

Statistical techniques for modeling of *Corylus*, *Alnus*, and *Betula* pollen concentration in the air

Jakub Nowosad · Alfred Stach · Idalia Kasprzyk · Kazimiera Chłopek · Katarzyna Dąbrowska-Zapart · Łukasz Grewling · Małgorzata Latałowa · Anna Pędziszewska · Barbara Majkowska-Wojciechowska · Dorota Myszkowska · Krystyna Piotrowska-Weryszko · Elżbieta Weryszko-Chmielewska · Małgorzata Puc · Piotr Rapiejko · Tomasz Stosik

Jakub Nowosad
Space Informatics Lab,
University of Cincinnati,
219 Braunstein Hall, Cincinnati, OH 45221, USA
E-mail: nowosad.jakub@gmail.com

Alfred Stach
Institute of Geoecology and Geoinformation,
Adam Mickiewicz University,
Krygowskiego 10, 61-680 Poznań, Poland

Idalia Kasprzyk
Department of Ecology and Environmental Biology,
University of Rzeszów,
Zelwerowicza 4, 35-601 Rzeszów, Poland

Kazimiera Chłopek
Faculty of Earth Sciences,
University of Silesia,
Będzińska 60, 41-200 Sosnowiec, Poland

Katarzyna Dąbrowska-Zapart
Faculty of Earth Sciences,
University of Silesia,
Będzińska 60, 41-200 Sosnowiec, Poland

Łukasz Grewling
Laboratory of Aeropalynology, Faculty of Biology,
Adam Mickiewicz University,
Umultowska 89, 61-614 Poznań, Poland

Małgorzata Latałowa
Department of Plant Ecology,
University of Gdańsk,
Wita Stwosza 59, 80-308 Gdańsk, Poland

Anna Pędziszewska
Department of Plant Ecology,
University of Gdańsk,
Wita Stwosza 59, 80-308 Gdańsk, Poland

Barbara Majkowska-Wojciechowska
Department of Immunology, Rheumatology and Allergy, Faculty of Medicine,
Medical University,
Pomorska 251, 92-215 Łódź, Poland

Dorota Myszkowska
Department of Clinical and Environmental Allergology,
Jagiellonian University Medical College,
Śniadeckich 10, 31-531 Kraków, Poland

Krystyna Piotrowska-Weryszko
Department of Botany,
University of Life Sciences in Lublin,
Akademicka 15, 20-950 Lublin, Poland

Elżbieta Weryszko-Chmielewska
Department of Botany,
University of Life Sciences in Lublin,
Akademicka 15, 20-950 Lublin, Poland

Małgorzata Puc

Abstract Prediction of allergic pollen concentration is one of the most important goals of aerobiology. Past studies have used a broad range of modeling techniques; however, the results cannot be directly compared owing to the use of different datasets, validation methods, and evaluation metrics. The main aim of this study was to compare nine statistical modeling techniques using the same dataset. An additional goal was to assess the importance of predictors for the best model. Aerobiological data for *Corylus*, *Alnus*, and *Betula* pollen counts were obtained from nine cities in Poland and covered between five and 16 years of measurements. Meteorological data from the AGRI4CAST project was used as a predictor variables. The results of 243 final models (3 taxa x 9 cities x 9 techniques) were validated using a repeated k-fold cross-validation and compared using relative and absolute performance statistics. Afterwards, the variable importance of predictors in the best models were calculated and compared. Simple models performed poorly. On the other hand, regression trees and rule-based models proved to be the most accurate for all of the taxa. Cumulative growing degree days proved to be the single most important predictor variable in the random forest models of *Corylus*, *Alnus*, and *Betula*. Finally, the study suggested potential improvements in aerobiological modeling, such as the application of robust cross-validation techniques and the use of gridded variables.

Keywords Allergenic pollen · Pollen concentration in the air · Betulaceae · Regression models · Predictive modeling · Machine learning

Introduction

Modeling and forecasting of pollen concentration and pollen season properties are among the most important goals of aerobiology. Models are used to provide better understanding and broaden the knowledge of pollen release and dispersion. Such models could also be used for prediction purposes; therefore, their results would be useful for allergists and their patients.

Two main groups of models—numerical and statistical—are used in aerobiological studies. Numerical models are based on mathematical equations and algorithms of atmospheric dispersion. They estimate pollen concentration using information about the distribution of pollen sources and phenological, aerobiological, and meteorological data (Vogel et al, 2008; Sofiev et al, 2013b). On the other hand, statistical models determine the relationship between dependent variables (such as pollen data) and one or more independent variables. Statistical models in aerobiology describe the numerical relations between pollen characteristics and explanatory variables, and they aim to predict the pollen concentration or pollen season properties. Importantly, statistical models do not require an understanding of the physical processes of pollen emission and dispersion.

Several studies using statistical modeling and forecasting of *Corylus*, *Alnus*, and *Betula* pollen concentrations properties were conducted in the past. Multiple regression was used by Bringfelt et al (1982); Ritenberga et al (2016) to predict daily pollen concentrations of *Betula*, and by Laaidi (2001); Emberlin et al (1993); Myszkowska (2013) to model *Betula* pollen season characteristics. Daily pollen concentration levels were predicted by Cotos-Yáñez et al (2004), who used a generalized additive model and a partially linear model on data from Vigo (Spain), and by Castellano-Méndez et al (2005), who used artificial neural networks on data from Santiago de Compostela (Spain). Puc (2012) used an artificial neural networks technique to model daily pollen concentrations of *Betula* in Szczecin (Poland). *Alnus* pollen concentration was predicted by Rodríguez-Rajo et al (2006) using ARIMA in four cities in northeastern Spain. Hilaire et al (2012) built models for daily pollen concentrations of *Alnus* and *Betula* using stochastic gradient boosting in Switzerland. Nowosad (2016); Nowosad et al (2016) created predictive models for *Corylus*, *Alnus*, and *Betula* using a random forest technique in Poland.

A validation was performed in most of these studies, with the exception of Bringfelt et al (1982). Different model validation techniques and different measures of models performance were used. Puc (2012) validated the results on 15% of randomly chosen days; Hilaire et al (2012) used the most recent 25% of the data; Nowosad (2016) used a stratified random split of 1/3 of the data; and Nowosad et al (2016) created two testing sets—for temporal and spatial validation. In the remaining studies, either one, two, or three years of data was used as a validation set. In addition, the set of independent variables differs between these studies. Overall, therefore, it is impossible to explicitly compare the performance of these models against each other.

Department of Botany and Nature Conservation, Faculty of Biology,
University of Szczecin,
Felczaka 3c, 71-412 Szczecin, Poland

Piotr Rapięjko
Allergen Research Center,
Kalinowej Łąki 8, 01-934 Warszawa, Poland

Tomasz Stosik
Department of Botany and Ecology,
University of Science and Technology,
Kaliskiego 7, 85-796 Bydgoszcz, Poland

The main goals of this study were to compare the predictive modeling techniques using one dataset for each taxa and to assess the variable importance of the best models.

Materials and methods

Data

Aerobiological data

Aerobiological sampling was performed in nine cities in Poland (Bydgoszcz, Gdańsk, Kraków, Łódź, Lublin, Poznań, Rzeszów, Sosnowiec, Szczecin) and covered between five and 16 years of measurement (Fig. 1). More information on the study area can be found in the Material and Methods section in Nowosad et al (2016).

A volumetric spore-trap of the Hirst design was used at all sites (Hirst, 1952). The pollen grains of *Corylus*, *Alnus*, and *Betula* were counted in accordance with the method recommended by the European Aerobiology Society's Working Group on Quality Control (Galán et al, 2014), and the values were expressed as the number of grains/m³ of air per 24 h.



Fig. 1 Sites used for the study of forecasting the daily pollen concentrations in Poland. A time period of aerobiological measurements for each site is shown in parentheses

Grid data

AGRI4CAST Interpolated Meteorological Data was used as a source of the daily meteorological data (Baruth et al, 2007). This contains meteorological parameters interpolated to a 25x25-km grid for the European Union member states, neighboring European countries, and the Mediterranean countries. Meteorological variables include temperature, vapor pressure, wind speed, precipitation, evaporation, radiation, and snow depth. For the purpose of this study, the meteorological data of the grid cells containing the sites analyzed were used.

Methods

One of the main goals of this study was to evaluate modeling techniques used for the prediction of *Corylus*, *Alnus*, and *Betula* pollen concentrations in the air. The workflow for each taxon and location was as follows:

1. Ten independent, meteorological variables were extracted from the grid cell where the monitoring station was located. In addition, an eleventh independent variable (cumulated growing degree days) was calculated for the same grid cell.
2. Dependent (aerobiological) and independent variables were combined into one dataset.
3. Nine modeling techniques were used to build models. Models were validated using a repeated k-fold cross-validation procedure.
4. Final models were compared in terms of predictive performance.
5. For the best models, cluster analysis was performed based on performance statistics.
6. For the best models, independent variable importance was calculated and compared.

All the calculations were performed using R (R Core Team, 2016) and R packages (Kuhn, 2016; Wickham, 2009; Golemund and Wickham, 2011). Models were built using pls (Mevik et al, 2015), elasticnet (Zou and Hastie, 2008), elmNN (Gosso, 2012), kernlab (Karatzoglou et al, 2004), earth (Milborrow, 2016), rpart (Therneau et al, 2015), randomForest (Liaw and Wiener, 2002), and Cubist (Kuhn et al, 2014) packages.

Predictor variables

Eleven meteorological parameters from the same day as pollen concentration values were used as independent variables and the daily pollen concentration was used as a dependent variable. Ten meteorological properties (maximum temperature, minimum temperature, average temperature, vapor pressure, wind speed, sum of precipitation, potential evaporation from a free water surface, potential evapotranspiration from a crop canopy, potential evaporation from a moist bare soil surface, and total global radiation) are available in the AGRI4CAST Interpolated Meteorological Data. An additional property, cumulated growing degree days (GDD), was calculated as follows:

$$\text{Daily GDD} = \frac{T_{max} + T_{min}}{2} - T_{base}$$

where T_{max} is the daily maximum temperature, T_{min} is the daily minimum temperature, and T_{base} is the base temperature. A value of 5°C was used as the base temperature. This value is the standard threshold temperature for growth in temperate species (Dahl et al, 2013). The cumulated GDD were calculated as the sum of degree days from January 1. If the daily mean temperature (calculate as $(T_{max} + T_{min})/2$) is higher than the base temperature, then degree days accumulate.

Regression models

Nine modeling techniques were used to predict the pollen concentrations of *Corylus*, *Alnus*, and *Betula*. These techniques can be divided into three groups: (i) linear regression models; (ii) nonlinear regression models; and (iii) regression trees and rule-based models:

Linear regression:

- Linear Model (LM) (Nelder and Wedderburn, 1972)
- Partial Least Square (PLS) (Wold et al, 1983)
- The Lasso (Tibshirani, 1996)

Nonlinear regression models:

- Neural Networks (NN) (Bishop, 1995)
- Support Vector Machines (SVM) (Drucker et al, 1997)
- Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991)

Regression trees and rule-based models:

- Basic Regression Tree (BRT) (Breiman et al, 1984)
- Random Forest (RF) (Breiman, 2001)
- Cubist (Kuhn and Johnson, 2013)

Model validation and comparison

A repeated k-fold cross-validation was used to obtain the best combinations of algorithms' parameters and to assess the accuracy of the models (Kuhn and Johnson, 2013). For each city, data was divided into yearly subsets. One of the yearly data subsets was omitted, while the other k-1 data was used to train the model. The omitted subset was predicted, and the prediction was summarized as the coefficient of determination (r^2), mean absolute error (MAE), and symmetric mean absolute percentage error (SMAPE). The validation procedure was repeated for each year of data and the k estimates of performance for each combination of parameters were averaged. Finally, the optimal model was determined as the one with the highest coefficient of determination (r^2).

The quality metrics were selected to describe different aspects of the models' performance. An r^2 value is the squared correlation coefficient between the observed and predicted value. It ranges between 0 and 1, and thus allows for comparison between models of different taxa. However, it does not describe the size of error. SMAPE also allows for comparison between models of different taxa, but focuses on the differences between predicted and actual values. It measures the performance of models in relative terms. A modified version of SMAPE (Makridakis, 1993) was calculated as follows:

$$SMAPE = \frac{\sum_{t=1}^n |F_t - A_t|}{\sum_{t=1}^n (F_t + A_t)}$$

where F_t is a predicted value for day t , and A_t is an observed value for day t . The MAE is an average of the absolute errors and therefore can be only used for comparison between models of the same taxon. The advantage of this metric is that it is on the same scale of data being measured.

Error analysis

Partitioning Around Medoids (PAM) (Kaufman and Rousseeuw, 2005) was used as a clustering algorithm. The L method (Salvador and Chan, 2004) was applied to determine the optimal number of clusters. For each taxon and using the best model:

1. Values of r^2 , MAE, and SMAPE were centered and scaled.
2. A distance matrix was computed for each pollen season in each city (combination of site and year) using Euclidean distance.
3. The optimal number of clusters were determined using the L method based on the total within-clusters sum of squares.
4. PAM clustered combinations of site and year in the number of clusters were given by the L method.
5. A medoid (the most representative object) was selected for each cluster and visualized by comparing the time-series of observed (measured) values and predicted values of pollen concentration.

Afterwards, a PERMANOVA (Anderson, 2001) test was used to verify if there was a difference in average values of meteorological parameters between clusters.

Variable importance

The general effect of independent variables on the *Corylus*, *Alnus*, and *Betula* pollen concentration models was determined using permutation importance (mean decrease in accuracy) (Breiman, 2001; Liaw and Wiener, 2002). Based on the best modeling technique, values of variable importance were obtained separately for each model and scaled to have a maximum value of 100. Afterwards, for each taxon the mean and standard errors of variable importance were calculated.

Results

Performance of the models

The performances of final models were compared using the r^2 , MAE, and SMAPE. The comparison revealed several patterns (Fig. 2). Firstly, random forest gave the overall highest average value of r^2 (0.39) and the lowest SMAPE (0.56). Random forest had the smallest value of SMAPE in 20 models and the highest value of r^2 in 19 models. Its result was comparable to the cubist models, which had an average r^2 of 0.35 (the highest r^2 in 5 models) and an average SMAPE of 0.57 (the lowest SMAPE in 6 models). Multivariate adaptive regression splines and basic regression tree average performances were moderate, with an r^2 of 0.33 and 0.29 and SMAPE of 0.83 and 0.59, respectively. Basic regression tree gave the highest r^2 and the smallest SMAPE in one model. Multivariate adaptive regression splines was the best in terms of r^2 in two models. The rest of the models

(neural networks, lasso, linear model, support vector machines, and partial least square) performed poorly. Their average r^2 values were 0.08–0.13, and their average SMAPE values were 0.89–0.94.

Random forest models gave the best average model performance for all of the taxa analyzed. *Corylus* random forest models had an average r^2 of 0.38; *Alnus* had an average r^2 of 0.36; and *Betula* had an average r^2 of 0.41. However, there were differences between model performances at the sites studied. The r^2 of *Corylus* random forest models varied between 0.12 at Bydgoszcz and 0.50 at Rzeszów and Kraków. *Alnus* random forest models gave r^2 between 0.22 at Bydgoszcz and 0.48 at Sosnowiec. The results of *Betula* models were more stable, with r^2 between 0.31 at Sosnowiec and 0.51 at Bydgoszcz (Table 1).

MAE is scale-dependent accuracy measures, and therefore its results are not comparable between taxa. However, its values can be used for a comparison of modeling techniques. MAE gave low values in all *Corylus* models. Average values of MAE in *Corylus* models were between 4.4 (linear model, partial least square, lasso) and 2.6 (cubist). On the other hand, values of MAE separated *Alnus* models into two groups: those with values of approximately 29 (linear model, partial least square, lasso, neural networks, and multivariate adaptive regression splines), and those with values of approximately 19 (neural networks, basic regression tree, random forest, cubist). *Betula* models followed a similar pattern. The linear model, partial least square, lasso, and neural networks had the highest MAE value (approx. 87), while the values of the cubist were the lowest (53.1).

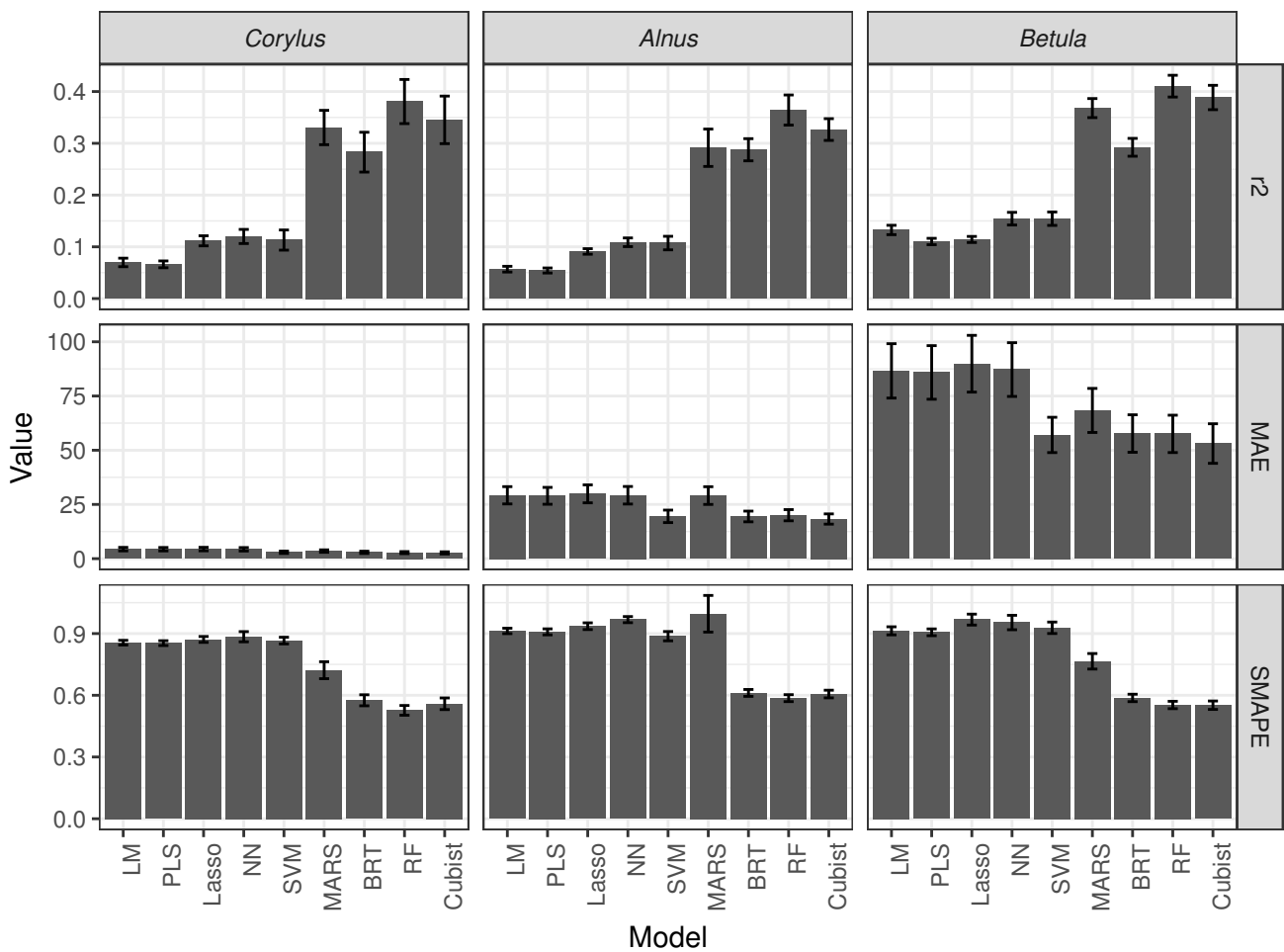


Fig. 2 A comparison of *Corylus*, *Alnus*, and *Betula* models' performance. The height of bars shows the mean value of the models performance statistic for all of the sites. Error bars represent one standard error

Error analysis

Model performance statistics (r^2 , MAE, SMAPE) of random forest models were clustered using the Partitioning Around Medoids (PAM) method (Fig. 3). The optimal number of clusters was chosen using the L-method for each taxon, based on the total within-cluster sum of squares.

Three clusters were extracted from the *Corylus* model results for each site and year (107 objects). The first cluster (23%) consists of situations with the lowest values of r^2 and the highest values of SMAPE. In this

Table 1 A summary of *Corylus*, *Alnus*, and *Betula* random forest models of pollen concentration in the air, at each location

Taxon	Site	R2	MAE	SMAPE
<i>Corylus</i>	Rzeszów	0.50	2.59	0.46
<i>Corylus</i>	Kraków	0.50	2.91	0.45
<i>Corylus</i>	Sosnowiec	0.47	3.12	0.44
<i>Corylus</i>	Lublin	0.46	5.99	0.51
<i>Corylus</i>	Łódź	0.43	1.90	0.51
<i>Corylus</i>	Poznań	0.38	1.76	0.53
<i>Corylus</i>	Gdańsk	0.31	1.32	0.62
<i>Corylus</i>	Szczecin	0.26	3.52	0.56
<i>Corylus</i>	Bydgoszcz	0.12	1.86	0.64
<i>Alnus</i>	Sosnowiec	0.48	13.55	0.54
<i>Alnus</i>	Rzeszów	0.45	19.66	0.56
<i>Alnus</i>	Kraków	0.41	10.87	0.56
<i>Alnus</i>	Lublin	0.40	32.27	0.55
<i>Alnus</i>	Poznań	0.39	30.54	0.60
<i>Alnus</i>	Łódź	0.37	23.89	0.63
<i>Alnus</i>	Szczecin	0.31	21.12	0.54
<i>Alnus</i>	Gdańsk	0.25	11.95	0.61
<i>Alnus</i>	Bydgoszcz	0.22	16.51	0.69
<i>Betula</i>	Bydgoszcz	0.51	28.93	0.47
<i>Betula</i>	Kraków	0.48	27.87	0.54
<i>Betula</i>	Poznań	0.42	63.20	0.51
<i>Betula</i>	Łódź	0.42	83.15	0.57
<i>Betula</i>	Gdańsk	0.41	32.68	0.54
<i>Betula</i>	Rzeszów	0.40	43.52	0.59
<i>Betula</i>	Lublin	0.39	94.43	0.54
<i>Betula</i>	Szczecin	0.34	58.50	0.56
<i>Betula</i>	Sosnowiec	0.31	85.58	0.66

cluster, the temporal scope of a pollen season is predicted with good agreement; however, models overestimate or underestimate pollen concentration. Cases in the second cluster (40%) show an average performance. The third cluster (37%) has the best values of model performance statistics. Its most representative object (medoid) is the model for Rzeszów in 1997, with an r^2 of 0.79, MAE of 2.1, and SMAPE of 0.38.

The performance of *Alnus* models of pollen concentration were more heterogeneous over five clusters. The first cluster (17%; medoid - Rzeszów 2011) consists of the lowest r^2 and the highest SMAPE. The second cluster (27%) has medium values of r^2 , but high values of SMAPE and MAE. Predicted values are underestimated or overestimated, but follow true changes in values. The third cluster (14%) has high values of r^2 and medium values of SMAPE; however, its values of MAE are high, and the predicted values are underestimated. Cases in the fourth cluster (13%) have medium values of r^2 and SMAPE and low values of MAE. They occurred primarily in seasons with low annual values of *Alnus* pollen concentration. The last cluster (29%; medoid - Rzeszów 1997) contains the cases with the highest model accuracy. In this cluster, predicted values follow true values closely, even in seasons with rapid changes in pollen concentration.

Four clusters were obtained for *Betula* models. The first cluster (23%) has the worst values of model performance statistics. The second (31%) and third (27%) clusters have similar r^2 and SMAPE values, although their MAE values differ greatly. Observations in the second cluster have extreme values of pollen concentration and therefore they are underestimated by the model, while small values are overestimated. The third cluster consists primarily of pollen seasons with low or medium annual pollen concentration values and only single extreme events. The last cluster (19%) has seasons with the highest prediction accuracy. Its medoid (Gdańsk) has an r^2 of 0.77, SMAPE of 0.4, and MAE of 32.1.

PERMANOVA was used to test for differences between average values of independent variables in clusters for each site and taxon. Tests showed significant differences in meteorological parameter values between clusters in four cities for *Corylus* models, in six cities for *Alnus* models, and in five cities for *Betula* models (Table 2).

Table 2 Results of PERMANOVA tests of average meteorological parameter value differences between the quality clusters of the models for each taxon and site. Statistically significant results ($p < 0.05$) are given in boldface

Site	<i>Corylus</i>	<i>Alnus</i>	<i>Betula</i>
Bydgoszcz	0.0026	0.1003	0.169
Gdańsk	0.0964	4e-04	1e-04
Kraków	0.1721	0.011	0.2699
Lublin	0.2726	0.0949	0.084
Łódź	0.0019	2e-04	0.0071
Poznań	0.5967	2e-04	0.0224
Rzeszów	0.0011	7e-04	0.0016
Sosnowiec	0.0898	0.0933	0.4428
Szczecin	0.0111	0.0136	0.004

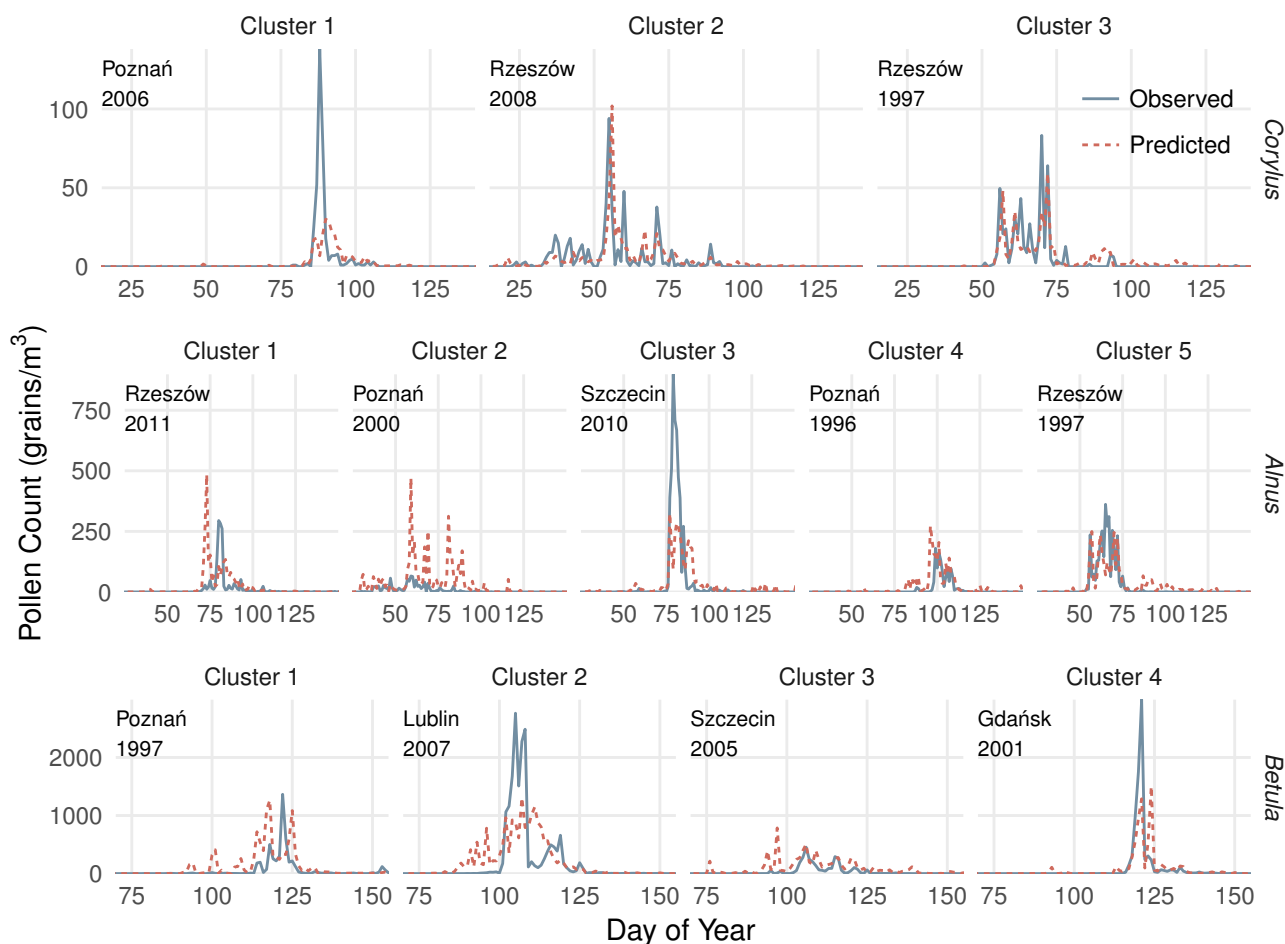


Fig. 3 Examples of observed and predicted pollen concentrations for each cluster of *Corylus*, *Alnus*, and *Betula* random forest models. Detailed description can be found in subsection

Variable importance

Random forest models had the highest average value of r^2 . Therefore, the variable importance for these types of models was obtained and averaged (Fig. 4). Cumulated GDD was the single most important variable in all of the *Corylus*, *Alnus*, and *Betula* models. The other variables had distinctly lower importance. In the *Corylus* models, the next important variables were maximum temperature, potential evapotranspiration from a crop canopy, and vapor pressure. Maximum temperature, potential evapotranspiration from a crop canopy, and total global radiation were next most important in the *Alnus* models. The influence of daily precipitation sum varied between models of different taxa. It had a low importance in *Corylus* models, higher importance in *Alnus* models, and was the fifth most important in *Betula* models. In the *Betula* models, the other important variables were vapor pressure and maximum temperature. In addition, wind speed was the least important variable in most of the *Corylus*, *Alnus*, and *Betula* models.

Discussion

Modeling of pollen concentration in the air is one of the main goals of aerobiological studies, and there are many potential benefits from accurate aerobiological modeling. It could help (i) to understand temporal and spatiotemporal changes of numbers of pollen grains in the air, (ii) to quantify relationships between pollen concentration in the air and external factors (such as spatial, environmental, weather), and (iii) to predict pollen concentration values. The decision on what kind of modeling technique should be used depends on the modeling purpose. In this study, nine different statistical modeling techniques were compared based on their ability to correctly predict pollen concentration values.

The Minimum Requirements for pollen monitoring networks (Galán et al, 2014) state that “[...] the sampler must be placed on a readily accessible, flat, horizontal surface” on the roof of a building. This requirement is vital to assure that pollen count is representative for a large region and is not affected by local factors. Therefore, it is also important in aerobiological modeling to use independent variables which are representative

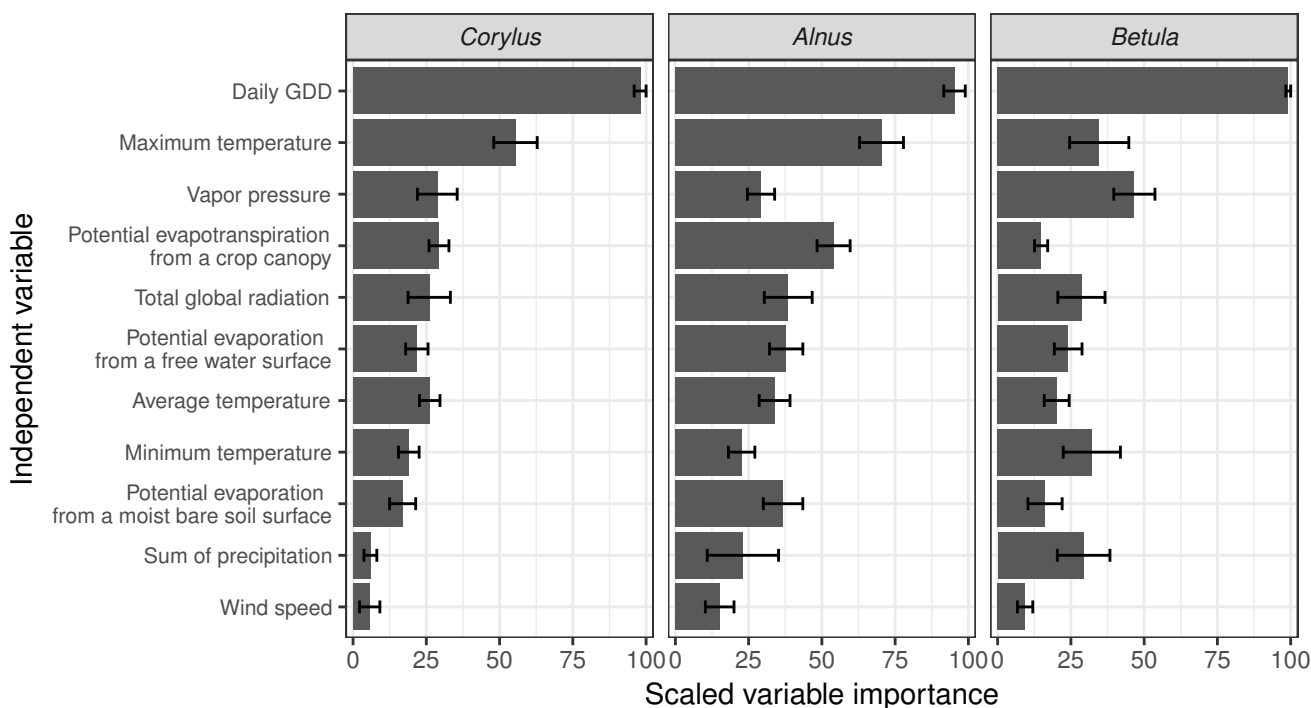


Fig. 4 Averaged scaled variable importance of each predictor for *Corylus*, *Alnus*, and *Betula* random forest models of pollen concentration in the air. Error bars represent one standard error

for an extensive area. The majority of past aerobiological studies relied on in situ meteorological measurements from one location/point. Meteorological instruments were located either in the same place as pollen traps (on a roof), in the close vicinity (at ground level), or even several kilometers further away (e.g. at a local airport). For this analysis, meteorological variables from a regular grid (25x25 km) were used to represent weather conditions over a large region. This approach is more appropriate for modeling of pollen concentration. However, it remains to be further clarified (i) what the optimal grid size is, and (ii) whether (and how) the optimal value varies among the taxa analyzed.

The *Corylus*, *Alnus*, and *Betula* model performances varied distinctly among modeling techniques. For each taxon, the linear model, partial least square, lasso, neural networks, and support vector machines were the least correct; multivariate adaptive regression splines and basic regression tree gave better results; random forest and cubist proved to be the most accurate. Random forest models showed similar values of relative performance statistics for *Corylus*, *Alnus*, and *Betula* (average r^2 between 0.38–0.41 and average SMAPE of 0.53–0.59). On the other hand, large differences in MAE could be observed among the models for each taxa. The mean value of MAE for all models was about 3.6 in *Corylus* models, 24.9 in *Alnus* models, and 71.5 in *Betula* models. Average values were smaller for random forest: 2.8, 20.0, and 57.5, respectively, for *Corylus*, *Alnus*, and *Betula* models. These results showed that while meteorological parameters have a similar influence on pollen concentration in the air, the absolute errors are connected with the abundance of pollen grains of the given taxon. In addition, values of MAE in *Corylus* and *Alnus* models were distinctly lower than thresholds, based on first-symptom values for patients allergic to each taxon (35 grains/m³ for *Corylus*, 45 grains/m³ for *Alnus* (Rapiejko et al, 2007)). The models' performances also differed among the sites studied. The most distinct examples were the models for Bydgoszcz, where data covered only five years of observations. Random forest models for *Corylus* and *Alnus* gave the lowest values of performance statistics, and models for *Betula* had the highest values of performance statistics in comparison with the other cities. This could be an indication of a low stability of models built on short time-series data. Variations of the models' performances among the sites could be also explained by differences in the relative position of samplers as well as a technician variability.

Final random forest model results were clustered based on the model performance statistics (r^2 , MAE, SMAPE). Three clusters were created for *Corylus* models, five for *Alnus* models, and four for *Betula* models. A PERMANOVA test was used to verify the impact of average values of meteorological parameters on models' performance. While significant differences were found for 55% (15 of 27) of the taxon/city pairs, some disagreements between clusters remained unexplained. Thus, the variation in model quality could also be explained by the differences of the other meteorological parameter characteristics, such as distribution and variability of values, or time course.

Cumulated GDD and maximum temperature proved to be the most important variables in *Corylus*, *Alnus*, and *Betula* random forest models. Potential evapotranspiration from a crop canopy and potential evaporation

from a moist bare soil surface were clearly important in *Alnus* models. Vapor pressure was the second most important variable in *Betula* random forest models. These results on variable importance are in accord with a previous predictive study of high pollen concentration levels (Nowosad, 2016) and can be explained by the biological requirements of these trees (Dahl et al, 2013). Precipitation scavenging affects deposition of pollen grains (Sofiev et al, 2013a). However, the impact of the daily sum of precipitation varied greatly among the taxa. This could be partially explained by the length and intensity of the pollen seasons of *Corylus*, *Alnus*, and *Betula*. *Betula* pollen seasons are relatively short and have a high pollen count. On the other hand, *Corylus* pollen seasons are usually longer but with a lower number of pollen grains in the air. Therefore, models supposedly could not detect the impact of precipitation on a pollen concentration. Moreover, the impact of precipitation could be delayed in time, with a greater importance of rainfall from one or two previous days. Finally, wind speed had the lowest impact on the models. This predictor is highly changeable during the course of a day, and thus daily averages can hide important information. In addition, wind impact on pollen concentration cannot be fully understood without a knowledge of wind direction.

The decision on which modeling techniques should be applied needs to be based on the final purpose of the model. Linear models or basic regression tree provide one with the ability to interpret results simply; however, they are not the best choice in cases of complicated, nonlinear predictive problems. This study showed that more complex models, such as random forest or cubist, can provide better predictions. These models are often falsely described as ‘black boxes.’ In fact, they have indirect methods for interpreting their results, such as measures of predictors’ importance and visualizations of relationships between output and independent variables.

Previous studies varied greatly in terms of their modeling techniques, predictor variables, and validation methods. Therefore, the model results in these studies cannot be directly compared. *Betula* linear models of Bringfelt et al (1982) gave correlation coefficient values up to 0.81, which corresponds to r^2 of 0.66. Rodríguez-Rajo et al (2006) predictions of *Alnus* pollen concentration using ARIMA lacked numerical information on the model’s quality. Authors reported only that “the estimated curve[s] accurately describe the *Alnus* pollen grains’ behaviour.” Stochastic gradient boosting models created by Hilaire et al (2012) showed values up to 0.78 (Geneva) and 0.87 (Locarno) of pseudo- R^2 based on deviance residuals for *Alnus* and *Betula*, respectively. However, their study did not provide any information about the lowest values of pseudo- R^2 or about the distribution of errors. The artificial neural network model of the relationship between *Betula* pollen and meteorological factors of Puc (2012) using raw variables showed an accuracy of R lower than 0.5. Transformed values of *Betula* pollen concentration using $\log(x+1)$ gave better results, with a root mean square error of 0.14. Nevertheless, the use of logarithmic transformation of dependent variables does not permit the use of these results for forecasting purposes. The Ritenberga et al (2016) model of *Betula* concentration gave an r^2 of 0.24, based on untransformed data. The performance of a predictive model is overestimated when determined simply on basis of the sample that was used to construct the model. The magnitude of overfitting depends on the modeling technique, on a number of predictors, and on the complexity of the relationship between output and predictors. There are several possible combinations of highly overfitted *Corylus*, *Alnus*, and *Betula* pollen concentration models. One combination consists of simple models with a small number of predictors. A linear model with only a few predictors could produce falsely high values of performance statistics. However, with an increased number of predictors, the quality of linear models will decrease—even without using validation. On the other hand, more complex models (e.g. random forest) could produce greatly overfitted results in both cases. Therefore, one of main challenges in predictive modeling is to determine the true quality of the model. Cross-validation must be used for this purpose. The majority of aerobiological studies use one or two years’ pollen data as a validation (testing) set (Emberlin et al, 1993; Laaidi, 2001; Cotos-Yáñez et al, 2004; Castellano-Méndez et al, 2005; Rodríguez-Rajo et al, 2006; Myszkowska, 2013). This can provide a wrong estimation of model performance in cases when the validation dataset consists of years with an average pollen season (model quality could be overestimated) as well as in cases when extreme years are in a validation set (model quality will therefore be underestimated). Thus, partitioning a sample of data into two subsets—one for training and the other for testing—is not recommended. There are many alternative re-sampling techniques whose purpose is to provide a more robust estimation of model performance, such as the bootstrap, leave-one-out cross-validation, Monte Carlo cross-validation, and k-fold cross-validation. Repeating k-fold cross-validation was used in this study as it increases the precision of model performance estimation (Molinario et al, 2005). Hyndman and Athanasopoulos (2013) proposed a cross-validation for time-series, which could be used for prediction of taxa with long pollen seasons. However, the short seasons of *Corylus*, *Alnus*, and *Betula* and the large number of days per year without pollen grains in the air make it difficult to decide on the proper parameters for time-series validation. Finally, there is a lack of robust techniques for spatiotemporal validation.

The goal of this study was to compare predictive techniques, not to build the best model possible. There are several aspects which should be taken into consideration in the predictive modeling of pollen concentration. Firstly, in this study only meteorological data from the same day as pollen concentration values were used as an independent variable. Although the results clearly showed the importance of meteorological variables, they did not explain all of the variability in pollen count values. Additional predictors could improve performance of pollen concentration models. Potential predictors include other meteorological parameters (e.g. wind direction,

humidity, snow occurrence), past pollen concentration characteristics (average pollen concentration values), and spatial variables (local land cover/land use, share of analyzed taxa in local flora, spatial distribution of flowering trees). Past pollen count values can also be used, but only if there is a possibility of obtaining pollen concentration in a relevant time. This data could be more accessible within a short time with the advancement in automatic pollen concentration measurements. Moreover, variables with different temporal scope (e.g. lagged data, monthly data) should improve pollen concentration models. In addition, predictive statistical models of pollen concentration for one site cannot explain nor properly predict the episodes of long-distance transport from remote sources. A potential solution to this problem might be a combination of many point models with a numerical forecast of air mass trajectories. It should also be noted that aerobiological data are available on genus level (*Alnus*, *Corylus*, *Betula*). Therefore, it is possible that a quality of models is lower when several species (for example *Alnus incana*, *Alnus alnobetula*, *Alnus glutinosa*) occupy the same area, but differ in terms of phenology. Finally, the results of modeling techniques substantially depend on model parameters; thus, the parameters for models should be very carefully chosen.

Conclusion

- Nine modeling techniques were compared in this study based on pollen concentrations of *Corylus*, *Alnus*, and *Betula* and on meteorological variables. The use of rigid cross-validation provided reliable assessment of quality for 243 final models.
- Linear regression and nonlinear regression models performed poorly. Regression trees and rule-based models proved to be the most accurate for all of the taxa analyzed.
- Cumulated GDD was the most important variable in the random forest models of *Corylus*, *Alnus*, and *Betula*. In addition, maximum temperature was an important variable for the models. The importance of precipitation varies between the models, with an average importance for *Betula* models and low importance for *Corylus* models. Wind speed was the least important for all of the models.
- The main goal of this study was to compare different predictive modeling techniques. However, it would be worthwhile to try to improve model results. Potential enhancements include the use of additional meteorological, aerobiological, or spatial variables. In addition, a combination of statistical models with numerical forecasts of air mass trajectories could improve the prediction of high pollen concentration influenced by long-distance transport.

Acknowledgements

This study was carried out within the framework of Project no. NN305 321936, financed by the Ministry of Science and Higher Education in Poland.

References

- Anderson MJ (2001) A new method for non parametric multivariate analysis of variance. *Austral ecology* 26(2001):32–46, DOI 10.1111/j.1442-9993.2001.01070.pp.x
- Baruth B, Genovese G, Leo O (2007) CGMS Version 9.2 - User Manual and Technical Documentation. Tech. rep., Office for Official Publications of the European Communities, Luxembourg, DOI 10.2788/37265
- Bishop CM (1995) *Neural Networks for Pattern Recognition*. Oxford University Press
- Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32, DOI 10.1023/A:1010933404324
- Breiman L, Friedman J, Stone CJ, Olshen RA (1984) *Classification and regression trees*. CRC press
- Bringfelt B, Engström I, Nilsson S (1982) An evaluation of some models to predict airborne pollen concentration from meteorological conditions in Stockholm, Sweden. *Grana* 21(1):59–64, DOI 10.1080/00173138209427680
- 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. *International Journal of Biometeorology* 49(5):310–316, DOI 10.1007/s00484-004-0247-x
- Cotos-Yáñez TR, Rodríguez-Rajo FJ, Jato MV (2004) Short-term prediction of *Betula* airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models. *International Journal of Biometeorology* 48(4):179–185, DOI 10.1007/s00484-004-0203-9
- Dahl A, Galán C, Hajkova L, Pauling A, Sikoparija B, Smith M, Vokou D (2013) The onset, course and intensity of the pollen season. In: Sofiev M, Bergmann KC (eds) *Allergenic Pollen: A Review of the Production, Release, Distribution and Health Impacts*, Springer Netherlands, Dordrecht, pp 29–70
- Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. *Advances in Neural Information Processing Systems* 1:155–161, DOI 10.1.1.10.4845
- Emberlin J, Savage M, Woodman R (1993) Annual variations in the concentrations of *Betula* pollen in the London area, 1961–1990. *Grana* 32(6):359–363, DOI 10.1080/00173139309428965
- Friedman J (1991) Multivariate Adaptive Regression Splines. *The Annals of Statistics* 19(1):1–67, DOI 10.2307/2241837
- Galán C, Smith M, Thibaudon M, Frenguelli G, Oteros J, Gehrig R, Berger U, Clot B, Brandao R (2014) Pollen monitoring: minimum requirements and reproducibility of analysis. *Aerobiologia* 30(4):385–395, DOI 10.1007/s10453-014-9335-5
- Gosso A (2012) elmNN: Implementation of ELM (Extreme Learning Machine) algorithm for SLFN (Single Hidden Layer Feedforward Neural Networks). URL <https://cran.r-project.org/package=elmNN>

- Grolemund G, Wickham H (2011) Dates and Time Made Easy with lubridate. *Journal of Statistical Software* 40(3):1–25, URL <http://www.jstatsoft.org/v40/i03>
- Hilaire D, Rotach MW, Clot B (2012) Building models for daily pollen concentrations: The example of 16 pollen taxa in 14 Swiss monitoring stations. *Aerobiologia* 28(4):499–513, DOI 10.1007/s10453-012-9252-4
- Hirst J (1952) An Automatic Volumetric Spore Trap. *Annals of Applied Biology* 39(2):257–265, DOI 10.1111/j.1744-7348.1952.tb00904.x
- Hyndman RJ, Athanasopoulos G (2013) *Forecasting: principles and practice*. OTexts
- Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software* 11(9):1–20, DOI 10.1016/j.csda.2009.09.023
- Kaufman L, Rousseeuw PJ (2005) *Finding Groups in Ordinal Data. An Introduction to Cluster Analysis*, vol 344. John Wiley & Sons
- Kuhn M (2016) Package 'caret': Classification and Regression Training. DOI 10.1053/j.sodo.2009.03.002, URL <https://github.com/topepo/caret/>
- Kuhn M, Johnson K (2013) *Applied Predictive Modeling*. Springer New York, DOI 10.1007/978-1-4614-6849-3
- Kuhn M, Weston S, Keefer C, Coulter N (2014) Cubist: Rule- and Instance-Based Regression Modeling. URL <https://cran.r-project.org/web/packages/Cubist/index.html>
- Laaidi M (2001) Regional variations in the pollen season of *Betula* in Burgundy: Two models for predicting the start of the pollination. *Aerobiologia* 17(3):247–254, DOI 10.1023/A:1011899603453
- Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R news* 2(December):18–22, DOI 10.1177/154405910408300516, URL <http://cran.r-project.org/doc/Rnews/>, 1609-3631
- Makridakis S (1993) Accuracy Measure: Theoretical and Practical Concerns. *International Journal of Forecasting* 9(1):527–529
- Mevik B, Wehrens R, Liland K (2015) pls: Partial Least Squares and Principal Component Regression. URL <https://cran.r-project.org/package=pls>
- Milborrow S (2016) Multivariate Adaptive Regression Splines. URL <http://cran.r-project.org/package=earth>
- Molinaro AM, Simon R, Pfeiffer RM (2005) Prediction error estimation: A comparison of resampling methods. *Bioinformatics* 21(15):3301–3307
- Myszkowska D (2013) Prediction of the birch pollen season characteristics in Cracow, Poland using an 18-year data series. *Aerobiologia* 29(1):31–44, DOI 10.1007/s10453-012-9260-4
- Nelder JA, Wedderburn RWM (1972) Generalized linear models. *Journal of the Royal Statistical Society Series A (General)* 135(3):370–384
- Nowosad J (2016) Spatiotemporal models for predicting high pollen concentration level of *Corylus*, *Alnus*, and *Betula*. *International Journal of Biometeorology* 60(6):843–855, DOI 10.1007/s00484-015-1077-8
- Nowosad J, Stach A, Kasprzyk I, Weryszko-Chmielewska E, Piotrowska-Weryszko K, Puc M, Grewling L, Pedziszewska A, Uruska A, Myszkowska D, Chlopek K, Majkowska-Wojciechowska B (2016) Forecasting model of *Corylus*, *Alnus*, and *Betula* pollen concentration levels using spatiotemporal correlation properties of pollen count. *Aerobiologia* 32(3):453–468, DOI 10.1007/s10453-015-9418-y
- Puc M (2012) Artificial neural network model of the relationship between *Betula* pollen and meteorological factors in Szczecin (Poland). *International Journal of Biometeorology* 56(2):395–401, DOI 10.1007/s00484-011-0446-1
- R Core Team (2016) *R: A Language and Environment for Statistical Computing*. DOI 10.1007/978-3-540-74686-7, URL <http://www.r-project.org>, arXiv:1011.1669v3
- Rapiejko P, Stankiewicz W, Szczygielski K, Jurkiewicz D (2007) Progowe stężenie pyłku roślin niezbędne do wywołania objawów alergicznych (Threshold pollen count necessary to evoke allergic symptoms). *Otolaryngologia Polska* 61(4):591–594, DOI 10.1016/S0030-6657(07)70491-2
- Ritenberga O, Sofiev M, Kirillova V, Kalnina L, Genikhovich E (2016) Statistical modelling of non-stationary processes of atmospheric pollution from natural sources: Example of birch pollen. *Agricultural and Forest Meteorology* 226-227:96–107, DOI 10.1016/j.agrformet.2016.05.016
- Rodríguez-Rajo FJ, Valencia-Barrera RM, Vega-Maray AM, Suárez FJ, Fernández-González D, Jato V (2006) Prediction of airborne *Alnus* pollen concentration by using ARIMA models. *Annals of Agricultural and Environmental Medicine* 13(1):25–32
- Salvador S, Chan P (2004) Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI* pp 576–584, DOI 10.1109/ICTAI.2004.50
- Sofiev M, Belmonte J, Gehrig R, Izquierdo R, Smith M, Dahl Å, Siljamo P (2013a) Airborne Pollen Transport Mikhail. In: Sofiev M, Bergmann KC (eds) *Allergenic Pollen: A Review of the Production, Release, Distribution and Health Impacts*, Springer Netherlands, Dordrecht, pp 127–159, DOI 10.1007/978-94-007-4881-1
- Sofiev M, Siljamo P, Ranta H, Linkosalo T, Jaeger S, Rasmussen A, Rantio-Lehtimäki A, Severova E, Kukkonen J (2013b) A numerical model of birch pollen emission and dispersion in the atmosphere. Description of the emission module. *International Journal of Biometeorology* 57(1):45–58, DOI 10.1007/s00484-012-0532-z
- Therneau T, Atkinson B, Ripley B (2015) Recursive partitioning and regression trees. URL <https://cran.r-project.org/package=rpart>
- Tibshirani R (1996) Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B* 58(1):267–288, DOI 10.2307/2346178, 11/73273
- Vogel H, Pauling A, Vogel B (2008) Numerical simulation of birch pollen dispersion with an operational weather forecast system. *International Journal of Biometeorology* 52(8):805–814, DOI 10.1007/s00484-008-0174-3
- Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer New York, New York, DOI 10.1007/978-0-387-98141-3
- Wold S, Martens H, Wold H (1983) The multivariate calibration problem in chemistry solved by the PLS method. In: *Matrix pencils*, 1981, Springer, pp 286–293, DOI 10.1017/CBO9781107415324.004
- Zou H, Hastie T (2008) elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA. URL <http://cran.r-project.org/package=elasticnet>