Correlating local palynological record and weather conditions for an individualized pollen alarm

Miloš Marjanović*	University of Belgrade, Serbia
Mirjana Mitrović Josipović	Environmental Protection Agency, Serbia
Bojana Božanić	Pharm-Olam, Serbia
Vít Pászto	Palacký University Olomouc, Czech Republic
Lukaš Marek	University of Canterbury, New Zealand

*milos.marjanovic@rgf.bg.ac.rs

GeoMLA conference, Belgrade, Serbia, June 21-24, 2016

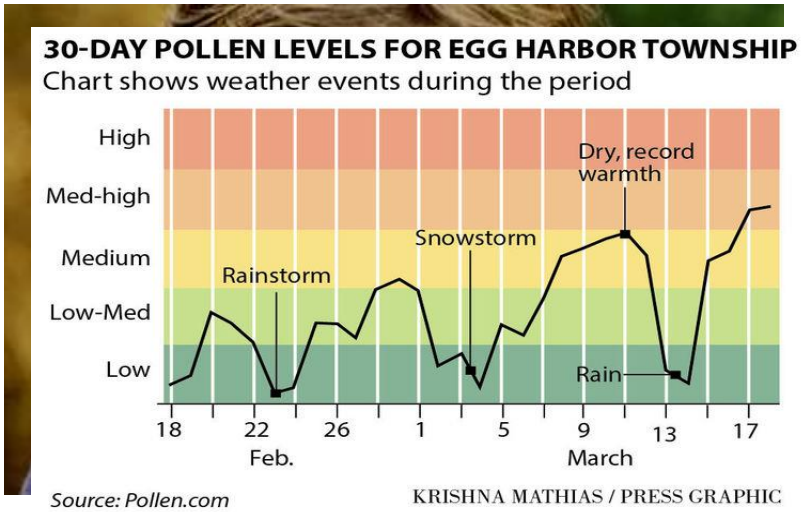
Outline

- Introduction
- Case study
- Data
- Methods
- Results and discussion
- Conclusion

Introduction

- Motivation
  - Growing number of allergies to pollen in Serbia (from 5.5% in 2006 to 8.8% in 2013) and worldwide, particularly in respect to specific allergens such as ambrosia (Höflich et al. 2016, Boričić, et al. 2014)
  - Possible relation to changing climate and shifted bio-ecological conditions
  - Multidisciplinary approach, fusing palynological measurements, with meteo data and personal symptom records
  - Personal condition

- Objective
  - Personalized pollen alarm, which will suggest which allergens are critical and in which weather conditions

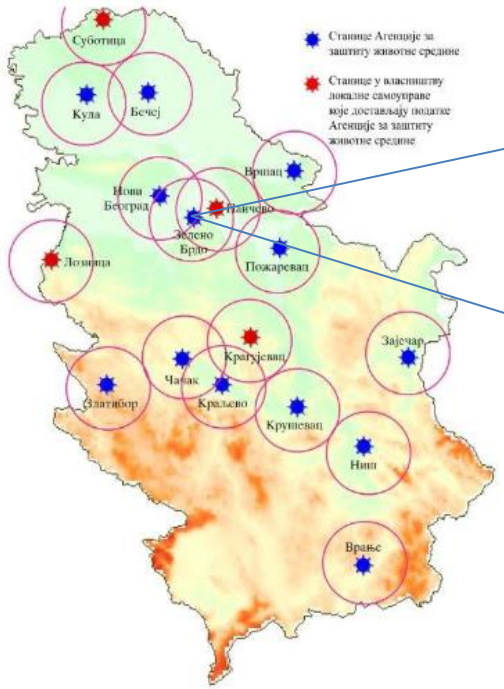


Introduction

- State of the art
  - Many general overviews and studies of the pollen distribution and phenological features, mostly in medical context
  - Several attempts to model and correlate pollen concentrations with meteo conditions using various tools
  - Few attempts to look into the symptoms and correlate them with weather conditions and pollen concentrations

Case study

- Case study area, Belgrade City, Serbia
- Period of observation 2012-14 (data limitations)



Select a location: Belgrade (Zvezdara) - Green Hill

Location: Beograd (Zemun) - Green Hill

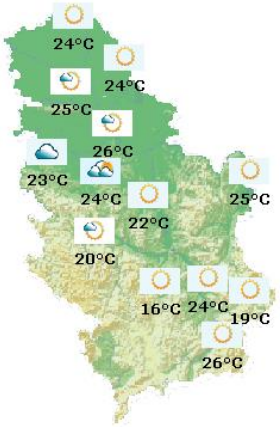
	Deut. 13-06	Uto. 14-06	Wed. 15-06	Thu. 16-06	Pet. 17-06	Sub. 18-06	Ned. 19-06	The tendency in the next week
ambrosia								—
Cinderella / styrene								—
Tisza / Čempresi								▼
Pines / Jela								—
plantain								—
Trave								▲
sorrel								—
lime								—
nettle								▲

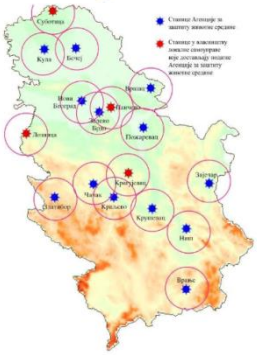
Legend:

allergenicity	concentration	trend
a big	high	Increasing ▲
median	median	No change —
Small	string	on the wane ▼

Data

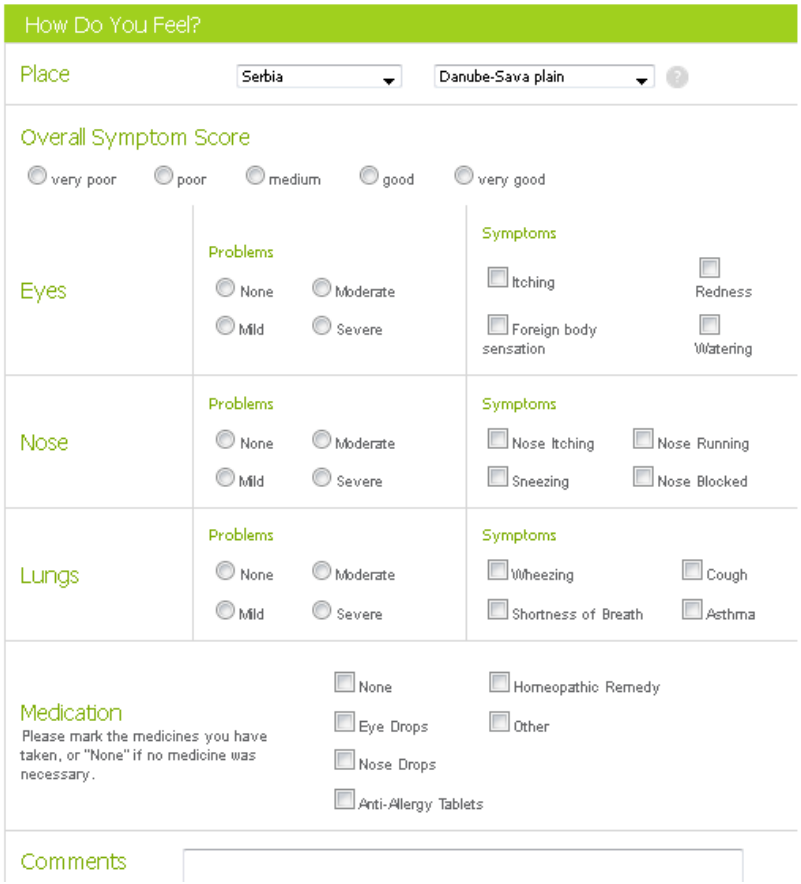
- Meteo data
  - collected from publically available archive for 2012-14 from www.hidmet.gov.rs
  - included predictors: averaged air pressure, temperature, relative humidity, wind speed, insolation, overcast and precipitation
  - these predictors are available in short-term (5-day) weather forecasts at www.hidmet.gov.rs
- Palynological data
  - pollen concentrations (measured per m^3) of 20 plant groups in Belgrade area from measuring station Zeleno Brdo for the same interval (2012-14)
  - final set included concentrations of: *Ambrosia*, *Betulaceae*, *Cupressacea*, *Moraceae*, *Poaceae* and *Urticaceae,* groups that were correlated with patient symptoms
  - available for preview at www.sepa.gov.rs
- Symptoms data
  - pollen symptoms of one volunteer from Belgrade over the same interval (2012-14), from www.pollendiary.com

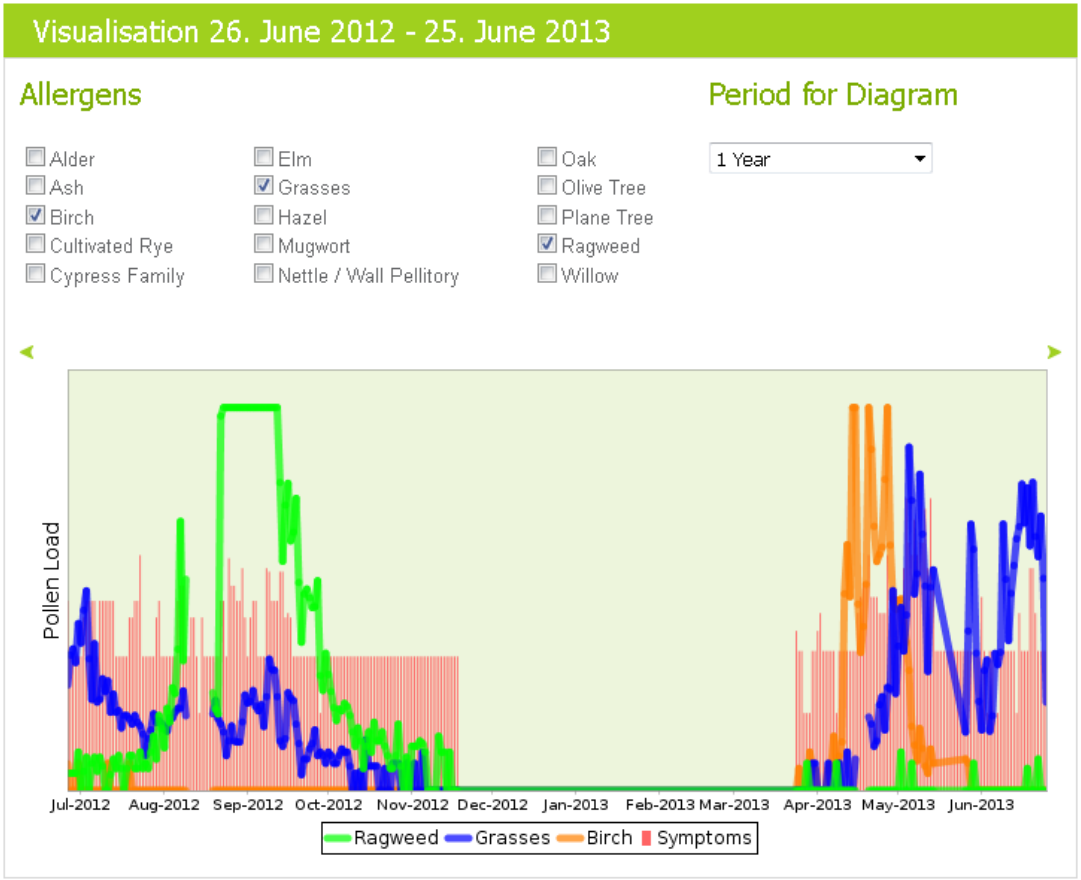




Data

- Pollendiary entry and overview





Methods

- Support Vector Machines Regression (step by step)
  - Error (loss) function controlled by ε: $|y_i-f(x_i)|_\varepsilon$
  - Regression (risk) function: $R(f)=0.5||w||2+c\Sigma(|y_i-f(x_i)|_\varepsilon)$
  - Minimize the regression function (limit the margin 2ε):

$$\min(0.5||w||2+c\Sigma(\xi_i+\xi_i^*))$$

  - Maximize the margin inside which the function exists:

$$\max(-0.5\Sigma(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)(x_i\cdot x_j)-\varepsilon\Sigma(\alpha_i^*+\alpha_i)+\Sigma y_i(\alpha_j^*-\alpha_j)$$

  - Obtain linear regression function: $f(x)=(\alpha_j^*-\alpha_j)(x_i-x_j)+b$
  - Apply kernel to transform the space in higher dimension:

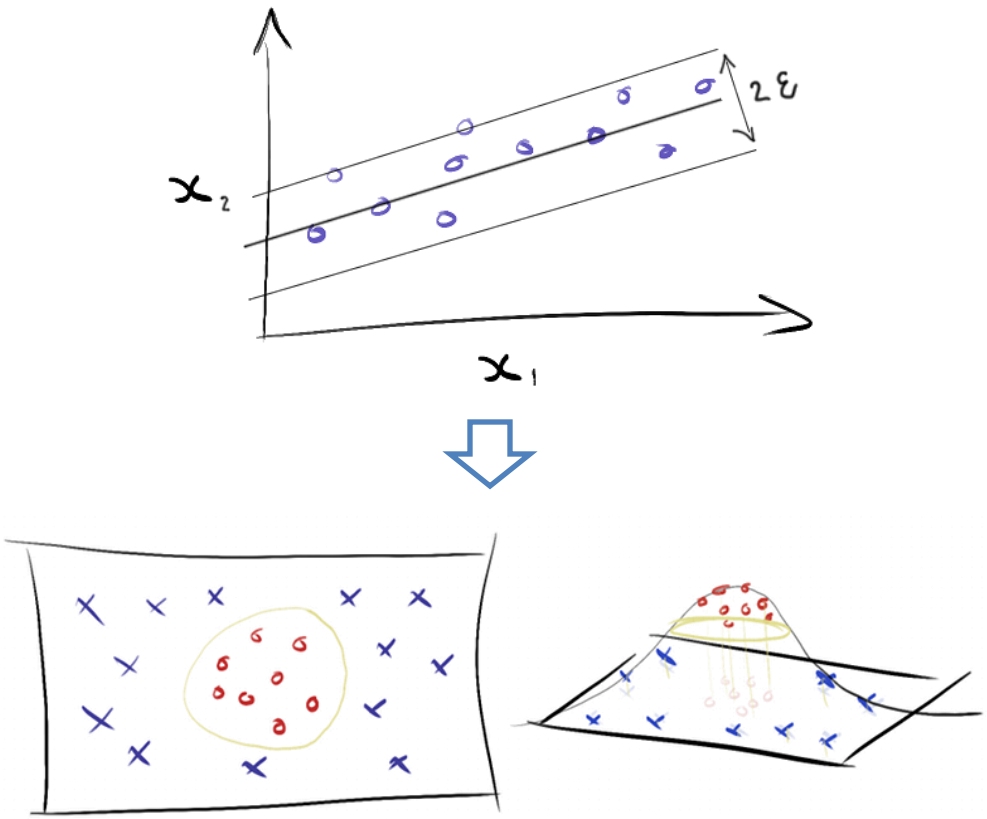
$$K(x_i\cdot x_j)=\phi(x_i)\cdot\phi(x_j)=\exp(-\gamma\cdot||x_i-x_j||2)$$

  - Obtain final regression function:

$$f(x)=(\alpha_j^*-\alpha_j)\exp(-\gamma\cdot||x_i-x_j||2)+b$$

Methods

- SVM regression is basically a linear regression with optimal margins using the kernel trick to solve non-linearity



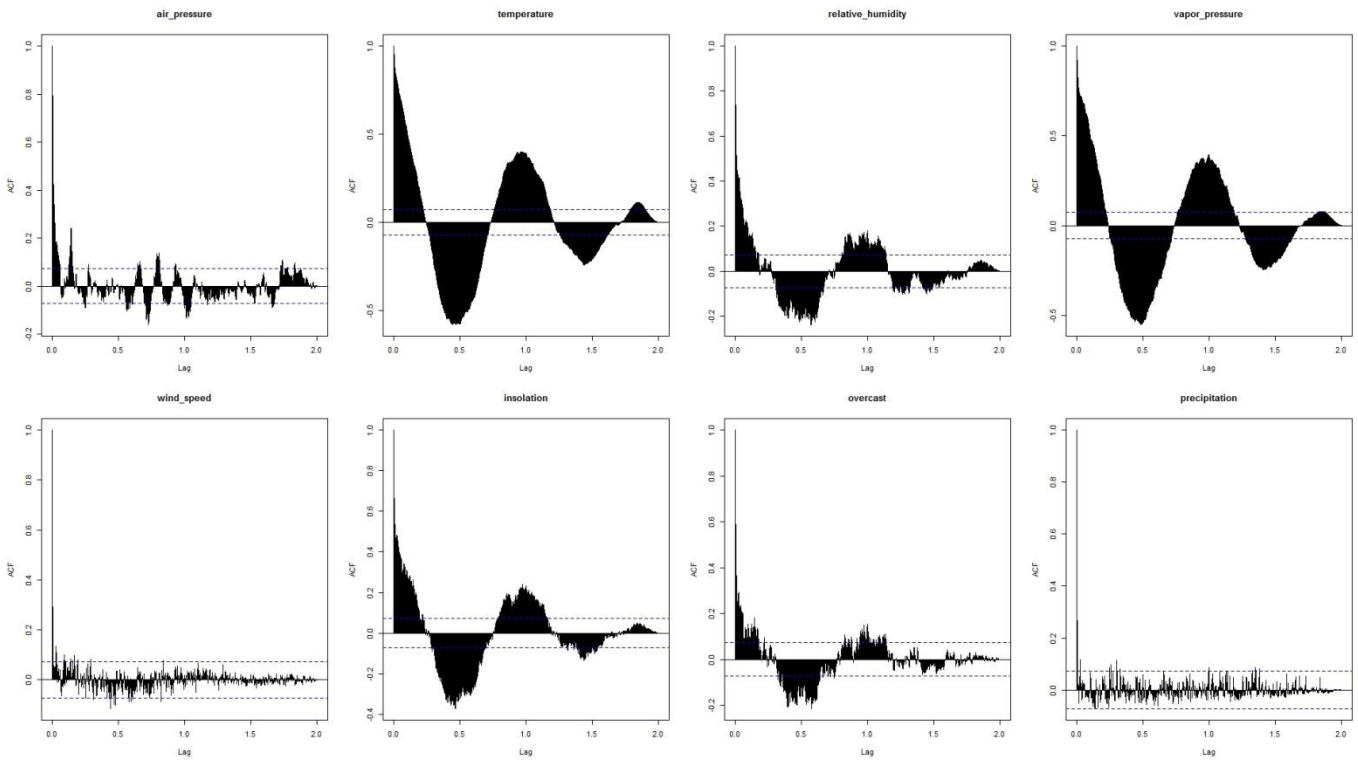
Methods

- Modeling concept
  - **Test1**
    1. autocorrelation as a predictor
       - included
       - excluded
    2. learning on 2012 data and validating on 2013 (50-50% split)
    3. predicting pollen concentrations of 6 target allergens
    4. predicting symptoms
       - step-wise (from 3)
       - directly from 2
  - **Test2**
    1. autocorrelation as a predictor
       - included
       - excluded
    2. learning on 2012-13 data and validating on 2013 (66-33% split)
    3. predicting pollen concentrations of 6 target allergens
    4. predicting symptoms
       - step-wise (from 3)
       - directly from 2

Results and discussion

- Autocorrelation and correlation

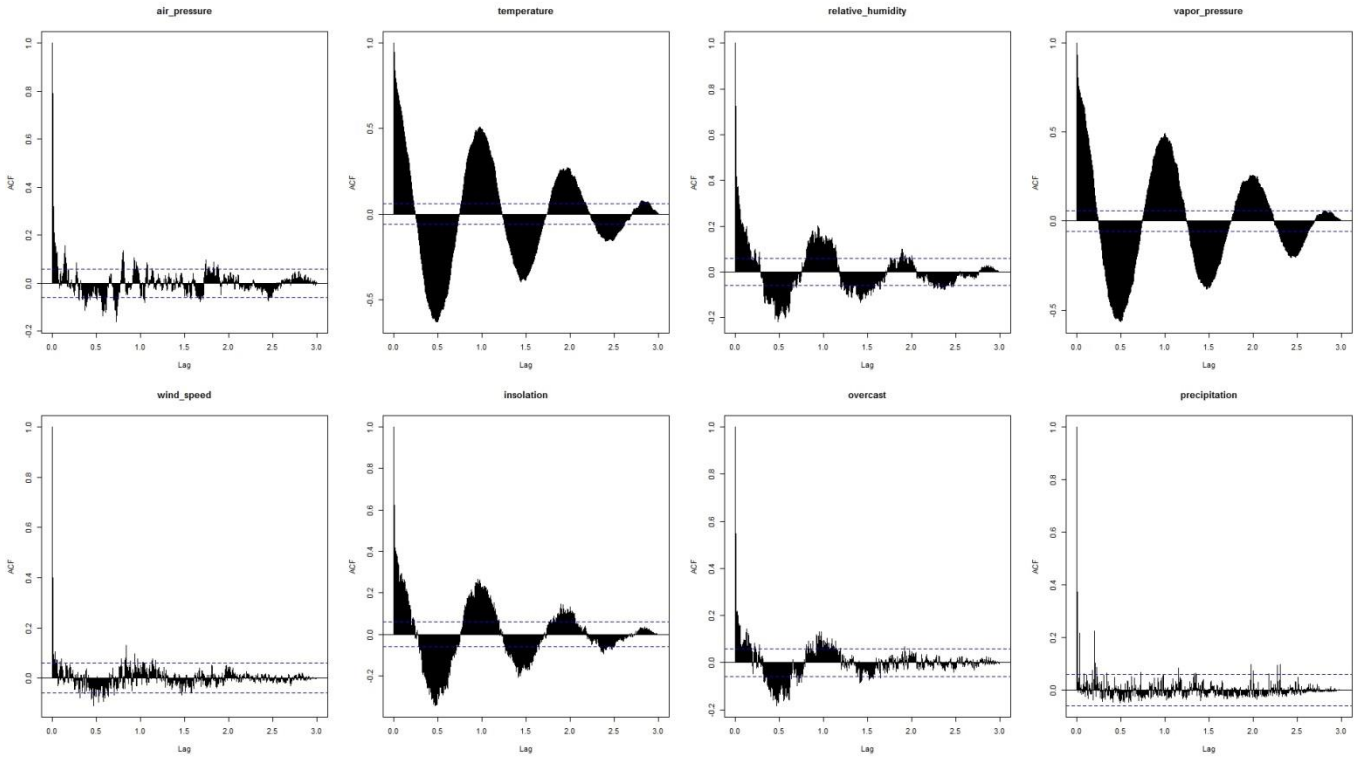
Test1



Results and discussion

- Autocorrelation and correlation

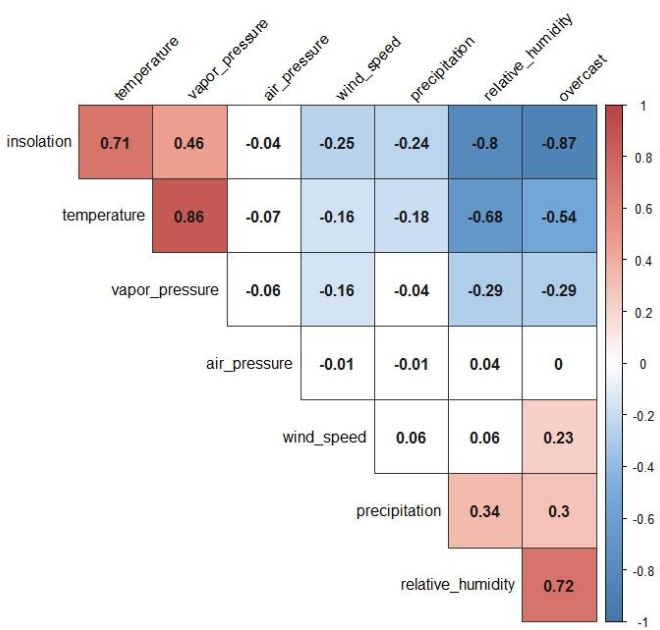
Test2

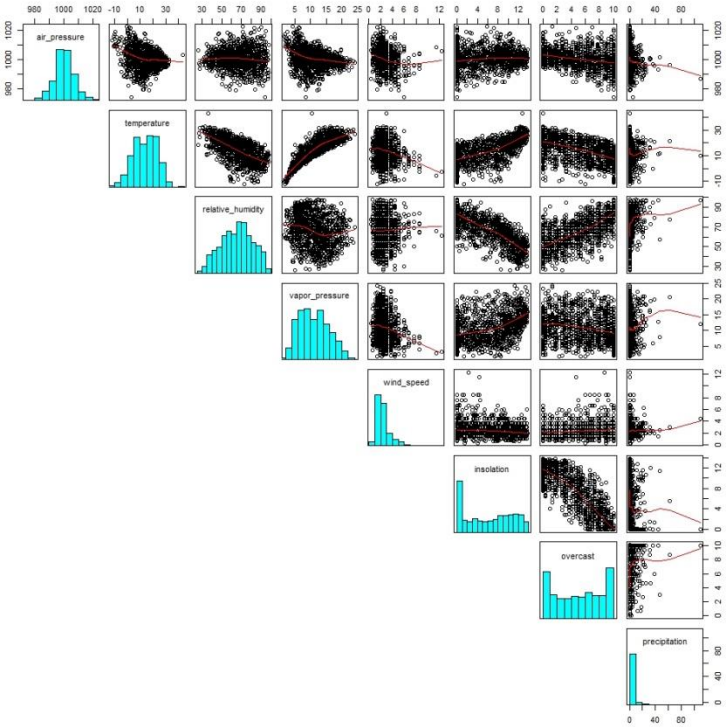


Results and discussion

- Autocorrelation and correlation

Test1

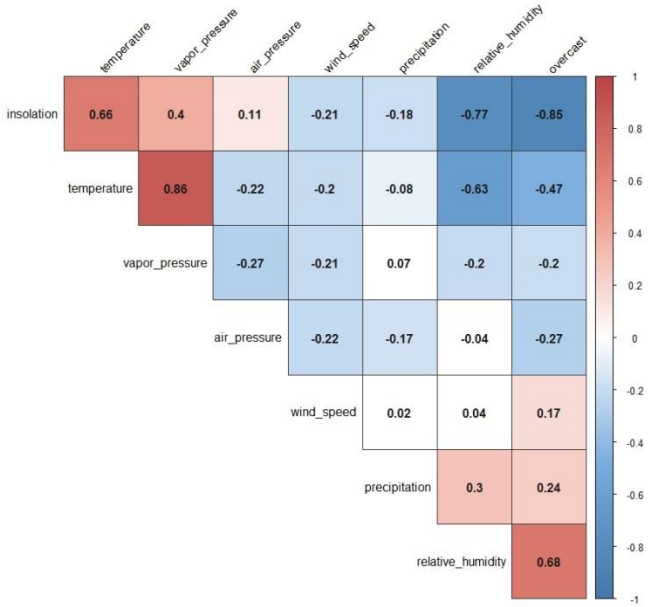


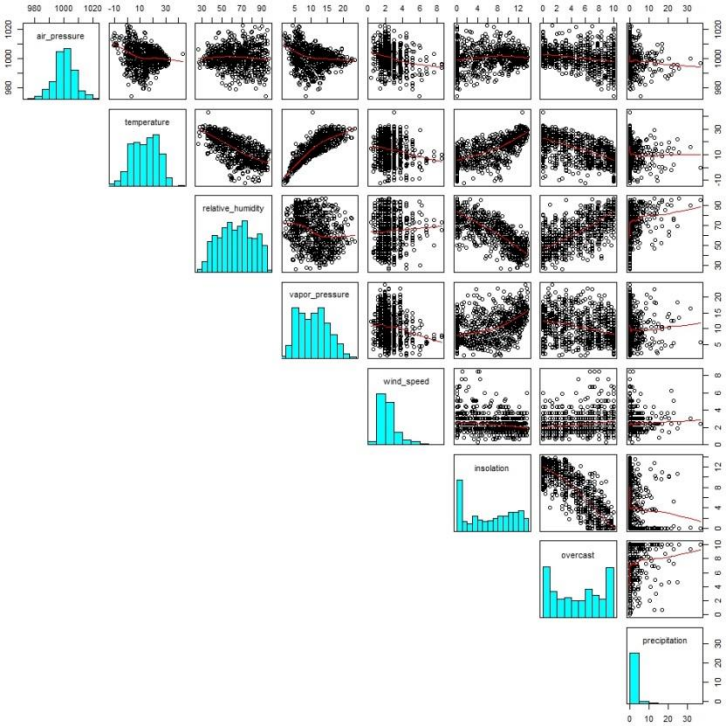


Results and discussion

- Autocorrelation and correlation

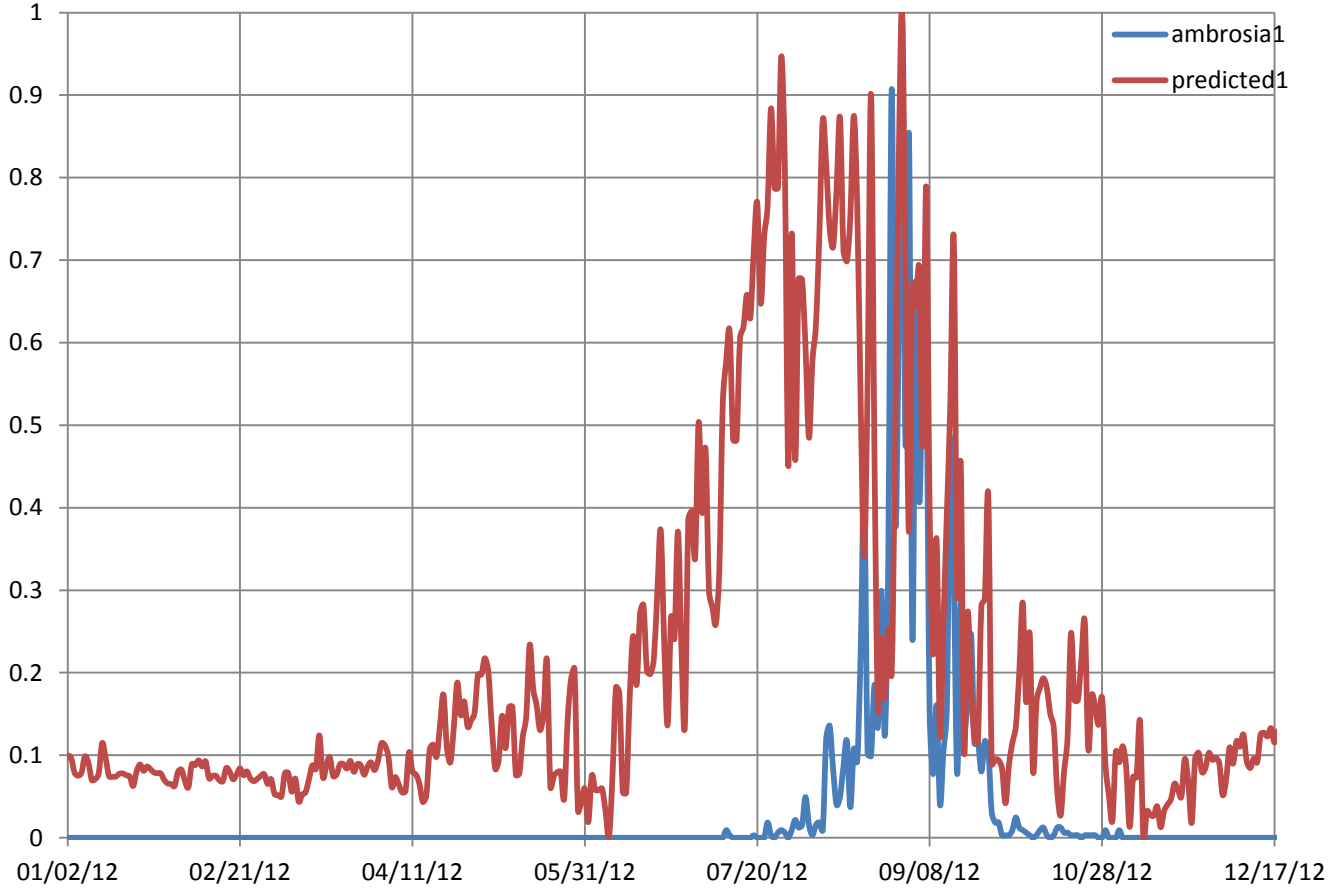
Test2





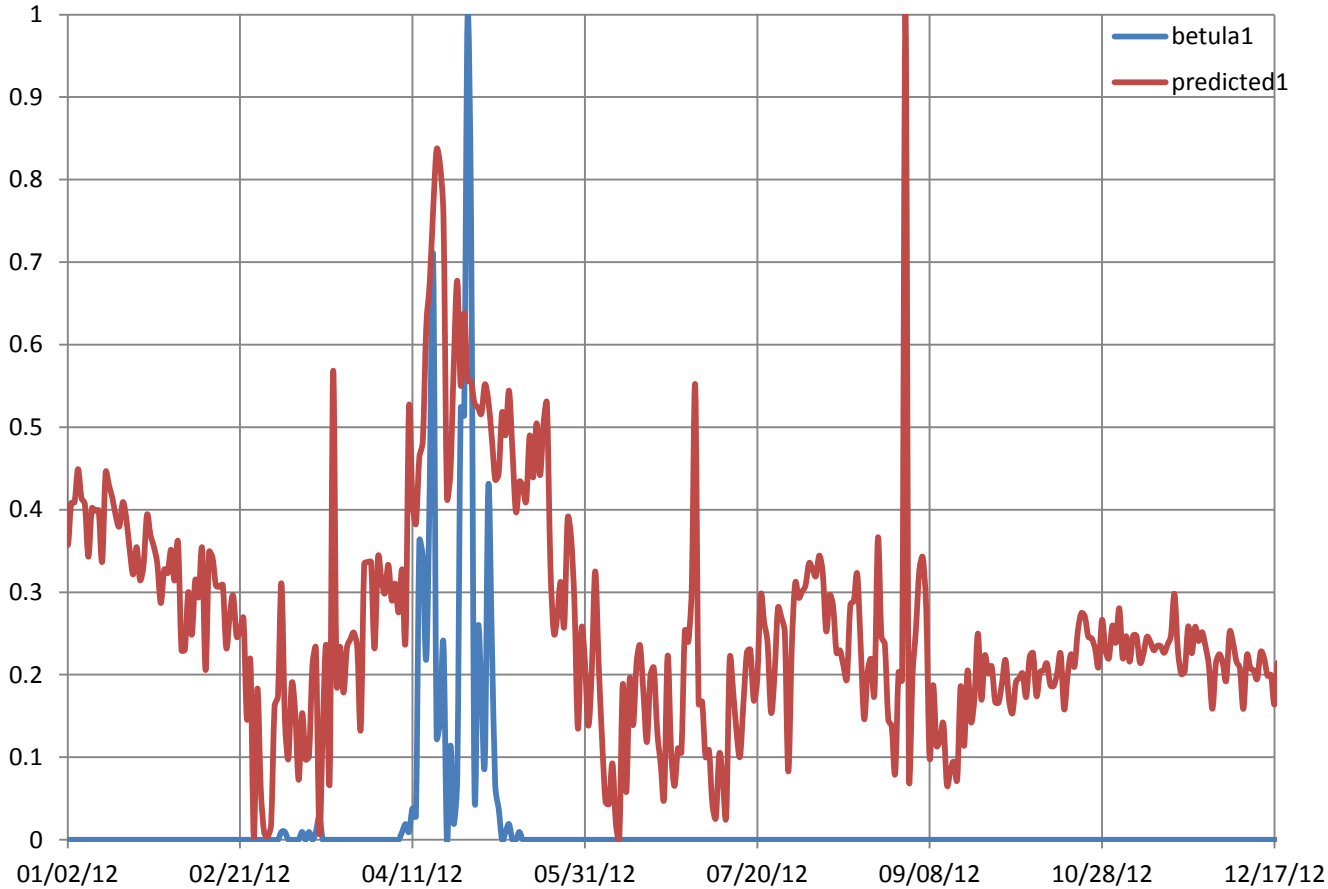
Results and discussion

- Test 1 pollen concentration prediction



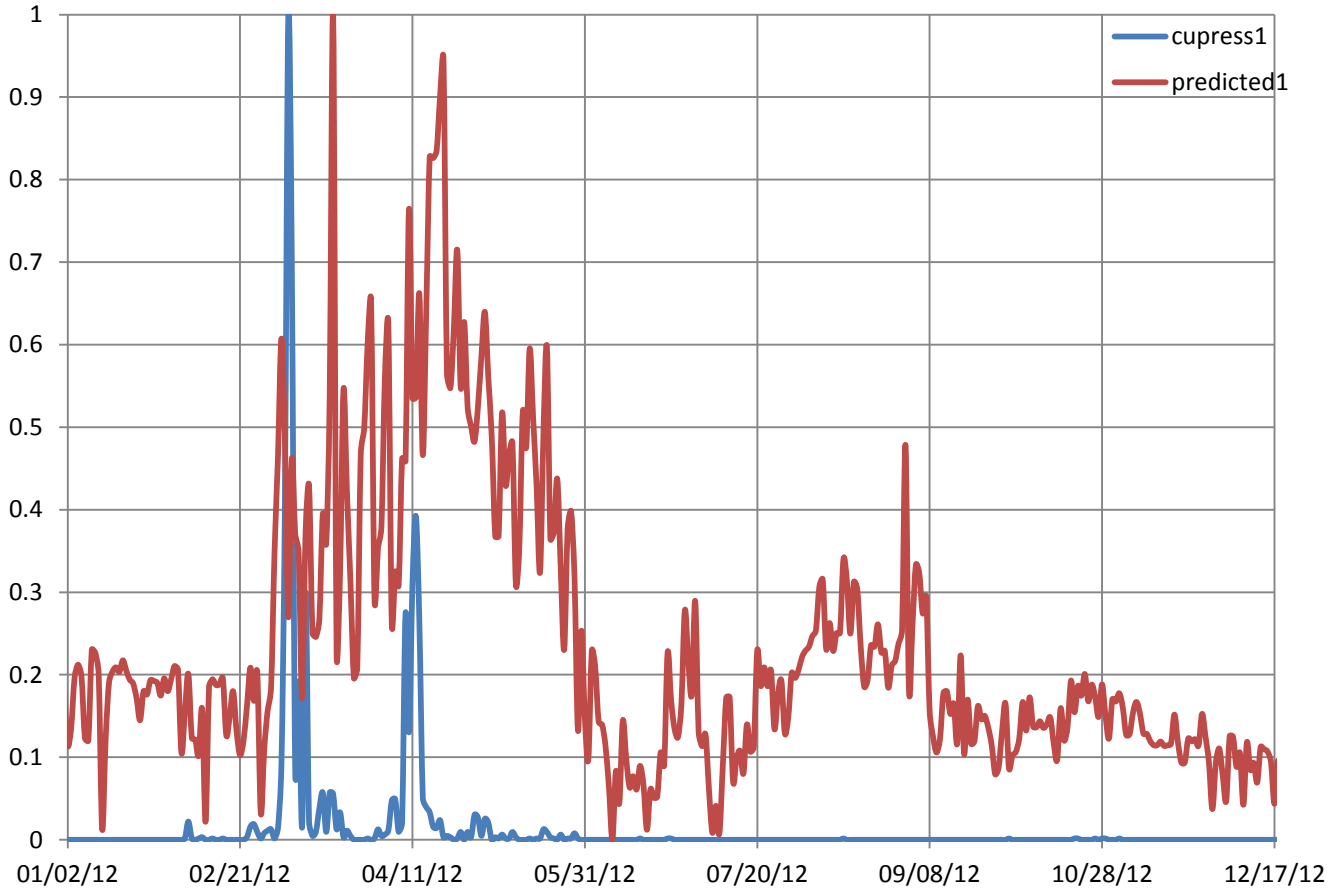
Results and discussion

- Test 1 pollen concentration prediction



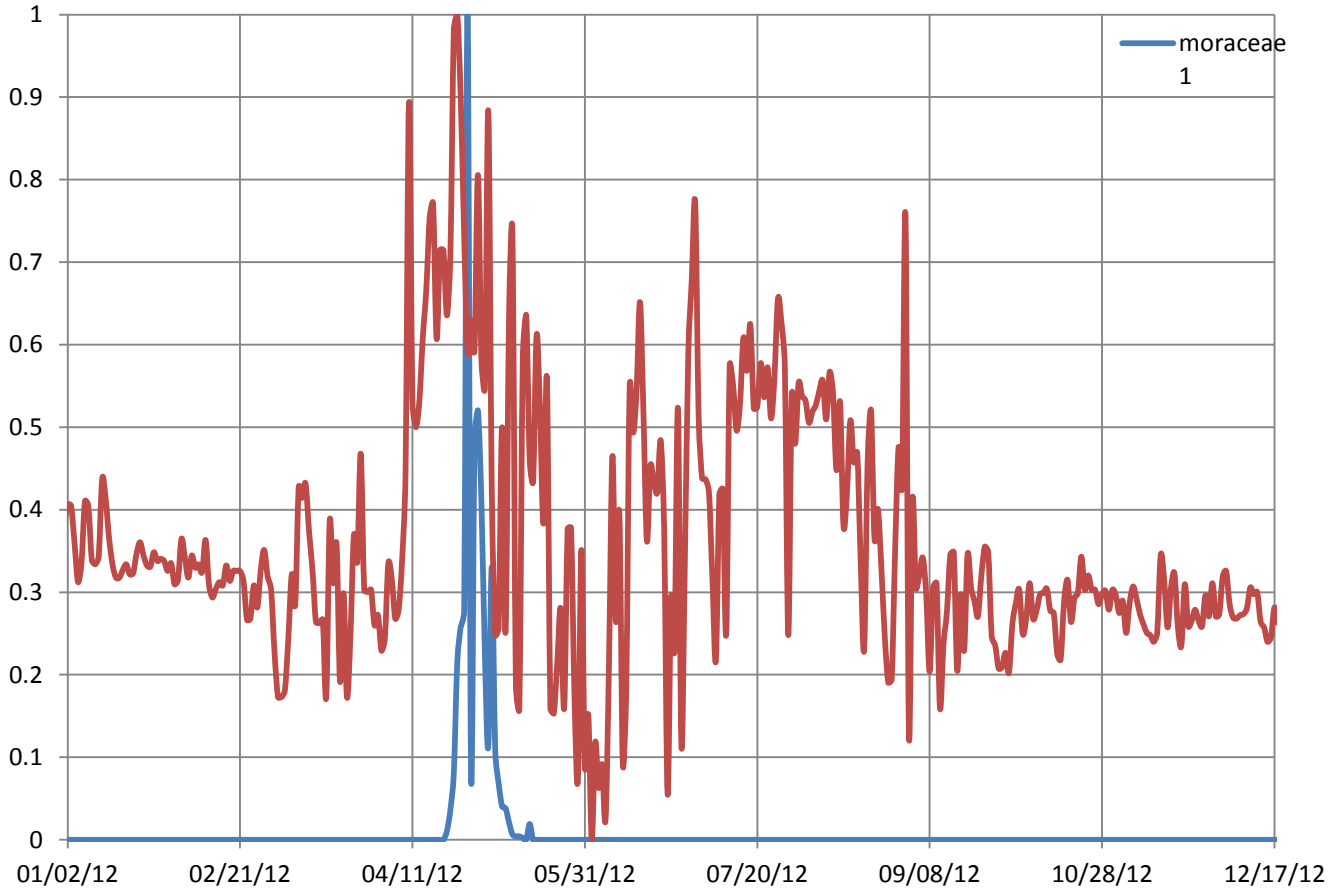
Results and discussion

- Test 1 pollen concentration prediction



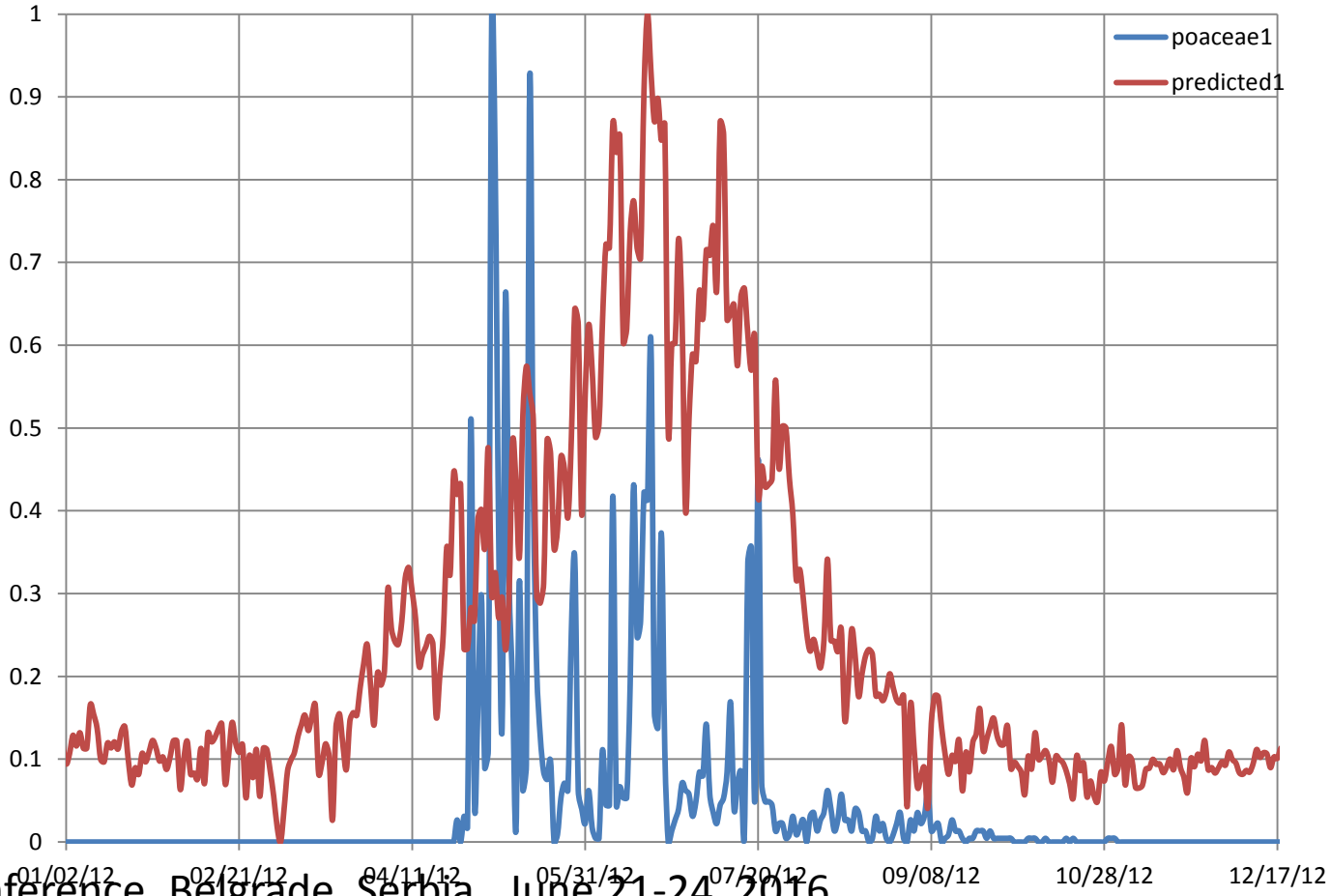
Results and discussion

- Test 1 pollen concentration prediction



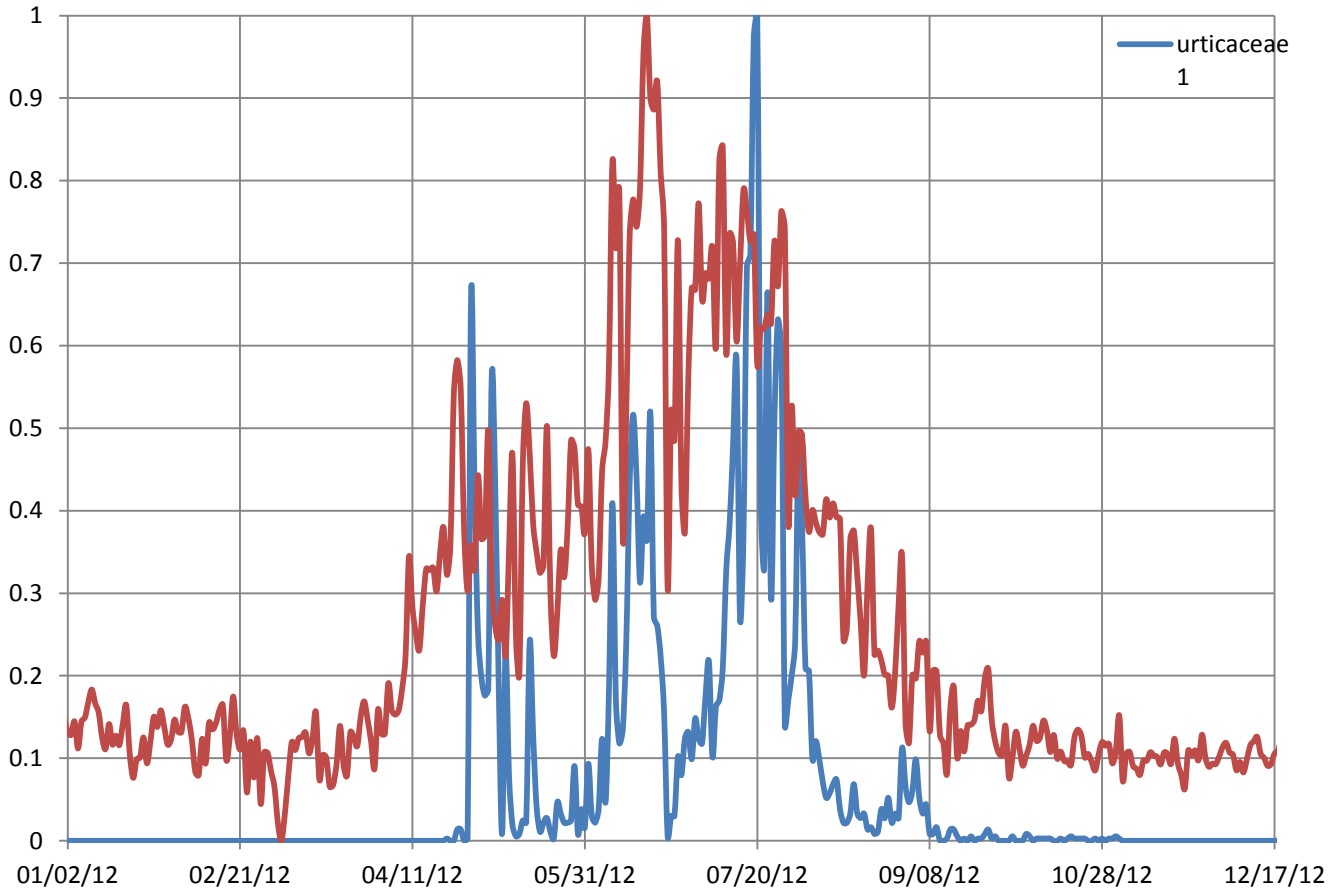
Results and discussion

- Test 1 pollen concentration prediction



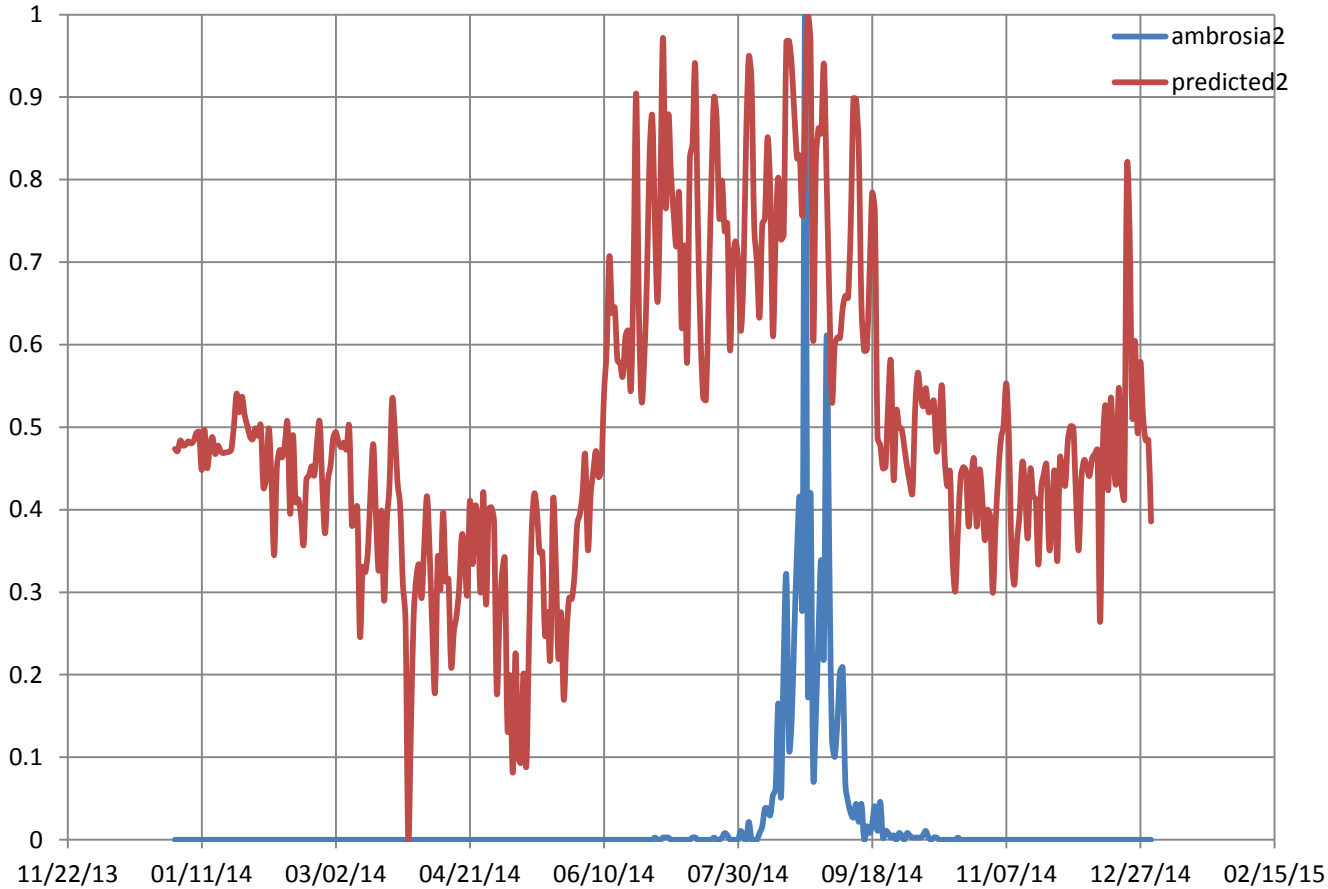
Results and discussion

- Test 1 pollen concentration prediction



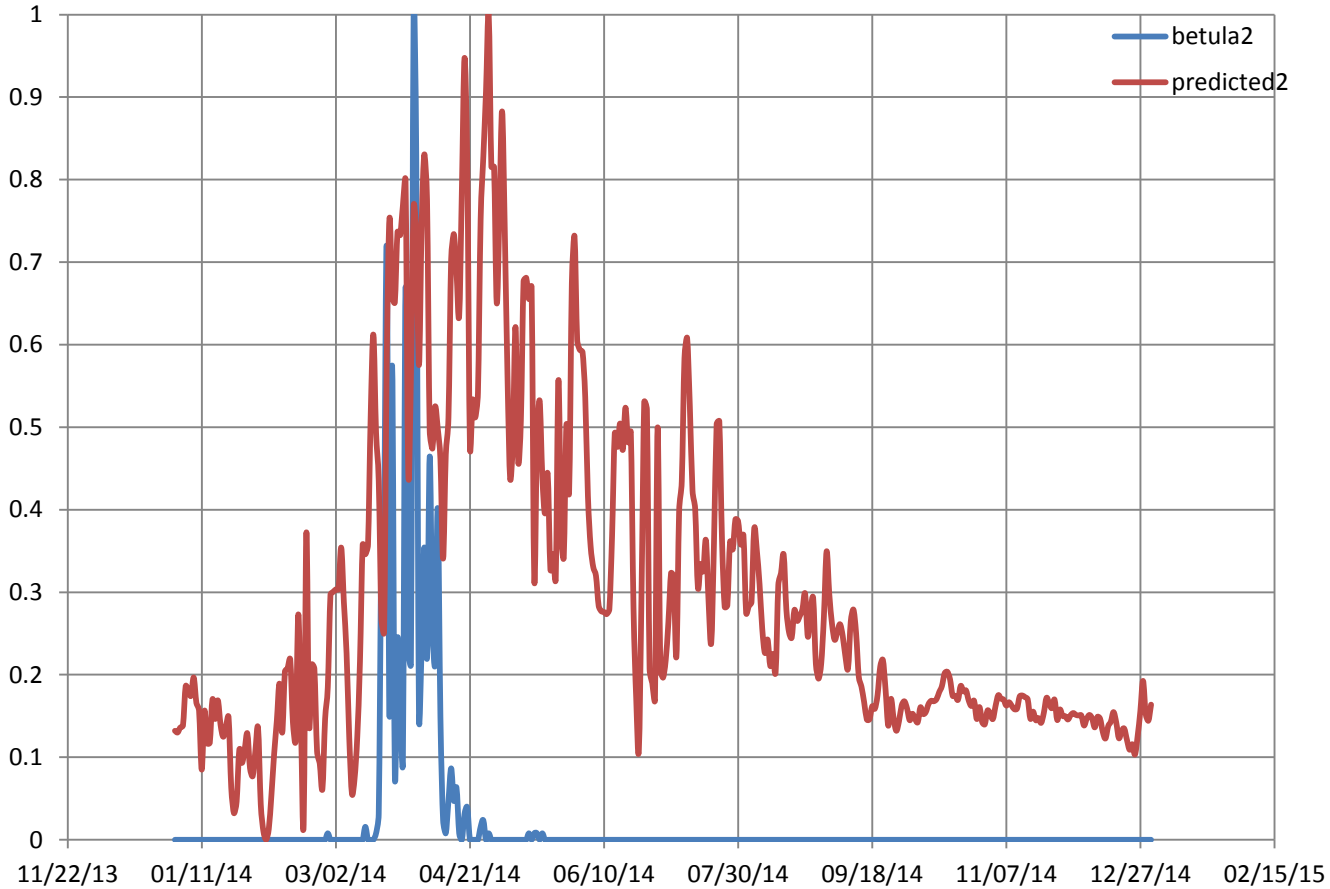
Results and discussion

- Test 2 pollen concentration prediction



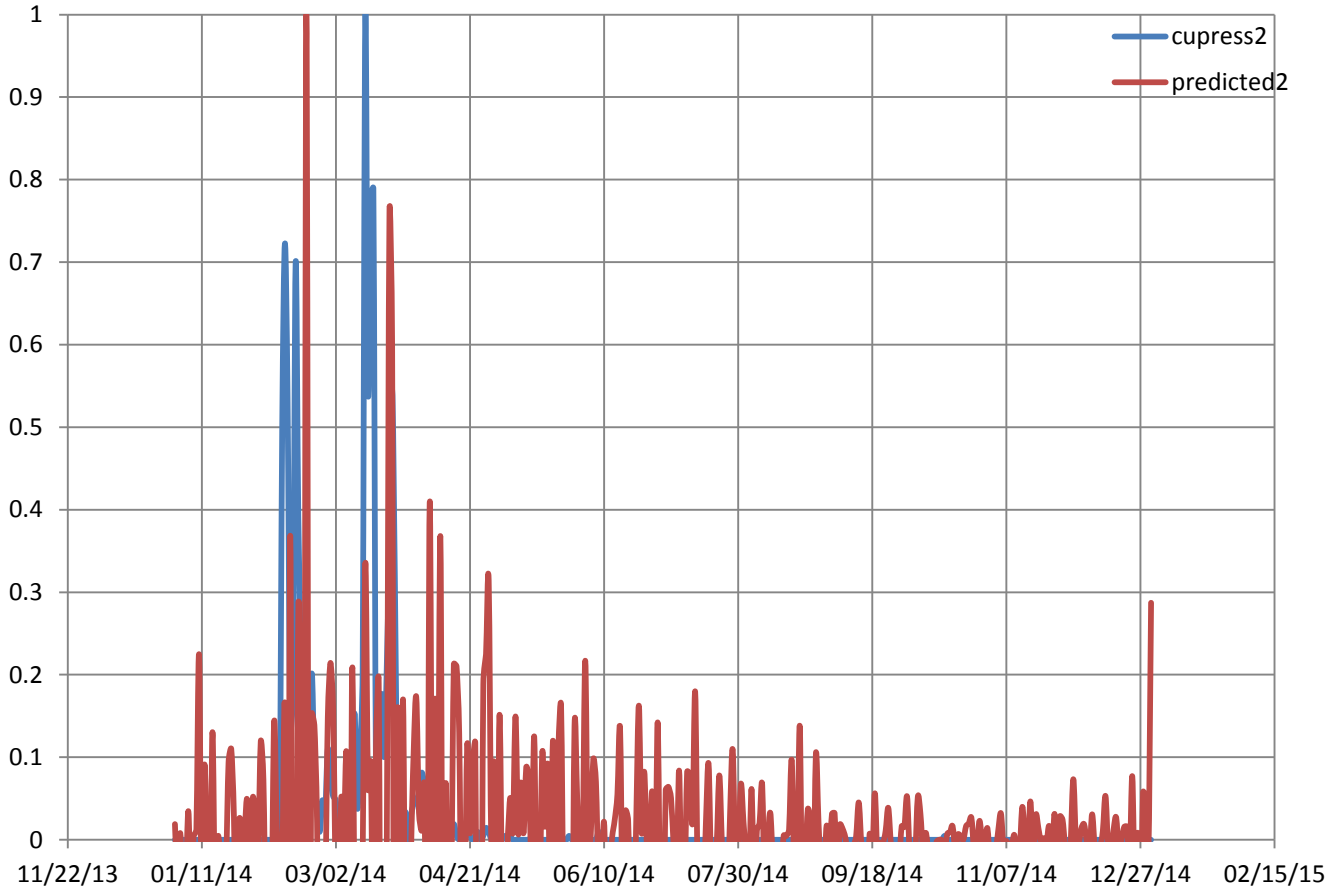
Results and discussion

- Test 2 pollen concentration prediction



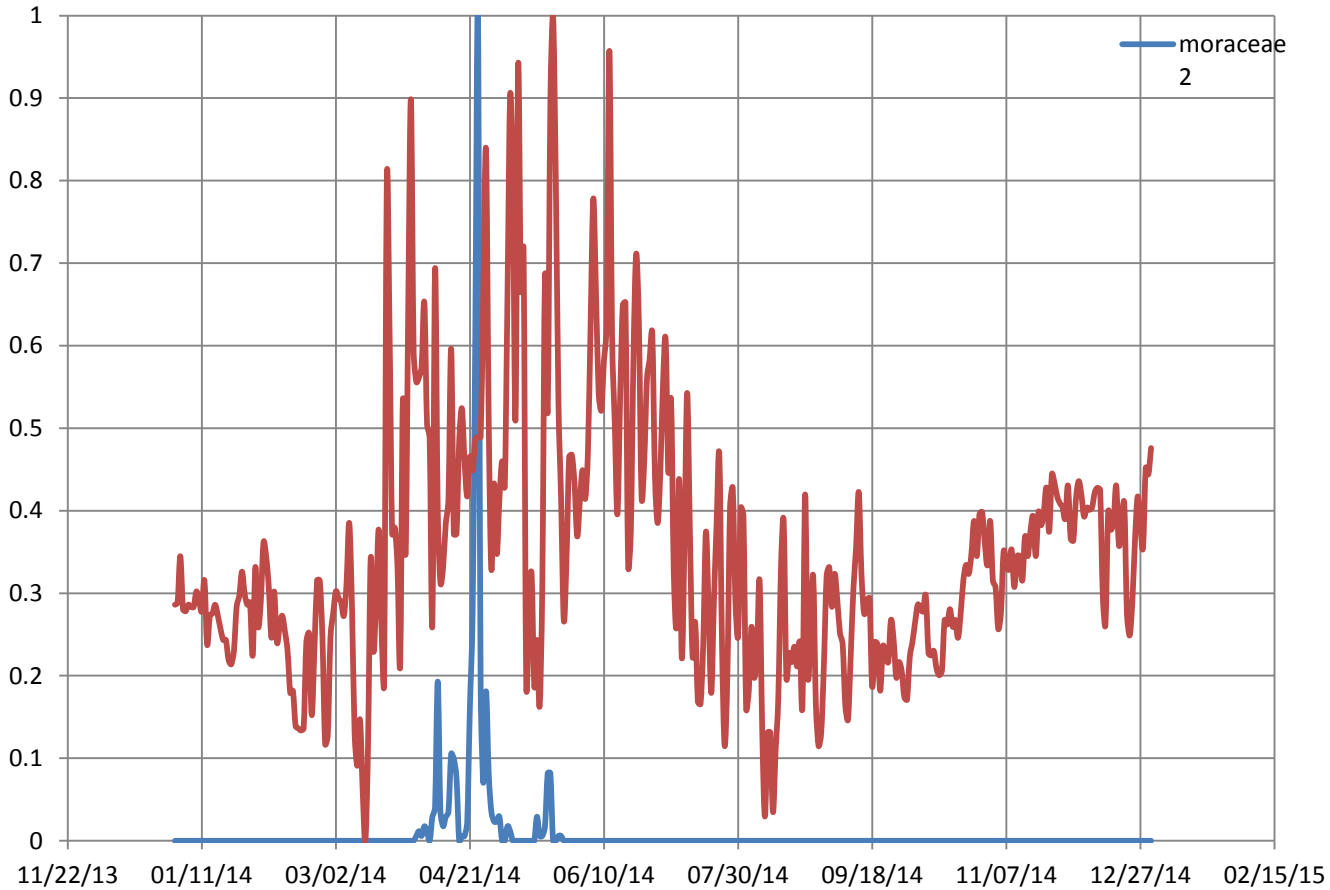
Results and discussion

- Test 2 pollen concentration prediction



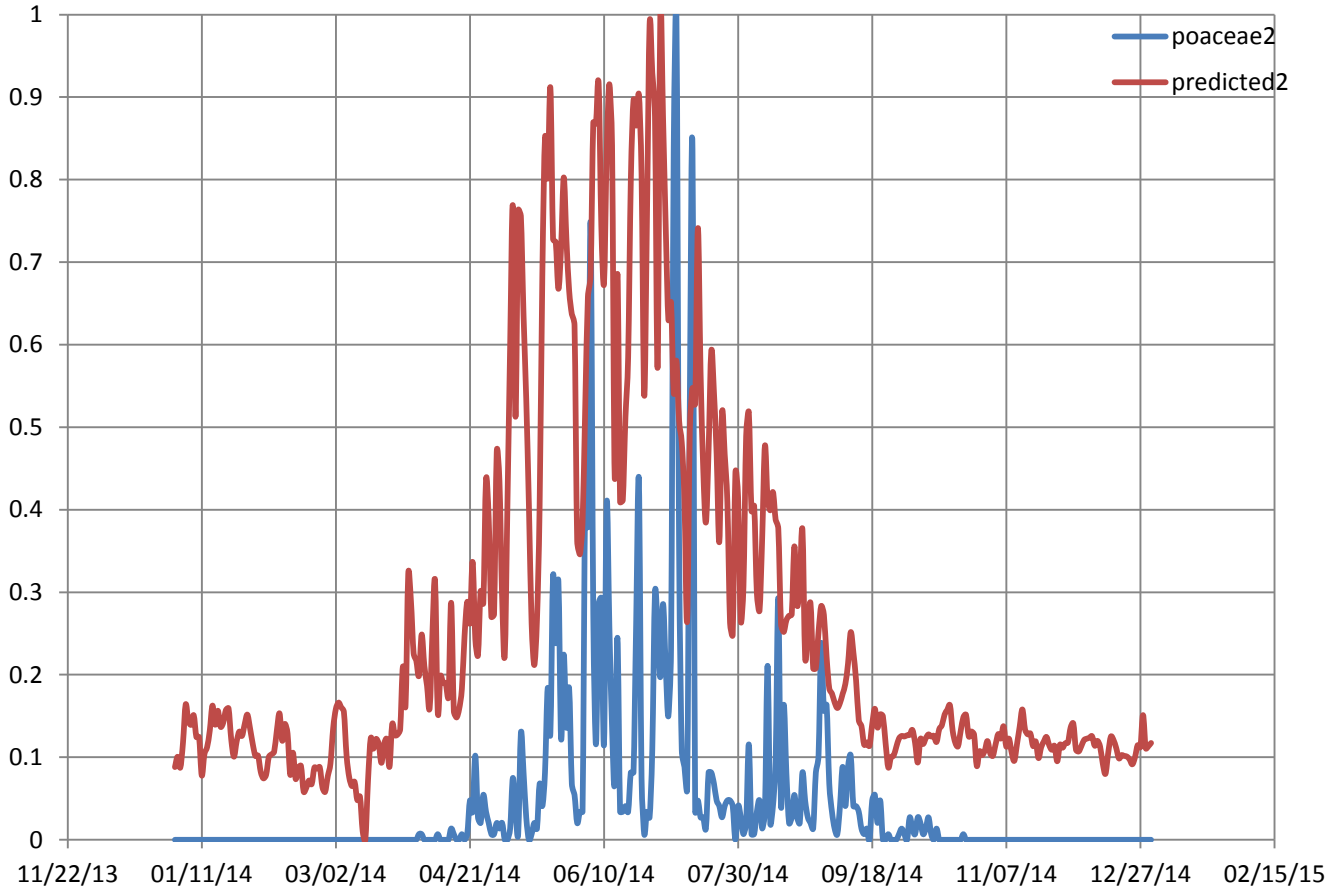
Results and discussion

- Test 2 pollen concentration prediction



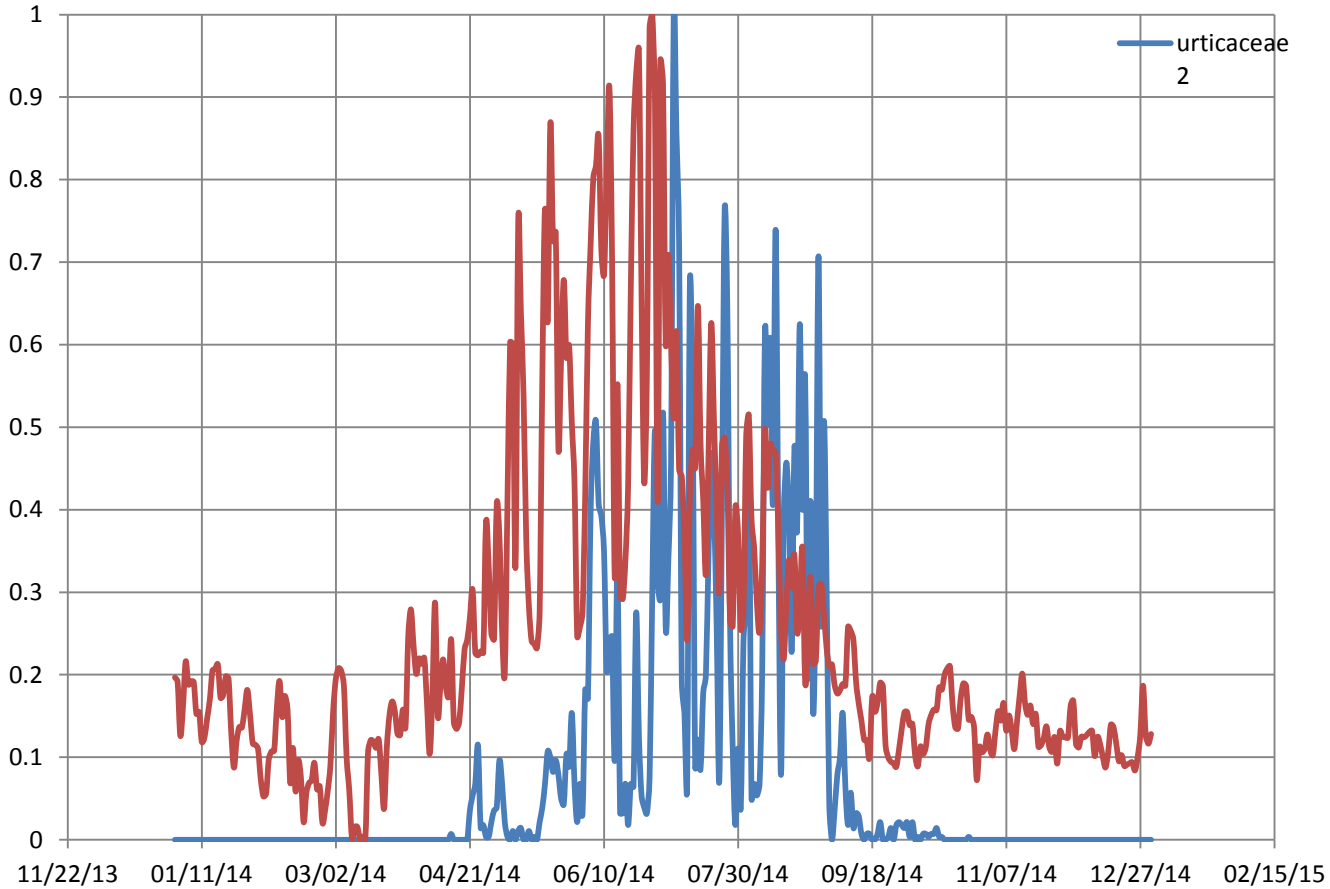
Results and discussion

- Test 2 pollen concentration prediction



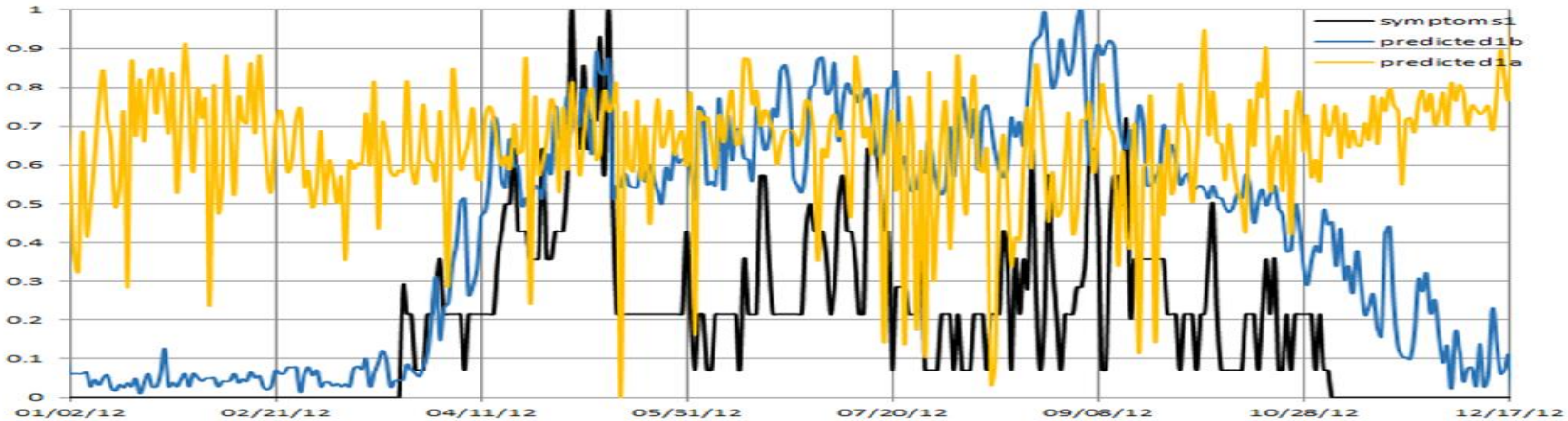
Results and discussion

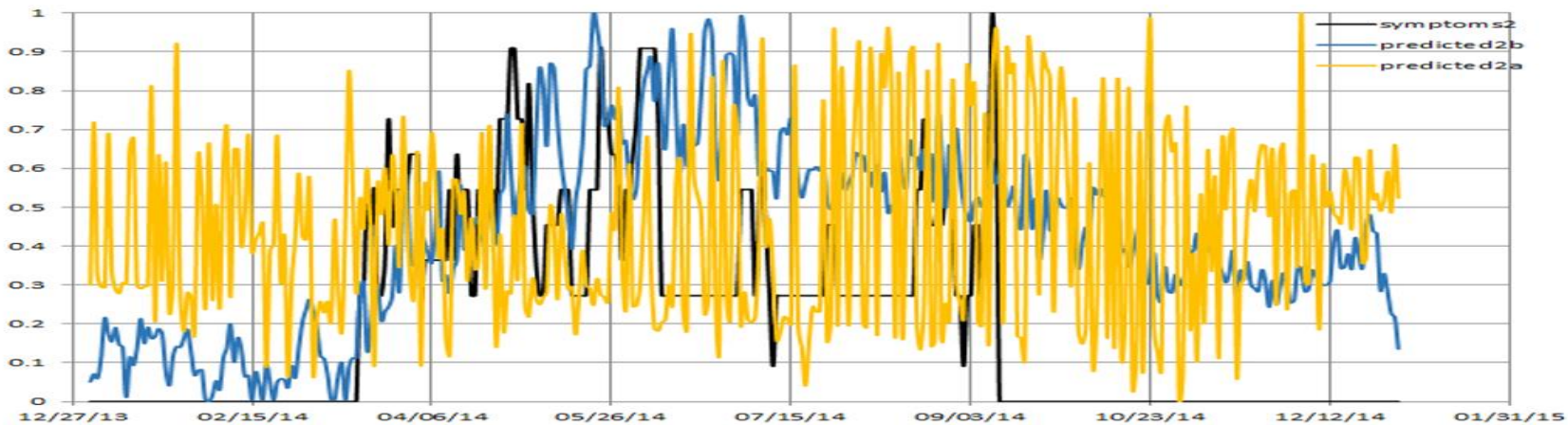
- Test 2 pollen concentration prediction



Results and discussion

- Test 1&2 symptoms prediction





Results and discussion

- Errors

Variant	Symptom prediction RMSE in %	
	without measured pollen	with measured pollen
TEST1	10.43	6.00
TEST2	39.00	37.03

Conclusion

- Work in progress…
  - better tune the SVMr hyper parameters
  - try other non-linear ML regression methods (ANN, Decision Trees)
  - Prolong training
  - detect noise and exclude it from training
  - transform the input data (log, PCA etc.)
  - include more voluntary data for allergy symptoms
  - include spatial context for wider region or entire country
  - use the same data for short-term forecast (5-days) of allergy simptoms

Thank you for your attention!

Miloš Marjanović*	University of Belgrade, Serbia
Mirjana Mitrović Josipović	Environmental Protection Agency, Serbia
Bojana Božanić	Pharm-Olam, Serbia
Vít Pászto	Palacký University Olomouc, Czech Republic
Lukaš Marek	University of Canterbury, New Zealand

*milos.marjanovic@rgf.bg.ac.rs

GeoMLA conference, Belgrade, Serbia, June 21-24, 2016