Modeli kratkoročne prognoze koncentracije peludi bazirani na strojnom učenju

Slobodan Jelić

University J. J. Strossmayer of Osijek
Department of Mathematics
Trg Ljudevita Gaja 6
31000 Osijek, Croatia
http://www.mathos.unios.hr

sjelic@mathos.hr

Joint work with:

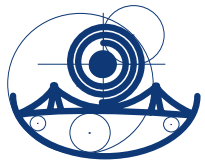
Kristian Sabo

Interreg – IPA CBC
Croatia - Serbia
RealForAll

HRZZ
Hrvatska zaklada za znanost

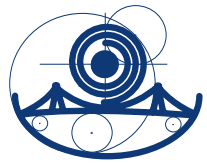
[Statistički seminar, Odjel za matematiku]

6. prosinca 2018.



Pregled predavanja

- upoznavanje s projektom i projektnim zadacima
- upoznavanje s podacima
- model kratkoročne prognoze
- analiza rezultata



Projekt i projektni zadaci

- rad nastao u okviru dva projekta:
  - Real-time measurements and forecasting for successful prevention and management of seasonal alergies in Croatia-Serbia cross-border region (RealForAll)
  - The optimization and statistical models and methods in recognizing properties of data sets measured with errors (OSMoMeSIP-IP-2016-06-6545)

Projekt i projektni zadaci

Interreg – IPA CBC
Croatia - Serbia
RealForAll

More information:
www.realforall.com

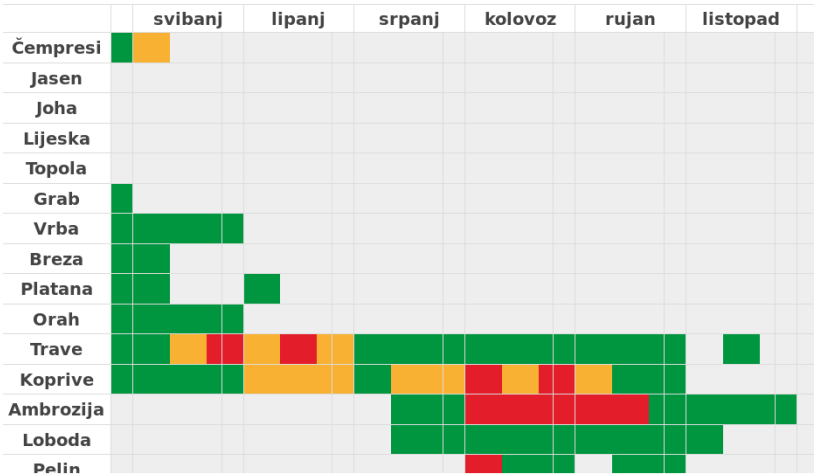
- **Project title:** *Real-time measurements and forecasting for successful prevention and management of seasonal allergies in Croatia-Serbia cross-border region*
- **Project Acronym:** *RealForAll*
- **Project ID:** *HR-RS151*
- **Beneficiary Lead:** *BioSense Institute*
- **Programme Co-financing:** *530.587 EUR*
- **Start of project:** *15.07.2017.*
- **End of project:** *14.01.2020.*

This project is co-financed by the European Union

BioSense INSTITUTE

Grad Osijek

Projekt i projektni zadaci



- projekt usmjeren na rješavanje problema javnog zdravlja
- izgradnja sustava za **mjerenje i predikciju koncentracije peludi u stvarnom vremenu**
- http://www.realforall.com
- postojeći sustav zasnovan je na **peludnom kalendaru**
- semafor za svaki mjesec u godini i svaku alergijsku biljku

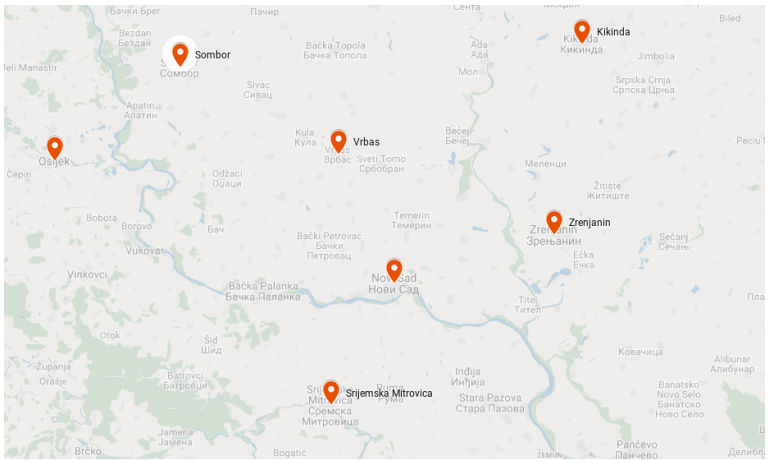
Projekt i projektni zadaci

- dugoročna prognoza peludi:
  - peludni kalendar[1]
  - daje predikcije nekoliko mjeseci prije početka sezone
  - **ne uzima** u obzir meteorološke čimbenike
- kratkoročna prognoza peludi:
  - daje predikcije nekoliko dana unaprijed
  - **uzima** u obzir meteorološke čimbenike
  - izuzetno važna za pacijente i njihovo planiranje aktivnosti tijekom i neposredno prije sezone

[1] B. Šikoparija et al. "How to prepare a pollen calendar for forecasting daily pollen concentrations of Ambrosia, Betula and Poaceae?" In: *Aerobiologia* (Jan. 2018).

Podaci

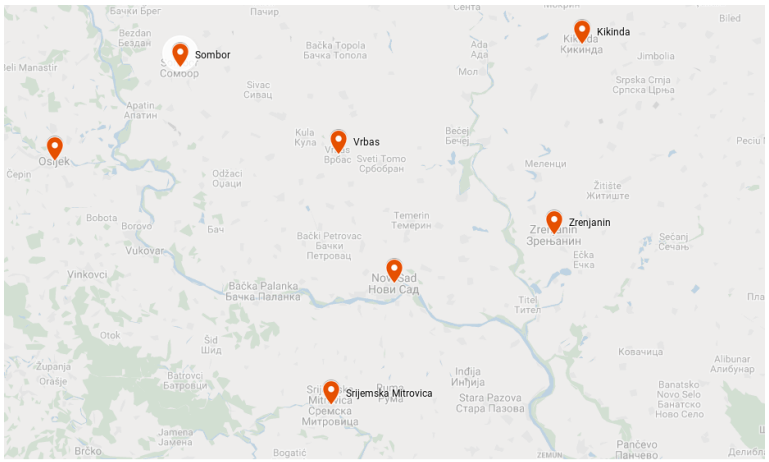
- vremenski period: Novi Sad (od 2000. godine), ostale lokacije (od 2008. godine)
- skup podataka sadrži **25620 mjerenja** prosječne koncentracije (u P/m^3) ambrozije (Ambrosia), breze (Betula) i trava (Poaceae) na 7 različitih lokacija



- Osijek (OS)
- Sombor (SO)
- Novi Sad (NS)
- Vrbas (VB)
- Zrenjanin (ZR)
- Sremska Mitorvica (SM)
- Kikinda(KI)

Projekt i projektni zadaci

- vremenski period: Novi Sad (od 2000. godine), ostale lokacije (od 2008. godine)
- dostupni kroz portal **TuTiempo.net**[2] na svim lokacijama mjerenja koncentracije peludi

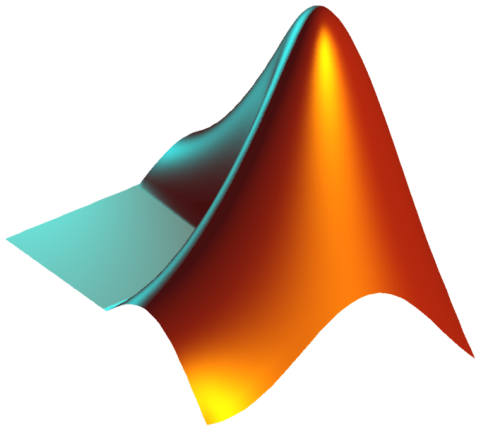


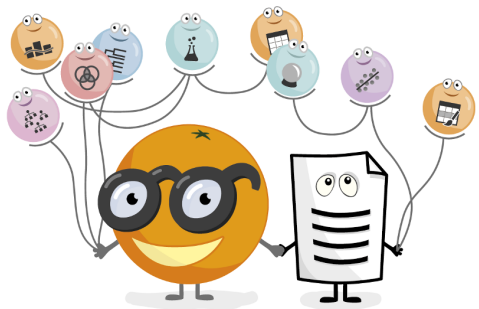
- Osijek (OS)
- Sombor (SO)
- Novi Sad (NS)
- Vrbas (VB)
- Zrenjanin (ZR)
- Sremska Mitorvica (SM)
- Kikinda(KI)

[2] https://en.tutiempo.net/climate

Priprema i rad s podacima

- priprema jedinstvene baze podataka za rad na projektu
- alati: MySQL, Matlab, Orange (Python)





Varijable

25620 instance(s), 21 feature(s), 0 meta attribute(s)
Data has no target variable.

Columns (Double click to edit)

	Name	Type	Role	Values
1	RA	C categorical	feature	0.0, 1.0
2	SN	C categorical	feature	0.0, 1.0
3	TS	C categorical	feature	0.0, 1.0
4	FG	C categorical	feature	0.0, 1.0
5	MBV	N numeric	feature	
6	SBV	N numeric	feature	
7	PAD	N numeric	feature	
8	VLZ	N numeric	feature	
9	ATT	N numeric	feature	
10	MNT	N numeric	feature	
11	MKT	N numeric	feature	
12	SRT	N numeric	feature	
13	GOD	N numeric	feature	
14	MSC	N numeric	feature	
15	DAN	N numeric	feature	
16	RBD	N numeric	feature	
17	LOK	C categorical	feature	KI, NS, OS, SM, SO, VB, ZR
18	PRAM	N numeric	feature	
19	PRBR	N numeric	feature	
20	PRTR	N numeric	feature	
21	DAT	T datetime	feature	

Vizualizacija podataka

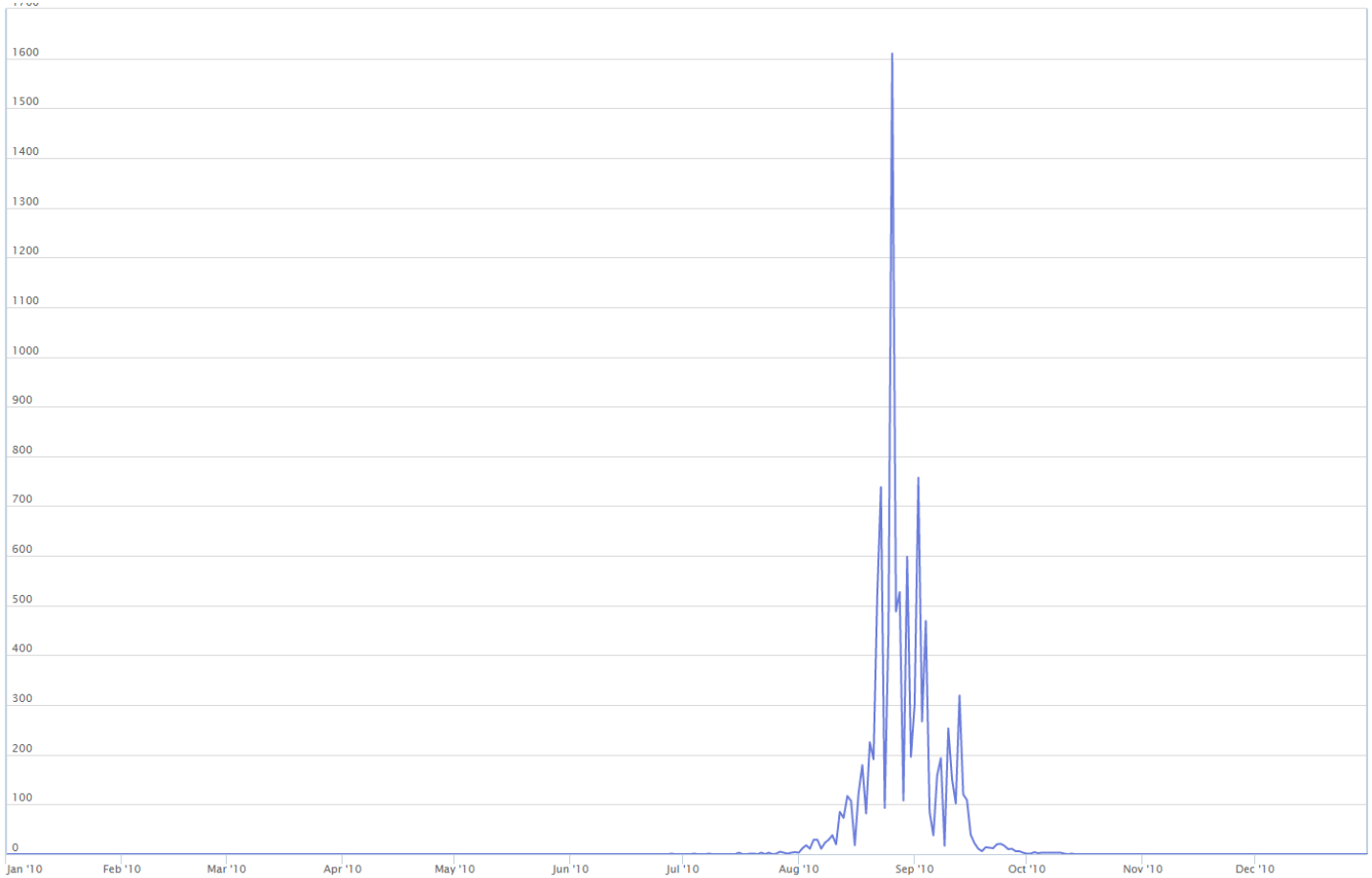


Figure: Podaci za koncentracije ambrozije u Osijeku, 2010. godina

Vizualizacija podataka

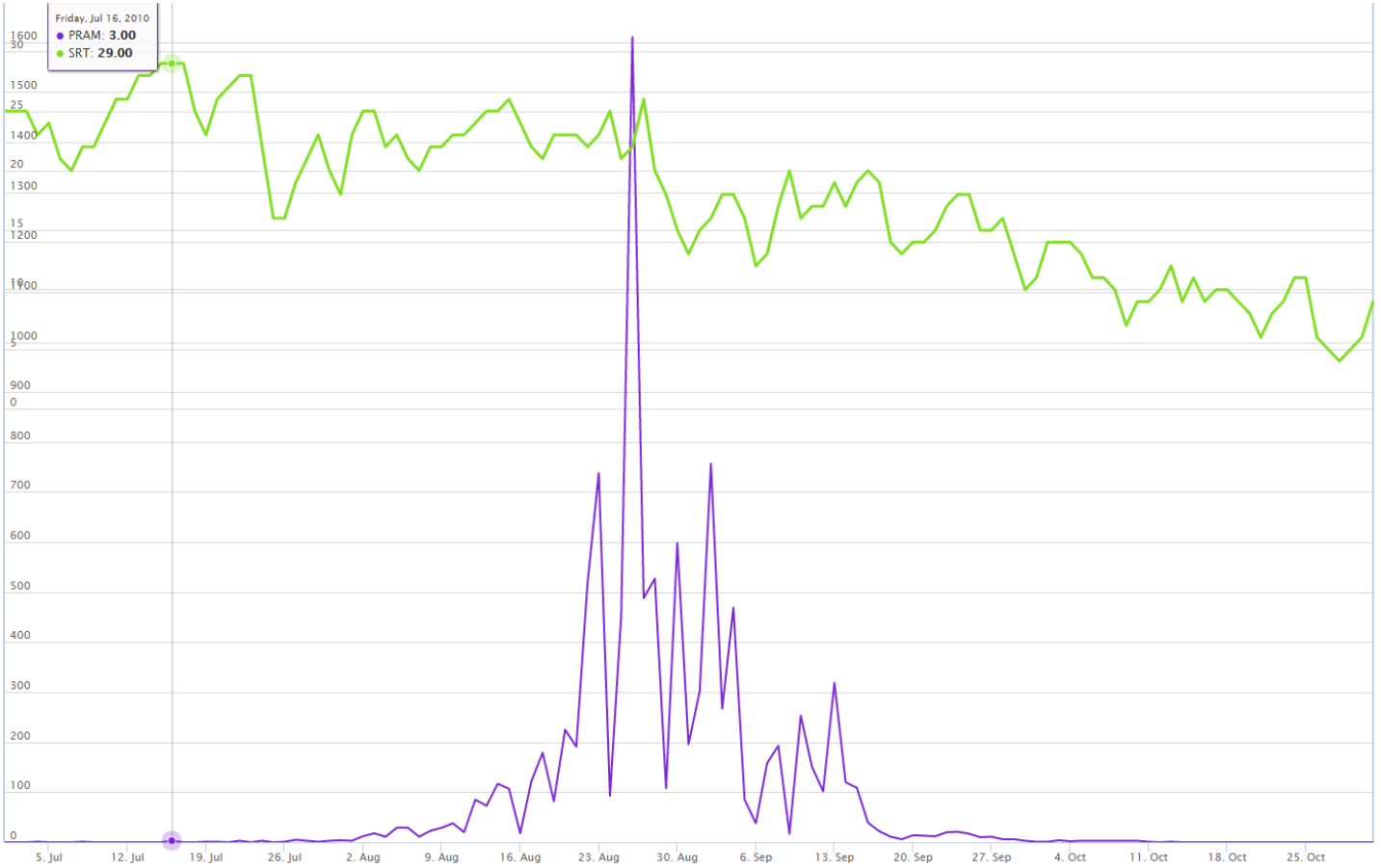


Figure: Podaci za koncentracije ambrozije i srednju dnevnu temepraturu u Osijeku (od 7. - 10. mjeseca), 2010. godina

Vizualizacija podataka

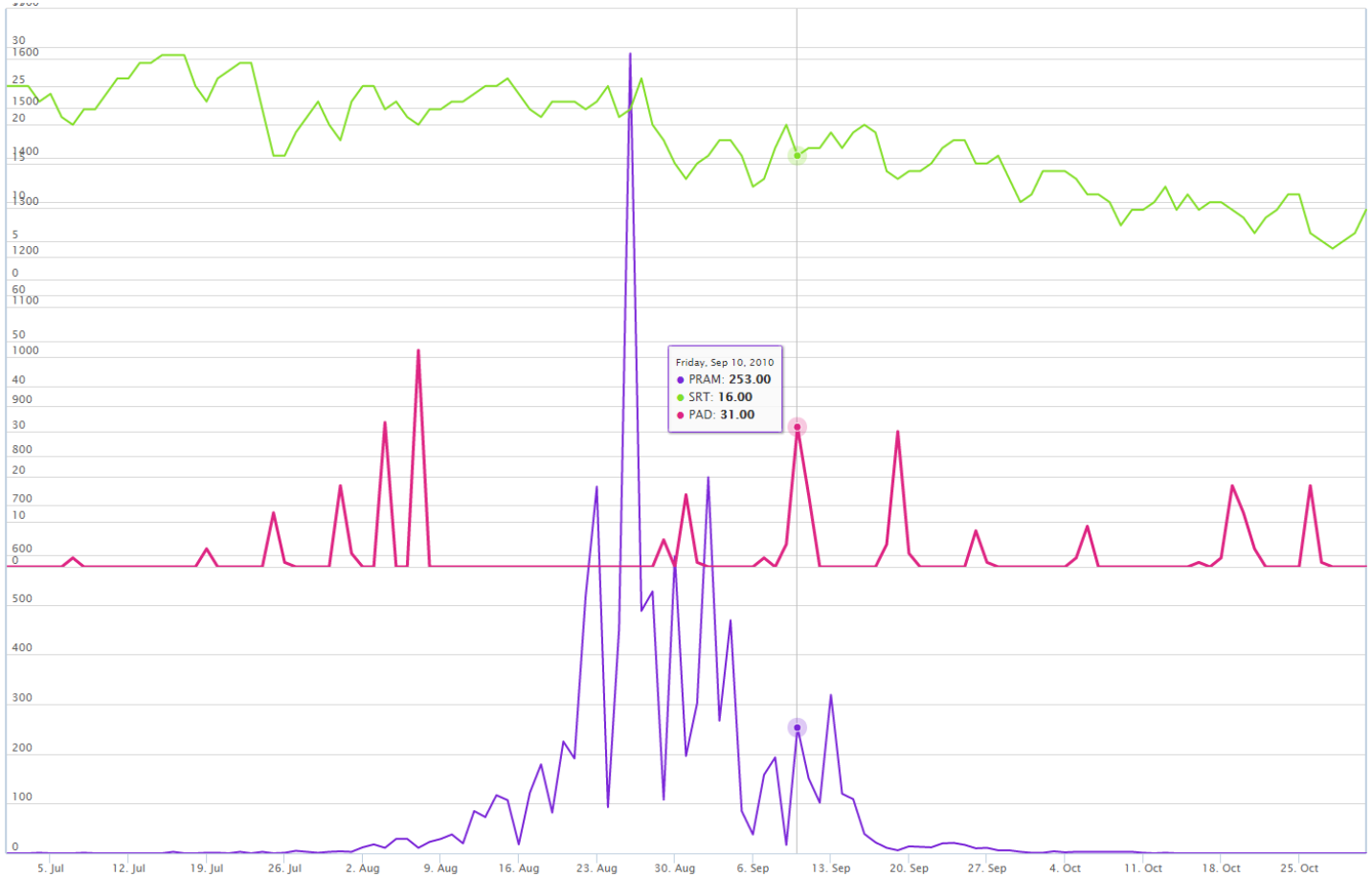


Figure: Podaci za koncentracije **ambrozije**, srednju dnevnu temepraturu i dnevnu količinu padalina u **Osijeku** (od 7. - 10. mjeseca), **2010. godina**

Vizualizacija podataka

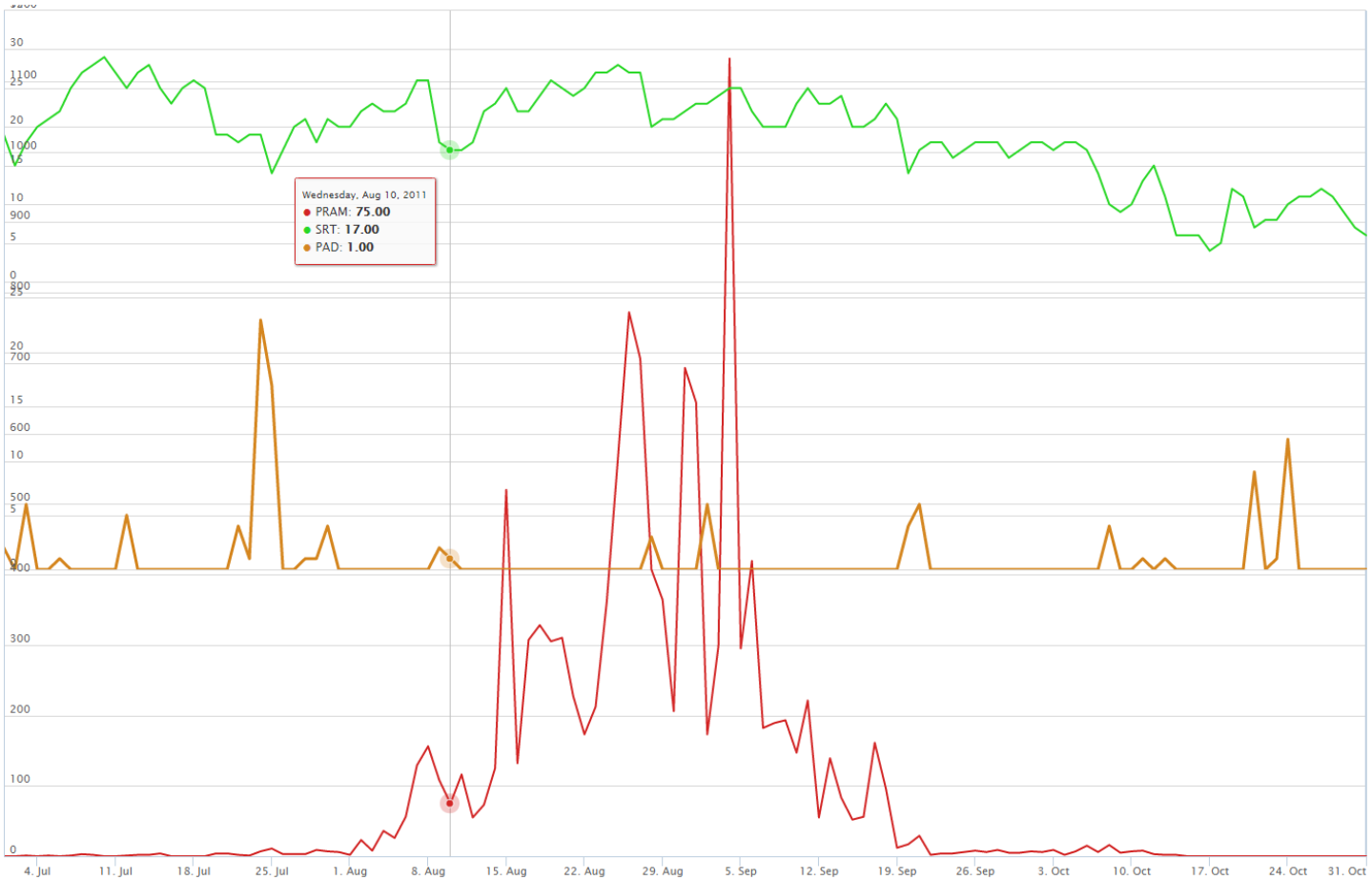


Figure: Podaci za koncentracije **ambrozije**, srednju dnevnu temepraturu i dnevnu količinu padalina u **Osijeku** (od 7. - 10. mjeseca), **2011. godina**

Vizualizacija podataka

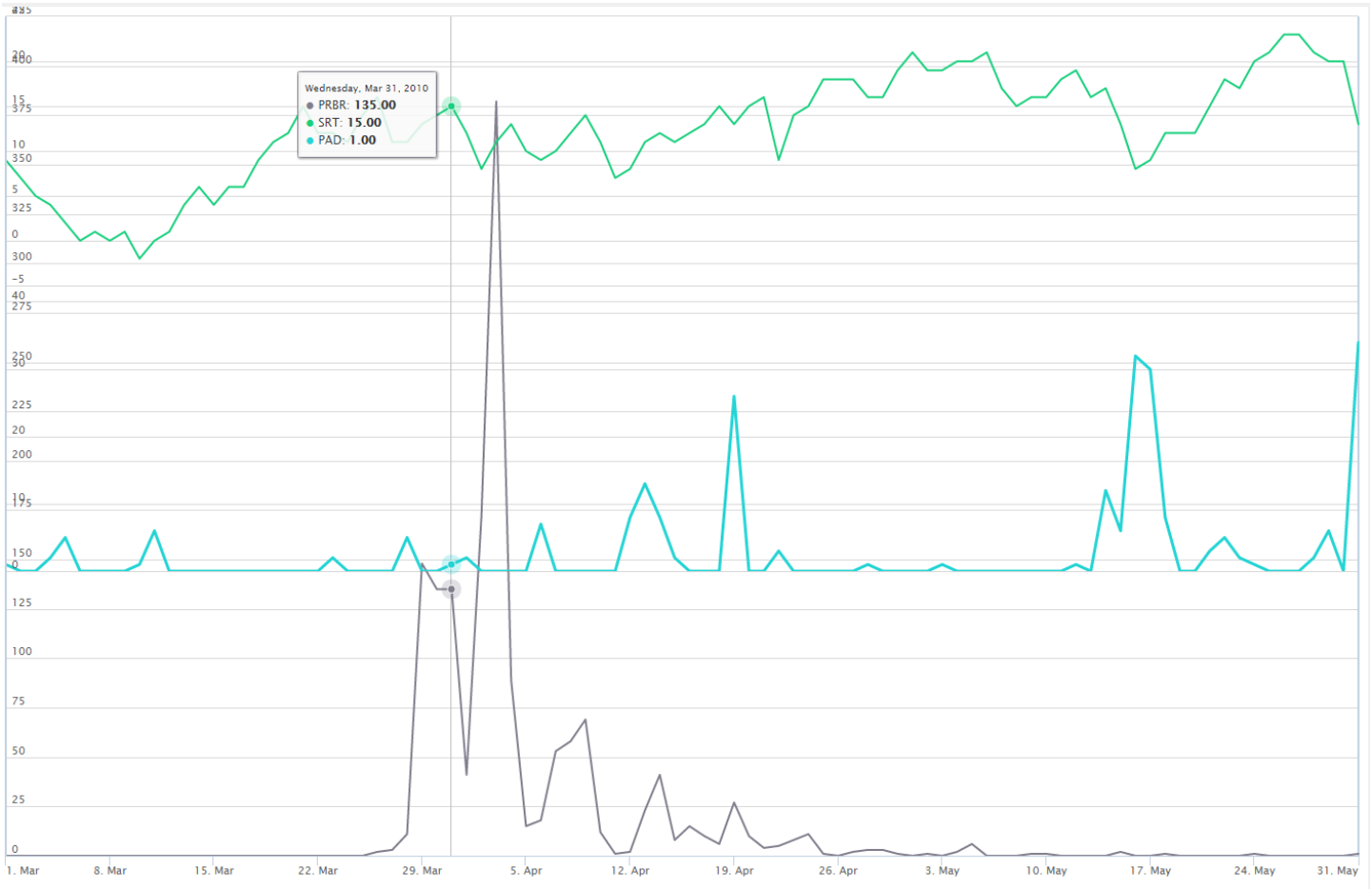


Figure: Podaci za koncentracije **breze**, srednju dnevnu temepraturu i dnevnu količinu padalina u **Osijeku** (od 3. - 5. mjeseca), **2010. godina**

Vizualizacija podataka

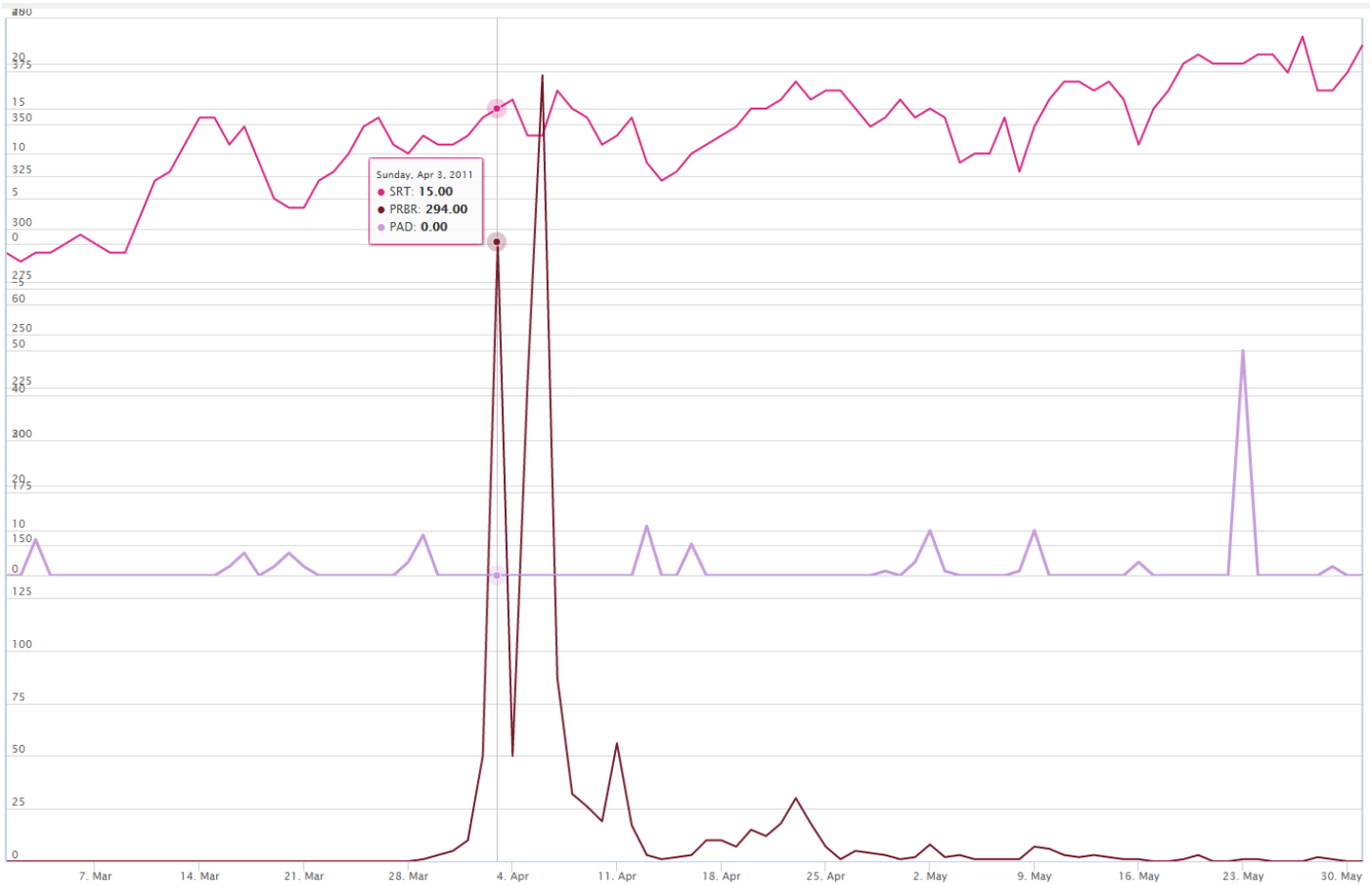


Figure: Podaci za koncentracije **breze**, srednju dnevnu temepraturu i dnevnu količinu padalina u **Osijeku** (od 3. - 5. mjeseca), **2011. godina**

Vizualizacija podataka

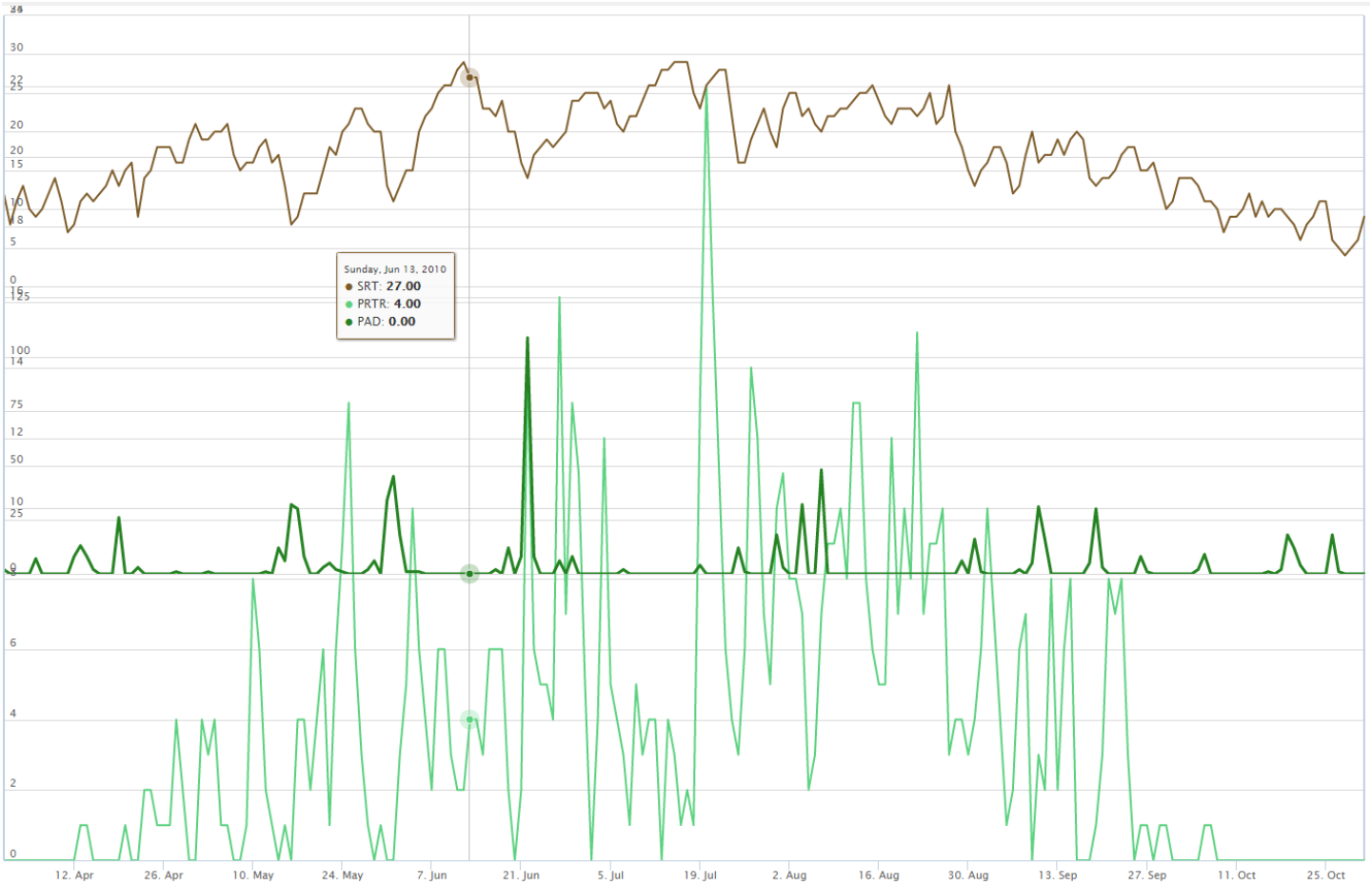


Figure: Podaci za koncentracije **trave**, srednju dnevnu temepraturu i dnevnu količinu padalina u **Osijeku** (od 4. - 10. mjeseca), **2010. godina**

Vizualizacija podataka

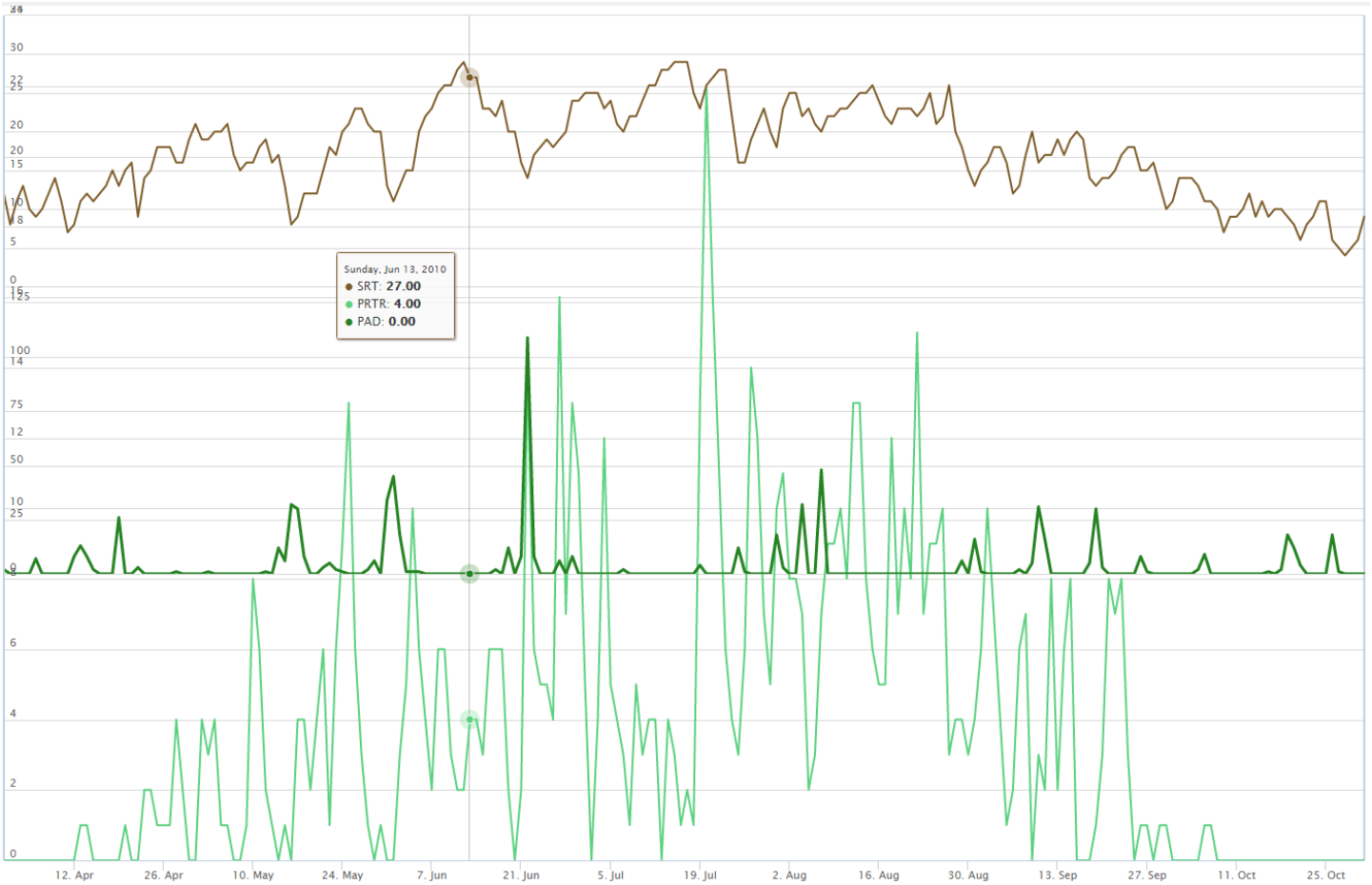


Figure: Podaci za koncentracije **trave**, srednju dnevnu temepraturu i dnevnu količinu padalina u Osijeku (od 4. - 10. mjeseca), **2011. godina**

Znanstveni radovi

- dugoročna prognoza peludi:
  - AVN - različiti alergeni[3]
- kratkoročna prognoza peludi:
  - AI tehnike - ambrozija[4]
  - regresijski modeli- breza[5]
  - Poissonov regresijski model - ambrozija[6]

[3] H. García-Mozo et al. "Statistical approach to the analysis of olive long-term pollen season trends in southern Spain". In: *Science of the Total Environment* 473-474 (2014), pp. 103–109.

[4] Zoltán Csépe et al. "Predicting daily ragweed pollen concentrations using Computational Intelligence techniques over two heavily polluted areas in Europe". In: *Science of the Total Environment* 476-477 (2014), pp. 542–552.

[5] Tomas R. Cotos-Yáñez, F. J. Rodríguez-Rajo, and M. V. Jato. "Short-term prediction of Betula airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models". In: *International Journal of Biometeorology* 48.4 (2004), pp. 179–185.

[6] P C Stark et al. "Using meterologic data to predict daily ragweed pollen levels". In: *Aerobiologia* 13 (1997), pp. 177–184.

Razmatranje podataka

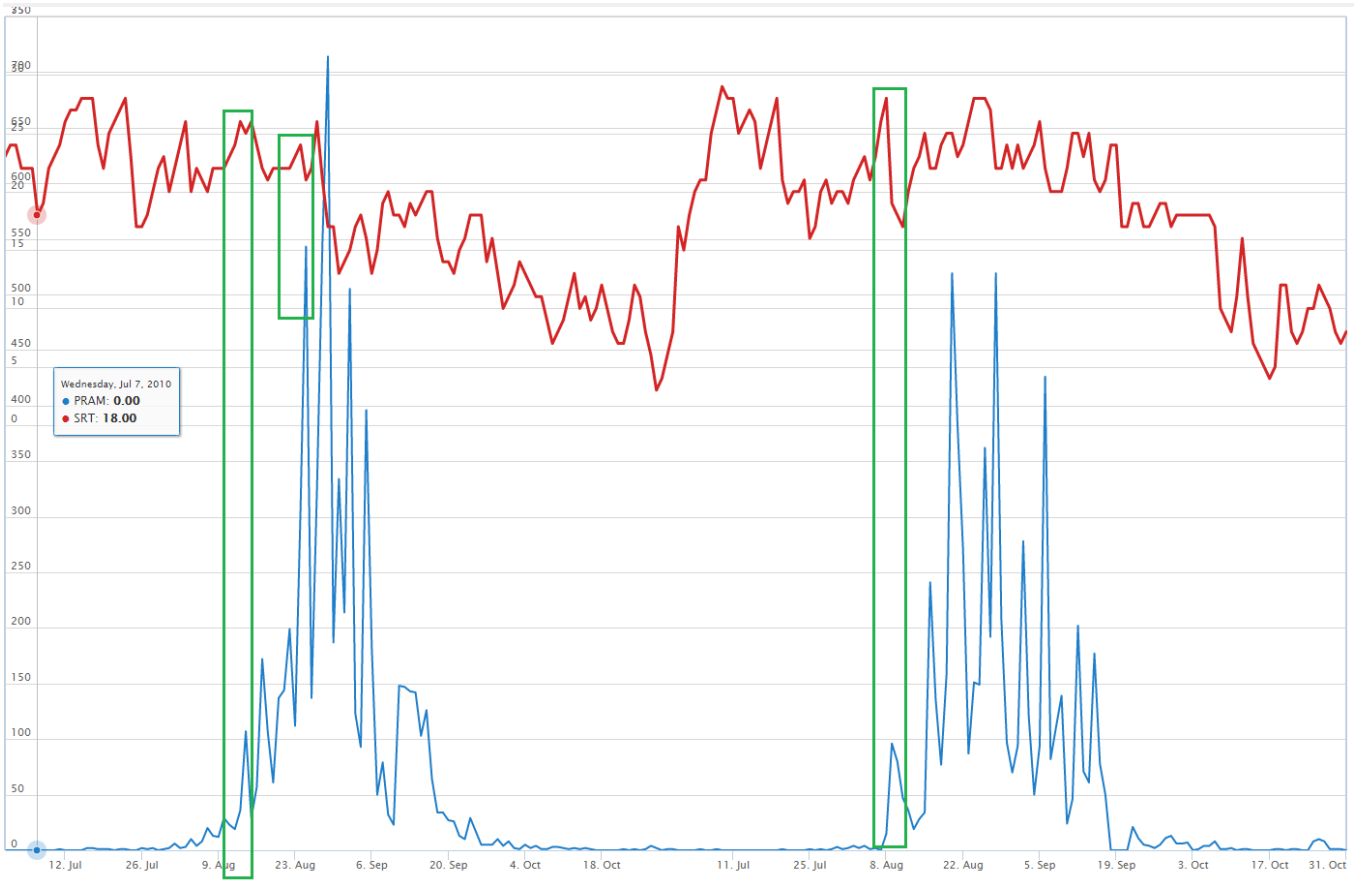
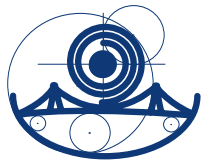


Figure: Uočavanje obrazaca na osnovu SRT i PRAM (Novi Sad, 2010. godina)



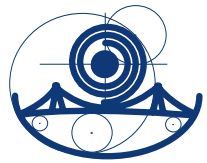
Model klasteriranja (k, l)-obrazaca

- standardni pristup za izradu kalendara:
  - za promatrani dan, predikcija koncentracije peludi je arit. sredina/median **koncentracija** na isti dan u svim prethodnim godinama
- Spieksma's pristup:
  - za promatrani dan, predikcija koncentracije peludi je arit. sredina/median 10**-dnevnih arit. sredina koncentracija** na isti dan u svim prethodnim godinama

Model klasteriranja (k, l)-obrazaca

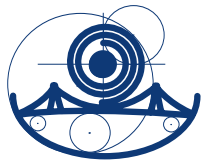
- napredni pristup[7]:
  - proširiti rezoluciju "prozora" za dnevnu koncentraciju u Spieksma pristupu (1 - 30 dana)
  - pre-procesuiranje početnog signala pomičnim arit. sredinama/medianima
  - za promatrani dan, predikcija koncentracije peludi je arit. sredina/median 10**-dnevnih arit. sredina koncentracija** na isti dan u svim prethodnim godinama
  - ukupno modela: MNMN, MNMD, MDMN, MDMD

[7] B. Šikoparija et al. "How to prepare a pollen calendar for forecasting daily pollen concentrations of Ambrosia, Betula and Poaceae?" In: *Aerobiologia* (Jan. 2018).



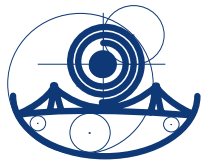
Model klasteriranja (k, l)-obrazaca

- skup podataka za treniranje: sastoji se od 13 godina (2000. - 2012.),
- **Zadatak:** potrebno je prognozirati koncentraciju za 1. kolovoza 2013.
- **Rješenje:** izračunati prosječnu koncentraciju peludi 1. kolovoza u svim prehodnim godinama u skupu za treniranje
- **Interpretacija:**
  - 13 povjesnih koncentracija (2000. - 2012.) pripadaju jednom klasteru
  - svaki dan u sezoni definira jedan klaster
- **Zaključak:** standardni/Spieksma/napredni pristup predviđa da su isti dani u prethodnim godinama **adekvatni** za prognozu.



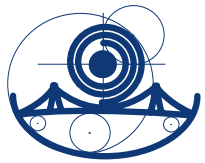
Model klasteriranja

- koristi meteorološke podatke kako bi pronašao adekvatnije klastere
- **Metoda:**
  - pronaći klastere sličnih dana s obzirom na meteorološke podatke i koncentracije peludi
  - za promatrani dan, pronaći najbliži klaster,
  - vjerujemo da taj klaster sadrži sve slične situacije u prethodnim godinama, **i to ne samo s obzirom na koncentracije peludi, nego i meteorološke uvjete**
  - unutar najbližeg klastera, prognoziramo koncentraciju kao arit. sredinu/median dana samo u tom klasteru



Model klasteriranja (k, l)-obrazaca

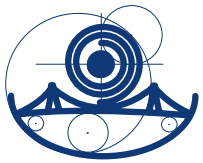
Model	sezonska varijabi-nost	sezonski kratkoročni pomaci	dnavna varijabil-nost	meteor. var.
Standardni	✓			
Spieksma	✓	✓		
Napredni	✓	✓	✓	
Klasteriranje	✓	✓	✓	✓



Model klasteriranja (k, l)-obrazaca

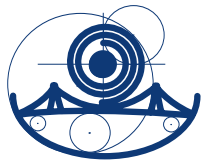
- skup podataka je podijeljen u 3 podskupa:
  - za treniranje
  - za kalibraciju
  - za evaluaciju
- pretpostavke:
  - svaka godina sadrži 366 dana
  - koncentracije, koje nedostaju, interpolirane su pomoću lin. interpolacijskog splinea[8]

[8] B. Šikoparija et al. "How to prepare a pollen calendar for forecasting daily pollen concentrations of Ambrosia, Betula and Poaceae?" In: *Aerobiologia* (Jan. 2018).



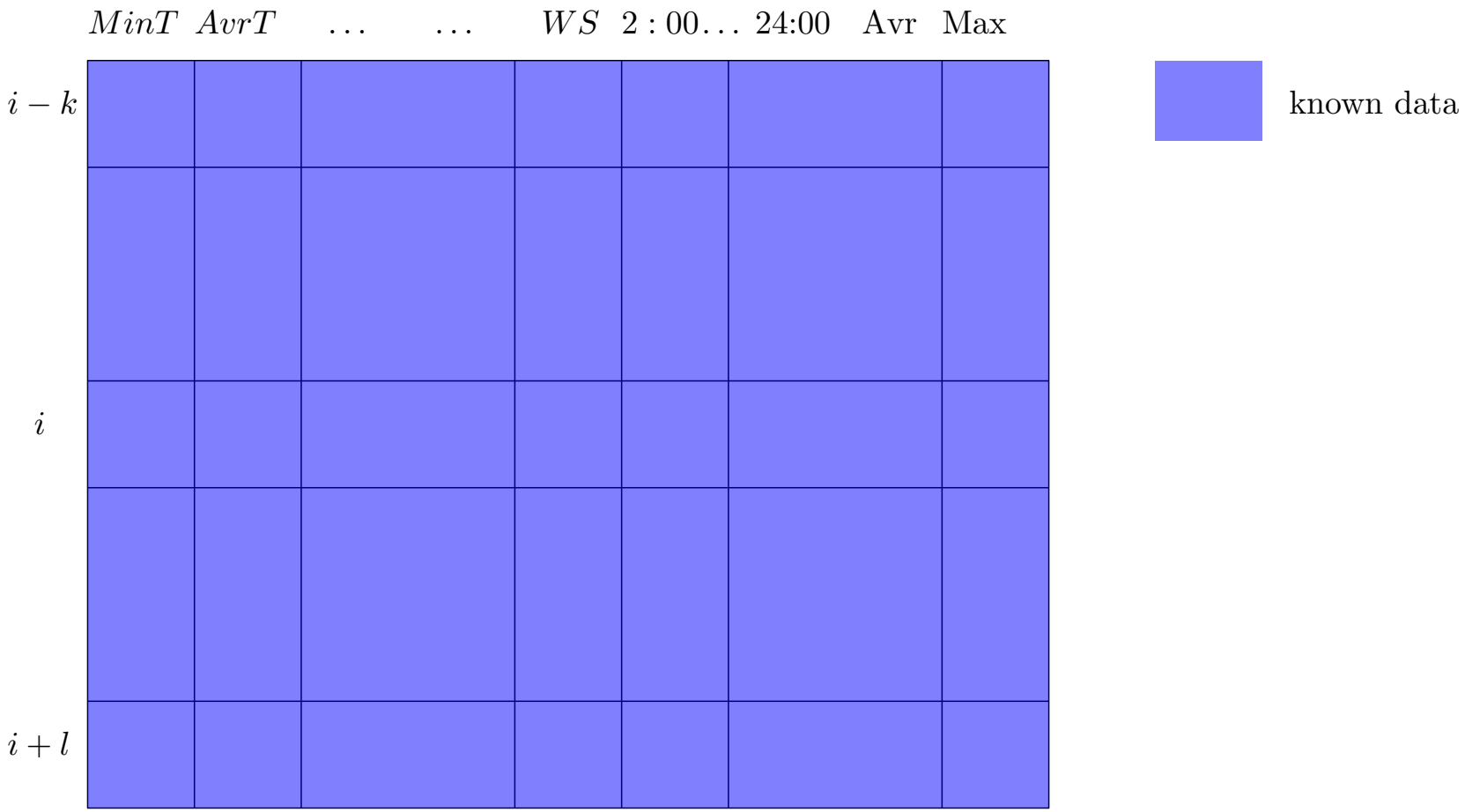
Model klasteriranja (k, l)-obrazaca

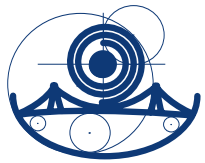
- **Ideja:** za opis i-tog dana koristimo ne samo podatke za taj dan, nego i podatke za k dana **prije** i l dana **poslije**
- podaci za svaki dan i, $i \in \{S_s, \ldots, S_e\}$, definiraju (k, l)-obrazac:
  - $S_s \in \{1, \ldots, 366\}$ - početak sezone (redni broj u godini)
  - $S_e \in \{1, \ldots, 366\}$ - kraj sezone
  - $k \in \mathbb{N}_0$ - broj dana prije i-tog dana
  - $l \in \mathbb{N}_0$ - broj dana poslije i-tog dana
- (k, l)-obrazac je vektor koji sadrži sve meteorološke podatke i koncentracije polena k dana prije i l dana poslije (ili prognozu za l dana poslije), dana i.



Model klasteriranja (k, l)-obrazaca

- i-ti (k, l)-obrazac dobivamo vektorizacijom podataka za i-ti dan zajedno s podacima za k dana prije i l dana poslije



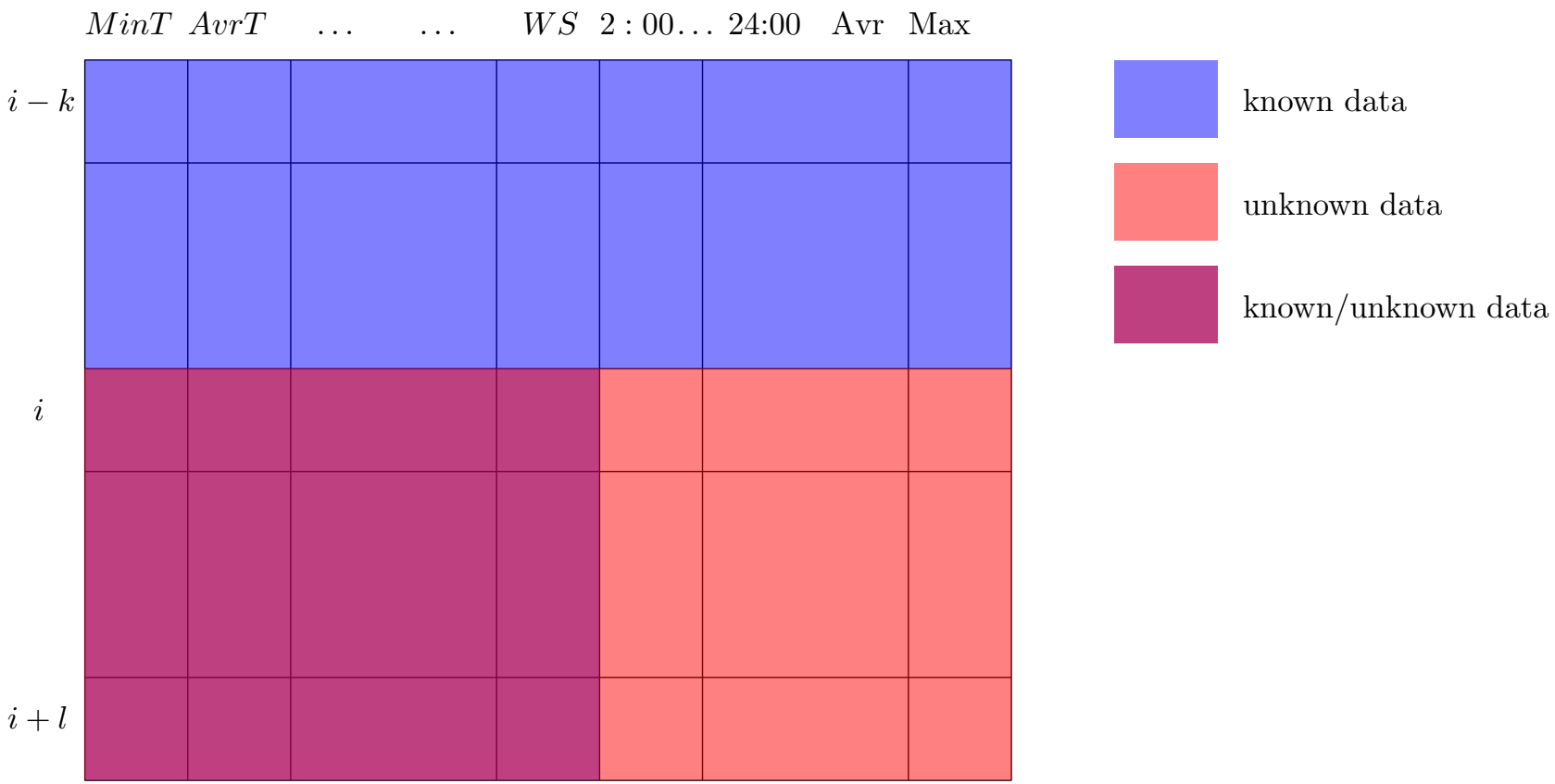


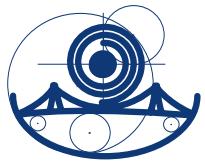
Model klasteriranja (k, l)-obrazaca

- pronaći K klastera u skupu od (k, l)-obrazaca na skupu podataka za treniranje
  - k-means metoda s replikacijama
- svaki klaster sadrži slične situacije u prethodnim godinama, **i to ne samo s obzirom na koncentracije peludi, nego i meteorološke uvjete**

Model klasteriranja (k, l)-obrazaca

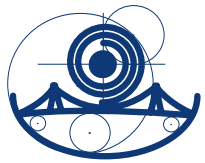
- **Zadatak:** prognozirati koncentraciju za i-ti dan
- za dani i-ti dan kreirati (k, l)-obrazac iz **poznatih** podataka





Model klasteriranja (k, l)-obrazaca

- pronaći najbliži centroid klastera (k, l)-obrascu za i-ti dan
- koristiti "neki" od procjenitelja za nepoznatu koncentraciju peludina samo na podaci u najbližem klasteru (average,median,max)



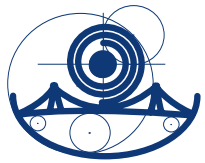
Model klasteriranja (k, l)-obrazaca

- **Zadatak:** za danu kombinaciju parametara modela k, l, K, (pod)skup od M meteoroloških varijabli izračunati $NRMSE(k, l, K, M)$ i Spearmanov koeficijent korelacije $\rho(k, l, K, M)$.
- želimo (k, l, K, M) s malim $NRMSE(k, l, K, M)$ i velikim $\rho(k, l, K, M)$.
- definirati mjeru kvalitete modela:

$$C(k, l, K, M) = \frac{1}{NRMSE(k, l, K, M)} + \rho(k, l, K, M)$$

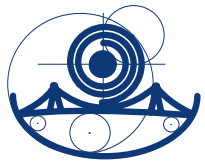
- pronaći (k^*, l^*, K^*, M^*) takav da je

$$(k^*, l^*, K^*, M^*) = \arg \max_{(k,l,K,M)} C(k, l, K, M)$$



Model klasteriranja (k, l)-obrazaca

- reduciranje domene parametara modela
- nepoznati parametri:
- $k \in \{1, 2, 3, 4, 5, 6, 7\}, l \in 0, 1, K \in \{2, \ldots, 300\}$
- razmatrane kombinacije meteoroloških varijabli:
  - prosječna dnevna temperatura (SRT) + prosj. kol. padalina (PAD)
  - prosječna dnevna temperatura (SRT) + vlažnost zraka (VLZ)
  - prosječna dnevna temperatura (SRT) + prosj. brzina vjetra (SBV)
- **skup za treniranje:** koncentracije peludi ambrozije u Novom Sadu, 2000 - 2011
- **skup za kalibraciju:** koncentracije peludi ambrozije u Novom Sadu, 2012 - 2014



Model klasteriranja (k, l)-obrazaca

k	l	M_1	M_2	$PolVar$	S_s	S_e	K	$NRMSE$	ρ	C
1	0	SRT	PAD	PRAM	183	274	4	0.0641	0.7880	16.3863
1	0	SRT	VLZ	PRAM	183	274	4	0.0641	0.7880	16.3863
1	0	SRT	SBV	PRAM	183	274	4	0.0641	0.7880	16.3863
2	0	SRT	VLZ	PRAM	183	274	3	0.0684	0.7232	15.3465
1	0	SRT	PAD	PRAM	183	274	3	0.0687	0.7078	15.2615
1	0	SRT	SBV	PRAM	183	274	3	0.0687	0.7077	15.2615
1	0	SRT	VLZ	PRAM	183	274	3	0.0687	0.7077	15.2614

Model klasteriranja (k, l)-obrazaca

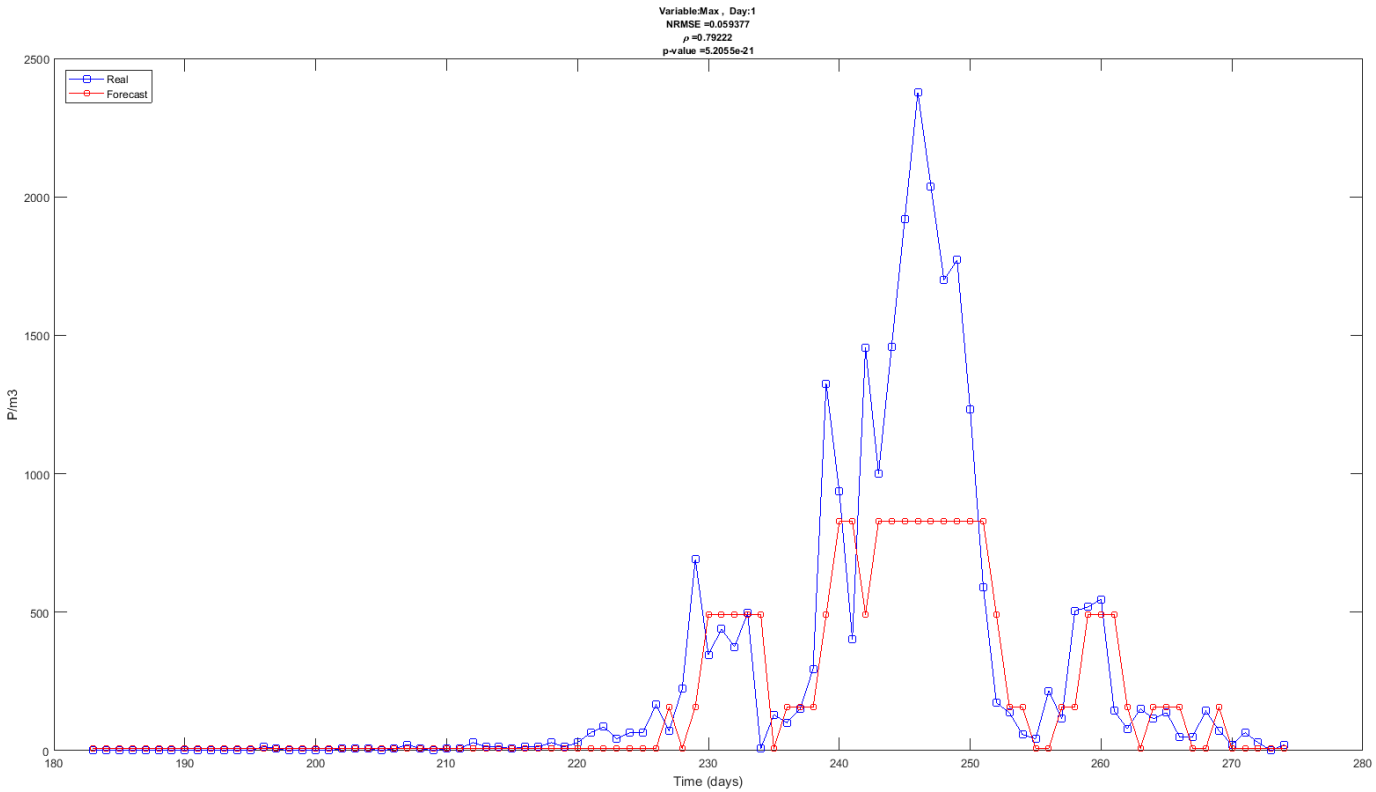


Figure: Observed vs. predicted maximum ragweed pollen concentrations in 2015 (Novi Sad)

Model klasteriranja (k, l)-obrazaca

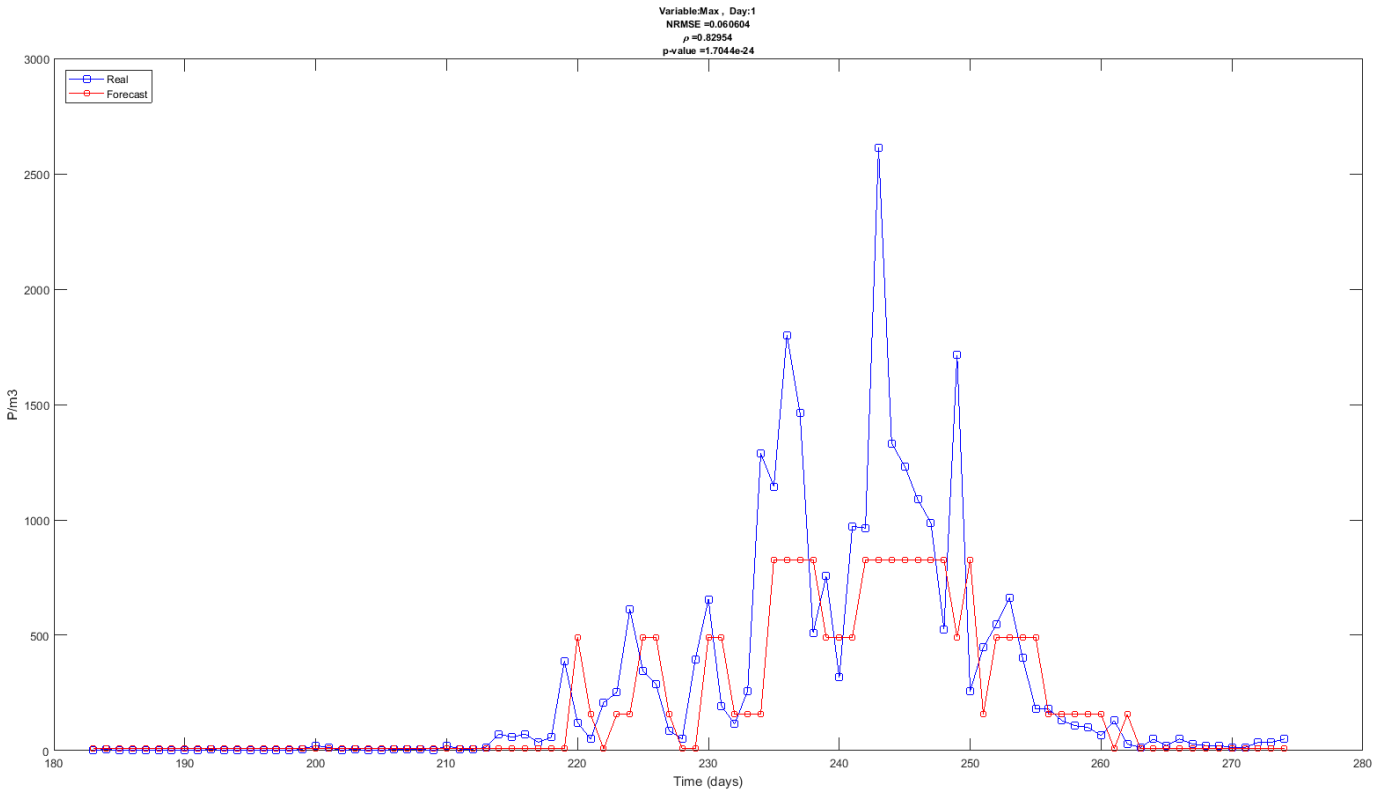
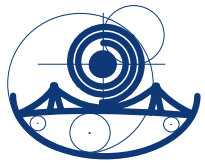


Figure: Observed vs. predicted maximum ragweed pollen concentrations in 2016 (Novi Sad)



Model klasteriranja (k, l)-obrazaca

k	l	M_1	M_2	$PolVar$	S_s	S_e	K	$NRMSE$	ρ	C
2	1	SRT	VLZ	PRAM	183	274	3	0.0694	0.7595	15.1596
2	1	SRT	VLZ	PRAM	183	274	5	0.0698	0.7648	15.0961
2	1	SRT	PAD	PRAM	183	274	3	0.0697	0.7479	15.0960
2	1	SRT	SBV	PRAM	183	274	3	0.0697	0.7479	15.0960

Model klasteriranja (k, l)-obrazaca

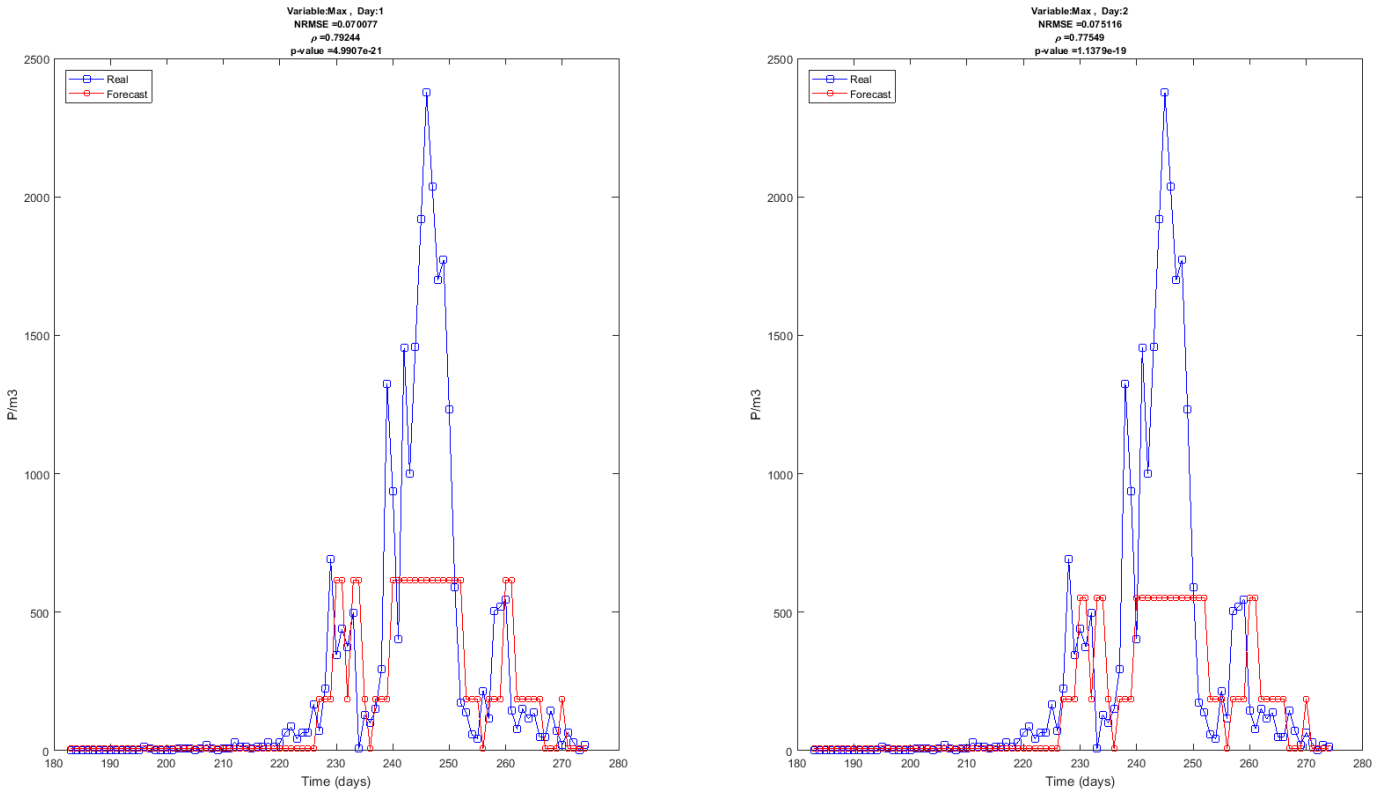


Figure: Izmjerene i prognozirane prosječne koncentracije ambrozije u 2015 (Novi Sad)

Model klasteriranja (k, l)-obrazaca

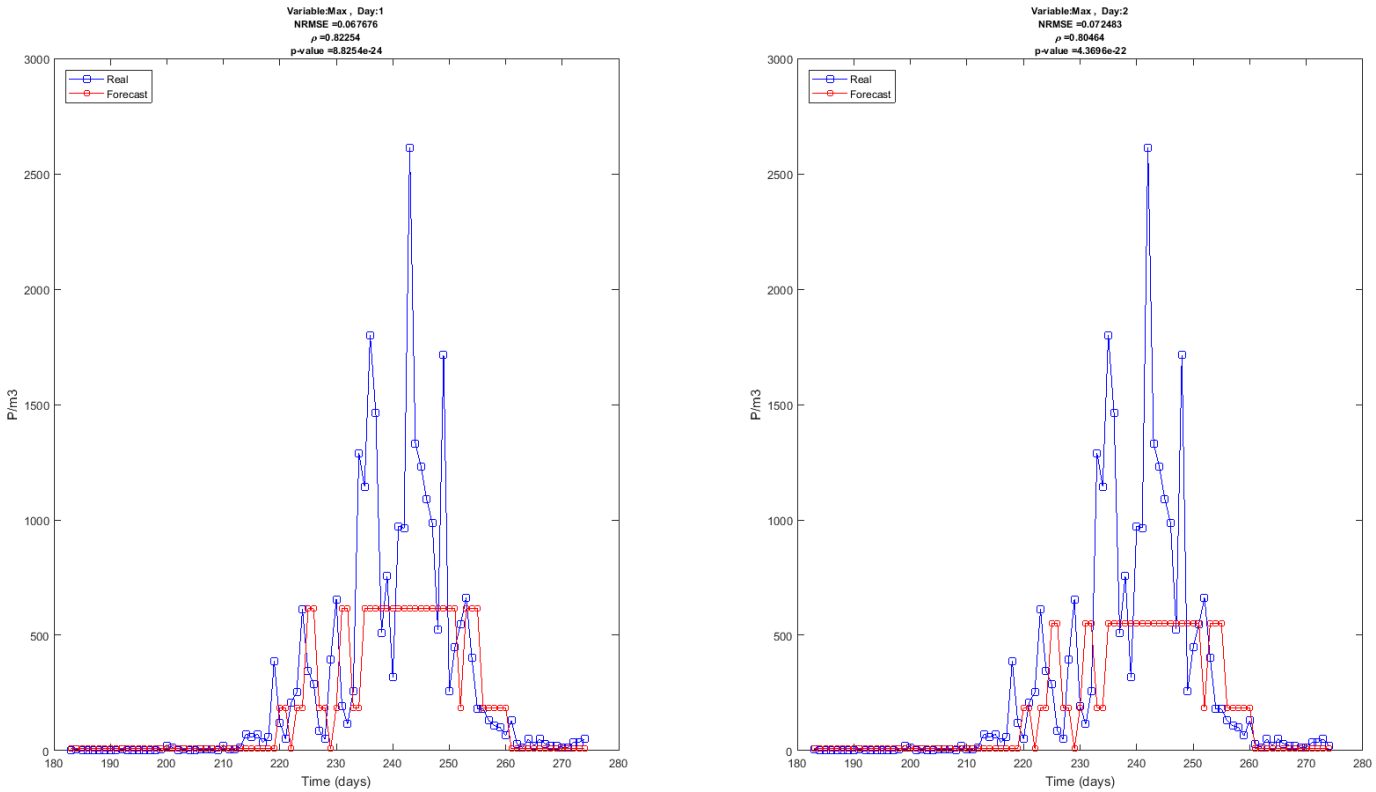
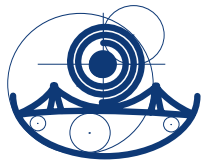


Figure: Izmjerene i prognozirane prosječne koncentracije ambrozije u 2016 (Novi Sad)



Model klasteriranja (k, l)-obrazaca

k	l	M_1	M_2	$PolVar$	S_s	S_e	K	$NRMSE$	ρ	C
2	2	SRT	PAD	22	183	274	3	0.0704	0.7318	14.9400
2	2	SRT	SBV	22	183	274	3	0.0704	0.7318	14.9400
6	2	SRT	SBV	22	183	274	13	0.0714	0.8039	14.8054
3	2	SRT	SBV	22	183	274	3	0.0736	0.7312	14.3207
3	2	SRT	PAD	22	183	274	3	0.0736	0.7323	14.3193
3	2	SRT	VLZ	22	183	274	3	0.0737	0.7266	14.2862

Model klasteriranja (k, l)-obrazaca

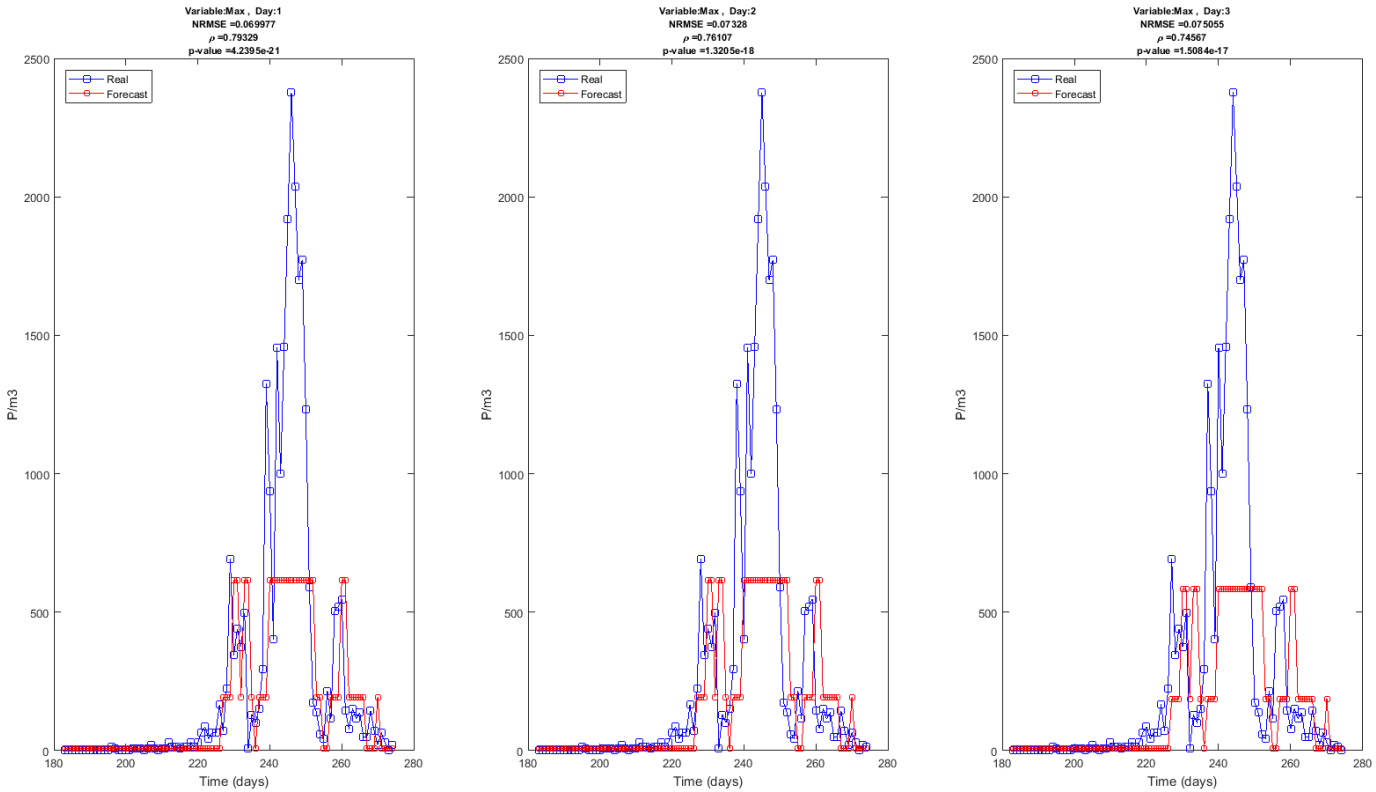


Figure: Izmjerene i prognozirane prosječne koncentracije ambrozije u 2015 (Novi Sad)

Model klasteriranja (k, l)-obrazaca

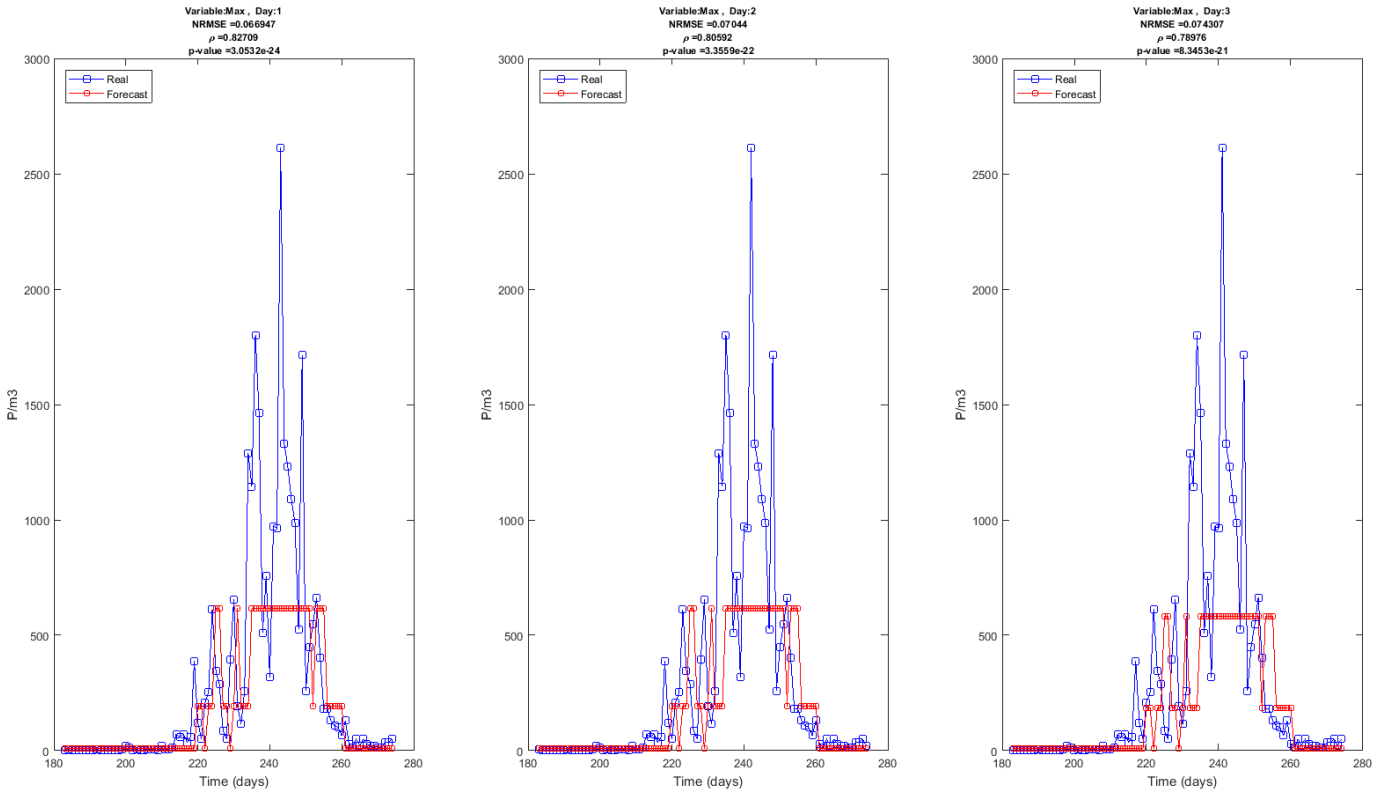
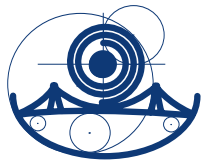


Figure: Izmjerene i prognozirane prosječne koncentracije ambrozije u 2016 (Novi Sad)



Model klasteriranja (k, l)-obrazaca

k	l	M_1	M_2	$PolVar$	S_s	S_e	K	$NRMSE$	ρ	C
2	3	SRT	PAD	PRAM	183	274	3	0.0716	0.7235	14.6960
2	3	SRT	SBV	PRAM	183	274	3	0.0716	0.7235	14.6960
1	3	SRT	VLZ	PRAM	183	274	3	0.0745	0.7040	14.1393
1	3	SRT	PAD	PRAM	183	274	3	0.0745	0.7040	14.1331
1	3	SRT	SBV	PRAM	183	274	3	0.0745	0.7028	14.1326
3	3	SRT	VLZ	PRAM	183	274	3	0.0749	0.7161	14.0737
3	3	SRT	PAD	PRAM	183	274	3	0.0749	0.7161	14.0723
3	3	SRT	SBV	PRAM	183	274	3	0.0749	0.7161	14.0723
4	3	SRT	PAD	PRAM	183	274	3	0.0751	0.7211	14.0372
4	3	SRT	VLZ	PRAM	183	274	3	0.0751	0.7211	14.0372
4	3	SRT	SBV	PRAM	183	274	3	0.0751	0.7211	14.0372

Model klasteriranja (k, l)-obrazaca

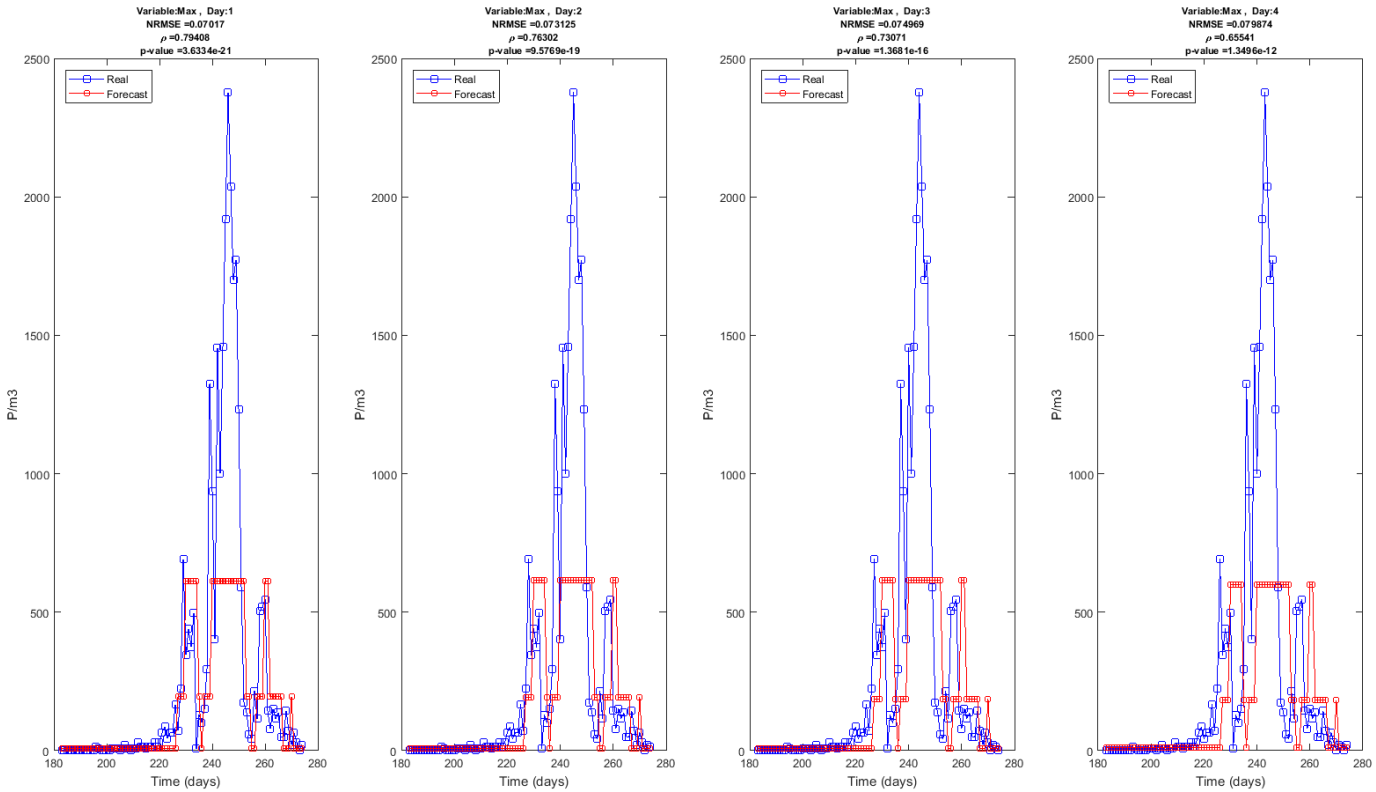


Figure: Izmjerene i prognozirane prosječne koncentracije ambrozije u 2015 (Novi Sad)

Model klasteriranja (k, l)-obrazaca

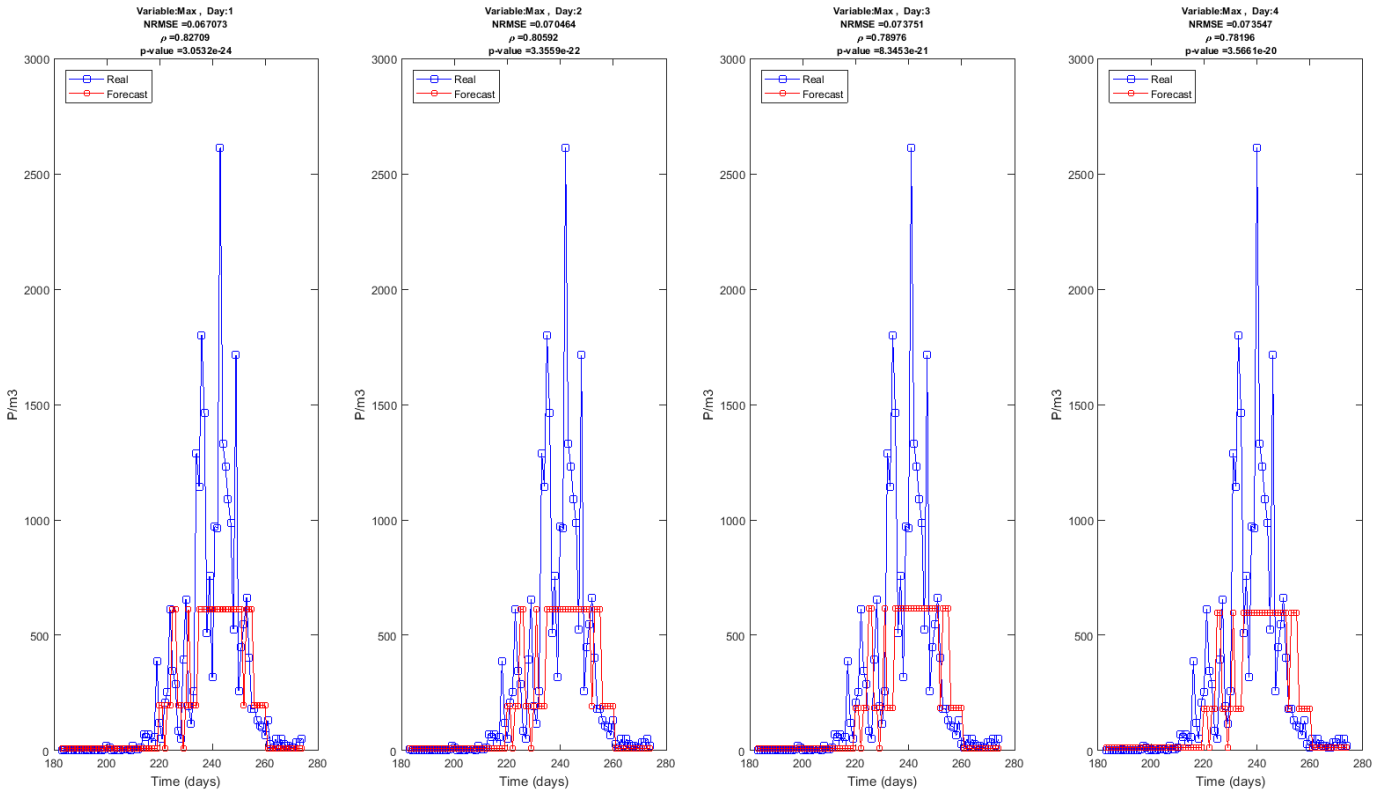


Figure: Izmjerene i prognozirane prosječne koncentracije ambrozije u 2016 (Novi Sad)

Model klasteriranja (k, l)-obrazaca

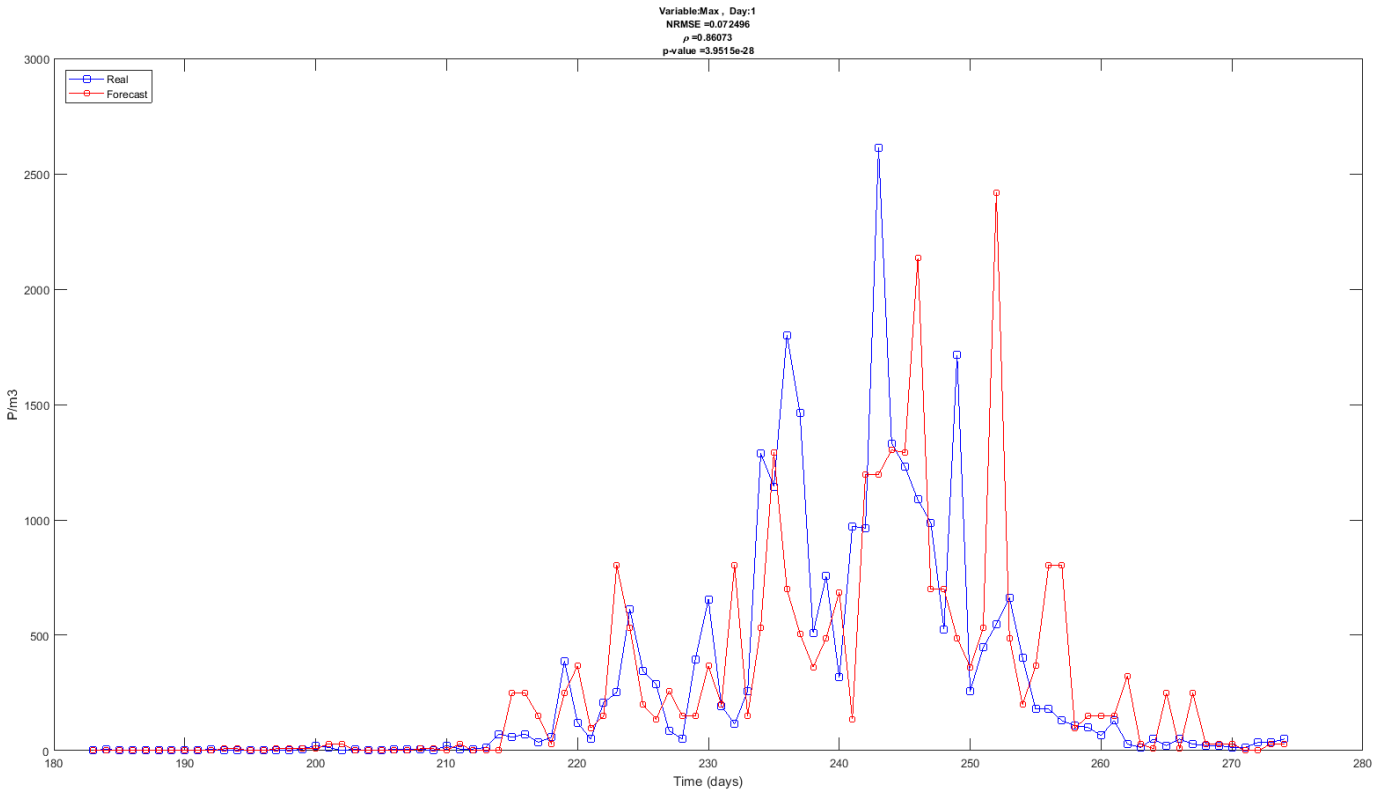


Figure: Izmjerene i prognozirane prosječne koncentracije ambrozije u 2016 $K = 30$ (Novi Sad)

Model klasteriranja (k, l)-obrazaca

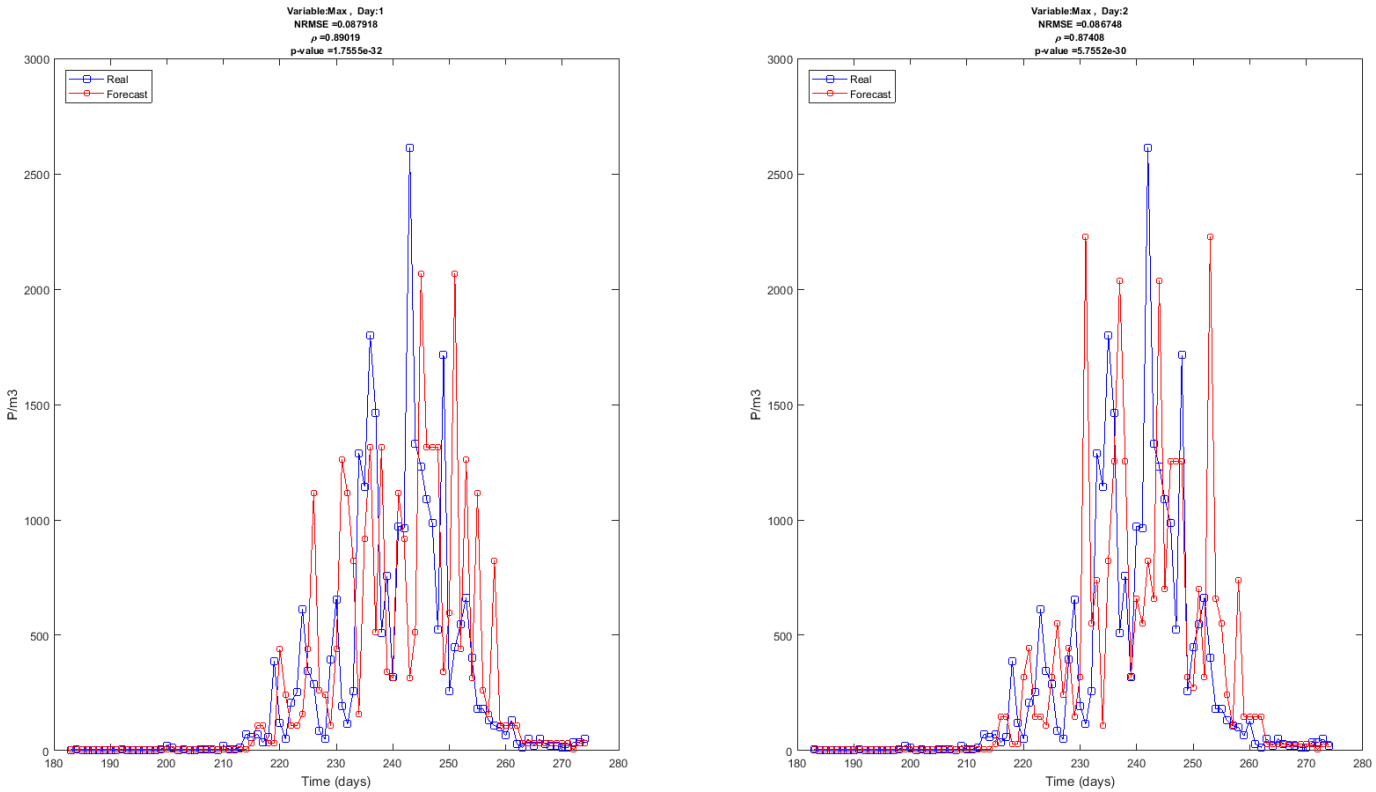
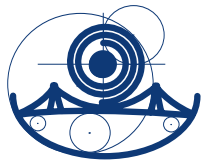


Figure: Izmjerene i prognozirane prosječne koncentracije ambrozije u 2016 $K = 30$ (Novi Sad)



Zaključak

- model pokazuje kašnjenja "reda veličine" l
- koncentracija danas utječe najviše na prognozu koncentracije za sutra
- trenutni model u razvoju:
  - uvođenje disjunktnih (k, l)-obrazaca
  - svaki dan svrstati u jedan od klastera disjunktnih obrazaca