

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

Diplomski rad

Student: Luka Milutinović

Mentor: prof. dr Predrag Tadić

Beograd, Oktobar 2025



Sadržaj

- 1 Motivacija
- 2 Podaci o polenu
- 3 Metodologija
- 4 Rezultati
- 5 Diskusija i zaključak

Motivacija

- Polen izaziva alergijske reakcije kod velikog broja ljudi — u Srbiji čak oko **trećine stanovništva**.
- Veća izloženost dovodi do jačih simptoma i smanjenog kvaliteta života (san, produktivnost, opšte stanje).
- Pravovremena **prognoza koncentracije polena** pomaže u smanjenju rizika:
 - omogućava izbegavanje visokih koncentracija,
 - olakšava preventivno lečenje.
- Postoje servisi poput **iqair.com** i **accuweather.com**, ali njihovi modeli nisu javno dostupni.

Podaci o polenu

Lokacije praćenja polena u Srbiji

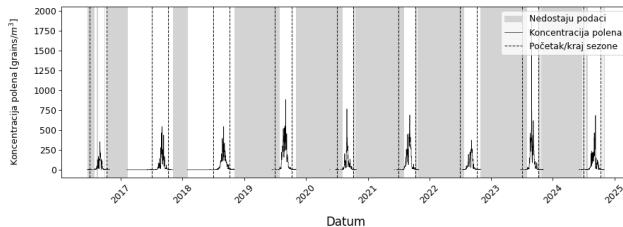


Podaci su prikupljeni sa sajta **Portala otvorenih podataka Republike Srbije**, iz baze "*Polen – objedinjeni podaci od 2016. godine*".

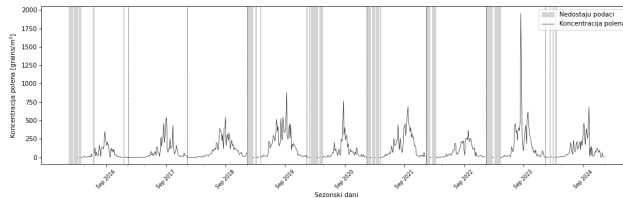
Ova baza sadrži podatke o koncentraciji polena za **29 mernih mesta** širom Srbije i obuhvata **26 razliĉitih alergena**.

Primer merenja – Ambrozija (Požarevac)

Dnevna koncentracija polena - AMBROZIJA (POŽAREVAC)

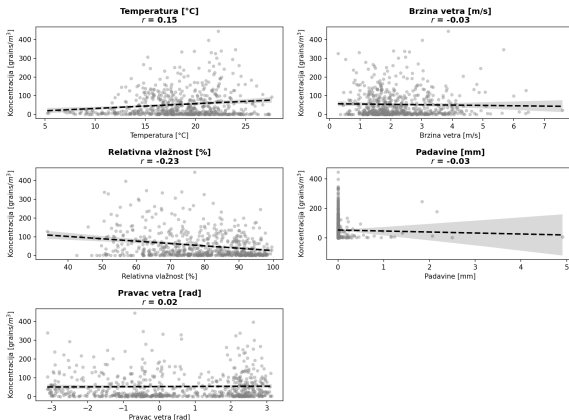


Sezonska koncentracija polena - AMBROZIJA (POŽAREVAC)



Meteorološki parametri

Zavisnost koncentracije polena od meteoroloških parametara
AMBROZIJA - POŽAREVAC



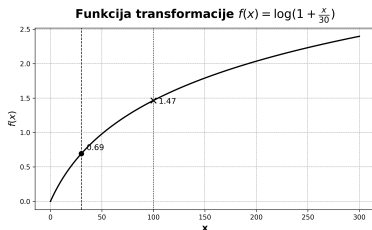
Uz polenske podatke, analizirani su i meteorološki parametri koji značajno utiču na koncentraciju polena u vazduhu:

- Temperatura vazduha (+).
- Brzina vetra (+).
- Vlažnost vazduha (-).
- Padavine (-).
- Pravac vetra.

Transformacije i evaluacija modela

Transformacije:

- **Log transformacija:** $y = \log(1 + x/30)$



- **Box-Cox transformacija:**

$$y = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases}$$

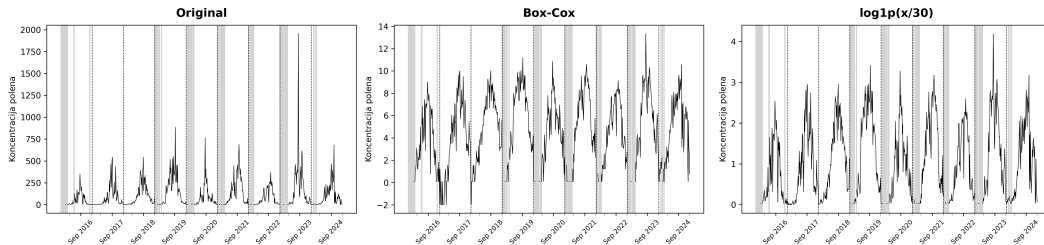
Evaluacija: RMSLE

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

- n – broj primera u skupu podataka
- y_i – stvarna koncentracija polena za i -ti primer
- \hat{y}_i – predviđena koncentracija polena za i -ti primer

Efekat transformacija na koncentraciju polena

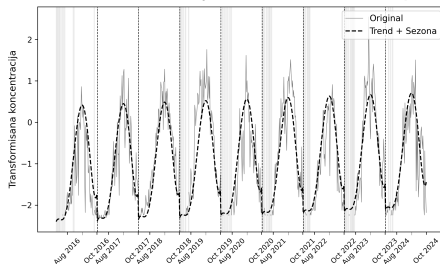
Sezonska koncentracija polena - AMBROZIJA (POŽAREVAC)



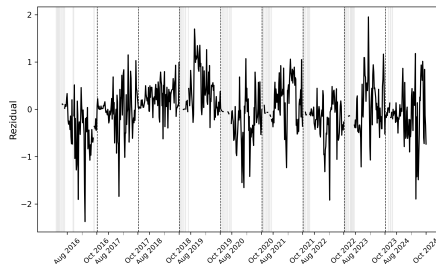
Imputacija podataka

Sezonska dekompozicija - AMBROZIJA, lokacija: POŽAREVAC

Transformisana koncentracija i trend sa sezonskom komponentom

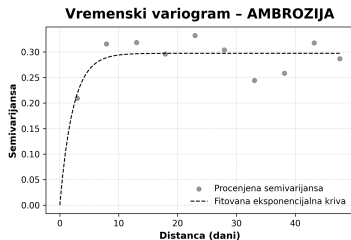
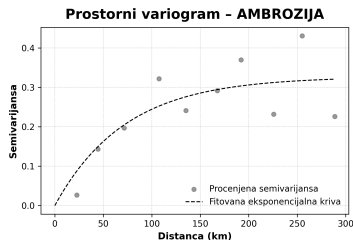


Reziduali



Slika: Signal nakon uklanjanja trenda

Kriging i prostorno-vremenski variogram



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 je:

- $\gamma(u, 0)$ – prostorni variogram
- $\gamma(0, v)$ – vremenski variogram
- k – parametar određen optimizacijom za najbolje zadovoljenje uslova separabilnosti

Modelovanje vremenskih serija: SARIMAX i Prophet

- **SARIMAX** — sezonski ARIMA model sa egzogenim promenljivama
Definisan parametrima $(p, d, q) \times (P, D, Q, s)$:

$$\Phi_P(L^s)\phi_p(L)(1-L)^d(1-L^s)^D y_t = \Theta_Q(L^s)\theta_q(L)\varepsilon_t + \beta^T \mathbf{x}_t$$

gde su:

- $\phi_p(L), \theta_q(L)$ — AR i MA operatori
 - $\Phi_P(L^s), \Theta_Q(L^s)$ — sezonski AR i MA operatori
 - d, D — red diferenciranja
 - s — dužina sezonskog perioda
 - \mathbf{x}_t — egzogene promenljive
- **Prophet** — model razvijen od strane kompanije *Meta* (bivši *Facebook*), pri čemu je u ovom radu primenjena njegova **multiplikativna varijanta**:

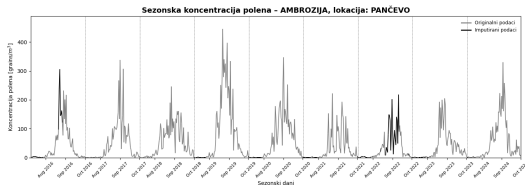
$$y(t) = g(t) \times (1 + s(t)) + \varepsilon_t$$

11/20 gde su: $g(t)$ — trend, $s(t)$ — sezonska komponenta (Furijeovi redovi).

Modelovanje vremenskih serija pomoću Random Foresta

- Model baziran na **atributima izvedenim iz vremenskih serija**:
 - Vrednosti koncentracije polena sa zaostatkom od nekoliko 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 i godina
 - Furijeovi redovi za modelovanje sezonalnosti
- Ovim pristupom model ne koristi vremensku zavisnost direktno, već model uči obrasce iz generisanih vremenskih karakteristika.

Rezultati imputacije podataka

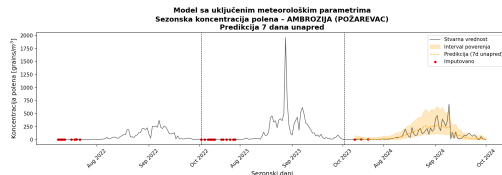
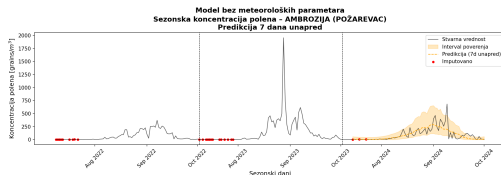


Slika: Primer imputacije vremenske serije

Metod	Bez meteo	Sa meteo
Naivna interpolacija	0.309	—
IDW interpolacija	1.085	—
Kriging + Box-Cox	0.260	0.262

Tabela: Uporedni rezultati imputacije (RMSLE).

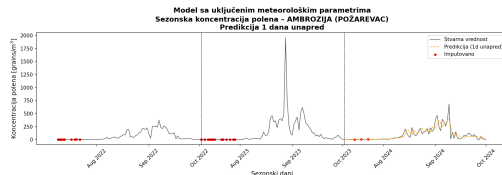
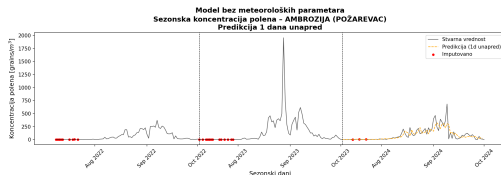
Predikcija 7 dana unapred — SARIMAX



Slika: Primer predikcije 7 dana unapred (SARIMAX model sa logaritamskom transformacijom).

Model	Bez meteo	Sa meteo
SARIMAX	2.7	2.6
Prophet	3.6	3.3
Random Forest	4.0	4.0

Predikcija 1 dan unapred — Random Forest



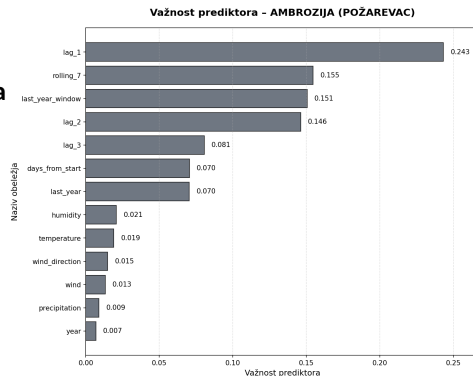
Slika: Primer predikcije 1 dan unapred (Random Forest model sa logaritamskom transformacijom).

Model	Bez meteo	Sa meteo
SARIMAX	0.912	0.918
Prophet	0.887	0.906
Random Forest	0.911	0.922

Značaj karakteristika u Random Forest modelu

Analiza značaja ulaznih karakteristika:

- Najuticajnije promenljive:
 - Vremensko kašnjenje koncentracije polena do 3 dana
 - Prosečna koncentracija u prethodnih 7 dana
 - Temperatura i relativna vlažnost vazduha
 - Furijeovi redovi
- Pokazuje koje meteorološke i istorijske komponente najviše utiču na predikciju.



Slika: Značaj karakteristika u Random Forest modelu.

Uporedna evaluacija modela — RMSLE

Model	1 dan unapred		7 dana unapred	
	Bez meteo	Sa meteo	Bez meteo	Sa meteo
SARIMAX	0.45	0.42	0.50	0.46
Prophet	0.50	0.46	0.59	0.59
Random Forest	0.46	0.41	0.58	0.55
Naivna predikcija	0.54	–	0.78	–

Tabela: RMSLE rezultati modela — ambrozija, Požarevac.

Zaključak — Glavni rezultati i doprinos

- Razvijen je i implementiran sistem za **predikciju koncentracija polena u Srbiji**, koji kombinuje:
 - **Prostorno-vremenski kriging** — za imputaciju nedostajućih podataka.
 - **SARIMAX, Prophet i Random Forest** modele — za predikciju koncentracije polena.
- **Kriging** uspešno smanjuje diskontinuitet u podacima i omogućava pouzdano modelovanje vremenskih serija.
- **SARIMAX** — najpogodniji za višednevne prognoze (iskorišćava sezonalnost i meteorološke faktore).
- **Random Forest** — najbolji za kratkoročne prognoze (1 dan unapred), jer prepoznaje nelinearne odnose.
- **Prophet** — jednostavan za implementaciju i brz za treniranje, pogodan za dugoročne vremenske serije.
- Meteorološki podaci doprinose tačnosti **kratkoročnih predikcija**, dok kod višednevnih dominira sezonska komponenta.

Zaključak — Značaj, ograničenja i budući rad

Praktični značaj:

- Pravovremeno informisanje i planiranje terapija kod alergičnih osoba.
- Podrška poljoprivredi i pčelarstvu kroz predikciju perioda cvetanja i polinacije.

Ograničenja:

- Ograničena prostorna rezolucija meteoroloških podataka.
- Nemogućnost modela da prate iznenadne promene koncentracija.
- Nedostatak real-time podataka prilikom treniranja.

Preporuke za budući rad:

- Integracija sa **real-time meteorološkim i satelitskim podacima**.
- Proširenje **prostorno-vremenskog kriging-a** na interpolaciju i rekonstrukciju serija.
- Razvoj **operativne platforme** za javno obaveštavanje o očekivanim koncentracijama polena.

Hvala na pažnji!

Pitanja i diskusija

Beograd, Oktobar 2025.