



What are the most important variables for Poaceae airborne pollen forecasting?

Ricardo Navares^a, José Luis Aznarte^{b,*}

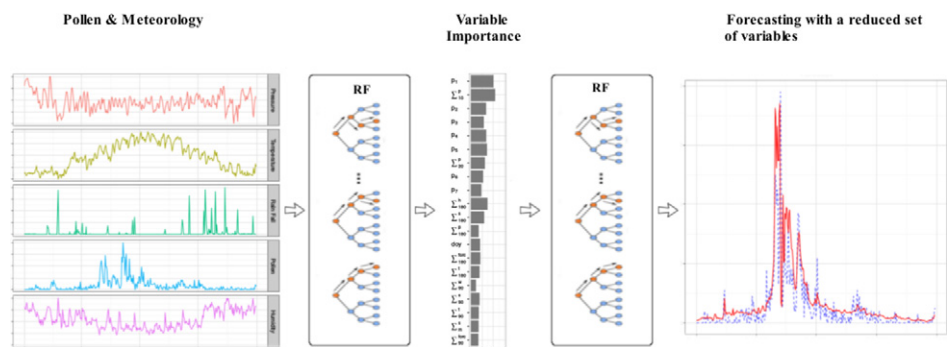
^aSuperior Technical School of Computer Engineering, UNED, Juan del Rosal, 16, Madrid 28040, Spain

^bDepartment of Artificial Intelligence, UNED, Juan del Rosal, 16, Madrid 28040, Spain

HIGHLIGHTS

- A new data-driven approach to forecast airborne pollen concentrations is presented.
- The relative importance of variables is estimated and used to build the models.
- The assumption-free findings are coherent with previous phenological-based results.
- Non-parametric hypothesis testing confirms the statistical validity of the results.
- A reduced set of important variables renders more simple and accurate models.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 3 October 2016

Received in revised form 9 November 2016

Accepted 15 November 2016

Available online 5 December 2016

Keywords:

Aerobiology

Prediction

Random forests

Nonparametric tests

Feature selection

Time series

Madrid

ABSTRACT

In this paper, the problem of predicting future concentrations of airborne pollen is solved through a computational intelligence data-driven approach. The proposed method is able to identify the most important variables among those considered by other authors (mainly recent pollen concentrations and weather parameters), without any prior assumptions about the phenological relevance of the variables. Furthermore, an inferential procedure based on non-parametric hypothesis testing is presented to provide statistical evidence of the results, which are coherent to the literature and outperform previous proposals in terms of accuracy. The study is built upon Poaceae airborne pollen concentrations recorded in seven different locations across the Spanish province of Madrid.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The ability to anticipate future values of pollen concentrations in the air is crucial both for the allergic population, which can use predictions to foresee and adapt their needs concerning their outdoor

presence, and for clinical institutions and public health organisms, which can prearrange resources before a predicted future outburst of pollen-related affections occurs. Many authors have faced the pollen forecasting problem in the last decades, with approaches that range from numerical models as SILAM (Sofiev et al., 2013), classic statistical time series analysis such as multivariate regression and nonlinear models (Cotos-Yáñez et al., 2004; Rodríguez-Rajo et al., 2004; Tassan-Mazzocco et al., 2015), to machine learning and computational intelligence, for example artificial neural networks

* Corresponding author.

E-mail address: jlaznarte@dia.uned.es (J. Aznarte).

(Astray et al., 2016; Aznarte et al., 2007; Iglesias-Otero et al., 2015). Other approaches use numerical method to forecast pollen dispersions based on operational weather forecasts such as COSMO-ART (Vogel et al., 2008) or HYSPLIT (de Water et al., 2003).

However, there is a common underlying question which is independent of the chosen approach: which past information should be used when forecasting future values of pollen concentrations? For example, in univariate time series analysis, the models try to extract information from the past behaviour of just the pollen concentrations data (Aznarte et al., 2007). Of course, botany tells us that meteorology plays a crucial role in the development of the plants and hence in the pollen emission, and thus many authors have included meteorological variables in their models. In fact, there are studies about the influential factors in the growth state of plant buds (and, consequently, airborne pollen atmospheric concentrations) based on a phenological point of view (Cannell and Smith, 1983; Kmenta et al., 2016; Ribeiro et al., 2007; Smith and Emberlin, 2006), or based on the relation with climate conditions (Andersen, 1991; Pauling et al., 2014; Rodríguez-Rajo et al., 1983), or both. However, there is no consensus over which meteorological variables are more relevant.

For example, some studies employ meteorological daily data in order to forecast pollen concentrations, such as previous daily precipitation (Castellano-Méndez et al., 2005; Iglesias-Otero et al., 2015) or the relative humidity, wind speed and radiation upon the surface (Jones and Harrison, 2004). Others prefer the use of autoregressive indices, as, for example, thermal indices during plant formation season, in order to capture climatological information prior to pollen emission (Andersen, 1991; Myszkowska, 2014; Otero et al., 2013; Pauling et al., 2014). Some studies combine both approaches by using daily data and cumulative meteorological indices (Matyasovszky et al., 2015) researching the relationships between past and current weather conditions and current pollen levels (Deák et al., 2013), assuming that the timing of flowering is mostly driven by the accumulated temperature during a certain time period as the pollen release model in SILAM (Sofiev et al., 2013).

On the other hand, automatic feature selection is an important research field in computational intelligence. Feature selection techniques are used, among other reasons, to simplify the models in order to make them more interpretable and to shorten the training times. The idea behind automatic feature selection is that usually the data contains many features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information.

An alternative to feature selection for pollen forecasting is factor analysis (Deák et al., 2013; Matyasovszky et al., 2015), which tries to find a latent representation of the observed variables that is good at explaining it. This technique can be used, for example, to avoid collinearity among the variables which might influence the predictive capability of the model. Given RF is robust enough against multicollinearity, its application would be sufficient to approach the problem.

The objective of this paper is to apply an automatic feature selection procedure to pollen forecasting, and validate it through statistical inference. By avoiding any *a priori* assumptions about the importance of the variables, neither based on the phenology of the plant nor on meteorological considerations nor on derived indices, we expect to question other author's assumptions and to provide new insight on the predictive power of the different available variables. Statistical inference through a nonparametric ranking-based statistical test (Friedman, 1937), along with a pairwise variable comparison in a post-hoc procedure, will allow to soundly establish the validity of the results.

As a case study, we chose to work on Poaceae airborne pollen concentrations in the Madrid region. Poaceae is the largest family of monocotyledonous flowering plants known as grasses and is

considered to be one of the most important aeroallergens in Europe (Sánchez-Mesa et al., 2003). Poaceae pollen not only is one of the most prolific in Madrid but also is known to be one of the most aggressive, accounting 94% of positive reaction in patients (Subiza et al., 1995). The increase of allergy cases and the severity of the reactions (de Weger et al., 2013) motivates the need for prediction of Poaceae concentrations.

2. Materials and methods

2.1. Data description

Daily Poaceae airborne concentrations were provided by Red Palinológica de la Comunidad de Madrid and were obtained following the standard methodology of the Spanish Aerobiological Network (Galán Soldevilla et al., 2007). They are measured in grains per cubic meter of air. The observations come from 7 locations around the region of Madrid (Alcalá de Henares, Alcobendas, Aranjuez, Faculty of Pharmacy of Complutense University of Madrid, Getafe, Leganés and Villalba) and span periods that go from 2000 to 2013. Fig. 1 shows the location of the measuring stations.

On the other hand, we used weather observations consisting of average daily temperature in Celsius degrees, solar radiation in W/m², wind speed measured in m/s, daily rainfall in mm/h, pressure in mbar and degree of humidity in percentage. Data sets for locations Alcalá de Henares, Alcobendas, Aranjuez, Getafe and Leganés consist of 5 years of observations from 2005 to 2009. At the Faculty of Pharmacy data is available from 2001 to 2013 while in Villalba only three years are available starting 2007 to 2009 as shown in Fig. 2.

The geography of the Autonomous Community of Madrid provides several peculiarities given the locations studied. Villalba, which is located at 903 m above the sea level, has a mountain climate with a yearly average temperature of 10–11°C and a yearly average rainfall of 1250–1500 mm. As opposed to Villalba, Aranjuez has an elevation of 495 m above the sea level with an average yearly temperature above 13°C and a yearly average rainfall below than 400 mm. The remaining locations consist of metropolitan areas located between 594 and 668 m above the sea level.

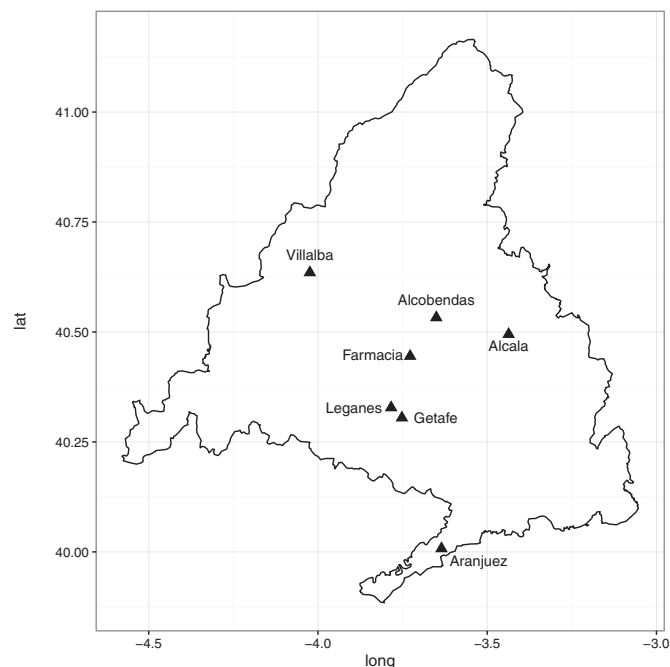


Fig. 1. Location of weather and pollen stations.

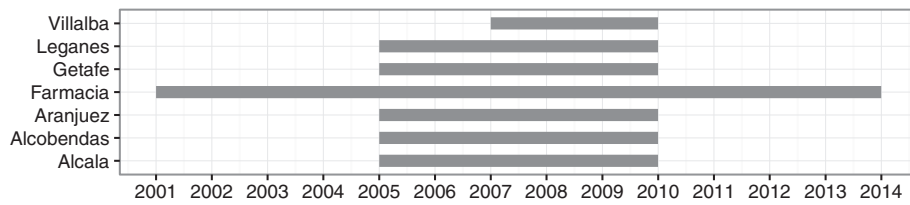


Fig. 2. Years available per observation station.

Very few missing data points were observed in the meteorological series so these were directly linearly interpolated to fill the gaps. On the other hand, pollen series contain missing observations in *a priori* critical months as February, March and April. In general, during these months pollen concentrations are meant to increase, thus the missing data is regressed within each year and across the years.

2.2. Features

In the pollen forecasting literature, the common approach is to include a set of independent variables which contain as much relevant information as possible. However, the core idea presented in this paper is to consider all the variables used by previous studies and compute a measure of variable importance to establish which set of independent variables are the most influential.

It is assumed that previous cumulative pollen observations serve as a proxy of the state of the plant (Ribeiro et al., 2007; Smith and Emberlin, 2006) and are considered as an indicator of plant development. Accordingly, cumulative sum of pollen concentrations within the year is included as an independent variable, along with the 10 and 30-day cumulative sums. As well, the presence of an autocorrelative pattern in the pollen time series suggests that previous daily concentrations have an influence in present values (Aznarte et al., 2007). In order to capture this inner relation between lagged airborne daily concentrations, previous single daily observations up to a week are also considered.

Weather conditions play a major role in determining the severity and length of the pollination season, as they are directly responsible for increases and decreases of the airborne pollen concentrations through its effect on the plants. The growth state of the buds is assumed to be linearly related to the amount of energy a plant has received (Cannell and Smith, 1983). This energy is usually represented by the sum of temperatures up to some point (Andersen, 1991; Cannell and Smith, 1983; Rodríguez-Rajo et al., 1983). Other recent studies (Pauling et al., 2014) introduce the concept of *chilling temperatures* and *forcing temperatures* instead, as they constitute a more optimal parameterization to represent the energy absorbed by the plants. Forcing temperatures are defined as the weighted sum of temperatures above a certain level for a fixed period:

$$F_{\text{sum}}(d) = \sum_{i=d-n}^d R_{\text{forc}}(i), \quad (1)$$

where

$$R_{\text{forc}}(i) = \begin{cases} 0 & \text{if } T(i) < T_{\text{forc}} \\ T(i) - T_{\text{forc}} & \text{if } T(i) \geq T_{\text{forc}} \end{cases}, \quad (2)$$

being d the forecast date, n the number of days which define the calculation period for the sum of forcing temperatures, $T(i)$ the temperature for day i , and T_{forc} the base temperature for forcing fixed at 8 °C. The same applies for chilling temperatures, but computing the temperatures below the threshold of 6 °C. To allow for more flexibility, our study does not predefine the chilling and forcing

periods, but chilling and forcing temperatures are calculated by accumulation of 30, 90 and 180 days prior to the forecast date.

Sudden rainfall during the pollination cleans the air, washing pollen concentrations away, whereas wind speed contributes to pollen dispersion (Palacios et al., 2000), motivating the selection of the prior 7 daily raw data observations for these two variables. Conversely, rainy and humid periods prior to bloom is known to boost plant development when combined with sunny days. Thus, as was the case for chilling/forcing temperatures, the cumulative sums of 30, 90 and 180 days prior to the forecasting date are considered for rainfall and wind.

This leads to the availability of a total of 79 features as detailed in Table 1, to which we added a dummy variable which represents the day of the year. This makes 144 features which are distributed in a matrix corresponding to the desired forecast horizon. In order to train the model for a forecast horizon of $t = 1$ (one day-ahead forecast), vectors of the form $(x_{1,t}, \dots, x_{144,t} | y_{t+1})$ where t is the time in which the forecast is done.

2.3. Random forests for regression

In the last years, there has been a growing interest in ensemble learning which aggregates the results of several independent models selected to boost their predictive performance. A well-known method is called *bagging* or bootstrap aggregating, proposed by Breiman (1996). Subsequently, Breiman (2001) presented a model called random forests (RF) which adds an additional layer of randomness to *bagging* providing robustness against overfitting with a limited number of parameters. These two characteristics favor RF against other computational intelligence methods such as neural networks.

The procedure combines several randomized regression trees generated over sample fractions of the data, and aggregates their prediction by averaging. This averaging process mitigates the influence of outlier data points giving RF advantage over other common

Table 1
Summary of features generated by variable.

	i	Σ_{10}	MA_i	Σ_{360}	Σ_{30}	Σ_{90}	Σ_{180}	Std
Pollen	7	1	4	1	1	1	1	–
Temperature	7	1	4	1	1	1	1	1
T_{forc}	–	–	–	1	1	1	1	–
T_{chill}	–	–	–	1	1	1	1	–
Humidity	7	1	4	1	1	1	1	1
Wind	7	1	4	1	1	1	1	1
Rain	7	1	4	1	1	1	1	1
Pressure	7	1	4	1	1	1	1	1
UV	7	1	4	1	1	1	1	1
Sun	7	1	4	1	1	1	1	1

i: previous $i \in [1, 7]$ day observation.

Σ_{10} : previous 10-day cumulative sum.

MA_i : max and min i -days moving average $i \in \{5, 15\}$.

Σ_{360} : year to date cumulative sum.

Σ_{30} : previous month cumulative sum.

Σ_{90} : previous 90-day cumulative sum.

Σ_{180} : previous 180-day cumulative sum.

Std: previous 15days standard deviation.

methods as support vector regression, which are highly sensitive in presence of outliers. As opposed to classification trees, the optimal split condition is the variance, which at the same time, is used to compute a measure of the importance of the independent variables.

The importance of a variable is estimated by measuring the increases in prediction error or variance when data from that variable is randomly permuted while the rest are left unchanged. The underlying idea is that if the variable is not important, then rearranging the values of that variable will not degrade prediction accuracy. For each tree, the prediction error is recorded before and after the permutation, the difference between both errors is averaged over all trees and normalized, thus providing the relative importances. The bigger the difference, the higher the importance of the permuted variable.

In order to check the performance of the forecasts, a general purpose error metric for numerical predictions as root mean squared error (RMSE), defined by Eq. (3), was used along with the coefficient of determination R^2 , which indicates the proportion of variance of the observed data was predicted.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (3)$$

where y_i is the observed i th data point, $\hat{y}_i$ the predicted and n the total number of data points in the test set.

2.4. Non-parametric hypothesis testing

Deciding if a variable is a better predictor than other is not a trivial task. Even if a measure of variable importance can be computed, it is not sufficient to evaluate it in a single case. In this work, we use statistical inference to investigate, in proper significance terms, which variables can be considered better than others.

When comparing a set of variables, the common statistical method for testing the differences between more than two related sample means is the repeated-measures ANOVA. Unfortunately, parametric hypothesis tests are based on assumptions such as normality or symmetry of the data distribution (Demšar, 2006). These assumptions are likely to be violated unless the data are well conditioned. On the other hand, nonparametric tests, whose main characteristic is that they are applied over nominal or ordinal data, are usually less restrictive (albeit also less robust). However, through a ranking-based transformation, they can be applied to continuous data.

The Friedman test (Friedman, 1937) is a multiple comparisons test to detect significant differences between a set of at least two samples. In our approach, the idea is to prove the existence of features which are more important than others across the years and the locations. The first step of the procedure is converting the original computed variable importance for each year and location to its correspondent rank within the set to obtain the average rank $R_j = \frac{1}{n} \sum_i r_{ij}^j$ (where j denotes the feature, i refers to each year and location and n is the total number of pairs {location, year}). Then the null hypothesis of equality of medians is tested through the statistic

$$F = \frac{12n}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right], \quad (4)$$

where k is the number of variables and $F \sim \chi_{k-1}^2$. This test only allows to detect significant differences in the whole variable space without comparing each one against each other. A conversion of the rankings can be computed to obtain the p -value of each pair (Conover, 1999). The main drawback is that these p -values are not suitable for multiple comparison as the probability error of a certain

comparison, does not take into account the remaining comparisons belonging to the family.

To overcome this limitation, it is needed to take into account that multiple tests are conducted via adjusted p -values which can be directly compared with a significance level α . A post-hoc test adjusts the value of α when dealing with multiple comparisons. One of the most commonly used of these adjustments is the Holm procedure (Holm, 1979) which adjusts the value of α by ordering, from smallest to largest, the p -values of each test. Then starting with the most significant p_i tests the hypothesis of $H_i : p_i > \alpha/(k-i)$, being k the total number of variables in our proposal. If H_i is rejected then allows to test H_{i+1} and so on.

An extension of Holm's step-down method was proposed by Shaffer (1986), which uses a logical relation between the combination of the hypotheses of all pairwise comparisons. For instance, if a variable v_1 is more/less important than v_2 , it is not possible that v_1 is as important as v_3 and v_2 has the same importance as v_3 . Based on this argument and following Holm's method, instead of rejecting $H_i : p_i \leq \alpha/(k-i)$, rejects $H_i \leq \alpha/t_i$, being t_i the maximum number of hypotheses which can be true given the number of false hypotheses in $j \in \{1, \dots, i\}$

In order to contrast the difference between the importance of two variables we can use as an estimator the medians of the differences of each computed variable importance across locations and years (García et al., 2010). Being i the number of sets composed by each pair {location, year}, the median of the difference of each pair of variables Z_{v_i, v_j} is computed, then for each variable the average of the medians where the variable is involved is calculated as follows:

$$m_{v_i} = \frac{\sum_{j=1}^k Z_{v_i, v_j}}{k}, \quad (5)$$

where k is the total number of variables. The estimator of each pair of variables is defined as $m_{v_i} - m_{v_j}$, which provides how far a pair variables are in terms of importance.

2.5. Experimental design

Algorithm 1. Prediction.

```

Require: series; iterations = 50
1: for  $l$  in locations do
2:   for  $y$  in years(location) do
3:      $X_{\text{test}}^l = X_y^l$ 
4:      $X_{\text{train}}^l = X^l - X_{\text{test}}^l$ 
5:     for  $i$  in iterations do
6:        $\text{RF}_i = \text{RF}(X_{\text{train}}^l)$ 
7:       importance[ $y$ ][ $i$ ] =  $\text{RF.importance}(X_{\text{train}}^l)$ 
8:       prediction[ $y$ ][ $i$ ] =  $\text{RF}_i(X_{\text{test}}^l)$ 
9:     end for
10:   end for
11: total.importance[ $l$ ] = average(importance[ $y$ ][ $i$ ])
12: total.prediction[ $l$ ] = average(prediction[ $y$ ][ $i$ ])
13: end for

```

Algorithm 2. Nonparametric test.

```

Require: series; iterations = 50;  $i = 0$ 
Require:  $M \leftarrow NULL$ 
1: for all  $[l, y]$  in {location, years} do
2:   var.imp  $\leftarrow NULL$ 
3:   for all iterations do
4:     var.imp = var.imp +  $\text{RF.imp}(\text{series}_{l,y})$ 
5:   end for
6:   var.imp = var.imp/iterations
7:    $M[i,] = \text{var.imp}$ 
8:    $i++$ 
9: end for
10:  $F = \text{friedman.test}(M)$ 
11: if  $F \leq \alpha$  then
12:   post.hoc( $M$ )
13: end if

```

In order to study the importance of the different available variables for pollen forecasting, the experiment is divided in three parts. Firstly, we use the full set of variables to build models designed to predict one day-ahead pollen concentrations for each location. From these models, we obtain a first insight of the most important variables in the different locations. Subsequently, in order to generalize and verify if the average rankings of variables per location are statistically significant, we apply statistical inference through non-parametric tests applied to the importance obtained per location and per year. Finally, we use the results of the hypothesis testing process to compare the models, in terms of forecasting precision and computational cost, with different subsets of the most important variables.

The first part of the experiment faces the one day-ahead pollen concentration forecasting problem in a standard fashion. By using a *leave-one-out* setup, we split the series into training and testing set at each location, saving one of the available years in each iteration and training the models with the rest. Given that one year can be seen in this case as one observation, and given the limited amount of observations, the differences from other cross validation techniques, such as 10-fold cross validation, are not expected to be substantial. The process is summarized in [Algorithm 1](#). Through this cross validation approach, we obtain results which are more independent of the particular characteristics of each year, in the form of point-forecasts and estimates of the relative variable importances for each location. From these estimates we expect to see already

some common patterns in the set of variables which are important across the different locations.

However, in order to produce a more rigorous and general result, independent from the location and the particular shape of the pollen curve of each year, the second part of the experiment employs a different partition of the data. The idea is distributing the importances resulting from the RF in a matrix of dimension $N \times M$ where N represents each pair {location, year} present in the dataset and M the available variables. Given the stochastic nature of RF, an iterative approach was designed to avoid the likelihood of overfitting, the final result consisting of the average of 50 instances of each model. Over the resulting matrix a Friedman test is conducted to give evidence on whether to proceed with the post-hoc analysis. This procedure is outlined in [Algorithm 2](#).

Finally, with the results of the statistical inference process, we want to compare the forecasts obtained using the full set of variables in the first part of the experiment with those obtained by using reduced sets of only the most important ones. Precisely we chose to select the 5 and the 15 most important variables.

3. Results

3.1. Predicting with all the variables

Firstly, we applied RF using a *leave-one-out* setup, building the models with the full set of variables at our disposal. The models

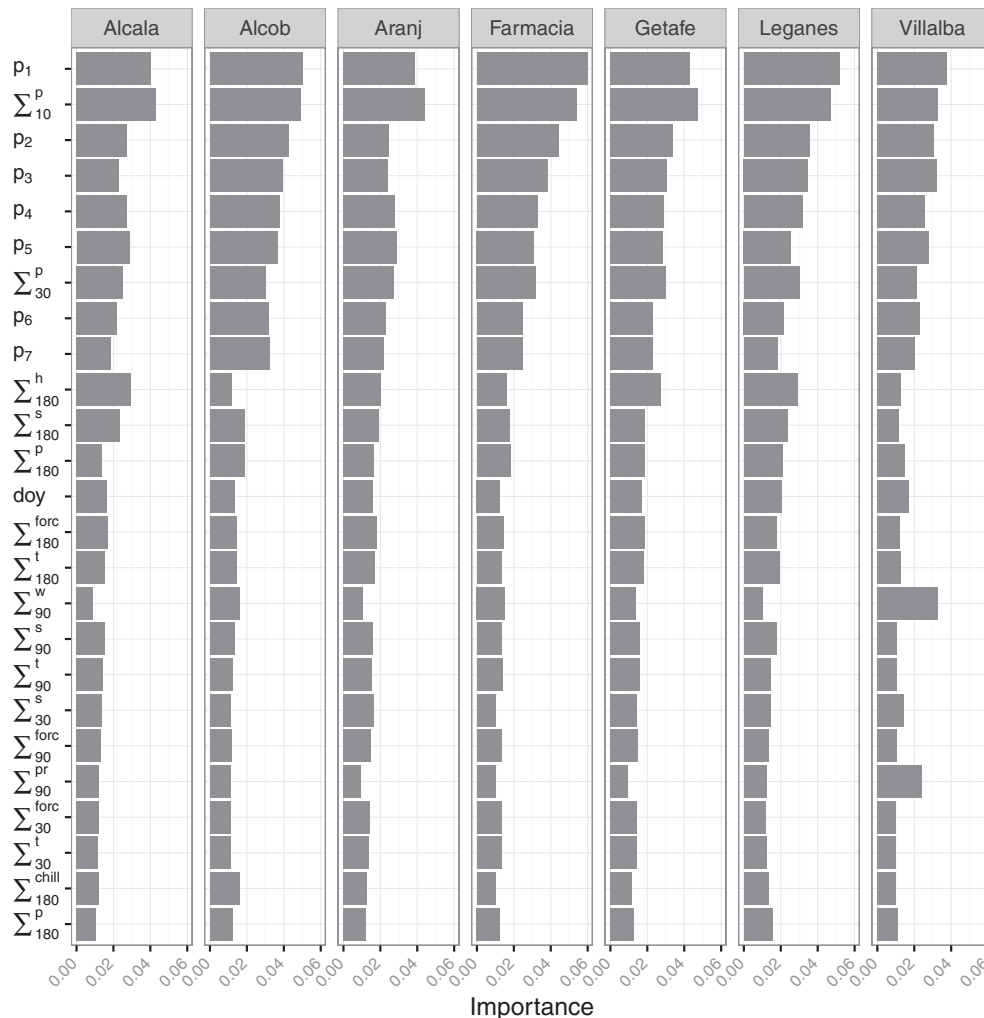


Fig. 3. Relative importance of the most important variables by location (variables are ordered by the sum of their importance across locations).

Table 2
Average ranking of the 16 most important features.

i	Variable	Ranking	i	Variable	Ranking
1	Σ_{10}^p	4.121	9	p ₅	8.682
2	p ₁	4.463	10	doy	8.682
3	p ₂	6.902	11	Σ_{90}^w	8.853
4	p ₃	6.829	12	p ₇	10.365
5	Σ_{180}^h	7.756	13	Σ_{90}^t	10.878
6	Σ_{30}^p	8.121	14	s ₃	11.048
7	p ₄	8.512	15	Σ_{30}^s	11.097
8	p ₆	8.658	16	MA ₅ ^t	12.195

were trained independently for all the years available at each location, leaving one different year on each iteration as a test set. At each location, the relative importance of the variables, averaged by iteration and test sets, is obtained and shown in Fig. 3.

From this figure, it seems clear that the relation among the four most important variables p₁, Σ_{10}^p , p₂ and p₃ is maintained across locations. In fact, the ordering of relative importances remains very similar for all of them except for Villalba, where wind plays a much more crucial role. At this location, the importance of daily wind speed accumulated during 90 days prior the forecast date (Σ_{90}^w) accounts for more than 3% of importance, which is much higher than in the rest of locations. However, as we can see in Fig. 1, Villalba is the most northern and western site. Furthermore, it is located in the Guadarrama mountains at 903 m above the sea level while for example Aranjuez and Getafe have an elevation of 495 m and 622 m respectively. The particular meteorological conditions of Villalba, related to mountain climate, produce a higher correlation of wind with pollen concentrations during the period of study. For instance, during those years Σ_{90}^w is correlated at 32.69% with the daily pollen concentration in Villalba, while in the same period at Alcalá is 12.74% and a 8.64% when the full study period (2005–2009) is considered. This explains the increase in its importance at this location with respect to the rest. Similar applies to Σ_{180}^p , which also gains importance due to the elevation, dropping average daily pressure and influencing flower formation and consequently pollen release.

Notwithstanding, the rest of the patterns of variable importance are quite stable across locations. Variable Σ_{10}^p keeps the best rank among them except for Farmacia, Villalba and Leganés, where its position is exchanged with the second most important variable from the test (p₁). On the other hand, ranks for p₂ and p₃ are perfectly maintained across all locations as well as their relation to the top 2 ranked features.

3.2. Nonparametric tests to compare variable importance

As stated above, the relative importance of the variables computed by location seems quite stable. However, in order to give statistical evidence of the existence of features which are more influential than others, a non-parametric test was used. To do so, a RF model was trained, in this case for each location and for each available year, and the relative importance for each variable was thus computed. Given the large number of variables, the pairwise computation is expensive, hence in this case the hypothesis test is applied on a reduced set of variables including those which represent more than 1% of the total variance, which leaves us with 16 variables.

Out of this setup the Friedman statistic obtained is $F = 148.09$ which is distributed according to chi-square with 15 degrees of freedom with a critical values of $\chi_{15}^2 = 24.99$ at $\alpha = 0.05$, leading to a computed p -value of $1.11e - 10$, which strongly suggests the existence of significant differences among the variables. Table 2 shows the average Friedman ranking of the variables.

Due to the fact that the null hypothesis for Friedman's test is rejected, a post-hoc test can be applied to detect the pairs which produce the differences. Table 3 shows the contrast estimation of medians of the importance in percentage, and it is noticeable how p₁ and Σ_{10}^p obtain, respectively, an average of around 6% and 5.5% more importance than other variables, supporting the conclusions from Friedman's test. These two variables are followed by p₂ and p₃ which outperform around 3% and 2.2%, respectively, the remaining variables. On the other hand, MA₅^t achieves the lowest relative importance across all stations and years.

Carrying out a post-hoc pairwise test will tell the evidence in the differences among pairs of variables. Table 4 and Fig. 4 show the rejected hypothesis (p -value $\leq \alpha$) with a significance level of $\alpha = 0.05$ for each compared pair. It can be seen how there is evidence that Σ_{10}^p and p₁ significantly differ from most of the other variables, although there is no evidence that they differ from each other. As shown above, the contrast estimation and the ranks from Friedman indicate that these two variables retain the highest importance among the whole set. Hence, there is statistical evidence that Σ_{10}^p and p₁ can be considered, in general, the most influential variables regardless the year and location.

There also exists evidence of differences between p₂ and p₃ and the group composed by {MA₅^t, Σ_{30}^s , s₃, Σ_{90}^t , p₇} whose higher rank is 10.36, separating the importance of these two variables from the lowest ranked group in this study.

Table 4 does not show clear distinction between { Σ_{10}^p , p₁} and {p₂, p₃} but, referring to the logical relation between the combination of the pairwise hypotheses proposed by Shaffer (1986), there is

Table 3
Contrast estimation in %.

	Σ_{90}^t	Σ_{10}^p	p ₆	doy	Σ_{30}^s	p ₁	Σ_{90}^w	s ₃	p ₂	Σ_{180}^h	p ₄	p ₃	p ₅	MA ₅ ^t	p ₇	Σ_{30}^p
Σ_{90}^t	0.00	-5.45	-0.95	-0.64	-0.03	-5.98	-0.72	-0.09	-3.06	-1.51	-1.26	-2.21	-1.30	0.12	-0.71	-1.19
Σ_{10}^p	5.45	0.00	4.50	4.81	5.43	-0.52	4.73	5.36	2.39	3.94	4.20	3.24	4.15	5.57	4.74	4.26
p ₆	0.95	-4.50	0.00	0.31	0.92	-5.03	0.23	0.86	-2.11	-0.56	-0.31	-1.26	-0.35	1.07	0.24	-0.24
doy	0.64	-4.81	-0.31	0.00	0.62	-5.33	-0.07	0.55	-2.42	-0.87	-0.61	-1.57	-0.66	0.77	-0.07	-0.54
Σ_{30}^s	0.03	-5.43	-0.92	-0.62	0.00	-5.95	-0.69	-0.07	-3.04	-1.48	-1.23	-2.18	-1.28	0.15	-0.68	-1.16
p ₁	5.98	0.52	5.03	5.33	5.95	0.00	5.26	5.88	2.91	4.47	4.72	3.77	4.67	6.10	5.27	4.79
Σ_{90}^w	0.72	-4.73	-0.23	0.07	0.69	-5.26	0.00	0.62	-2.34	-0.79	-0.54	-1.49	-0.58	0.84	0.01	-0.47
s ₃	0.09	-5.36	-0.86	-0.55	0.07	-5.88	-0.62	0.00	-2.97	-1.42	-1.16	-2.12	-1.21	0.22	-0.62	-1.09
p ₂	3.06	-2.39	2.11	2.42	3.04	-2.91	2.34	2.97	0.00	1.55	1.81	0.85	1.76	3.19	2.35	1.87
Σ_{180}^h	1.51	-3.94	0.56	0.87	1.48	-4.47	0.79	1.42	-1.55	0.00	0.25	-0.70	0.21	1.63	0.80	0.32
p ₄	1.26	-4.20	0.31	0.61	1.23	-4.72	0.54	1.16	-1.81	-0.25	0.00	-0.95	-0.05	1.38	0.55	0.07
p ₃	2.21	-3.24	1.26	1.57	2.18	-3.77	1.49	2.12	-0.85	0.70	0.95	0.00	0.91	2.33	1.50	1.02
p ₅	1.30	-4.15	0.35	0.66	1.28	-4.67	0.58	1.21	-1.76	-0.21	0.05	-0.91	0.00	1.42	0.59	0.11
MA ₅ ^t	-0.12	-5.57	-1.07	-0.77	-0.15	-6.10	-0.84	-0.22	-3.19	-1.63	-1.38	-2.33	-1.42	0.00	-0.83	-1.31
p ₇	0.71	-4.74	-0.24	0.07	0.68	-5.27	-0.01	0.62	-2.35	-0.80	-0.55	-1.50	-0.59	0.83	0.00	-0.48
Σ_{30}^p	1.19	-4.26	0.24	0.54	1.16	-4.79	0.47	1.09	-1.87	-0.32	-0.07	-1.02	-0.11	1.31	0.48	0.00

evidence of difference between Σ_{10}^p and p_1 and p_4 , something which does not exist for p_2 and p_3 . This leads to conclude that Σ_{10}^p and p_1 are more important than p_4 , but there is no evidence that p_2 and p_3 are more important than p_4 so the conclusion is that Σ_{10}^p and p_1 are more influential than p_2 and p_3 .

In summary, the non-parametric tests prove the existence of features which are more relevant than others. Among them, there are groups of features which significantly differ from other groups, while there is no statistical evidence of differences between group members. This means that, within a group, it is expected that their members maintain or alternate their ranks within the bounds of the rank of the group. For instance, we have seen that Σ_{10}^p and p_1 differ from the rest of the variables but do not differ from one another, constituting the top ranked group. As shown in Table 2, these variables have a Friedman's rank of 4.121 and 4.463 respectively, which is translated to position 1 and 2 in importance. It is expected that Σ_{10}^p and p_1 maintain their correspondent position or that p_1 takes position 1, dragging Σ_{10}^p to position 2 as they do not differ from each other.

3.3. Predicting with reduced sets of variables

Once we investigated the relative importance of the variables through statistical inference, we moved on to empirically verify how does selecting a subset of the most important variables affect the precision of the models and their computational efficiency.

In order to do so, we repeated the experiment in Section 3.1 for two reduced set of variables. First we selected the 15 variables which were proven to be the most important according to the tests of Section 3.2, and secondly we repeated the experiment reducing

Table 4

Pairwise rejected hypothesis at $\alpha = 0.05$ with unadjusted p -value and adjusted Holm and Shaffer p -values.

i	Hypothesis	p	p_{holm}	p_{shaff}
1	Σ_{10}^p vs MA_5^t	1.62e-14	1.94e-12	1.94e-12
2	p_1 vs MA_5^t	1.94e-13	2.31e-11	2.04e-11
3	Σ_{10}^p vs Σ_{30}^s	3.27e-11	3.86e-09	3.43e-09
4	Σ_{10}^p vs s_3	4.47e-11	5.23e-09	4.70e-09
5	Σ_{90}^w vs Σ_{10}^p	1.32e-10	1.53e-08	1.38e-08
6	Σ_{30}^s vs p_1	2.81e-10	3.23e-08	2.95e-08
7	p_1 vs s_3	3.78e-10	4.31e-08	3.97e-08
8	p_2 vs MA_5^t	5.09e-10	5.75e-08	5.34e-08
9	Σ_{90}^w vs p_1	1.06e-09	1.19e-07	1.11e-07
10	Σ_{10}^p vs p_7	2.89e-09	3.20e-07	3.03e-07
11	p_1 vs p_7	1.99e-08	2.18e-06	2.08e-06
12	Σ_{30}^s vs p_2	2.31e-07	2.52e-05	2.42e-05
13	s_3 vs p_2	2.96e-07	3.19e-05	3.11e-05
14	p_3 vs MA_5^t	4.82e-07	5.16e-05	5.06e-05
15	Σ_{90}^w vs p_2	6.91e-07	7.33e-05	7.26e-05
16	Σ_{10}^p vs Σ_{90}^w	6.80e-06	0.00	0.00
17	p_2 vs p_7	7.58e-06	0.00	0.00
18	Σ_{10}^p vs doy	1.44e-05	0.00	0.00
19	Σ_{10}^p vs p_5	1.44e-05	0.00	0.00
20	Σ_{10}^p vs p_6	1.60e-05	0.00	0.00
21	Σ_{180}^h vs MA_5^t	2.43e-05	0.00	0.00
22	Σ_{10}^p vs p_4	2.98e-05	0.00	0.00
23	p_1 vs Σ_{90}^w	2.98e-05	0.00	0.00
24	doy vs p_1	6.00e-05	0.01	0.01
25	p_1 vs p_5	6.00e-05	0.01	0.01
26	p_6 vs p_1	6.62e-05	0.01	0.01
27	Σ_{30}^s vs p_3	6.62e-05	0.01	0.01
28	s_3 vs p_3	8.04e-05	0.01	0.01
29	MA_5^t vs Σ_{30}^s	0.00	0.01	0.01
30	p_1 vs p_4	0.00	0.01	0.01
31	Σ_{10}^p vs Σ_{30}^s	0.00	0.01	0.01
32	Σ_{90}^w vs p_3	0.00	0.01	0.01
33	p_4 vs MA_5^t	0.00	0.04	0.04
34	p_1 vs Σ_{30}^s	0.00	0.04	0.04
35	Σ_{10}^p vs Σ_{180}^h	0.00	0.05	0.04
36	p_6 vs MA_5^t	0.00	0.07	0.06

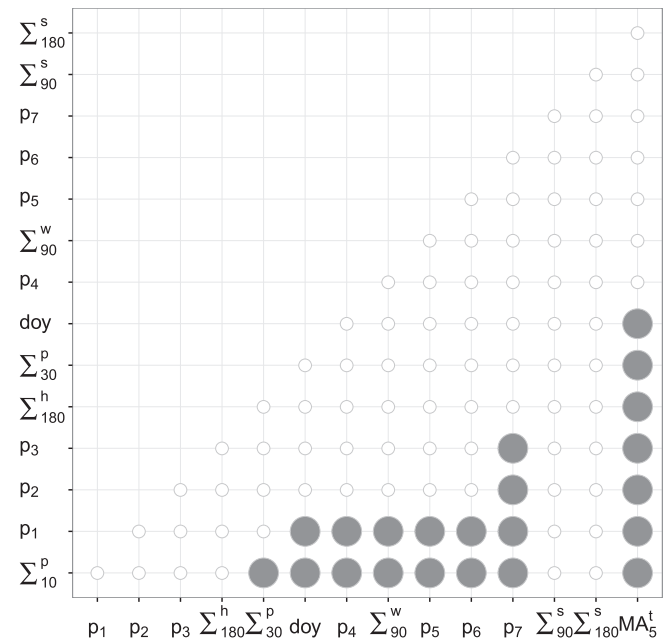


Fig. 4. Pairwise rejected hypothesis at $\alpha = 0.05$ with adjusted Shaffer p -values. The dots indicate pairs of variables which are not equally important.

the set to just the 5 most important variables. Then, for both reduced sets and for each location, 50 instances of a RF model were trained to produce one day-ahead forecasts. Again, the models were trained through the LOO cross-validation approach by entire years.

The predictive performance of the models in terms of RMSE and R^2 , for the setup using all variables (setup A), the reduced sets of 15 (setup B) and 5 (setup C) most important variables is shown in Table 5. Their respective execution times (for a single instance) are shown in Table 6.

From the study of both tables, we see that the reduction from the setup A using 144 variables to just the 15 most important of them (setup B) implies a reduction of almost a third of the execution time. Most importantly, models built with the reduced set of 15 variables yield better RMSE results in average (only in Alcobendas there is a slightly worse error result). In terms of R^2 , both setups are approximately equal. Therefore, a first conclusion is that removing redundant non-important variables helps the models to converge to better results.

Regarding setup C, which uses just the 5 most important variables, the execution time is cut to one half with respect to setup A. However, the RMSE and R^2 values are worst in average, which indicates that such a reduced set is not enough to capture the inner characteristics of the pollen time series. In other words, the reduction leaves some important variables out, resulting in worse models.

By closely studying the results, we have identified a pattern which could explain some of the difficulties of the model in capturing the

Table 5

Average RMSE and R^2 of the test years studied at each location.

Station	RMSE _A	RMSE _B	RMSE _C	R_A^2	R_B^2	R_C^2
Alcalá	19.05	18.06	18.64	0.62	0.57	0.50
Alcobendas	15.42	15.57	14.99	0.70	0.68	0.59
Aranjuez	16.64	16.41	18.05	0.64	0.58	0.48
Farmacia	18.66	17.64	18.16	0.63	0.59	0.50
Getafe	16.34	16.11	21.87	0.69	0.70	0.49
Leganés	18.43	16.41	20.22	0.71	0.71	0.54
Villalba	17.46	16.43	20.61	0.58	0.69	0.55
Average	17.43	16.66	18.90	0.65	0.64	0.52

Table 6

Average execution time (in seconds) for the models built using different subsets of variables.

Setup	# variables	Time
A	144	29.00
B	15	20.29
C	5	14.75

inner behaviour of the data series. Precisely, the best performing test year across locations (2007) obtain an average RMSE equal to 3.84 grains/m³, while the average RMSE in the worst performing year (2010) goes up to 28.16 grains/m³. Concretely, this situation is related to the usual appearance of sudden extreme pollen concentration peaks during the pollination season.

These findings indicate that the models consistently fail to forecast the whole height of the peaks, producing an increase of the error: the higher the peak, the bigger the error. Another example can be seen in Fig. 5, where it is shown how, for the Alcobendas site in 2008, there are a number of sudden peaks, which result in a high average RMSE of 31.02 grains/m³, while, in 2007 for the same location, the RMSE obtained was of 6.89 grains/m³. This is most probably related to the lack of concentration peaks over 100 grains/m³ for that year.

Finally, further evidence of this can be obtained from Table 5, where we can see that the model obtains low precision when forecasting concentrations exceeding 150 grains/m³, as is the case for the Faculty of Pharmacy (Farmacia) and Alcalá where peak concentrations up to 400 grains/m³ were observed.

4. Discussion

As we have seen, there is statistical evidence of the existence of relevant groups of variables when forecasting one day-ahead airborne pollen concentrations. The proposed approach succeeds in ranking and identifying these influential features, finding that previous pollen daily observations and 10-day cumulative airborne concentrations (which have both been proved to serve as indicators of the state of development of the plant (Ribeiro et al., 2007; Smith and Emberlin, 2006)) are the two top ranked features. These results, entirely based on the data and free from *a priori* assumptions are supported by nonparametric hypothesis tests and *post-hoc* procedures, and are coherent with previous phenological studies.

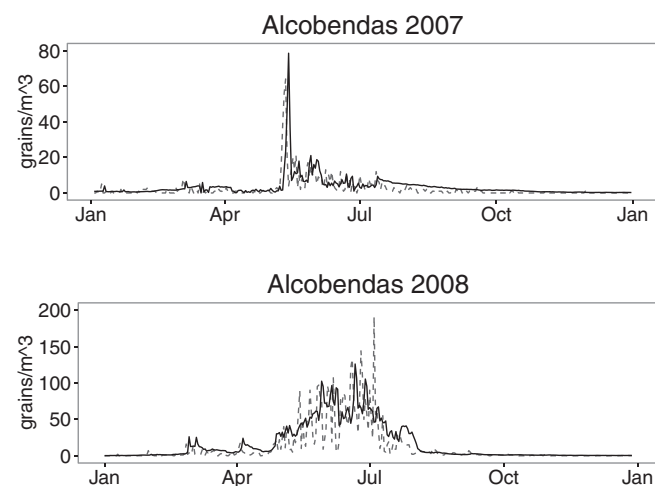


Fig. 5. Forecast (solid line) and observed (dashed) pollen concentration values for two years in the site of Alcobendas.

They have immediate practical application, as proven by the fact that subsets of top-ranked variables produce better results while reducing computational complexity.

However, the appearance of sudden high airborne concentrations increases the error on the performance of the model, as it is not able to capture extreme pollen levels given the limited amount of observations of this class. This phenomenon is common to other approaches, and could be investigated in the framework of the extreme value theory. However, it does not seem to be a particularly worrying issue, as the metric could be improved by limiting the observed concentrations to avoid the high peaks without hampering the usefulness of the proposal. For example, it could make sense to impose an upper limit to the values of the series as in Navares and Aznarte (2016). This limit should be related to the concentration levels which are considered to be risky for human health, for instance, symptoms are reported to appear over 30 grains/m³ in Finland and Croatia (Peternel et al., 2005; Rantio-Lehtimäki et al., 1991), while in Spain the first symptoms are reported at 25 grains/m³ (Rodríguez-Rajo et al., 1983). Another option is to transform the data into a logarithmic scale (or similar) as in Aznarte et al. (2007).

However, the results presented in this work are comparable in terms of accuracy to other studies such as Iglesias-Otero et al. (2015), which reports an average R² of 0.66 compared to an average 0.65 in our setup B with 15 variables. Furthermore, the approach proposed by Iglesias-Otero et al. (2015) uses artificial neural networks (ANN) which, as shown in the research, is strongly dependent on the configuration of the network. RF is considered as a more robust approach to mitigate overfitting, limiting the assumptions to be taken by the practitioner, avoiding decisions such as network structures and requiring minimal preprocessing of the data. The same applies to Astray et al. (2016) who achieves a RMSE of 22.56 on the best topology of the ANN. In average, our models achieve an RMSE of 17.43 across all configurations, being the setup B with 15 variables the best configuration: it obtains an RMSE of 16.66. These values clearly outperform the best regression models proposed by Csépe et al. (2014) which achieves a RMSE of 28.26 on its best performing algorithm (M5P), which is an implementation for regression trees. The *bagging* technique used in RF gives an edge over M5P algorithm reducing estimation variance and consequently prediction error. RF are relatively complex in computational terms, hence the focus on variable selection, which has shown that setup B (15 variables) as the most optimal among the configurations tested.

Additionally, the use of the LOO approach across multiple locations produces more generalized results as opposed to a predefined test set, on which the results heavily rely on the specific characteristics of the years selected to define the test set.

5. Conclusions

This paper presents a new approach to forecast airborne Poaceae pollen concentrations in the region of Madrid by identifying the most influential among the set of available variables. It provides statistical evidence, through non-parametric hypothesis tests, of the benefits of selecting the most influential variables for a one day-ahead forecast horizon.

Concretely, from all the considered pollen and meteorological variables, the data indicate that previous days pollen concentrations and cumulative sums of recent pollen concentrations are among the most important. They are followed by cumulative sums of humidity and solar radiation. These findings support, from a pure data-based point of view, the conclusions of previous phenology-based studies.

However, the statistical significance of the results from a single experiment are always to be questioned, and for this reason, in this paper, a sound statistical procedure based on Friedman tests and *post-hoc* analysis is used to support and validate the conclusions.

Finally, the usefulness of the proposal is shown by training the model with a reduced set of the most important variables, eliminating redundancies. This reduced model has been proven to increase the accuracy while limiting the computational burden.

The results clearly outperform those obtained by other authors, although there is still room for improvement and further research. For example, the models have difficulties in predicting extremely high concentrations, something that could be addressed by imposing upper limits according to risk levels for allergy patients. Also, a systematic exploration of the different combinations that can be drawn from the set of important variables should be studied, together with the optimization of the trade-off between the number of variables and the accuracy.

Acknowledgments

This work has been partially funded by the Ministerio de Economía y Competitividad, Gobierno de España, through a Ramón y Cajal grant (RYC-2012-11984).

The authors would like to thank Patricia Cervigón (Comunidad de Madrid) and Montserrat Gutiérrez Bustillo (Universidad Complutense de Madrid) for their assistance in obtaining the data for this study.

References

- Andersen, T.B., 1991. A model to predict the beginning of the pollen season. *Grana* 30, 269–275.
- Astray, G., Fernández-González, M., Rodríguez-Rajo, F., López, D., Mejuto, J., 2016. Airborne castanea pollen forecasting model for ecological and allergological implementation. *Sci. Total Environ.* 548–549, 110–121.
- Aznarte, J.L., Benítez Sánchez, J.M., Lugilde, D.N., de Linares Fernández, C., de la Guardia, C.D., Sánchez, F.A., 2007. Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. *Expert Syst. Appl.* 32 (4), 1218–1225.
- Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 25, 123–140.
- Breiman, L., 2001. Random forest. *Mach. Learn.* 45, 5–32.
- Cannell, M., Smith, R., 1983. Thermal time, chill days and prediction of budburst in *Picea sitchensis*. *J. Appl. Ecol.* 20, 269–275.
- Castellano-Méndez, M., Aira, M.J., 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. Biometeorology* 49, 310–316.
- Conover, W.J., 1999. Nonparametric methods. In: Wiley, B., O'Sullivan, M. (Eds.), *Practical Nonparametric Statistics*. John Wiley and Sons., pp. 233–305. <http://dx.doi.org/10.1002/bimj.19730150311>.
- Cotos-Yáñez, T., Rodríguez-Rajo, F., Jato, M., 2004. Short-term prediction of *Betula* airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models. *Int. J. Biometeorol* 48, 179–185.
- Csépe, Z., Makra, L., Voukantis, D., Matyasovszky, I., Tusnády, G., Karatzas, K., Thibaudon, M., 2014. Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in Europe. *Sci. Total Environ.* 542–552, 476–477.
- de Water, P.K.V., Keever, T., Main, C.E., Levetin, E., 2003. An assessment of predictive forecasting of *Juniperus ashei* pollen movement in the Southern Great Plains, USA. *Int. J. Biometeorol* 48, 74–82.
- de Weger, L.A., Bergmann, K.C., Rantio-Lehtimäki, A., Dahl, A., Buters, J., Déchamp, C., Belmonte, J., Thibaudon, M., Cecchi, L., Besancenot, J.-P., Galán, C., Waisel, Y., 2013. Impact of pollen. In: Sofiev, M., Bergmann, K.-C. (Eds.), *Allergenic Pollen*. Springer Netherlands., pp. 161–215. http://dx.doi.org/10.1007/978-94-007-4881-1_6.
- Deák, A., Makra, L., Matyasovszky, I., Csépe, Z., Muladi, B., 2013. Climate sensitivity of allergenic taxa in Central Europe associated with new climate change related forces. *Sci. Total Environ.* 442, 36–47.
- Demšar, J., 2006. Dec. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* 7, 1–30.
- Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* 32, 674–701.
- Galán Soldevilla, C., Cariñanos González, P., Alcázar Teno, P., Domínguez Vilches, E., 2007. *Manual de Calidad y Gestión de la Red Española de Aerobiología*. Universidad de Córdoba.
- García, S., Fernández, A., Luengo, J., Herrera, F., 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. *Inf. Sci.* 180, 2044–2064.
- Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.
- Iglesias-Otero, M.A., Fernández-González, M., Rodríguez-Caride, D., Astray, G., Mejuto, J.C., Rodríguez-Rajo, F.J., 2015. A model to forecast the risk periods of *Plantago* pollen allergy by using ANN methodology. *Aerobiologia* 31, 201–211.
- Jones, A., Harrison, R., 2004. The effects of meteorological factors on atmospheric bioaerosol concentrations—a review. *Sci. Total Environ.* 326, 151–181.
- Kmenta, M., Bastl, K., Kramer, M., Hewings, S., Mwange, J., Zetter, R., Berger, U., 2016. The grass pollen season 2014 in Vienna: a pilot study combining phenology, aerobiology and symptom data. *Sci. Total Environ.* 566–567, 1614–1620.
- Matyasovszky, I., Makra, L., Csépe, Z., Sümegehy, Z., Deák, A., Pál-Molnár, E., Tusnády, G., 2015. Plants remember past weather: a study for atmospheric pollen concentrations of *Ambrosia*, *Poaceae* and *Populus*. *Theor. Appl. Climatol.* 122, 181–193.
- Myszkowska, D., 2014. Predicting tree pollen season start dates using thermal conditions. *Aerobiologia* 30, 307–321.
- Navares, R., Aznarte, J., 2016. Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. *Int. J. Biometeorol* <http://dx.doi.org/10.1007/s00484-016-1242-8>.
- Otero, J., García-Mozo, H., Hervás, C., Galán, C., 2013. Biometeorological and autoregressive indices for predicting olive pollen intensity. *Int. J. Biometeorol* 57, 307–316.
- Palacios, I.S., Molina, R.T., Rodríguez, A.F.M., 2000. Influence of wind direction on pollen concentration in the atmosphere. *Int. J. Biometeorology* 44, 128–133.
- Pauling, A., Gehrig, R., Clot, B., 2014. Toward optimized temperature sum parametrizations for forecasting the start of the pollen season. *Aerobiologia* 30, 45–57.
- Peternel, R., Srncic, L., Culig, J., Hrga, I., Hercog, P., 2005. Poaceae pollen in the atmosphere of Zagreb (Croatia), 2002–2005. *Grana* 45, 130–136.
- Rantio-Lehtimäki, A., Koivikko, A., Kupias, R., Mäkinen, Y., Pohjola, A., 1991. Significance of sampling height of airborne particles for aerobiological information. *Allergy* 46, 68–76.
- Ribeiro, H., Cunha, M., Abreu, I., 2007. Definition of main pollen season using logistic model. *Ann. Agric. Environ. Med.* 14, 259–264.
- Rodríguez-Rajo, F., Dopazo, A., Jato, V., 2004. Environmental factors affecting the start of pollen season and concentrations of airborne *Alnus* pollen in two localities of Galicia (NW Spain). *Ann. Agric. Environ. Med.* 11, 35–44.
- Rodríguez-Rajo, F., Frenguelli, G., Jato, M., 1983. Effect of air temperature on forecasting the start of the *Betula* pollen season at two contrasting sites in the south of Europe (1995–2001). *Int. J. of Biometeorology* 47, 117–125.
- Sánchez-Mesa, J., Smith, M., Emberlin, J., Allitt, U., Caulton, E., Galán, C., 2003. Characteristics of grass pollen seasons in areas of southern Spain and the United Kingdom. *Aerobiologia* 19, 243–250.
- Shaffer, J., 1986. Modified sequentially rejective multiple test procedures. *J. Am. Stat. Assoc.* 81, 826–831.
- Smith, M., Emberlin, J., 2006. A 30-day-ahead forecast model for grass pollen in north London, UK. *Int. J. Biometeorology* 50, 233–242.
- Sofiev, M., Siljamo, P., Ranta, H., Linkosalo, T., Jaeger, S., Rasmussen, A., Rantio-Lehtimäki, A., Severova, E., Kukkonen, J., 2013. A numerical model of birch pollen emission and dispersion in the atmosphere. Description of the emission module. *Int. J. Biometeorol* 57, 45–58.
- Subiza, J., Jerez, M., Jiménez, J., Narganes, M., Cabrera, M., Varela, S., Subiza, E., 1995. Allergenic pollen pollinosis in Madrid. *J. Allergy Clin. Immunol.* 96, 15–23.
- Tassan-Mazzocco, F., Felluga, A., Verardo, P., 2015. Prediction of wind-carried Gramineae and Urticaceae pollen occurrence in the Friuli Venezia Giulia region (Italy). *Aerobiologia* 31, 559–574.
- Vogel, H., Pauling, A., Vogel, B., 2008. Numerical simulation of birch pollen dispersion with an operational weather forecast system. *Int. J. Biometeorol* 52, 805–814.