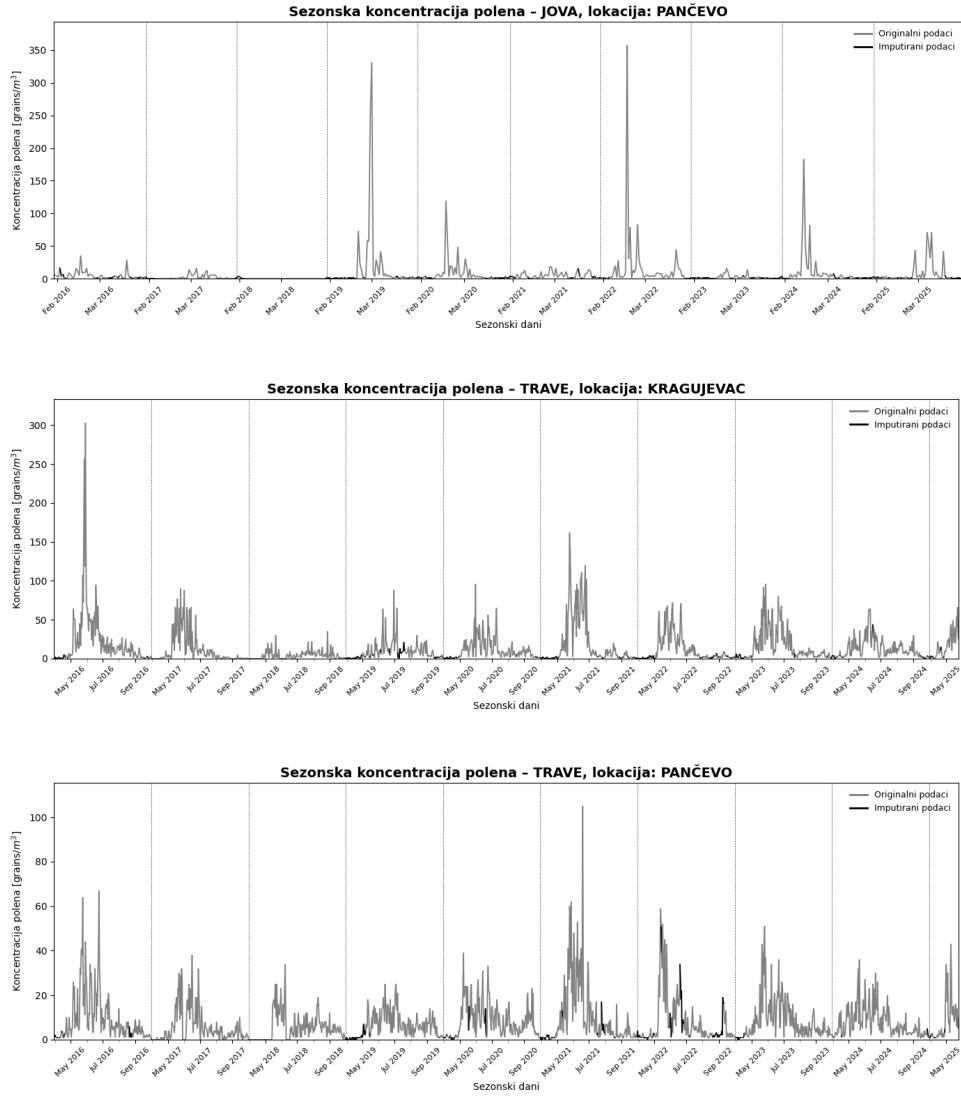
Jova. Za jovu, najbolji rezultat postignut je standardizovanim *ST Kriging* modelom sa Box-Cox transformacijom, sa RMSLE od **0.244** ($\sigma = 0.017$). Najlošiji rezultat imao je IDW metod sa RMSLE od 0.567 ($\sigma = 0.024$), dok je naivna metoda dala RMSLE od 0.349 ($\sigma = 0.020$).

Trave. Kod trava, najniži RMSLE postignut je standardizovanim *ST Kriging* modelom sa Box-Cox transformacijom, iznoseći **0.201** ($\sigma = 0.008$). Naivna metoda dala je RMSLE od 0.231 ($\sigma = 0.004$), dok je IDW metoda pokazala značajno lošiju tačnost sa RMSLE od 0.437 ($\sigma = 0.004$).

Tabela 1: Rezultati imputacije nedostajućih vrednosti koncentracije polena (RMSLE sa standardnom devijacijom) za različite metode i alergene.

Metoda	Ambrozija	Jova	Trave
<i>ST Kriging</i> (Box-Cox)	0.336 (0.125)	0.290 (0.081)	0.238 (0.063)
<i>ST Kriging</i> (log)	0.373 (0.155)	0.289 (0.017)	0.206 (0.002)
Standardizovani <i>ST Kriging</i> (Box-Cox)	0.260 (0.011)	0.244 (0.017)	0.201 (0.008)
Standardizovani <i>ST Kriging</i> (log)	0.307 (0.010)	0.300 (0.013)	0.223 (0.002)
Standardizovani <i>ST Kriging + exog</i> (Box-Cox)	0.262 (0.011)	0.244 (0.017)	0.201 (0.008)
Standardizovani <i>ST Kriging + exog</i> (log)	0.322 (0.013)	0.302 (0.013)	0.225 (0.003)
<i>Naive Temporal Nearest</i>	0.309 (0.007)	0.349 (0.020)	0.231 (0.004)
<i>IDW Temporal Fallback</i>	1.085 (0.009)	0.567 (0.024)	0.437 (0.004)





Slika 12: Vizuelni prikaz imputacije koncentracija polena za ambroziju, jovu i trave na odabranim lokacijama.

4.1.3 Diskusija

Na osnovu dobijenih rezultata, može se zaključiti da je **standardizovani prostorno-vremenski kriging model sa Box-Cox transformacijom** pokazao najniže RMSLE vrednosti za sve analizirane alergene, što ukazuje na njegovu superiornost u imputaciji nedostajućih podataka o koncentraciji polena. Ovaj pristup omogućava stabilizaciju varijanse i bolje prilagođavanje raspodeli podataka, dok standardizacija dodatno uklanja uticaj različitih opsega koncentracija između lokacija.

Analizom uticaja egzogenih promenljivih, konkretno temperature i vlažnosti vazduha, uočeno je da modeli sa i bez ovih kovarijata daju gotovo identične rezultate. S obzirom na to, za praktičnu implementaciju preporučuje se korišćenje modela bez egzogenih promenljivih zbog jednostavnije primene i manjih zahteva za eksternim podacima. Ovakav rezultat može se objasniti činjenicom da su prostorne i vremenske zavisnosti koncentracija polena, koje *kriging* model uspešno hvata, delimično već povezane sa meteorološkim faktorima, zbog čega eksplicitno uključivanje temperature i vlažnosti vazduha ne doprinosi poboljšanju tačnosti predikcija.

Sa druge strane, metoda **IDW Temporal Fallback** pokazala je značajno slabije performanse u

poređenju sa ostalim modelima. Razlog za to leži u njenom konceptualnom ograničenju, jer se oslanja isključivo na prostornu interpolaciju inverznim ponderisanjem po udaljenosti unutar istog dana, a u slučaju nedostatka prostorno bliskih merenja pribegava *fallback* strategiji uzimanja vrednosti sa najbližeg datuma iste lokacije. Ovakav pristup ne uspeva da adekvatno modeluje složene prostorno-vremenske obrasce distribucije polena, što dovodi do znatno većih grešaka.

Važno je istaći i specifičnost načina prikupljanja podataka – merenja su vršena kontinuirano, sa uzastopnim dnevnim uzorkovanjima koncentracije polena. Zbog toga postoji izražena vremenska autokorelacija između dva uzastopna dana, što objašnjava relativno dobre rezultate **naivne metode**, koja imputaciju zasniva na najbližem vremenskom susedu. Ipak, rezultati pokazuju da **prostorno-vremenski kriging** uspeva da nadmaši i ovu metodu, zahvaljujući sposobnosti da istovremeno uvaži i prostornu i vremensku zavisnost u predikciji koncentracija polena.

4.2 Predikcija

U ovom poglavlju predstavljeni su rezultati predikcije koncentracije polena korišćenjem tri različita modela: **SARIMAX** [14, 35], **Prophet** [21], i **Random Forest** [22], uz poređenje sa **naivnom metodom** kao baznom linijom. Rezultati poređenja prikazani su u tabelama 22 i 23.

4.2.1 Metodologija evaluacije

Za evaluaciju je korišćen *rolling forecast* pristup [37], kojim se simulira realni scenario operativnog predviđanja, gde su podaci dostupni samo do trenutka predikcije. Modeli su trenirani na podacima do kraja 2023. godine, dok je evaluacija sprovedena na podacima iz 2024. godine.

Kao transformacije podataka razmatrane su Box-Cox [17], $\log(1 + \frac{x}{30})$ i scenario bez transformacije.¹ Predikcije su pravljene sa i bez korišćenja meteoroloških promenljivih, kako bi se procenio njihov uticaj na performanse modela.

Evaluacija performansi modela sprovedena je korišćenjem metrika definisanih u odeljku 3.8 [18, 35].

4.2.2 SARIMAX

Optimalni parametri. Tabela 4.2.2 prikazuje optimalne SARIMAX parametre za Kragujevac, za ambroziju, za obe transformacije (Box-Cox i $\log(1 + \frac{x}{30})$) i za modele sa i bez meteoroloških podataka. Važno je napomenuti da **AICc kriterijum nije direktno uporediv između različitih transformacija**, tako da se obe transformacije posebno testiraju, pa se naknadno razmatra prikladniji model.

Tabela 2: Optimalni SARIMAX parametri za Požarevac (ambrozija) sa različitim transformacijama, meteorološkim parametrima i Furijeovim redom.

Transformacija	Meteo	<i>p</i>	<i>d</i>	<i>q</i>	<i>P</i>	<i>D</i>	<i>Q</i>	<i>s</i>	Fourier(<i>k</i>)	AICc	T	H
Logaritamska	Da	2	0	1	1	0	1	85	3	361.56	0.000	0.371
Logaritamska	Ne	2	0	2	1	0	0	85	3	388.70	–	–
Box-Cox	Da	2	0	1	1	0	1	85	3	1801.26	0.767	0.000
Box-Cox	Ne	2	0	2	1	0	1	85	3	1805.45	–	–

T = temperatura, H = vlažnost vazduha. P-vrednosti označavaju statističku značajnost meteoroloških kovarijata u modelu.

¹Kod SARIMAX modela transformacija je obavezna.

Performanse modela po transformaciji. Za detaljniju analizu, RMSLE vrednosti su prikazane posebno za logaritamsku i Box-Cox transformaciju, uz razdvajanje po gradovima, vrstama polena (ambrozija i jova) i po tome da li su meteorološki parametri uključeni u model ili ne. Tabela 3 prikazuje rezultate za logaritamsku transformaciju, dok tabela 4 sadrži RMSLE vrednosti za Box-Cox transformaciju. Na ovaj način može se jasno uočiti kako izbor transformacije i prisustvo meteoroloških kovarijata utiče na preciznost predikcije, kao i razlike u grešci između predikcije za 1 i 7 dana unapred.

Tabela 3: RMSLE vrednosti za logaritamsku transformaciju (1 i 7 dana unapred, sa i bez meteo)

Grad	Ambrozija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana
KRAGUJEVAC	0.18	0.29	0.19	0.31	0.15	0.26	0.16	0.28
POŽAREVAC	0.42	0.46	0.45	0.50	0.29	0.37	0.28	0.38
PANČEVO	0.27	0.35	0.28	0.40	0.29	0.37	0.33	0.39

Tabela 4: RMSLE vrednosti za Box-Cox transformaciju (1 i 7 dana unapred, sa i bez meteo)

Grad	Ambrozija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana
KRAGUJEVAC	0.19	0.30	0.20	0.33	0.16	0.25	0.17	0.28
POŽAREVAC	0.43	0.51	0.48	0.54	0.30	0.41	0.29	0.40
PANČEVO	0.26	0.35	0.31	0.41	0.31	0.41	0.31	0.41

Budući da transformacijom $\log(1 + \frac{x}{30})$ se favorizuje greška RMSLE, za potpuniju usporedbu datih su i sledeći rezultati za metrike RMSE i MAE.

Tabela 5: Najbolji modeli po RMSLE, sa prikazom RMSLE, RMSE i MAE

Transformacija	RMSLE	RMSE	MAE
Sa meteorološkim parametrima			
Logaritamska	0.422	72.61	41.57
Box-Cox	0.431	74.42	43.49
Bez meteoroloških parametara			
Logaritamska	0.447	81.46	43.65
Box-Cox	0.479	86.16	47.24

Početak sezone polena. Za dodatnu analizu performansi modela, izračunat je početak sezone polena za svaku kombinaciju grada i vrste alergena (ambrozija i jova). Početak sezone definisan je kao prvi dan u godini kada koncentracija polena prvi put premaši zadatu granicu tri dana zaredom (30 za ambroziju i 60 za jovu). Pored stvarnog početka sezone, prikazana je i predikcija koju model generiše

7 dana unapred, kako bi se procenila preciznost ranog upozorenja. Za analizu je korišćena logaritamska transformacija, a rezultati su prikazani za modele sa i bez meteoroloških parametara.

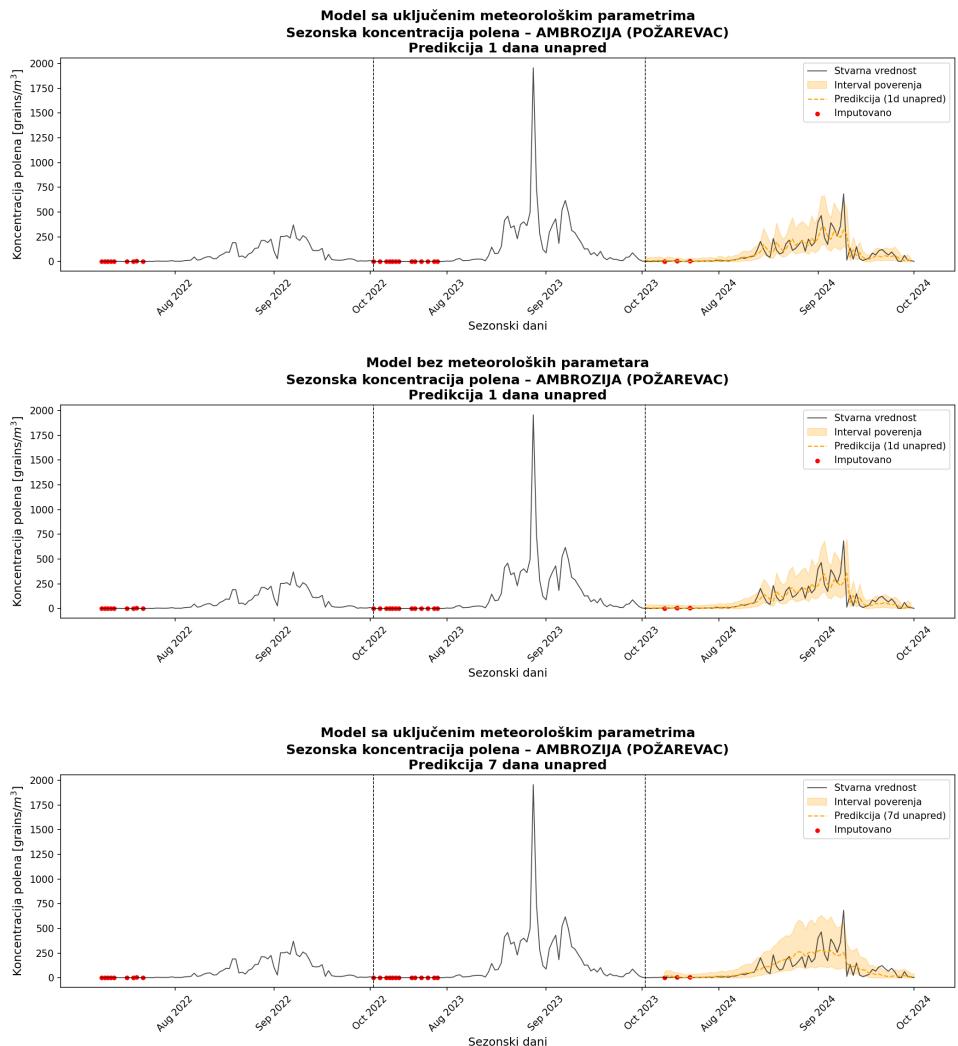
Tabela 6: Početak sezone polena: stvarni datum i predikcija 7 dana unapred (sa i bez meteo).

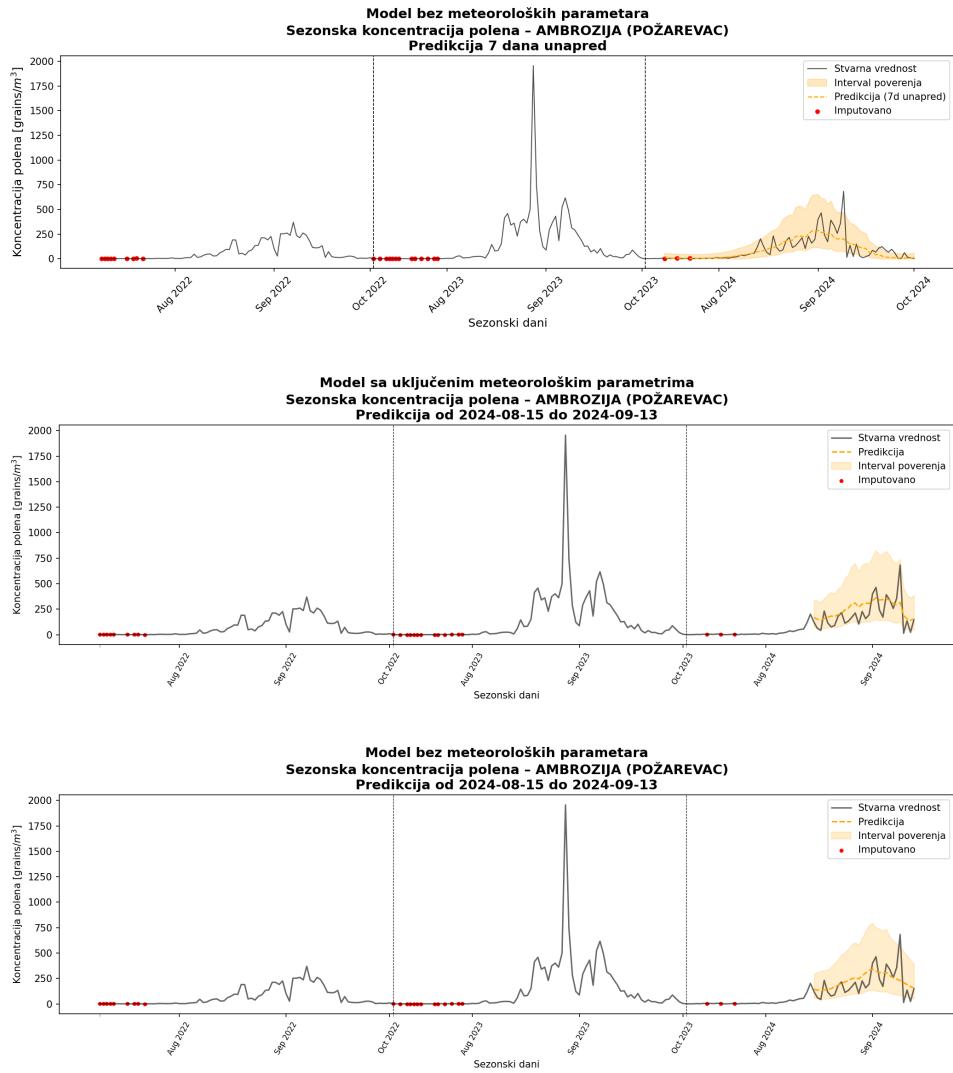
Grad	Ambrožija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	Stvarni datum	Predikcija 7d						
KRAGUJEVAC	22.08.2024	18.08.2024	22.08.2024	17.08.2024	/	/	/	/
POŽAREVAC	16.08.2024	19.08.2024	16.08.2024	19.08.2024	/	/	/	/
PANČEVO	14.08.2024	15.08.2024	14.08.2024	14.08.2024	17.02.2024	/	17.02.2024.	/

Primer predikcije. Na slici 13 prikazane su predikcije koncentracije ambrozije u Požarevcu pomoću SARIMAX modela i primenom logaritamske transformacije. Svaka slika predstavlja pojedinačnu predikciju, prikazanu naizmenično sa meteorološkim kovarijatima i bez njih.

Redovi odgovaraju različitim horizontima predikcije:

- Predikcija 1 dan unapred – prve dve slike (sa i bez meteo).
- Predikcija 7 dana unapred (*rolling forecast*) – naredne dve slike (sa i bez meteo).
- Prognoza za narednih 30 dana – poslednje dve slike (sa i bez meteo).





Slika 13: Predikcije koncentracije ambrozije u Požarevcu pomoću SARIMAX modela. Predikcije su prikazane naizmenično sa meteorološkim kovarijatima i bez njih. Prve dve slike: predikcija 1 dan unapred; sledeće dve: predikcija 7 dana unapred (*rolling forecast*); poslednje dve: prognoza za narednih 30 dana.

Klasifikacija nivoa koncentracije. Koncentracije polena su podeljene u tri nivoa: niska (<30), srednja (<100), visoka (>100). Klasifikacija se odnosi na predikciju koncentracije za posmatrani dan, određenu 7 dana unapred. U tabelama 7 prikazane su konfuzione matrice za klasifikaciju nivoa ambrozije u Požarevcu, posebno za modele sa i bez meteoroloških kovarijata.

Tabela 7: Konfuzione matrice za klasifikaciju nivoa ambrozije u Požarevcu, sa i bez meteoroloških kovarijata.

		SA meteo			BEZ meteo				
		Predikcija	Nisko	Srednje	Visoko	Predikcija	Nisko	Srednje	Visoko
Stvarno	Predikcija								
	Nisko	28	3	2		Nisko	27	1	5
Srednje	5	8	4		Srednje	5	9	3	
Visoko	1	3	24		Visoko	1	4	23	

Diskusija. Prilikom odabira modela za predikciju koncentracije polena, analiza je pokazala da logaritamska transformacija $\log(1 + \frac{x}{30})$ daje konzistentno bolje rezultate u poređenju sa Box-Cox transformacijom, ne samo kod RMSLE metrike već i kod RMSE i MAE, što je razlog da se smatra superiornijom za potrebe predikcije polena.

Iako je testirano uključivanje dodatnih meteoroloških parametara, poput padavina i brzine vетра, utvrđeno je da ne postoji značajna linearna zavisnost između ovih parametara i koncentracije polena. Njihovo uključivanje u SARIMAX model u nekim slučajevima dovodilo je do preobučavanja i pogoršanja performansi. Stoga se za SARIMAX pokazalo da su ključni parametri temperatura i vlažnost vazduha, dok ostali vremenski faktori mogu biti zanemarljivi za predikciju u ovom kontekstu.

Predikcije sa meteorološkim uslovima omogućavaju veću **varijabilnost** i realističniji prikaz promena koncentracije polena kroz sezonom. Nasuprot tome, modeli bez meteoroloških podataka generišu gotovo izgladjenu krivu, što je verovatno posledica efekta Furijeove analize i modelovanja sezonskih komponenti. Ovo naglašava značaj vremenskih kovarijata za hvatanje kratkoročnih oscilacija u koncentraciji polena.

Analiza početka sezone pokazuje da za ambroziju praktično nema značajne razlike između modela sa i bez meteoroloških podataka – oba pristupa predviđaju sličan datum kada koncentracija prvi put prelazi zadatu granicu tri dana zaredom. To sugerise da predikcija ambrozije zavisi pre svega od sezonskog trenda, dok meteorološki faktori imaju manji uticaj na određivanje početka sezone. Nasuprot tome, koncentracija jove retko prelazi definisani prag, što se jasno vidi i u rezultatima SARIMAX modela.

4.2.3 Prophet model

Optimalni parametri. Tabela 4.2.3 prikazuje pregled transformacija koje su testirane (logaritamska, Box-Cox i bez transformacije), kao i da li su meteorološki parametri uključeni. Rezultati su dobijeni na testirajućem skupu, a predstavljeni su najbolji modeli prema RMSLE metrici.

Performanse modela po transformaciji. U tabelama 9–11 prikazane su RMSLE vrednosti za sve gradove, posebno za logaritamsku, Box-Cox i bez transformacije, uz razdvajanje po tome da li su korišćeni meteorološki podaci i za koliko dana unapred se vrši predikcija (1 ili 7).

Najbolji modeli i metrike. Tabela 12 prikazuje najbolje rezultate po RMSLE metrikama, uz dodatne RMSE i MAE vrednosti, kako sa, tako i bez meteoroloških kovarijata.

Tabela 8: Prophet – transformacije, sezonalnost i parametri za Požarevac (ambrozija).

Transformacija	Meteo	Sezonalni mod	RMSE	MAE	RMSLE
None	Da	Multiplikativni	199.98	78.56	0.464
Box-Cox	Da	Multiplikativni	191.39	67.43	0.400
Logaritamska	Da	Multiplikativni	179.40	70.33	0.400
None	Ne	Multiplikativni	207.40	77.52	0.452
Box-Cox	Ne	Multiplikativni	211.84	75.54	0.447
Logaritamska	Ne	Multiplikativni	201.16	73.80	0.446

Tabela 9: Prophet – RMSLE vrednosti za logaritamsku transformaciju (1 i 7 dana unapred, sa i bez meteo).

Grad	Ambrozija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana
KRAGUJEVAC	0.28	0.37	0.29	0.38	0.30	0.37	0.31	0.39
POŽAREVAC	0.46	0.59	0.50	0.59	0.35	0.45	0.36	0.44
PANČEVO	0.26	0.29	0.31	0.33	0.39	0.67	0.37	0.66

Tabela 10: Prophet – RMSLE vrednosti za Box-Cox transformaciju (1 i 7 dana unapred, sa i bez meteo).

Grad	Ambrozija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana
KRAGUJEVAC	0.32	0.42	0.32	0.42	0.29	0.32	0.28	0.35
POŽAREVAC	0.41	0.46	0.46	0.49	0.38	0.43	0.36	0.40
PANČEVO	0.26	0.28	0.30	0.33	0.37	0.41	0.37	0.40

Tabela 11: Prophet – RMSLE vrednosti bez transformacije (1 i 7 dana unapred, sa i bez meteo).

Grad	Ambrozija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana
KRAGUJEVAC	0.31	0.37	0.32	0.36	0.32	0.42	0.32	0.43
POŽAREVAC	0.63	0.69	0.57	0.63	0.37	0.51	0.31	0.46
PANČEVO	0.28	0.32	0.31	0.34	0.45	0.69	0.44	0.69

Početak sezone polena. Tabela 13 prikazuje stvarni i predviđeni početak sezone polena za ambroziju i jovu, analogno kao kod SARIMAX modela.

Tabela 12: Prophet – najbolji modeli po RMSLE, sa prikazom RMSLE, RMSE i MAE.

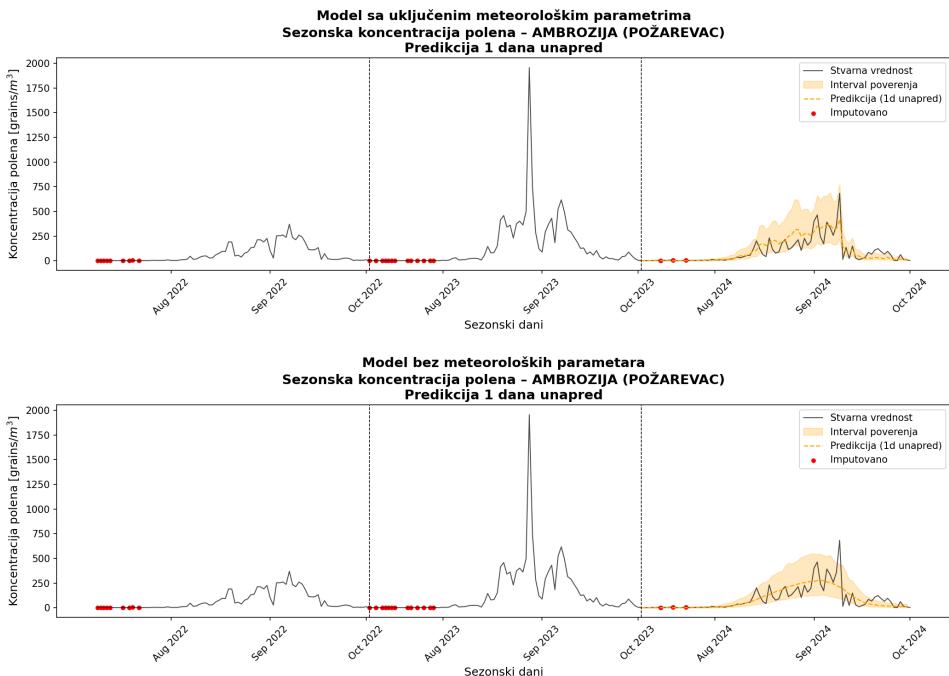
Transformacija	RMSLE	RMSE	MAE
Sa meteorološkim parametrima			
Logaritamska	0.462	85.06	51.06
Box-Cox	0.413	66.98	42.60
<i>None</i>	0.630	105.22	71.85
Bez meteoroloških parametara			
Logaritamska	0.501	91.60	54.40
Box-Cox	0.459	80.15	44.92
<i>None</i>	0.573	100.78	64.67

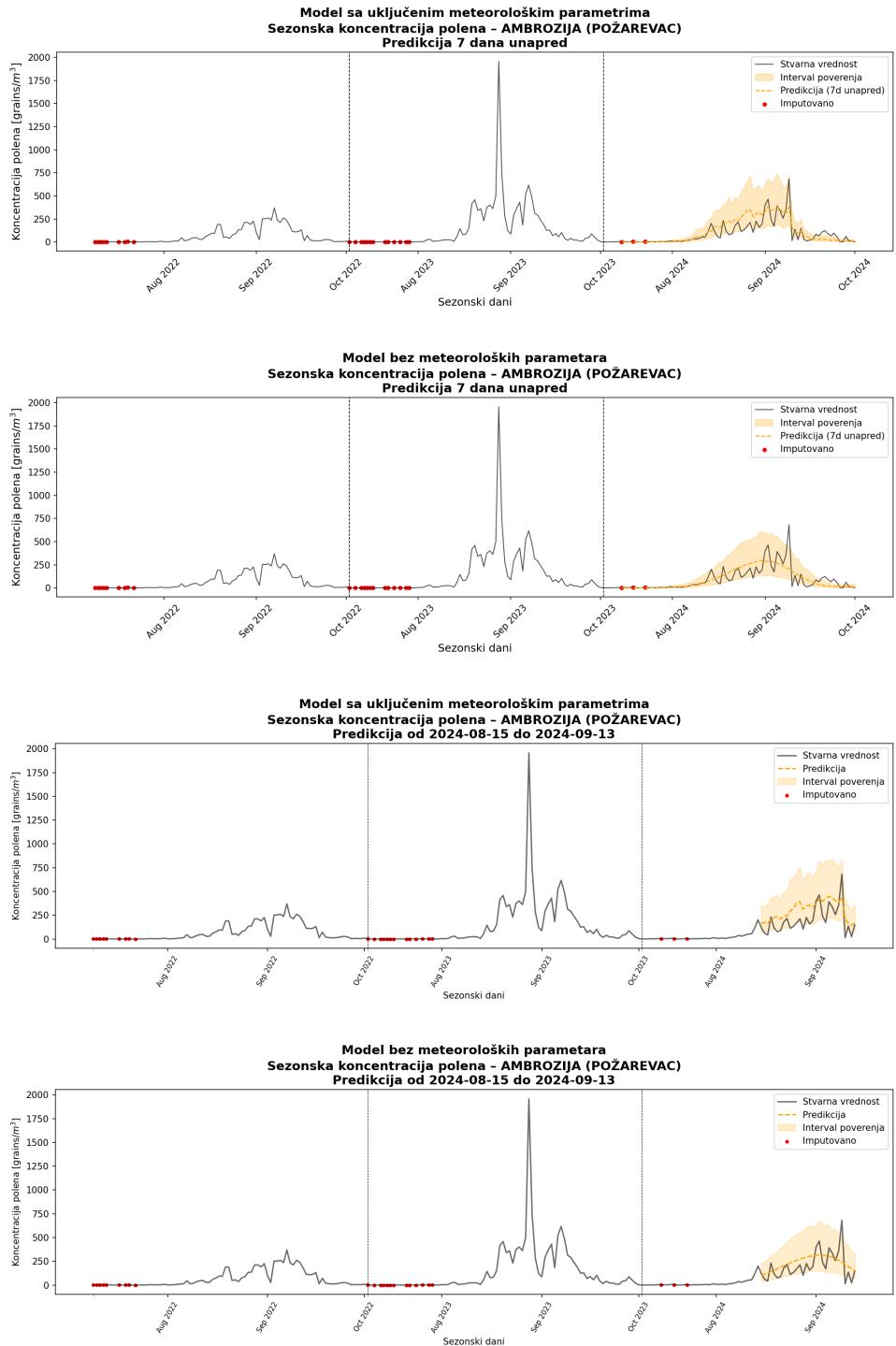
Tabela 13: Prophet – početak sezone polena: stvarni datum i predikcija 7 dana unapred.

Grad	Ambrožija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	Stvarni datum	Predikcija 7d						
KRAGUJEVAC	22.08.2024	15.08.2024	22.08.2024	15.08.2024	/	/	/	/
POŽAREVAC	16.08.2024	19.08.2024	16.08.2024	20.03.2024	/	/	/	/
PANČEVO	14.08.2024	13.08.2024	14.08.2024	12.03.2024	17.02.2024	/	17.02.2024.	/

Primer predikcije. Na slici 14 prikazane su predikcije koncentracije ambrozije u Požarevcu pomoću Prophet modela i primenom Box-Cox transformacije. Predikcije su prikazane naizmenično sa meteorološkim kovarijatima i bez njih. Redovi odgovaraju različitim horizontima predikcije:

1. Predikcija 1 dan unapred – prve dve slike (sa i bez meteoroloških kovarijata).
2. Predikcija 7 dana unapred (*rolling forecast*) – sledeće dve slike (sa i bez meteoroloških kovarijata).
3. Prognoza za narednih 30 dana – poslednje dve slike (sa i bez meteoroloških kovarijata).





Slika 14: Predikcije koncentracije ambrozije u Požarevcu pomoću Prophet modela (Box-Cox transformacija). Predikcije su prikazane naizmenično sa meteorološkim kovarijatima i bez njih. Redovi: predikcija 1 dan unapred, predikcija 7 dana unapred (*rolling forecast*), prognoza za narednih 30 dana.

Klasifikacija nivoa koncentracije. Klasifikacija se odnosi na predikciju koncentracije za posmatrani dan, određenu 7 dana unapred. U tabeli 14 prikazane su konfuzione matrice za modele sa i bez meteoroloških kovarijata.

Tabela 14: Prophet – konfuzione matrice za klasifikaciju nivoa ambrozije u Požarevcu.

		SA meteo			BEZ meteo				
		Predikcija	Nisko	Srednje	Visoko	Predikcija	Nisko	Srednje	Visoko
Stvarno	Predikcija	Nisko	28	4	1	Nisko	28	2	3
		Srednje	7	6	4	Srednje	5	8	4
Visoko	Visoko	0	2	26	Visoko	1	4	23	

Diskusija. Rezultati Prophet modela pokazuju da ovaj pristup uspešno prepoznaje dominantne sezonske obrasce, ali uz određena ograničenja kada su u pitanju kratkoročne oscilacije. Primena **muliplikativnog moda sezonalnosti** bila je očekivana, jer se sa većim amplitudama javlja i izraženija heteroskedastičnost — što znači da veće koncentracije polena prate i snažnije oscilacije. Na taj način model može bolje da prati sezonske vrhove.

Što se tiče transformacija, Box-Cox se pokazala najstabilnijom i najpreciznijom. Njena prednost ogleda se u kontrolisanju ekstremnih vrednosti, što doprinosi nižim vrednostima greške (RMSLE, ali i RMSE i MAE). Logaritamska transformacija $\log(1 + \frac{x}{30})$ transformacija je u pojedinim slučajevima dovodila do prevelikog „izravnavanja“ serije, dok se bez transformacije često gubi stabilnost predikcija.

U pogledu meteoroloških kovarijata, rezultati pokazuju da **njihovo prisustvo ne mora nužno da poboljša performanse modela**. Iako su temperatura i vlažnost najkorisniji prediktori, u nekim konfiguracijama dodavanje dodatnih meteoroloških parametara (padavine, vetar) nije dovelo do manjeg RMSLE, što sugerise da Prophet ne koristi uvek ove informacije na optimalan način.

Analiza predikcije početka sezone polena pokazuje da modeli u celini pouzdano prepoznaju početak cvetanja ambrozije, čak i u odsustvu meteoroloških kovarijata. Razlike između modela sa i bez meteoroloških podataka su minimalne, što sugerise da za rano upozorenje o ambroziji nije presudno uključivanje vremenskih faktora. Za jovanu, koncentracije retko prelaze definisani prag, pa predikcije početka sezone često nisu generisane ili su neprecizne.

Važna osobina Prophet modela jeste da **generalno prati sezonski trend**, ali pri naglim promenama koncentracije ne reaguje dovoljno brzo. To dovodi do toga da je manje pouzdan za dnevne prognoze sa naglašenim oscilacijama, ali se pokazao **adekvatan za prognoze nekoliko dana unapred**, gde je sezonska komponenta dominantna.

Sveukupno, Prophet predstavlja robustan model za hvatanje sezonske strukture i procenu nivoa polena nekoliko dana unapred, dok mu ograničenja ostaju u kratkoročnim dnevnim oscilacijama.

4.2.4 Random Forest

Optimalni parametri. Tabela 4.2.4 prikazuje pregled transformacija koje su testirane (logaritamska, Box-Cox i bez transformacije), kao i uključivanje meteoroloških parametara i Furijeovih redova. Rezultati su dobijeni na testirajućem skupu, a predstavljeni su najbolji modeli prema RMSLE metrici.

Performanse modela po transformaciji. U tabelama 16–18 prikazane su RMSLE vrednosti za sve gradove, posebno za logaritamsku, Box-Cox i bez transformacije, uz razdvajanje po tome da li su korišćeni meteorološki parametri i za koliko dana unapred se vrši predikcija (1 ili 7 dana).

Najbolji modeli i metrike Tabela 19 prikazuje najbolje rezultate po RMSLE metrici, uz dodatne RMSE i MAE vrednosti, sa i bez meteoroloških kovarijata.

Tabela 15: Random Forest – transformacije, sezonalnost i parametri za Požarevac (ambrozija).

Transformacija	Meteo	Fourier (k)	Broj kašnjenja	RMSE	MAE	RMSLE
None	Da	3	3	188.65	66.33	0.358
Box-Cox	Da	3	3	193.18	65.43	0.371
Logaritamska	Da	0	3	191.30	64.27	0.365
None	Ne	3	5	79.65	46.49	0.470
Box-Cox	Ne	3	5	77.01	43.21	0.439
Logaritamska	Ne	3	5	79.43	44.43	0.443

Tabela 16: Random Forest – RMSLE vrednosti za logaritamsku transformaciju (1 i 7 dana unapred, sa i bez meteo).

Grad	Ambrozija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana
KRAGUJEVAC	0.20	0.33	0.21	0.36	0.16	0.24	0.17	0.27
POŽAREVAC	0.43	0.55	0.46	0.58	0.30	0.36	0.30	0.40
PANČEVO	0.25	0.33	0.28	0.42	0.29	0.33	0.30	0.36

Tabela 17: Random Forest – RMSLE vrednosti za Box-Cox transformaciju (1 i 7 dana unapred, sa i bez meteo).

Grad	Ambrozija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana
KRAGUJEVAC	0.20	0.31	0.20	0.35	0.19	0.29	0.20	0.32
POŽAREVAC	0.42	0.55	0.47	0.57	0.33	0.43	0.32	0.44
PANČEVO	0.24	0.34	0.28	0.44	0.32	0.38	0.33	0.39

Tabela 18: Random Forest – RMSLE vrednosti bez transformacije (1 i 7 dana unapred, sa i bez meteo).

Grad	Ambrozija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana	1 dan	7 dana
KRAGUJEVAC	0.23	0.42	0.22	0.40	0.16	0.25	0.17	0.26
POŽAREVAC	0.46	0.65	0.48	0.62	0.34	0.46	0.32	0.46
PANČEVO	0.26	0.40	0.28	0.41	0.31	0.40	0.31	0.37

Početak sezone polena Tabela 20 prikazuje stvarni i predviđeni datum početka sezone polena za ambroziju i jovu u Požarevcu.

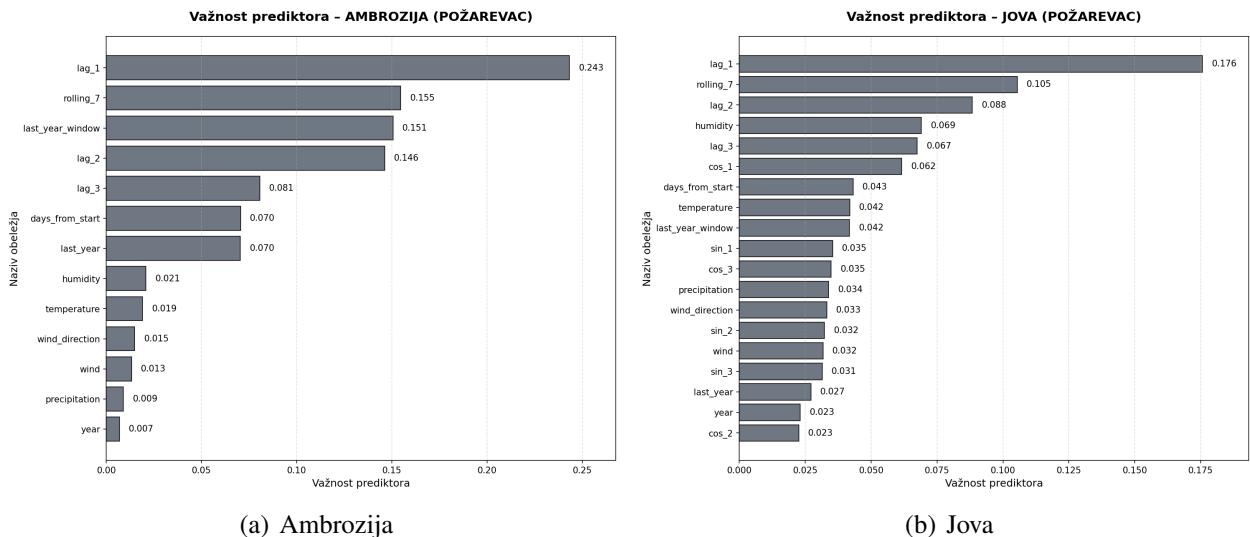
Tabela 19: Random Forest – najbolji modeli po RMSLE, sa prikazom RMSE i MAE.

Transformacija	RMSLE	RMSE	MAE
Sa meteorološkim parametrima			
Box-Cox	0.418	70.10	42.69
Logaritamska	0.426	69.67	41.83
<i>None</i>	0.457	73.76	45.44
Bez meteoroloških parametara			
Box-Cox	0.439	77.15	43.10
Logaritamska	0.444	78.99	44.12
<i>None</i>	0.470	79.65	46.49

Tabela 20: Prophet – početak sezone polena: stvarni datum i predikcija 7 dana unapred.

Grad	Ambrožija				Jova			
	Sa meteo		Bez meteo		Sa meteo		Bez meteo	
	Stvarni datum	Predikcija 7d						
KRAGUJEVAC	22.08.2024	20.08.2024	22.08.2024	22.08.2024	/	/	/	/
POŽAREVAC	16.08.2024	21.08.2024	16.08.2024	22.03.2024	/	/	/	/
PANČEVO	14.08.2024	19.08.2024	14.08.2024	20.03.2024	17.02.2024	/	17.02.2024.	/

Važnost prediktora Random Forest omogućava identifikaciju važnosti prediktora. Na slikama 15 prikazani su grafici važnosti prediktora za ambroziju i jovu u Požarevcu.

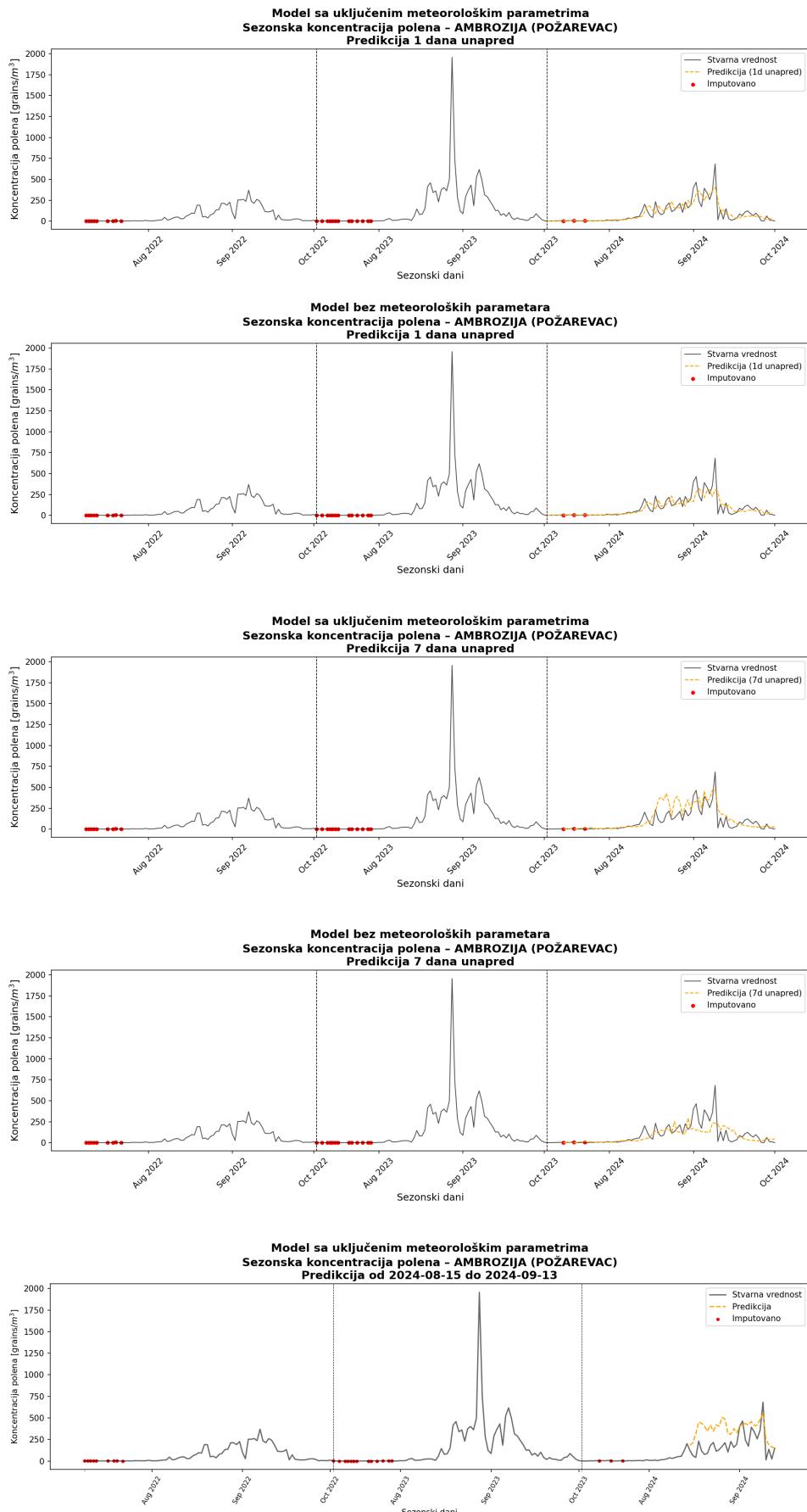


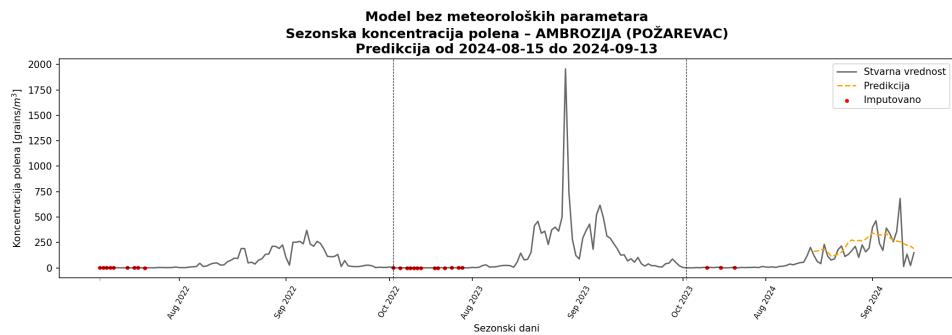
Slika 15: Važnost prediktora u RF modelu za ambroziju i jovu u Požarevcu.

Primer predikcije Na slici 16 prikazane su predikcije koncentracije polena ambrozije u Požarevcu pomoću Random Forest modela (logaritamska transformacija). Predikcije su prikazane naizmenično sa meteorološkim kovarijatima i bez njih. Redovi odgovaraju različitim horizontima predikcije:

1. Predikcija 1 dan unapred – prve dve slike (sa i bez meteoroloških kovarijata).
2. Predikcija 7 dana unapred (*rolling forecast*) – sledeće dve slike (sa i bez meteoroloških kovarijata).

3. Prognoza za narednih 30 dana – poslednje dve slike (sa i bez meteoroloških kovarijata).





Slika 16: Predikcije koncentracije ambrozije u Požarevcu pomoću Random Forest modela (logaritamska transformacija). Predikcije su prikazane naizmenično sa meteorološkim kovarijatima i bez njih. Redovi: predikcija 1 dan unapred, predikcija 7 dana unapred (*rolling forecast*), prognoza za narednih 30 dana.

Klasifikacija nivoa koncentracije Klasifikacija se odnosi na predikciju koncentracije za posmatrani dan, određenu 7 dana unapred. U tabeli 21 prikazane su konfuzione matrice za RF modele sa i bez meteoroloških kovarijata.

Tabela 21: Random Forest – konfuzione matrice za klasifikaciju nivoa ambrozije u Požarevcu.

		Predikcija						
		Nisko	Srednje	Visoko				
Stvarno	Predikcija							
	Nisko	26	3	4				
Nisko		26	3	4	Nisko	26	2	5
Srednje		5	9	2	Srednje	8	6	3
Visoko		0	5	23	Visoko	0	5	23

Diskusija. Rezultati Random Forest modela pokazuju da ovaj pristup uspešno hvata kratkoročne oscilacije u koncentraciji polena, posebno kada se koriste meteorološki kovarijati. Primena logaritamske transformacije $\log(1 + \frac{x}{30})$ dala je najbolje rezultate, jer omogućava kontrolu ekstremnih vrednosti i stabilizuje seriju, što se ogleda u nižim vrednostima RMSLE, RMSE i MAE. Box-Cox transformacija ili rad bez transformacije često dovode do prevelikog „izravnavanja“ serije ili gubitka stabilnosti predikcija.

Kada se posmatra **predikcija narednog dana**, model veoma dobro koristi meteorološke informacije, čime su dnevne oscilacije koncentracije polena precizno praćene. Sa druge strane, predikcija za nekoliko dana unapred (npr. 7 ili 30 dana) je u većini slučajeva znatno slabija, što može biti posledica preobučavanja modela na vremenskim podacima i previše specifičnih obrazaca u kratkom periodu.

Analiza broja prethodnih dana koji se koriste u modelu pokazuje da kada su dostupni meteorološki podaci, za postizanje dobre predikcije potrebna je manja istorija vrednosti polena. Suprotno tome, u situacijama kada meteorološki podaci nisu uključeni, model mora koristiti veći broj prethodnih dana kako bi kompenzovao nedostatak informacija o vremenskim faktorima.

Analiza važnosti prediktora pokazuje da temperatura i vlažnost imaju najveći uticaj na tačnost predikcija, dok dodatni parametri poput padavina i vetra u većini slučajeva ne doprinose značajno poboljšanju performansi modela. Prevelik broj vremenskih parametara može dovesti do prekomernog

učenja specifičnih dnevnih obrazaca, što smanjuje sposobnost modela da precizno predviđa koncentracije polena za nekoliko dana unapred.

Random Forest se pokazao kao pouzdan model za predikciju dnevnih vrednosti koncentracije polena, naročito kada se koriste meteorološki podaci. Njegova snaga leži u preciznom praćenju kratkoročnih dnevnih oscilacija, dok predikcije za više dana unapred imaju ograničenu tačnost.

4.3 Uporedna analiza modela

Za predikciju koncentracije polena korišćeni su modeli **SARIMAX**, **Prophet** i **Random Forest regresija**, pri čemu se svaki od njih pokazao pogodnijim u određenim uslovima i vremenskim horizontima prognoze.

Model **Random Forest** izdvojio se kao najpouzdaniji izbor za predikciju koncentracije polena *narednog dana*, naročito kada su dostupni meteorološki podaci. Analizom je utvrđeno da su dovoljna samo tri vremenska kašnjenja serije ukoliko su vremenski parametri poznati, što je u skladu i sa rezultatima SARIMAX modela gde su optimalni parametri p i q takođe ≤ 3 . Random Forest jasno prepoznaće složene nelinearne veze i pokazuje da su temperatura i vlažnost ključni faktori, što je potvrđeno i analizom linearne zavisnosti ovih parametara. Kada su meteorološki podaci dostupni, Random Forest vrlo precizno prati njihove promene i daje znatno bolje rezultate u predikciji za jedan dan unapred u odnosu na SARIMAX i Prophet. Međutim, pri pokušajima predikcije nekoliko dana unapred performanse Random Foresta naglo opadaju, što ukazuje na to da je ovaj model primarno pogodan za kratkoročne prognoze. Dodatno, značaj godine u Random Forest modelu se pokazao vrlo mali, što opravdava činjenicu da je kod SARIMAX modela optimalno bilo uzeti $D = 0$.

Model **SARIMAX** pokazao se kao stabilniji u predikcijama koje obuhvataju više dana. On uspešno koristi sezonske obrasce i egzogene faktore, a naročito je koristan kada su meteorološki parametri dostupni, jer konzistentno prati njihove promene. Ipak, ograničenje SARIMAX-a je u tome što modelira isključivo linearne zavisnosti, pa može davati pogrešne rezultate u situacijama kada se javljaju nelinearne interakcije, npr. kada je van sezone zabeležena neuobičajeno visoka temperatura.

Prophet model je pokazao najmanju osetljivost na dostupnost meteoroloških podataka. On dominantno oslanja svoju strukturu na sezonalnost i dugačke trendove, dok meteorološki faktori uglavnom samo blago utiču na oblik serije u blizini sezonskih maksimuma. Prednost ovog modela je njegova jednostavna implementacija i vrlo brzo treniranje, što ga čini pogodnim za situacije kada je potrebno često ažuriranje modela ili rad sa višegodišnjim podacima.

Za sve primenjene modele, pokazalo se da je za Prophet najpouzdanija Box-Cox transformacija, dok su za SARIMAX i Random Forest najbolje rezultate u pogledu stabilnosti i smanjenja varijanse greške davala $\log(1 + \frac{x}{30})$ transformacija.

Na osnovu analize može se zaključiti da je izbor modela uslovljen ciljem prognoze i dostupnošću podataka. Ako je cilj *kratkoročna prognoza za jedan dan unapred*, a meteorološki podaci su dostupni, preporuka je **Random Forest**. Ako je cilj *prognoza za više dana unapred*, bolji izbor je **SARIMAX**. U situacijama kada je neophodno brzo dobiti procenu, bez obzira na vremenski horizont, izbor je **Prophet**.

Performanse po modelima i meteo konfiguracijama U Tabelama 22 i 23 prikazane su RLMSLE vrednosti za logaritamsku transformaciju za ambroziju i jovu u Požarevcu, sa i bez meteoroloških podataka, po različitim horizontima predikcije.

Tabela 22: RLMSLE vrednosti za predikciju koncentracije ambrozije u Požarevcu po modelima i konfiguracijama meteoroloških podataka.

Model	Meteo	Horizont predikcije (dani)									
		1	2	3	4	5	6	7	8	9	10
SARIMAX	Da	0.42	0.42	0.44	0.44	0.46	0.46	0.46	0.47	0.47	0.46
	Ne	0.45	0.45	0.47	0.47	0.49	0.50	0.50	0.50	0.50	0.49
Prophet	Da	0.46	0.48	0.50	0.52	0.55	0.57	0.59	0.61	0.62	0.64
	Ne	0.50	0.52	0.53	0.54	0.56	0.58	0.59	0.60	0.61	0.63
Random Forest	Da	0.43	0.45	0.48	0.48	0.53	0.56	0.55	0.56	0.56	0.56
	Ne	0.46	0.49	0.51	0.51	0.57	0.56	0.58	0.58	0.58	0.58
Naive Forecasting	/	0.54	0.57	0.62	0.60	0.70	0.77	0.78	0.83	0.87	0.87

Tabela 23: RLMSLE vrednosti za predikciju koncentracije jove u Požarevcu po modelima i konfiguracijama meteoroloških podataka.

Model	Meteo	Horizont predikcije (dani)									
		1	2	3	4	5	6	7	8	9	10
SARIMAX	Da	0.29	0.33	0.34	0.34	0.35	0.36	0.37	0.37	0.37	0.38
	Ne	0.28	0.32	0.33	0.34	0.35	0.37	0.38	0.38	0.39	0.41
Prophet	Da	0.35	0.37	0.38	0.39	0.41	0.43	0.45	0.47	0.48	0.49
	Ne	0.36	0.37	0.38	0.39	0.41	0.42	0.44	0.45	0.46	0.48
Random Forest	Da	0.30	0.34	0.34	0.33	0.34	0.36	0.36	0.36	0.37	0.40
	Ne	0.30	0.34	0.35	0.35	0.37	0.39	0.40	0.39	0.42	0.44
Naive Forecasting	/	0.31	0.41	0.43	0.43	0.42	0.47	0.51	0.51	0.54	0.58

5 Zaključak

U ovom diplomskom radu razvijen je i implementiran **sistem za predikciju koncentracija polena u Srbiji**, zasnovan na integraciji **geostatističkih metoda, modela vremenskih serija i mašinskog učenja**. Prvi korak podrazumevao je primenu metode prostorno-vremenskog *kriging-a* za popunjavanje nedostajućih vrednosti u dostupnim skupovima podataka. Na taj način formirane su potpune vremensko serije za sve analizirane lokacije i alergene. Rezultati su pokazali da je ova metoda uspešno smanjila problem diskontinuiteta u podacima i obezbedila pouzdanu osnovu za dalje modelovanje i predikciju.

Za predikciju koncentracije polena korišćeni su modeli **SARIMAX, Prophet i Random Forest regresija**. Model **SARIMAX** pokazao se najpogodnijim za predikcije nekoliko dana unapred, jer uspešno koristi sezonske obrasce i egzogene faktore poput temperature i vlažnosti vazduha. **Random Forest** se izdvojio kao najbolji izbor za predikciju koncentracije polena narednog dana, zahvaljujući sposobnosti da prepozna složene nelinearne veze između meteoroloških parametara i aktuelnog stanja u seriji. **Prophet model** se istakao svojom jednostavnom implementacijom i vrlo brzim treniranjem, što ga čini praktičnim za rad sa višegodišnjim podacima i čestim ponovnim ažuriranjima.

Kada je reč o ulozi meteoroloških podataka, oni mogu značajno doprineti tačnosti predikcije u kratkom roku (npr. jedan dan unapred), dok u predikcijama koje obuhvataju više dana ipak dominantnu ulogu ima sezonska komponenta, pa meteorološki faktori imaju ograničeniji uticaj.

Praktični značaj ovog istraživanja ogleda se u mogućnosti pravovremenog informisanja i planiranja terapija kod osoba koje pate od alergija, čime se može značajno smanjiti intenzitet simptoma i poboljšati kvalitet života. Takođe, dobijeni modeli mogu doprineti unapređenju poljoprivredne proizvodnje i pčelarstva, kroz bolju organizaciju aktivnosti u periodima cvetanja biljaka značajnih za opravljanje i prikupljanje polena.

Kao **glavna ograničenja** rada identifikovana su: ograničena rezolucija meteoroloških podataka, nemogućnost modela da u potpunosti obuhvate iznenadne promene koncentracija usled neočekivanih meteoroloških fenomena, kao i nedostatak real-time podataka u trenutku treniranja modela.

Preporuke za budući rad uključuju:

- Primenu naprednih *dubokih neuralnih mreža* poput LSTM i GRU, koje su posebno dizajnirane za sekvensijalne podatke i mogu bolje modelovati kompleksne vremenske zavisnosti.
- Povezivanje sistema sa *real-time meteorološkim i satelitskim podacima* radi poboljšanja prostorne i vremenske rezolucije predikcija.
- Proširenje primene prostorno-vremenskog *kriging-a* ne samo u svrhu imputacije nedostajućih vrednosti, već i za prostorno-vremensku interpolaciju, u skladu sa [40], čime bi se omogućila rekonstrukcija vremenskih serija za gotovo svaku lokaciju u Srbiji, a zatim njihova predikcija pomoću modela kao što su SARIMAX, Prophet ili Random Forest.
- Razvoj i implementaciju **operativne platforme** za javno obaveštavanje o očekivanim koncentracijama polena, u saradnji sa zdravstvenim i meteorološkim institucijama, što bi doprinelo **javnom zdravlju, poljoprivredi i ekologiji**.

Zaključno, rad potvrđuje da integracija **geostatistike, vremenskih serija i mašinskog učenja** predstavlja moćan i efikasan pristup za rešavanje problema predikcije koncentracija polena i da ima široku praktičnu primenu u različitim sferama društva.

Literatura

- [1] Bionette. Alergija na polen - simptomi i lečenje, 2023.
- [2] Drugačiji Pristup. Alergija — sve što treba da znate, 2025.
- [3] Y. Zhang et al. Omalizumab is effective in the preseasonal treatment of seasonal allergic rhinitis. *Clinical and Translational Allergy*, 12:e12018, 2022.
- [4] M. Worm et al. Efficacy and safety of birch pollen allergoid subcutaneous immunotherapy. *Journal of Allergy and Clinical Immunology*, 143:1234–1242, 2019.
- [5] R. Minić et al. Impact of tree pollen distribution on allergic diseases in serbia. *PMC*, 2020.
- [6] M. Sofiev et al. Designing an automatic pollen monitoring network for europe. *Science of the Total Environment*, 859:160292, 2023.
- [7] J. Smith et al. Biology and ecology of pollen grains. *Journal of Palynology*, 57:23–45, 2021.
- [8] K. Piotrowska et al. The effect of meteorological factors on airborne betula pollen concentration in lublin. *PMC*, 2012.
- [9] J. Chico-Fernández et al. Relationship of meteorological variables with the concentration of airborne pollen. *MDPI Applied Sciences*, 15:692, 2025.
- [10] Met Museum. Telling time in ancient egypt, 2025.
- [11] Historical Law Archive. Laws restricting divination and prediction, 2025.
- [12] G. Udny Yule. On the method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226:267–298, 1927.
- [13] Eugen Slutsky. The summation of random causes as the source of cyclic processes. *Econometrica*, 5:105–146, 1937.
- [14] George E. P. Box and Gwilym M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, 1970.
- [15] D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74:427–431, 1979.
- [16] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54:159–178, 1992.
- [17] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26:211–252, 1964.
- [18] Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer, New York, 2002.
- [19] James D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.

- [20] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, 2002.
- [21] Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.
- [22] Leo Breiman. *Random Forests*, volume 45. Springer, 2001.
- [23] Noel A.C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, New York, 1993.
- [24] Edward H. Isaaks and R. Mohan Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 1989.
- [25] Jean-Paul Chiles and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, Hoboken, NJ, 2012.
- [26] K. R. Gabriel and P. J. Diggle. Spatio-temporal variograms. *Biometrika*, 67(2):411–420, 1980.
- [27] Republički hidrometeorološki zavod Srbije. Polen – objedinjeni podaci od 2016. godine. [https://data.gov.rs/sr/datasets/polen-objelinjeni-podatsi-od-2016-godine/](https://data.gov.rs/sr/datasets/polen-objedinjeni-podatsi-od-2016-godine/), 2023. Pristupljeno: 2025-09-07.
- [28] G. D’Amato, L. Cecchi, S. Bonini, C. Nunes, I. Annesi-Maesano, H. Behrendt, G. Liccardi, T. Popov, and P. van Cauwenberge. Allergenic pollen and pollen allergy in europe. *Allergy*, 62(9):976–990, 2007.
- [29] World Health Organization. *Health aspects of air pollution with particulate matter, ozone and nitrogen dioxide: Report on a WHO working group*. WHO Regional Office for Europe, Copenhagen, 2003.
- [30] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellán, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J. N. Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.
- [31] M. Sofiev, P. Siljamo, H. Ranta, and A. Rantio-Lehtimäki. Towards numerical forecasting of long-range air transport of birch pollen: theoretical considerations and a feasibility study. *International Journal of Biometeorology*, 50(6):392–402, 2006.
- [32] Ł. Grewling, B. Jackowiak, M. Nowak, A. Uruska, M. Ziemianin, and M. Smith. Combining phenological observations and pollen transport modeling for management of ambrosia pollen sources. *International Journal of Biometeorology*, 60(11):1493–1506, 2016.
- [33] B. Šikoparija, P. Radišić, T. Pejak-Šikoparija, M. Smith, and C. Galán. The pannonian plain as a source of ambrosia pollen in the balkans. *International Journal of Biometeorology*, 61(10):1697–1709, 2017.
- [34] Jean-Paul Chilès and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, Hoboken, NJ, 2nd edition edition, 2009.

- [35] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3rd edition edition, 2018.
- [36] M. C. Jones. On winsorizing and trimming for robust estimation of location. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(4):311–322, 1996.
- [37] Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.
- [38] Greta M. Ljung and George E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.
- [39] Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 1968 23rd ACM national conference*, pages 517–524, 1968.
- [40] Aleksandar Sekulić, Milan Kilibarda, Gerard B. M. Heuvelink, Milena Nikolić, and Branislav Bajat. Random forest spatial interpolation. *Remote Sensing*, 12(10), 2020.