ACTA CLIMATOLOGICA ET CHOROLOGICA
Universitatis Szegediensis, Tomus 47-48, 2014, 17-32

PREDICTING DAILY RAGWEED POLLEN CONCENTRATIONS USING NEURAL NETWORKS AND TREE ALGORITHMS OVER LYON (FRANCE) AND SZEGED (HUNGARY)

Z CSÉPE[1], L MAKRA[1], D VOUKANTSIS[2], I MATYASOVSZKY[3], G TUSNÁDY[4], K KARATZAS[2] and M THIBAUDON[5]

[1]*Department of Climatology and Landscape Ecology, University of Szeged, P.O.B. 653, 6701, Szeged, Hungary*
[2]*Department of Mechanical Engineering, Informatics Systems & Applications Group, Aristotle University, P.O. Box 483, 54124 Thessaloniki, Greece*
[3]*Department of Meteorology, Eötvös Loránd University, Pázmány Péter st. 1/A, 1117 Budapest, Hungary*
[4]*Mathematical Institute of the Hungarian Academy of Sciences, P.O.B. 127, 1364 Budapest, Hungary*
[5]*RNSA (Aerobiology Network of France), La Parličre, 69610 Saint Genis l'Argentière, France*
E-mail: csepe.zoltan@gmail.com

Summary: Forecasting ragweed pollen concentration is a useful tool for sensitive people in order to prepare in time for high pollen episodes. The aim of the study is to use methods of Computational Intelligence (CI) (Multi-Layer Perceptron, M5P, REPTree, DecisionStump and MLPRegressor) for predicting daily values of *Ambrosia* pollen concentrations and alarm levels for 1-7 days ahead for Szeged (Hungary) and Lyon (France), respectively. Ten-year daily mean ragweed pollen data (within 1997-2006) are considered for both cities. 10 input variables are used in the models including pollen level or alarm level on the given day, furthermore the serial number of the given day of the year within the pollen season and altogether 8 meteorological variables. The study has novelties as (1) daily alarm thresholds are predicted in the aerobiological literature for the first time; (2) data-driven modelling methods including neural networks have never been used in forecasting daily *Ambrosia* pollen concentration; (3) algorithm J48 has never been used in palynological forecasts; (4) we apply a rarely used technique, namely factor analysis with special transformation, to detect the importance of the influencing variables in defining the pollen levels for 1-7 days ahead. When predicting pollen concentrations, for Szeged Multi-Layer Perceptron models deliver similar results with tree-based models 1 and 2 days ahead; while for Lyon only Multi-Layer Perceptron provides acceptable result. When predicting alarm levels, the performance of Multi-Layer Perceptron is the best for both cities. It is presented that the selection of the optimal method depends on climate, as a function of geographical location and relief. The results show that the more complex CI methods perform well, and their performance is case-specific for ≥2 days forecasting horizon. A determination coefficient of 0.98 (*Ambrosia,* Szeged, one day and two days ahead) using Multi-Layer Perceptron ranks this model the best one in the literature.

Key words: ragweed pollen allergy, forecasting, factor analysis with special transformation, neural networks, multi-layer perceptron, tree based methods

1. INTRODUCTION

The warming of the climate system is obvious, as it is now evident from observations of increases in global average air and ocean temperatures, the widespread melting of snow and ice, and rising global average sea level (IPCC 2013). Recent climate warming is associated with the modification of the distribution areas of plants producing allergenic pollen (Laaidi et al. 2011, Ziska et al. 2011), furthermore, with an earlier onset (Frei 2008, Rodríguez-Rajo et al. 2011), and earlier end dates (Stach et al. 2007, Recio et al. 2010), a longer pollen season (Stach et al. 2007, Ariano et al. 2010), an increase in the total annual pollen load (Cristofori et al. 2010, Ariano et al. 2010, Laaidi et al. 2011), as well as an increase of patient number sensitized to pollen throughout the year (Ariano et al. 2010).

The genus of ragweed (*Ambrosia spp)* comprises 42 species. They are the best known weeds for the most severe and widespread allergies caused by its pollen (Béres et al. 2005). However, in Europe, common ragweed *(Ambrosia artemisiifolia)* is predominant of all *Ambrosia* species (Makra et al. 2005, Bullock et al. 2010, Vinogradova et al. 2010). The most important habitat areas of common ragweed in Europe are the Rhône valley in France (Chauvel et al. 2006, Gladieux et al. 2011), north-western Milan and south Varese (Lombardy, Po River valley) in Italy (Bonini et al. 2012), the Pannonian Plain including Hungary and some parts of Serbia, Croatia, Slovenia, Slovakia and Romania (Kiss and Béres 2006, Makra et al. 2005), furthermore Ukraine (Rodinkova et al. 2012) and the south-western part of European Russia (Reznik 2009).

Advanced techniques such as neural networks, multi-layer perceptron and the support vector regression learning methods have been useful procedures for forecasting air quality parameters (Juhos et al. 2009, Vlachogianni et al. 2011, Voukantsis et al. 2011, Kassomenos et al. 2013). However, methods of Computational Intelligence (CI) have only been scarcely applied in airborne pollen related studies. They were used for forecasting (a) daily pollen concentrations (Delaunay et al. 2004, cedar pollen; Aznarte et al. 2007, olive pollen; Rodríguez-Rajo et al. 2010, Poaceae pollen; Voukantsis et al. 2010, Oleaceae, Poaceae and Urticaceae pollen; Puc 2012; *Betula* pollen), (b) pollen-induced symptoms (Voukantsis et al. 2013), (c) risk level of *Betula* pollen in the air (Castellano-Méndez et al. 2005) and (d) the severity of the Poaceae pollen season (Sánchez Mesa et al. 2005). Furthermore, Aznarte et al. (2007) used neuro-fuzzy models for forecasting olive pollen concentrations. The above applications of neural networks and neuro-fuzzy models produced better results than traditional statistical methods (Sánchez Mesa et al. 2005).

These methods of Computational Intelligence 1) can deal with the complexity of the mechanisms concerning the release and dispersion of the airborne pollen, 2) can be applied for different tasks (e.g. optimization and forecasting), 3) are computationally efficient and can be easily integrated into the operational use of the models (Voukantsis et al. 2010).

In this paper we use factor analysis with special transformation, a technique for detecting the importance of the influencing variables in defining the pollen levels for 1-7 days ahead. Furthermore, data-oriented models are applied for (1) predicting daily concentration of ragweed pollen that shows the highest allergenicity of all taxa and (2) comparing the efficiency of different prediction techniques over two heavily polluted areas in Europe, i.e. over Lyon (France) and Szeged (Hungary), respectively. The main objectives are: i) development of accurate forecasting models for operational use, ii) evaluation of CI methods that have not been previously applied for *Ambrosia* pollen, such as Multi-Layer Perceptron and regression trees and iii) obtaining a forecast of highest accuracy among CI methods based on input data of former prediction algorithms. Note that (1) data-driven modeling methods including neural networks have never been used in forecasting daily *Ambrosia* pollen concentration, (2) daily alarm thresholds are predicted in the aerobiological literature for the first time; furthermore (3) algorithm J48 has never been used in palynological forecasts.

2. MATERIALS

2.1. Location and data

2.1.1. Study area

Two European cities, namely Lyon (Rhône Valley, France) and Szeged (Pannonian Plain, Hungary) were selected as they show high ragweed pollen levels in Europe.

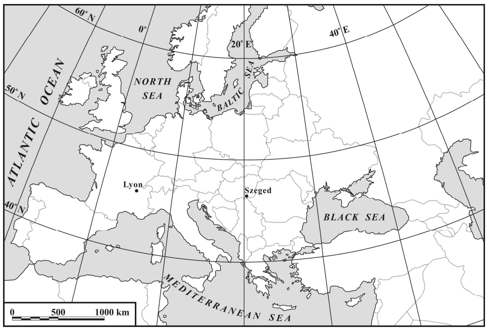


Fig. 1 The geographical positions of Lyon and Szeged

These cities differ in their topography and climate as well as in ragweed pollen characteristics. Szeged (46.25°N, 20.10°E), the largest settlement in South-eastern Hungary, is located at the confluence of the rivers Tisza and Maros (Fig. 1). The city is the centre of the Szeged region with 203,000 inhabitants. In the Köppen system the climate of Szeged is the Ca type (warm, temperate climate), with relatively mild and short winters and hot summers (Köppen, 1931). Lyon (45.77°N, 4.83°E) lies in the Rhône-Alpes of France.

The city is located in the Rhône valley at the confluence of the Rhône and Saône rivers with a population of 1.8 million (Fig. 1). In the Köppen system its climate is of the Cbf type. That is, it has a temperate oceanic climate with mild winters and cool-to-warm summers, as well as a uniform annual precipitation distribution (Köppen, 1931).

2.1.2. Pollen and meteorological data

Ten-year (1997-2006) daily mean ragweed pollen data were considered for both Szeged and Lyon. Ragweed pollen concentrations or ragweed pollen alarm threshold values for 1, 2, …, 7 days after the given day were used as resultant variables. Ragweed pollen levels or ragweed pollen alarm thresholds on the given day; furthermore, the serial number of the given day of the year within the pollen season and altogether 8 meteorological variables on the given day were selected as influencing variables. The meteorological variables include daily values of mean temperature (T_{mean},°C), minimum temperature (T_{min},°C) and maximum temperature (T_{max},°C), daily temperature range ($\Delta T = T_{max} - T_{min}$,°C), daily mean relative humidity (RH, %), daily total radiation (TR, $W \cdot m^{-2}$), daily means of air pressure (P, mm) and wind speed (WS, $m \cdot s^{-1}$). For Lyon, daily data of total radiation were absent hence they were replaced with daily sunshine duration (SD, hour).

Alarm levels of *Ambrosia* pollen used in Hungary are as follows (Mányoki et al. 2011). Level 0: there is no *Ambrosia* pollen in the air. Level 1: (1-9 pollen grains m^{-3} of air): (very low pollen concentration, it produces no symptoms. Level 2: (10-29 pollen grains m^{-3} of air): low pollen concentration, it may cause symptoms. Level 3: (30-49 pollen grains m^{-3} of air): medium pollen concentration, it may generate symptoms even for less sensitive people. Level 4: (50-99 pollen grains m^{-3} of air): medium high pollen concentration, it may induce medium strong reactions even for less sensitive people. Level 5: (100-199 pollen grains m^{-3} of air): high pollen concentration, it may provoke strong or very strong symptoms for all sensitive people. Level 6: (200-499 pollen grains m^{-3} of air): very high pollen concentration, health state of sensitive people may turn critical, asthmatic symptoms may also occur. Level 7: (500-999 pollen grains m^{-3} of air): exceptionally high pollen concentration, it may provoke acute symptoms inducing serious deterioration in the quality of life. Level 8: (>1000 pollen grains m^{-3} of air): extreme pollen concentration, excessively strong symptoms (Mányoki et al. 1011). The data were separated into two parts: the training set (1997-2004) to develop forecasting models, and the test set (2005-2006) to validate these models.

2.2. Methods

The study applies the factor analysis with special transformation. Furthermore, the following CI methods are evaluated for the task. Multi-layer perceptron (MLP) (Haykin 1999) models are artificial neural network models capable of modelling complex and highly nonlinear processes. Two types of neural networks are applied: a complex (MLP with more than one hidden layer) and a less complex (MLPRegressor with only one hidden layer) version. For predicting both the daily pollen concentrations and daily alarm levels of ragweed, several tree algorithms (M5P, REPTree, DecisionStump and J48) are used. These algorithms have not yet been used for the above tasks. The models have been developed in Matlab with WEKA implementation of the above algorithms, found in Hall et al. (2009).

2.2.1. Factor analysis with special transformation

Factor analysis identifies linear relationships among examined variables and thus helps to reduce the dimensionality of the initial database without substantial loss of information. Factor analysis was applied to our initial datasets consisting daily values of 11 correlated variables [10 explanatory variables including the serial number of the days in the year, 8 meteorological and 1 pollen variable (*Ambrosia* pollen level or alarm level) and 1 resultant variable (*Ambrosia* pollen level or alarm level for 1-7 target days, respectively)] in order to transform the original variables into fewer uncorrelated variables. These new variables, called factors, can be viewed as latent variables explaining the joint behaviour of the day in the year, furthermore the meteorological elements and the pollen variables. The number of retained factors can be determined by different criteria. The most common and widely accepted one is to specify a least percentage (80%) of the total variance of the original variables that has to be explained (Jolliffe 1993) by the factors. After performing the factor analysis, a special transformation of the retained factors was made to discover to what degree the above-mentioned explanatory variables affect the resultant variable and to give a rank of their influence (Jahn and Vahle 1968). When performing factor analysis on the standardized variables, factor loadings are correlation coefficients between the factors and the original variables. Consequently, if the resultant variable is strongly correlated with a factor and an explanatory variable is highly correlated with this factor, then the explanatory variable is also highly correlated with the resultant variable. Hence, it is advisable to combine all the factors together with the resultant variable into one new factor. It is effective to do so that only one factor has big contribution to the resultant variable and the remaining factors are uncorrelated with the resultant variable. This latter procedure is called special transformation (Jahn and Vahle 1968).

2.2.2. Multi-layer Perceptron (MLP)

MLP (Haykin 1999) is the most successful implementation of feedforward artificial neural networks and have been widely applied in the field of environmental science for classification, regression and function approximation problems. MLP can model complex and highly non-linear processes through the topology of the network. Multi-Layer Perceptron comprises an input and an output layer with one or more hidden layers of nonlinearly-activation functions. These capabilities have already been successfully utilized in previous studies in order to forecast pollen concentrations (e.g. Voukantsis et al. 2010), therefore MLP is an important procedure and this is the first occasion for using this method for predicting daily concentrations and daily alarm thresholds of ragweed pollen.

In the study, the MLP model always has more than one hidden layer and MLP has several parameters that need to be set. They are training time, learning rate, hidden layers and neurons in the layers. Training time was 1500, learning rate started from 0.3 and it was reduced in each step. This helps to stop the network from diverging from the target output as well as improve the general performance. The number of hidden layers is generated automatically by WEKA. MLP was applied with the same options for predicting both the daily pollen concentrations and daily alarm thresholds of ragweed.

2.2.3. MLPRegressor and MLPClassifier

Both classes are built-in WEKA modelling softwares (Hall et al. 2009). These algorithms are special parts of Multi-Layer Perceptrons. They always have only one hidden layer, where the number of neurons is user specific. Both use optimization by minimizing the squared error plus a quadratic penalty with the BFGS method. All parameters are standardized, including the target variable. The activation function is a logistic function. MLPRegressor and MLPClassifier are applied for predicting the daily pollen concentrations and daily alarm thresholds of ragweed, respectively.

2.2.4. Tree-based algorithms

This procedure is a reproduction of Quinlan's M5 algorithm (Quinlan 1992) being a combination of decision trees and multivariate regression models. Contrary to other regression trees the leaves of the M5P tree structure consist of MLR models. So, it is possible to model local linearity within the data similarly to piecewise linear functions. This is the first study applying M5P to model daily ragweed pollen data.

DecisionStump builds a decision tree with a single split point. It makes (1) regression based on mean-squared errors or (2) classification based on entropy depending on the data type to be forecasted.

REPTree is a fast decision tree learner. It builds a decision tree using information gain or makes a regression tree from the variance. It applies pruning with backfitting for reducing error.

J48 is an implementation of C4.5 algorithm in the WEKA data mining pool. C4.5 builds decision trees from a set of training data in the same way as ID3 using the concept of information entropy. J48 classifier achieves fast execution times and adequate scales of large datasets (Quinlan 1993).

3. RESULTS AND DISCUSSION

3.1. Performance evaluation

The importance of the serial number of the day in the year, furthermore daily values of eight meteorological variables and *Ambrosia* pollen level were analysed in determining a future day pollen level for 1-7 days ahead using factor analysis with special transformation (Tables 1-2). When comparing the results very little similarity was received for the two cities. The importance of the serial number of the day of the year shows a tendency of higher weights towards increasing target days for both Szeged and Lyon; however, this effect is more remarkable for Szeged. From the meteorological influencing variables, only TR and *Ambrosia* pollen level showed similarly significant positive weights with values of the same magnitude in determining a future day pollen level (Tables 1-2). The weights of actual day *Ambrosia* pollen level emerge extraordinarily from all variables indicating its high significance for both cities. This confirms former findings according to which the most decisive influencing variable of all is the actual day *Ambrosia* pollen level for assigning pollen levels 1-7 days ahead (Makra et al. 2011, Makra and Matyasovszky 2011).

For Szeged, T_{mean}, T_{max} and ΔT indicate significant and substantially higher positive weights compared to Lyon. While the importance of RH and WS can be negligible for Szeged, these parameters show highly relevant negative associations in the formation of pollen levels 1-7 days ahead for Lyon. P shows notable negative and positive weights for Szeged and Lyon, respectively. The here-mentioned definite difference in the weights and signs of the influencing variables for the two cities can be explained by their different climate and relief. The temperate oceanic climate of Lyon with cool-to-warm summers confirms the role of humidity parameters (RH) here, while the location of the city in the Rhone valley on the foothills of High Alps emphasizes the weight of the wind (WS). The warm, temperate climate of Szeged highlights the importance of the temperature parameters (T_{mean}, T_{max}, T_{min} and ΔT) and shows insignificant weights for the humidity (RH), while the central location of the city in the Pannonian Plain makes negligible the role of the wind (WS) (Tables 1-2).

3.2. Performance of the forecasting models

The following statistical indices were used to compare the performance of the models: (1) correlation coefficient as a measure of the strength; (2) Root Mean Square Error (RMSE) and (3) Mean Absolute Error (MAE) as measures of the error in the forecast.

For Szeged, MLP provides the best results for the forecasting horizon (1-7 days) that is confirmed by former studies (Sánchez-Mesa et al. 2002, Voukantsis et al. 2010). 1-day forecast indicates the best performance. This can be explained by the close association between the pollen concentrations of consecutive days and the predominant role of local pollen release in the measured pollen concentration in Szeged (Makra et al. 2010). The efficiency of MLPRegressor declines intensely when forecasting more than 2 days ahead due to its simpler construction (Table 3, Fig. 2). Considering decision trees, performance of REPTree decreases for >1-day forecasts, while DecisionStump provides an overall weak result for the forecasting horizon. MLPRegressor serves the best performance for 1 and 2-day ahead forecasts; however, when the forecasting horizon exceeds 2 days, the accuracy of the predictions sharply decreases. High values of RMSE and MAE can be attributed to the very high variability of the daily ragweed pollen concentrations. There are no periods in the pollen season that can be approximated linearly with high confidence.

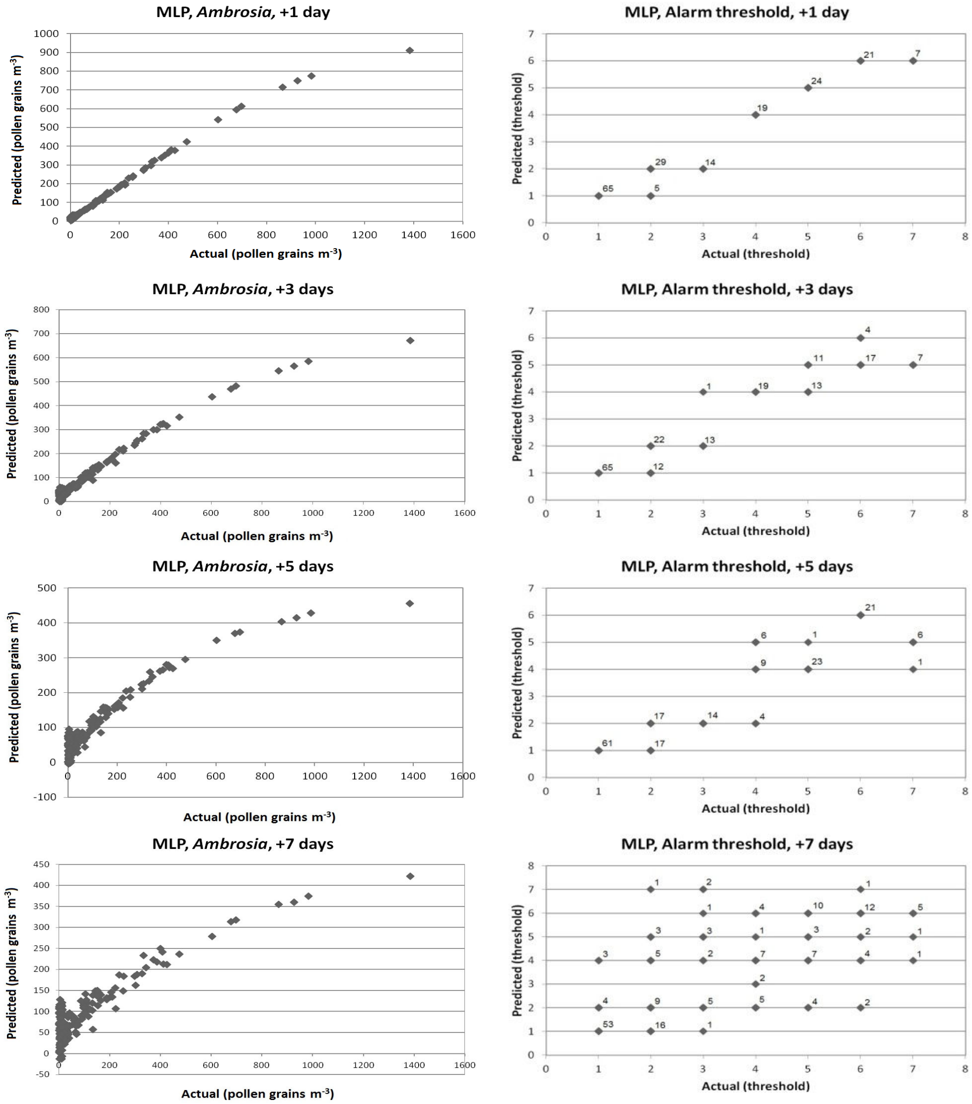


Fig. 2 Scatter plots, Szeged. Selected scatter plots of actual and predicted *Ambrosia* pollen concentrations (MLP), as well as alarm thresholds (MLP). The forecasting horizon is given in days

Table 1 Special transformation, Szeged. Relevance of the influencing variables in defining the resultant variable (pollen levels) 1-7 days ahead and the rank of importance of the influencing variables for determining the resultant variable (thresholds of significance: *italic:* $x_{0.05} = 0.064$; **bold:** $x_{0.01} = 0.084$)

	Influencing variables																				
Day[1]	Day[2]		T_{mean}[3]		T_{max}[4]		T_{min}[5]		ΔT[6]		RH[7]		TR[8]		P[9]		WS[10]		*Ambrosia*	
	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank
+1	0.00	7	**0.13**	7	**0.14**	7	0.06	7	**0.10**	4	-0.01	6	**0.11**	7	*-0.07*	4	*0.07*	1	**0.96**	1
+2	-0.02	6	**0.16**	6	**0.17**	6	*0.07*	6	**0.11**	3	-0.02	3	**0.13**	6	-0.06	6	0.06	2	**0.93**	2
+3	-0.04	5	**0.18**	5	**0.19**	5	**0.09**	4	**0.12**	2	-0.02	2	**0.15**	5	*-0.08*	3	0.05	3	**0.90**	3
+4	-0.06	4	**0.20**	4	**0.21**	4	**0.09**	5	**0.15**	1	-0.03	1	**0.17**	1	*-0.07*	5	0.02	4	**0.87**	4
+5	*-0.08*	3	**0.21**	3	**0.21**	3	**0.14**	3	**0.09**	5	-0.01	5	**0.15**	4	-0.05	7	-0.01	5	**0.84**	5
+6	**-0.13**	2	**0.26**	2	**0.25**	2	**0.19**	2	*0.07*	6	0.00	7	**0.16**	3	**-0.10**	1	0.01	6	**0.81**	6
+7	**-0.17**	1	**0.29**	1	**0.27**	1	**0.25**	1	0.03	7	0.01	4	**0.16**	2	**-0.09**	2	0.00	7	**0.78**	7

[1]: target day of the forecast; [2]: serial number of the day in the year; [3]: daily mean temperature (°C); [4]: daily maximum temperature (°C); [5]: daily minimum temperature (°C); [6]: daily temperature range (°C); [7]: daily relative humidity (%); [8]: daily total radiation ($W \cdot m^{-2}$); [9]: daily mean air pressure (hPa); [10]: daily wind speed ($m \cdot s^{-1}$)

Table 2 Special transformation, Lyon. Relevance of the influencing variables in defining the resultant variable (pollen levels) 1-7 days ahead and the rank of importance of the influencing variables for determining the resultant variable (thresholds of significance: *italic:* $x_{0.05} = 0.064$; **bold:** $x_{0.01} = 0.084$)

	Influencing variables																				
Day[1]	Day[2]		T_{mean}[3]		T_{max}[4]		T_{min}[5]		ΔT[6]		RH[7]		TR[8]		P[9]		WS[10]		*Ambrosia*	
	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank	weight	rank
+1	0.05	2	**0.12**	1	**0.11**	1	**0.09**	5	*0.08*	1	**-0.14**	5	**0.17**	5	0.03	7	-0.02	7	**0.87**	1
+2	0.04	4	0.06	5	0.05	2	*0.07*	7	0.03	6	**-0.17**	3	**0.24**	2	**0.16**	2	**-0.11**	4	**0.71**	2
+3	0.01	6	0.05	7	0.04	4	*0.08*	6	0.01	7	**-0.18**	1	**0.26**	1	**0.16**	1	**-0.13**	2	**0.69**	4
+4	-0.01	7	0.06	6	0.03	7	**0.13**	4	-0.03	5	**-0.17**	2	**0.18**	4	**0.10**	3	**-0.09**	6	**0.70**	3
+5	-0.03	5	*0.07*	4	0.03	6	**0.14**	3	-0.03	4	**-0.15**	4	**0.19**	3	**0.10**	4	**-0.11**	3	**0.66**	5
+6	-0.05	3	*0.08*	3	0.04	5	**0.17**	2	-0.03	2	**-0.12**	6	**0.17**	6	**0.11**	5	**-0.10**	5	**0.62**	7
+7	*-0.07*	1	**0.09**	2	0.04	3	**0.17**	1	-0.03	3	**-0.11**	7	**0.10**	7	0.05	6	**-0.14**	1	**0.62**	6

[1]: target day of the forecast; [2]: serial number of the day in the year; [3]: daily mean temperature (°C); [4]: daily maximum temperature (°C); [5]: daily minimum temperature (°C); [6]: daily temperature range (°C); [7]: daily relative humidity (%); [8]: daily total radiation ($W \cdot m^{-2}$); [9]: daily mean air pressure (hPa); [10]: daily wind speed ($m \cdot s^{-1}$)

Table 3 Statistical evaluation of the *Ambrosia* pollen concentration forecasting models for Szeged in terms of the correlation coefficient (r), the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). T indicates the forecasting horizon (in days). (MLP: Multi-Layer Perceptron model, M5P: Regression tree model, REPTree: regression tree model, DecisionStump: decision tree model and MLPRegressor: Multi-Layer Perceptron model)

T	MLP			M5P			REPTree			DecisionStump			MLPRegressor		
(day)	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE
+1	0.99	45.08	17.11	0.98	52.34	18.60	0.95	63.69	24.25	0.75	128.86	78.95	0.97	60.89	20.28
+2	0.99	66.66	27.54	0.83	115.81	38.30	0.85	110.29	38.11	0.64	150.69	64.11	0.97	76.02	34.72
+3	0.98	80.08	33.79	0.82	117.62	46.30	0.80	123.15	45.26	0.63	152.21	64.09	0.61	153.23	56.85
+4	0.96	94.02	40.69	0.78	126.04	54.91	0.71	138.79	58.79	0.63	153.05	64.97	0.54	162.09	61.61
+5	0.94	111.30	50.43	0.59	153.11	73.12	0.69	143.99	59.74	0.62	154.44	65.28	0.42	175.27	73.71
+6	0.92	121.51	58.88	0.53	161.07	81.92	0.49	166.75	77.98	0.60	157.77	66.24	0.65	149.11	71.15
+7	0.90	127.13	63.34	0.43	172.45	83.79	0.43	174.15	75.35	0.60	157.97	66.34	0.54	161.88	80.57

Table 4 Statistical evaluation of the *Ambrosia* pollen alarm level forecasting models for Szeged in terms of the correlation coefficient (r), the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). T indicates the forecasting horizon (in days). (MLP: Multi-Layer Perceptron model, J48: decision tree model, REPTree: decision tree model, DecisionStump: decision tree model and MLPClassifier: Multi-Layer Perceptron model)

T	MLP			J48			REPTree			DecisionStump			MLPClassifier		
(day)	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE
+1	0.98	0.37	0.14	0.94	0.70	0.44	0.98	0.40	0.16	0.74	1.32	0.92	0.98	0.37	0.14
+2	0.95	0.67	0.40	0.91	0.88	0.52	0.96	0.60	0.32	0.74	1.32	0.92	0.96	0.62	0.37
+3	0.96	0.67	0.38	0.87	1.01	0.63	0.90	0.93	0.58	0.74	1.32	0.93	0.90	0.94	0.53
+4	0.80	1.19	0.73	0.85	1.08	0.73	0.91	0.85	0.54	0.74	1.32	0.93	0.81	1.22	0.70
+5	0.94	0.77	0.46	0.79	1.26	0.87	0.92	0.78	0.46	0.73	1.32	0.94	0.82	1.14	0.78
+6	0.82	1.18	0.78	0.73	1.39	0.95	0.83	1.10	074	0.73	1.32	0.94	0.86	1.05	0.67
+7	0.76	1.36	0.88	0.72	1.41	0.93	0.77	1.32	0.92	0.73	1.32	0.95	0.74	1.39	1.02

Table 5 Statistical evaluation of the *Ambrosia* pollen concentration forecasting models for Lyon in terms of the correlation coefficient (r), the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). T indicates the forecasting horizon (in days). (MLP: Multi-Layer Perceptron model, M5P: Regression tree model, REPTree: regression tree model, DecisionStump: decision tree model and MLPRegressor: Multi-Layer Perceptron model)

T	MLP			M5P			REPTree			DecisionStump			MLPRegressor		
(day)	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE
+1	0.96	33.53	12.73	0.97	28.26	11.62	0.70	48.99	15.89	0.81	45.88	18.39	0.36	62.09	22.44
+2	0.91	48.31	21.67	0.68	52.14	21.05	0.42	60.85	23.63	0.43	60.13	23.33	0.59	56.29	20.08
+3	0.81	53.59	24.74	0.64	55.12	22.76	0.57	56.42	20.21	0.43	60.88	25.20	0.33	62.86	25.72
+4	0.74	63.17	29.13	0.29	63.06	24.80	0.41	60.85	24.47	0.43	60.63	24.19	0.01	70.02	29.37
+5	0.64	58.82	26.67	0.19	65.15	26.80	0.42	59.91	22.83	0.35	62.18	25.65	-0.01	73.23	32.35
+6	0.78	55.92	24.81	0.43	59.93	23.81	0.33	62.05	24.75	0.35	62.15	25.54	0.01	72.99	32.16
+7	0.92	51.67	22.47	0.80	52.29	21.84	0.34	61.75	23.94	0.34	62.06	25.33	0.12	69.23	30.04

Table 6 Statistical evaluation of the *Ambrosia* pollen alarm level forecasting models for Lyon in terms of the correlation coefficient (r), the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). T indicates the forecasting horizon (in days). (MLP: Multi-Layer Perceptron model, J48: decision tree model, REPTree: decision tree model, DecisionStump: decision tree model and MLPClassifier: Multi-Layer Perceptron model)

T	MLP			J48			REPTree			DecisionStump			MLPClassifier		
(day)	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE
+1	0.91	1.12	0.53	0.80	1.05	0.46	0.84	0.75	0.29	-	1.48	0.69	0.73	0.97	0.40
+2	0.65	1.31	0.62	0.35	1.67	0.85	0.52	1.22	0.48	-	1.48	0.70	0.51	1.20	0.58
+3	0.26	1.60	0.80	0.17	1.66	0.90	–	1.48	0.70	-	1.48	0.70	0.44	1.29	0.60
+4	0.39	1.41	0.67	0.63	1.07	0.49	0.47	1.33	0.63	-	1.48	0.70	0.45	1.35	0.60
+5	0.26	1.45	0.70	0.37	1.32	0.72	0.65	1.11	0.51	-	1.48	0.70	0.59	1.23	0.54
+6	0.46	1.40	0.68	0.38	1.28	0.61	0.52	1.23	0.52	-	1.48	0.70	0.31	1.30	0.68
+7	-	1.48	0.71	0.38	1.49	0.75	0.48	1.32	0.62	-	1.48	0.70	0.14	1.43	0.74

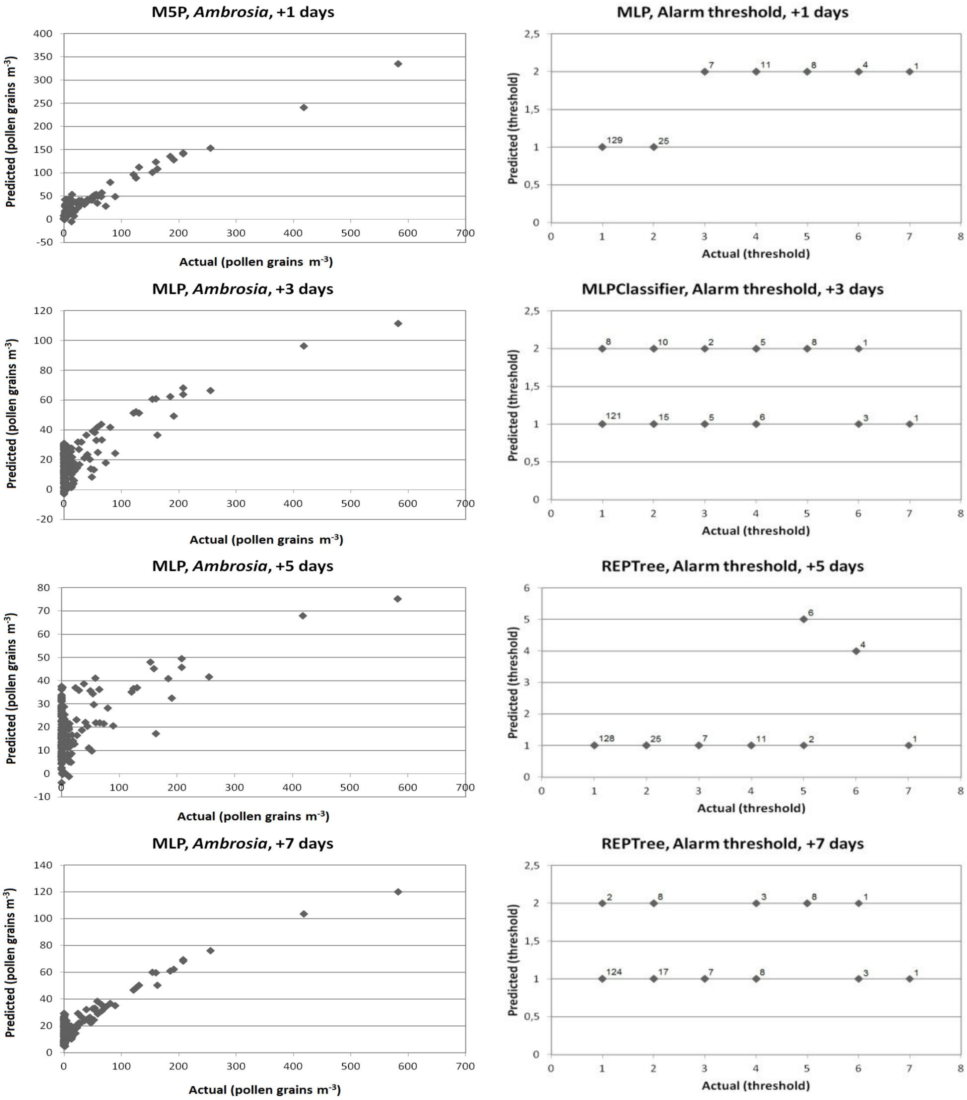


Fig. 3 Scatter plots, Lyon. Selected scatter plots of actual and predicted *Ambrosia* pollen concentrations (M5P, MLP), as well as alarm thresholds (MLP, MLPClassifier, REPTree). The forecasting horizon is given in days

This is why M5P is not a reliable method for >2 days forecasts. Based on the scatter plots, when the forecasting horizon expands, (1) the accuracy of the forecast weakens and (2) the best method (MLP) increasingly underestimates the pollen concentration (Fig. 2). Note that for the remaining methods, under- and overestimation may occur at both the beginning and end of the pollen season. However, MLP underestimates consistently regardless the day of the pollen season and the length of the forecasting horizon. On the whole, all the methods analysed in the study (except for the simplest DecisionStump) perform well for 1 and 2-day ahead forecasts for Szeged.

Note, however, that MLP provides a correlation coefficient 0.96 even for the 4-day forecast and the efficiency of the prediction does not decrease below r=0.90 even for 7-day forecast. For the remaining methods the accuracy of the forecasts for >2 days ahead indicate sharp decrease (Table 3, Fig. 2).

Predicting alarm levels is another area of pollen forecasts. Their fast and efficient prediction serves a simple and easily traceable tool for sensitive people in preparing for days of high pollen load. In order to better predict *Ambrosia* pollen alarm levels introduced for Hungary (Mányoki et al. 2011), the original 0-1 and 7-8 categories were aggregated. In the scatter plots of forecasting alarm levels for both Szeged and Lyon, the horizontal axis indicates the observed alarm level, while the vertical axis shows the forecasted alarm level. Starting from the actual day several alarm levels can be expected on the target day depending on the initial day, and the forecasts for the target day can result in different alarm levels. Note that with the increase of the forecasting horizon the uncertainty of the alarm level increases. The numbers beside the forecasted alarm levels indicate their total occurrences for the data set (Figs. 2-3).

MLP shows the best results for the alarm levels of Szeged. The decision tree based REPTree model provides better or similarly good performance compared to the MLP since alarm levels form classes for which RAPTree is very sensitive. Besides these methods the simply constructed MLPClassifier, that has a faster run-time compared to MLP, is also capable of predicting alarm levels with good performance. When forecasting 1-day alarm level, three methods (MLP, REPTree and MLPClassifier) show the same efficacy (Table 4). 1, 2 and 3-day ahead predictions of alarm levels perform well, while forecasts for >3 days ahead indicate substantial decrease for all the methods applied. Note that MLP provides good result even for a 5-day forecast, as well; whereas, the performance of DecisionStump is the worst due to the construction of the method: it carries out only one single split (Table 4, Fig. 2).

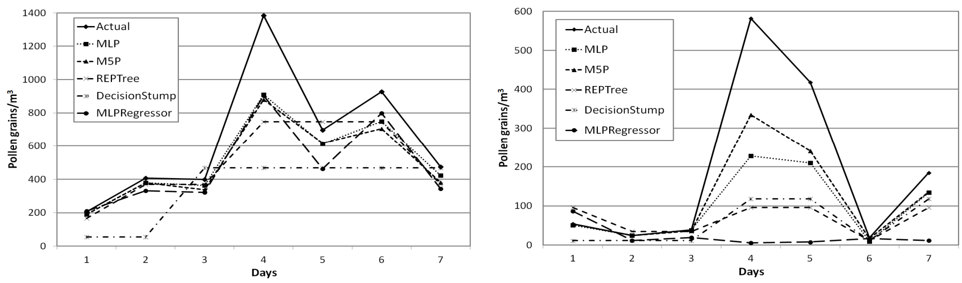
For Lyon, MLP provides the best performance of all the procedures. One-layer MLPRegressor is the least efficient and, similarly to the case of Szeged, DecisionStump is not capable of predicting alarm levels. As wind speed shows significant negative associations with the measured pollen concentrations for 1-7 days ahead (Table 2), this parameter strongly degrades the performance of the methods (Tables 5-6, Fig. 3).

The procedures perform well for Szeged, but they are not really efficient for Lyon. For the latter case, neither pollen concentrations nor alarm levels show a definite annual course, due to the substantially smaller pollen concentrations, furthermore different climate and relief in Lyon compared to those of Szeged (Tables 5-6). The predictability of alarm levels for Lyon is quite weak that can be explained with the following reasons: (1) alarm levels introduced for Hungary cannot be applied well for Lyon due to the different distribution of pollen concentrations for the two cities, (2) the structure of the association between the influencing and resultant variables are different for Szeged and Lyon (Tables 1-2, Tables 5-6, Fig. 3).

Uncertainties in the accuracy of the forecasts can be explained by the lack of sufficient number of influencing variables including the fact that environmental associations of ragweed pollen level have not been fully discovered yet. For example, high air pollutant concentrations are likely to have either short or long term impact on pollen levels (Minero et al. 1998; Jäger et al. 1991), especially in a polluted urban environment like Szeged and Lyon. The results show that the learning strategies of the algorithms can perform well, but the really good model is MLP for predicting both pollen concentrations and alarm levels for each city. Based on the results for Szeged and Lyon we can perform accurate forecasts of the daily pollen concentrations and alarm levels for several days ahead. The efficiency of the models belongs to the best ones compared to those reported in the literature. When forecasting, the following values of r^2 (i.e. squared correlations) of one day ahead forecasts were received: 0.60 for Poaceae using neural networks (Sánchez-Mesa et al. 2002); 0.93 again for Poaceae using neural networks (Rodríguez-Rajo et al. 2010); 0.45 for grass pollen (whole season) using correlation analysis (Stach et al. 2008) and 0.79 for Poaceae using Multiple Linear Regression (Voukantsis et al. 2010). Our study provides a coefficient of determination of 0.98 (*Ambrosia,* Szeged, one day and two days ahead) using Multi-Layer Perceptron that ranks this model the best one in the literature.

3.3. Model fitting on the days of the highest pollen levels

Pollen concentrations on the days exhibiting the highest pollen levels during a 7-day period were predicted and analysed for both cities (Fig. 4).



Szeged Lyon

Fig. 4 One-day forecasts for a seven-day period encompassing the day of the highest pollen load of *Ambrosia* (Actual: measured pollen concentrations, MLP: Multi-Layer Perceptron model, M5P: regression tree model, REPTree: decision tree model, DecisionStump: decision tree model, MLPRegressor: Multi-Layer Perceptron model)

For example, regarding the absolute maximum pollen counts within the 10-year period examined, for Szeged and Lyon the best 1-day forecast is provided by MLP (actual value: 1385 pollen grains m^{-3}; forecasted value: 910 pollen grains m^{-3}) and M5P (actual value: 582 pollen grains m^{-3}; forecasted value: 335 pollen grains m^{-3}), respectively. However, all methods underestimate the pollen concentrations in these episodic situations.

The message of the above experiment is that MLP, M5P and MLPRegressor follow well the annual course of the pollen concentration. This is important information as the usefulness of a good forecast is much higher for the days of the highest pollen concentrations than for those of small pollen levels at the beginning and end of the pollen season. Accordingly, these methods can help in developing personalized information services that could improve the overall quality of life for sensitized people.

4. CONCLUSIONS

We applied Computational Intelligence procedures in order to predict daily values of *Ambrosia* pollen concentrations and alarm levels for Szeged (Hungary) and Lyon (France). Despite the difficulties in the availability of daily pollen levels (they are at disposal only after a week), forecasts of daily ragweed pollen concentrations and alarm levels were successful for 1-7 days ahead for both cities. The importance of the influencing variables (the serial number of the day in the year, meteorological and pollen variables) in forming the resultant variable (pollen levels or alarm levels for 1-7 days ahead) was analysed. The weights of *Ambrosia* pollen level emerge extraordinarily from all variables indicating its high significance in determining pollen levels (alarm levels) for 1-7 days ahead for both cities. The weights of the rest of influencing variables are different for the two cities. For instance, the most important variables are temperature-related ones for Szeged, while relative humidity and wind speed have the most important role in forming pollen concentrations in Lyon.

For Szeged, Multi-Layer Perceptron models provide results similar with tree-based models for predicting pollen concentration 1 and 2-days ahead, while for more than two days ahead they deliver better results than tree-based models. For Lyon, only the Multi-Layer Perceptron gives acceptable result for predicting pollen levels 1 and 2-days ahead. Concerning the alarm levels, the efficiency of the procedures differs substantially.

When fitting the models to the days of the highest pollen levels the more complex CI methods proved better for both cities. MLP and M5P methods provided the best results for Szeged and Lyon, respectively. We have shown that the selection of the optimal method depends on climate as a function of geographical location and relief.

Results received can be utilized by the national pollen information services. Total medical costs of ragweed pollen can be substantially reduced if sensitized people can be prepared in time for serious ragweed pollen episodes. Decision-makers are responsible for introducing regulations and actions in order to facilitate the problem caused by ragweed pollen. Furthermore, it is the responsibility of aero-biologists to develop personalized information services in order to improve the overall quality of life of sensitized people. Note however, that due to the restrictions of the pollen sampling procedure (daily pollen counts are available after a 7-day period) the applicability of the present or any other statistical models for operative pollen forecast is limited in time. This problem can only be solved if instruments based on a totally new principle will be introduced measuring “in situ” pollen counts.

The methods applied are sensitive to the number of the influencing parameters. A further aim is to use much more influencing parameters (including further meteorological parameters, in addition chemical air pollutants, land use, relief, etc.) in order to develop a general model for different locations.

Acknowledgements: The authors would like to thank Gábor Motika (Environmental Conservancy Inspectorate, Szeged, Hungary) for providing meteorological data of Szeged, Miklós Juhász (University of Szeged) for providing daily pollen concentration data of Szeged and Zoltán Sümeghy for the digital mapping in Fig. 1. This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/2-11-1-2012-0001 ‘National Excellence Program’.

REFERENCES

Ariano R, Canonica GW, Passalacqua G (2010) Possible role of climate changes in variations in pollen seasons and allergic sensitizations during 27 years. Ann Allerg Asthma Im 104:215-222

Aznarte JL, Sánchez JMB, Lugilde DN, Fernández CDL, de la Guardia CD, Sánchez FA (2007) Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. Expert Syst Appl 32:1218-1225

Béres I, Novák R, Hoffmanné Pathy Zs, Kazinczi G (2005) Az ürömlevelű parlagfű (Ambrosia artemisiifolia L.) elterjedése, morfológiája, biológiája, jelentősége és a védekezés lehetőségei. [Distribution, morphology, biology and importance of common ragweed (Ambrosia artemisiifolia L.) and protection facilities. (in Hungarian)] Gyomnövények, Gyomirtás 6:1-48

Bonini M, Albertini R, Brighetti MA, Ugolotti M, Travaglini A (2012) RIMA-Italian Monitoring Network in Aerobiology. Ragweed pollen spreading in Italy. Second International Ragweed Conference, Lyon, France

Bullock JM, Chapman D, Schafer S, Roy D, Girardello M, Haynes T, Beal S, Wheeler B, Dickie I, Phang Z, Tinch R, Čivić K, Delbaere B, Jones-Walters L, Hilbert A, Schrauwen A, Prank M, Sofiev M, Niemelä S, Räisänen P, Lees B, Skinner M, Finch S, Brough C (2010) Assessing and controlling the spread and the effects of common ragweed in Europe. Final report: ENV.B2/ETU/2010/0037, Natural Environment Research Council, UK, 456

Castellano-Méndez M, Aira MJ, Iglesias I, Jato V, González-Manteiga W (2005) Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. Int J Biometeorol 49:310-316

Chauvel B, Dessaint F, Cardinal-Legrand C, Bretagnolle F (2006) The historical spread of Ambrosia artemisiifolia L. in France from herbarium records. J Biogeogr 33:665-673

Cristofori A, Cristofolini F, Gottardini E (2010) Twenty years of aerobiological monitoring in Trentino (Italy): assessment and evaluation of airborne pollen variability. Aerobiologia 26:253-261

Delaunay JJ, Seymour C, Fouillet V (2004) Investigation of short-range cedar pollen forecasting. Phys Rev E 70: Article No. 066214

Frei T (2008) Climate change and its impact on airborne pollen in Basel, Switzerland 1969-2007. Allergologie 31:165-169

Gladieux P, Giraud T, Kiss L, Genton BJ, Jonot O, Shykoff JA (2011) Distinct invasion sources of common ragweed (Ambrosia artemisiifolia) in Eastern and Western Europe. Biol Invasions 13:933-944

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explorations 11:10-18

Haykin S (1999) Neural Networks: a Comprehensive Foundation, 2nd ed. Upper Saddle River, Prentice Hall, NJ

IPCC (2013) Summary for policymakers. Climate Change 2013. Stocker TF, Qin D,Plattner GP, Tignor MMB, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, (Cambridge University Press, Cambridge, UK).

Jahn W, Vahle H (1968) Die Faktoranalyse und ihre Anwendung. Verlag die Wirtschaft, Berlin

Jäger S, Spieksma FTM, Nolard N (1991) Fluctuations and trend in airborne concentrations of some abundant pollen types, monitored at Vienna, Leiden and Brussels. Grana 30:309-312

Jolliffe IT (1993) Principal component analysis: A beginner's guide – II. Pitfalls, myths and extensions. Weather 48:246-253

Juhos I, Makra L, Tóth B (2009) The behaviour of the multi-layer perceptron and the support vector regression learning methods in the prediction of NO and NO_2 concentrations in Szeged, Hungary. Neural Comput Appl 18:193-205

Kiss L, Béres I (2006) Anthropogenic factors behind the recent population expansion of common ragweed (*Ambrosia artemisiifolia L.*) in Eastern Europe: is there a correlation with political transitions? J Biogeogr 33**:**2156-2157

Köppen W (1931) Grundriss Der Klimakunde. Walter De Gruyter & Co, Berlin

Laaidi M, Chinet T, Aegerter P (2011) Pollen allergies, pollution and climate: Literature review. Revue Française D'Allergologie 51:622-628

Makra L, Juhász M, Béczi R, Borsos E (2005) The history and impacts of airborne Ambrosia (Asteraceae) pollen in Hungary. Grana 44:57-64

Makra L, Sánta T, Matyasovszky I, Damialis A, Karatzas K, Bergmann KC, Vokou D (2010) Airborne pollen in three European cities: Detection of atmospheric circulation pathways by applying three-dimensional clustering of backward trajectories. J Geophys Res-Atmos 115:D24220

Makra L, Matyasovszky I, Thibaudon M, Bonini M (2011) Forecasting ragweed pollen characteristics with nonparametric regression methods over the most polluted areas in Europe. Int J Biometeorol 55:361-371

Makra L, Matyasovszky I (2011) Assessment of the daily ragweed pollen concentration with previous-day meteorological variables using regression and quantile regression analysis for Szeged, Hungary. Aerobiologia 27:247-259

Mányoki G, Apatini D, Magyar D, Páldy A (2011) A parlagfűpollen becsült országos eloszlása a Parlagfű Pollen Riasztási Rendszer (PPRR) szerint. (Assessed incidence of ragweed in Hungary according to the Ragweed Pollen Alarm System (RPAS). In: Apatini D (ed) Az ÁNTSZ Aerobiológiai Hálózatának tájékoztatója, éves jelentés, kézirat. [Report of the Aerobiological Network of ÁNTSZ, annual report (in Hungarian)], manuscript. OKI, Budapest, 81

Minero FJG, Iglesias I, Jato V, Aira MJ, Candau P, Morales J, Tomas C (1998) Study of the pollen emissions of Urticaceae, Plantaginaceae and Poaceae at five sites in western Spain. Aerobiologia 14:117-129

Puc M (2012) Artificial neural network model of the relationship between Betula pollen and meteorological factors in Szczecin (Poland). Int J Biometeorol 56:395-401

Quinlan RJ (1992) Learning with continuous classes. In: Proc. of the 5th Australian Joint Conference on Artificial Intelligence, Singapore, 343-348

Quinlan RJ (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA

Recio M, Docampo S, García-Sánchez J, Trigo MM, Melgar M, Cabezudo B (2010) Influence of temperature, rainfall and wind trends on grass pollination in Malaga (western Mediterranean coast). Agr Forest Meteorol 150:931-940

Reznik S (2009) Common ragweed (Ambrosia artemisiifolia L.) in Russia: spread, distribution, abundance, harmfulness and control measures. Ambroisie, The first international ragweed review 26

Rodinkova V, Palamarchuk O, Kremenska L (2012) The most abundant Ambrosia pollen count is associated with the southern, eastern and the northern-eastern Ukraine. Alergologia et Immunologia 9:181

Rodríguez-Rajo FJ, Astray G, Ferreiro-Lage JA, Aira MJ, Jato-Rodriguez MV, Mejuto JC (2010) Evaluation of atmospheric Poaceae pollen concentration using a neural network applied to a coastal Atlantic climate region. Neural Networks 23:419-425

Rodríguez-Rajo FJ, Aira MJ, Fernandez-Gonzalez M, Seijo C, Jato V (2011) Recent trends in airborne pollen for tree species in Galicia, NW Spain. Clim Res 48:281-291

Sánchez-Mesa JA, Galán C, Martínez-Heras JA, Hervás-Martínez C (2002) The use of a neural network to forecast daily grass pollen concentration in a Mediterranean region: the southern part of the Iberian Penisula. Clin Exp Allergy 32:1606-1612

Sánchez Mesa JA, Galán C, Hervás C (2005) The use of discriminant analysis and neural networks to forecast the severity of the Poaceae pollen season in a region with a typical Mediterranean climate. Int J Biometeorol 49:355-362

Stach A, García-Mozo H, Prieto-Baena JC, Czarnecka-Operacz M, Jenerowicz D, Silny W, Galán C (2007) Prevalence of Artemisia species pollinosis in western Poland: Impact of climate change on aerobiological trends, 1995-2004. J Invest Allerg Clin 17:39-47

Stach A, Smith M, Baena JCP, Emberlin J (2008) Long-term and short-term forecast models for Poaceae (grass) pollen in Poznań, Poland, constructed using regression analysis. Environ Exp Bot 62:323-332

Vinogradova YR, Majorov SR, Khorun LV (2010) Black Book of Central Russia: Alien Species of flora of Central Russia (in Russian). Moscow: GEOS

Vlachogianni A, Kassomenos P, Karppinen A, Karakitsios S, Kukkonen J (2011) Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. Sci Total Environ 409:1559-1571

Voukantsis D, Niska H, Karatzas K, Riga M, Damialis A, Vokou D (2010) Forecasting daily pollen concentrations using data-driven modeling methods in Thessaloniki, Greece. Atmos Environ 44:5101-5111

Voukantsis D, Karatzas K, Kukkonen J, Rasanen T, Karppinen A, Kolehmainen M (2011) Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci Total Environ 409:1266-1276

Voukantsis D, Karatzas K, Jaeger S, Berger U, Smith M (2013) Analysis and forecasting of airborne pollen-induced symptoms with the aid of computational intelligence methods. Aerobiologia 29:175-185

Ziska L, Knowlton K, Rogers C, Dalan D, Tierney N, Elder MA, Filley W, Shropshire J, Ford LB, Hedberg C, Fleetwood P, Hovanky KT, Kavanaugh T, Fulford G, Vrtis RF, Patz JA, Portnoy J, Coates F, Bielory L, Frenz D (2011) Recent warming by latitude associated with increased length of ragweed pollen season in central North America. P Natl Acad Sci USA 108:4248-4251