

Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features

Ricardo Navares¹ · José Luis Aznarte² 

Received: 24 May 2016 / Revised: 24 August 2016 / Accepted: 25 August 2016 / Published online: 16 September 2016
© ISB 2016

Abstract In this paper, we approach the problem of predicting the concentrations of Poaceae pollen which define the main pollination season in the city of Madrid. A classification-based approach, based on a computational intelligence model (random forests), is applied to forecast the dates in which risk concentration levels are to be observed. Unlike previous works, the proposal extends the range of forecasting horizons up to 6 months ahead. Furthermore, the proposed model allows to determine the most influential factors for each horizon, making no assumptions about the significance of the weather features. The performance of the proposed model proves it as a successful tool for allergy patients in preventing and minimizing the exposure to risky pollen concentrations and for researchers to gain a deeper insight on the factors driving the pollination season.

Keywords Poaceae · Pollen · Random forest · Forecasting · Time series

Introduction

Continuously increasing allergy symptoms in developed countries, and the clinic and socioeconomic relevance of

this problem, have boosted recent research around some of the issues dealt with by aerobiology, especially concerning predictive models. The fact is that not only has the number of cases increased, but also the severity and the prevalence of the reactions (de Weger et al. 2013). In order to enable preventive measures and reduce the exposure for patients, this study focuses on the prediction of pollen concentration levels which imply high risk for allergic population.

The main pollination season is defined as the period where high pollen concentrations are measured. In the literature, several definitions of what it is considered a pollen season have been established (Jato et al. 2006). It is possible to classify them into two main approaches, those based on the cumulative daily atmospheric concentrations (Andersen 1991; Galán et al. 1995; Nilsson and Persson 1981) and those based on a predefined threshold level over which the season is defined to start and end Sánchez-Mesa et al. (2003).

Weather plays a major role in the severity and length of the pollination season, as it is the cause for increases and decreases of the pollen concentration levels through its effect on the plants. For example, a mild winter usually implies an early pollen season, as it influences the plant development stages prior to the flowering (Cannell and Smith 1983; Pauling et al. 2014). On the other hand, a dry and windy weather spreads the airborne quickly, leading to higher distributions (Myszkowska 2014; Rodríguez-Rajo et al. 1983). In this study, we investigate the meteorological effects which determine the season of Poaceae pollen in Madrid, Spain. In our approach, the forecasting problem is cast to a binary classification problem with attention to the thresholds considered risk levels for the appearance of allergy reactions.

Several research teams have established models to predict the pollination season based on assumptions about

✉ José Luis Aznarte
jlaznarte@dia.uned.es

¹ Superior Technical School of Computer Engineering, UNED, Juan del Rosal, 16, 28040, Madrid, Spain

² Department of Artificial Intelligence, UNED, Juan del Rosal, 16, 28040, Madrid, Spain

the influence of meteorological conditions (Andersen 1991; Myszkowska 2014; Pauling et al. 2014; Rodríguez-Rajo et al. 1983) or previous pollen concentrations (Castellano-Méndez et al. 2005). The aim of this research is to provide an assumption-free predictive model using a computational intelligence technique known as random forests (RF) (Breiman 2001). The study lets the RF select the most influential features from a purely data point of view according to their predictive significance and provides this information allowing for interpretability of the results. Earlier applications of computational intelligence methods can be found, for example, in Aznarte et al. (2007).

Very few of the previous predictive studies for pollen were able to provide this type of information about the relevance of the variables. And most of them dealt with forecasts horizons ranging from 1 to 10 days (Andersen 1991; Castellano-Méndez et al. 2005; Myszkowska 2014). The procedure presented in this work provides long-term predictions, up to 180 days, expanding their usefulness to prevent allergy symptoms.

The aim of this study is to provide a framework to forecast and identify the main factors which influence high pollen concentrations, and do this from a purely data-driven point of view. These long-term predictions could help research centers and clinical institutions to plan in advance the implications of high airborne concentrations and their duration, as well as allergy patients to be able to limit their exposure to risky pollen levels. Furthermore, this study is also aimed to provide support to phenological studies by identifying the relevant pollination factors from the information obtained from the data.

Materials and methods

Data description

Weather data Meteorological data are provided by Ayuntamiento de Madrid for the weather stations located in Casa de Campo, Plaza de España and Cuatro Caminos. Weather observations consist of average daily temperature in Celsius degrees, hours of sunlight per day, wind speed measured in m/s, daily rainfall in mm/h, pressure in mbar, degree of humidity in percentage, and ultraviolet radiation in mW/m². Very few missing observations appear in the meteorological series, and these were linearly interpolated.

Pollen data Pollen observations correspond to daily Poaceae concentrations registered at the Faculty of Pharmacy of Complutense University of Madrid, Spain (located at 40°26′52.1″ N, 3°43′41.1″ W) from 2000 to 2013. These data have been kindly provided by Red Palinológica de la Comunidad de Madrid and were obtained following the standard methodology of the Spanish Aerobiological Network. They are measured in grains per cubic meter of air.

Missing values in the pollen time series may lead to an artificial delay of the season start, especially when those appear in the critical months of February, March, and April, as it is when the daily concentrations are expected to increase. Table 1 shows, for instance, high presence of consecutive missing values on March 2001 and August 2009 compared to other months. These are a priori critical months as the season might start and end on those periods. Using the standard 'last observation carried forward' (LOCF) method

Table 1 Maximum number of consecutive days of missing data per month and year

Year	Month											
	1	2	3	4	5	6	7	8	9	10	11	12
2000	–	1	2	–	–	–	–	–	–	–	4	–
2001	–	–	8	–	2	–	–	–	–	–	–	–
2002	–	–	–	–	–	–	–	3	–	–	–	–
2003	–	–	–	–	–	–	–	–	–	–	–	–
2004	–	–	–	–	–	–	–	–	–	–	–	–
2005	2	1	3	1	–	–	–	–	–	3	–	–
2006	–	–	–	–	–	–	–	–	–	–	–	–
2007	–	–	–	–	–	–	–	–	–	–	–	–
2008	–	–	–	–	–	–	–	–	–	–	–	–
2009	–	–	–	–	1	–	–	18	2	1	–	11
2010	7	–	–	–	–	–	1	–	–	–	–	–
2011	–	–	–	–	–	–	–	–	–	–	–	–
2012	4	–	–	–	–	2	–	–	–	–	–	–

to estimate the missing observations does not fully solve this problem. Thus, we applied a redistribution of the data into a matrix of dimensions $N \times 365$, being N the number of available years. (No data were missing for the 29th of February in any year, so that day was removed from all the years to make the matrix dimensions match).

Out of this set up, two new matrices are generated to regress the missing data points by rows (within each year) and by columns (by years). Data suggests that concentration levels for the same day in different years do not imply similar levels in another, and hence, the resulting matrices are weighted to give more relevance to most recent data (within each year), as in:

$$p_t = \beta \cdot r_{\text{row}} + (1 - \beta) \cdot r_{\text{col}}, \quad (1)$$

where r_{row} is a linear regression within the year, r_{col} is a linear regression across years, and $\beta = 0.6833$ is estimated from the data.

Season definition

In literature, the main pollination season is defined according to two different approaches (Jato et al. 2006). The first one is based on daily cumulative airborne concentrations

and the second considers the season started when a pollen concentration threshold is consistently surpassed.

There is no general consensus about the definition of the pollination season, and hence, season dates might differ according to their definition. Table 2 shows the differences between approaches on selected years and authors with their corresponding definition of the main pollination season. Threshold-based approaches such as Feher and Jarai-Komlodi (1997) and Sánchez-Mesa et al. (2003) tend to limit the season where peak concentrations appear, and this implies a high sensitivity to isolated peak concentrations. In contrast, cumulative approaches widen the pollination period being sensitive to early moderate concentrations, as is the case for 2002 in the table. Figure 1 shows how restrictive the proposal of Sánchez-Mesa et al. (2003) is compared to Nilsson and Persson (1981) and how the season period varies by reducing the threshold to 15 grains/m³.

However, in order to forecast the season start as defined by the cumulative approaches, it would first be necessary to forecast the expected total yearly accumulation, which determines the percentages to define the pollination season. Of course, this is unfeasible as it implies forecasting one quantity (the yearly sum) in order to forecast the other (a quantile). Hence, this study will be restricted to threshold-based season definitions. In what follows, if u is a fixed

Table 2 Sample start and end of the pollination season according to different definitions

Approach	Definition	Year	Start	End
Nilsson and Persson 1981	The day in which the sum of daily pollen concentration reaches a value over 5 % (start) and 95 % (end) of the total yearly sum.	2002	09 Feb	03 Nov
		2004	26 Feb	11 Jul
		2009	04 Apr	09 Sep
		2012	17 May	31 Aug
(Galán et al. 1995)	The day in which the sum of daily pollen concentration reaches a value over 1 % (start) and 99 % (end) of the total yearly sum.	2002	20 Jan	27 Dec
		2004	11 Jan	15 Sep
		2009	08 Mar	30 Oct
		2012	7 Feb	30 Nov
(Andersen 1991)	The day in which the sum of daily pollen concentration reaches a value over 2.5 % (start) and 97.5 % (end) of the total yearly sum.	2002	26 Jan	01 Dec
		2004	21 Jan	03 Aug
		2009	14 Mar	29 Sep
		2012	03 Mar	21 Sep
(Sánchez-Mesa et al. 2003)	The first day in which the daily pollen concentration reaches values over (start) and below (end) 30 grains/m ³	2002	17 May	03 Jun
		2004	11 Jan	15 Sep
		2009	07 May	30 Oct
		2012	25 May	18 Jun
(Feher and Jarai-Komlodi 1997)	The first day in which the (start) threshold reaches values over and below (end) 3 grains/m ³ for 4 consecutive days	2002	05 Feb	19 Jun
		2004	11 Apr	03 Jul
		2009	26 Sep	30 Oct
		2012	22 Aug	12 Oct

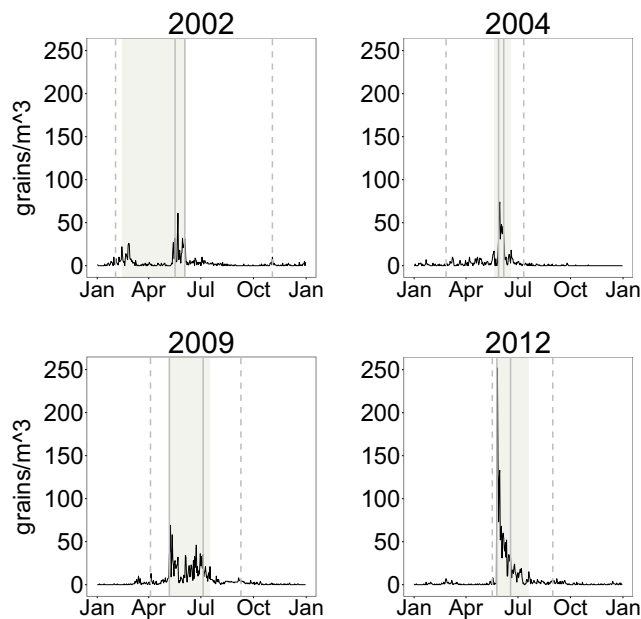


Fig. 1 Pollen concentrations for years 2002, 2004, 2009 and 2012 and definition of the season according to Nilsson and Persson (1981) (vertical dashed line) and Sánchez-Mesa et al. (2003) (vertical solid line). The shaded rectangle represents the latter approach relaxing the threshold to 15 grains/m³

daily pollen concentration threshold, then the pollen season starts (ends) at the first (last) day that surpasses u .

In literature, pollen concentration levels show regional variations on pollen reactivity. For instance, according to Peternel et al. (2005) and Rantio-Lehtimäki et al. (1991), symptoms appear over 30 grains/m³ in Finland and Croatia, while in Spain the first symptoms are observed between 25 grains/m³ (Rodríguez-Rajo et al. 1983) and studies such as Sánchez-Mesa et al. (2003) use 30 grains/m³. By far, the most common threshold level found in the literature is 30 grains/m³ (Castellano-Méndez et al. 2005; Green et al. 2004; Sánchez-Mesa et al. 2003) which corresponds to the concentration at which the first allergy symptoms appear. Therefore, this level is selected as a representative in this study.

Features

In the pollen forecasting framework, the set of independent variables should contain relevant meteorological data as well as past pollen levels, as all of them are known to play a crucial role in predicting pollen concentrations. At the same time, due to the “curse of dimensionality” and to ease the computational burden, it is important to avoid including features which might not influence the pollen production at a certain time frame as it is. An example would

be the rainfall registered 3 years before the forecast date: it will hardly be of interest to forecast the pollen season for that date. In our approach, feature relevance will be considered under different forecast horizons, thus enabling the proposed model to tell which set of independent variables are more influential for each horizon.

Cumulative pollen observations prior to the forecast date have been proved to serve as an indicator of the development stage of a plant (Ribeiro et al. 2007; Smith and Emberlin 2006). Correspondingly, 10- and 30-day cumulative sums of daily atmospheric concentrations prior to the forecast date are included as independent variables, along with the prior week daily concentrations for each date. Additionally, pollen accumulation within the year is also used as a proxy of the state of the plant.

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). Sum of temperatures up to some point are usually considered as a good representation of this absorbed energy (Cannell and Smith 1983; Andersen 1991; Rodríguez-Rajo et al. 1983). Other authors (Pauling et al. 2014) however, use the concept of *chilling temperatures* and *forcing temperatures*, which are defined as the weighted sum of temperatures below or above certain levels for a fixed period. 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 and 60 days prior the forecast date:

$$F_{\text{sum}}(d) = \sum_{i=d-n}^d R_{\text{forc}}(i), \quad (2)$$

where

$$R_{\text{forc}}(i) = \begin{cases} 0 & \text{if } T(i) < T_{\text{forc}} \\ T(i) - T_{\text{forc}} & \text{if } T(i) \geq T_{\text{forc}} \end{cases}, \quad (3)$$

being d is the forecast date, n is the number of days which define the calculation period for the sum of forcing temperatures, $T(i)$ is the temperature for day i , and T_{forc} is the base temperature for forcing (all temperatures are in degrees Celsius). The same applies for chilling. Base forcing and chilling temperatures for a determined threshold are derived using geometrical relations from the reference of {1°C, 16°C} for the forcing period and {−6°C, 8°C} for the chilling period at thresholds of 10 grains/m³ and 50 grains/m³ respectively, as in Pauling et al. (2014).

The cumulative approach introduced for temperatures is also used to capture rainy and humid periods. Humidity and rain prevent pollen spread during pollination, and

humid and rainy weather causes grass species to become more abundant during the growing period of the plant urging to include short and long term periods prior the forecast date.

Pollen dispersion being a fundamental aspect of the problem, wind speed is recognized as an important influential factor (Palacios et al. 2000). Hence a 30-days cumulative sum of wind speed features is generated. For all climate data, similar as for the pollen concentrations, the prior 7 daily raw data observations are also included.

This leads to the availability of 70 features as detailed in Table 3, to which we added a dummy variable which represents the day of the year. This makes 71 features which are distributed in a matrix corresponding to the desired forecast horizon and the discretized class:

$$\begin{bmatrix} x_{1,1} & \dots & x_{1,71} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,71} \end{bmatrix} \begin{vmatrix} p_1 \\ \vdots \\ p_n \end{vmatrix} \rightarrow \begin{bmatrix} x_{1,1} & \dots & x_{1,71} \\ \vdots & \ddots & \vdots \\ x_{n-t,1} & \dots & x_{n-t,71} \end{bmatrix} \begin{vmatrix} c_{1+t} \\ \vdots \\ c_n \end{vmatrix} \quad (4)$$

$$c_i = \begin{cases} 0 & \text{if } p_i < u \\ 1 & \text{if } p_i \geq u \end{cases}, \quad (5)$$

where p_i is the daily pollen observation at time i , t is the forecast horizon in number of days, and u is the threshold.

Table 3 Number of features generated by variable

	i	10	30	y	m	q	std
Pollen	7	1	1	1	—	—	—
Temperature	7	—	—	—	—	—	—
T _{forc}	—	—	—	—	1	1	—
T _{chill}	—	—	—	—	1	1	—
Humidity	7	—	—	—	1	—	—
Wind	7	—	—	—	1	—	—
Rain	7	—	—	—	1	—	1
Pressure	7	—	—	—	1	—	—
UV	7	—	—	—	1	—	—
Sun	7	—	—	—	1	—	—

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

10: previous 10-day cumulative sum

30: previous 30-day cumulative sum

y: year to date cumulative sum

m: previous month cumulative sum

q: previous 90-day cumulative sum

std: previous 15 days standard deviation

Random forest

Proposed for the first time in Breiman (2001), a random forest is an ensemble approach which leverages the performance of many simple decision trees that can be used to produce predictive models. It is a supervised learning procedure which combines several randomized decision trees and aggregates their predictions by averaging. The procedure operates over sample fractions of the data, grows a randomized tree predictor on each one and aggregate these predictors together.

The motivation to favor RF against other methods, like logistic regression (LR) is to avoid a correlation-based feature selection. It is known that LR is highly sensitive to variable collinearity and, as some features were generated from others, the parameterization of LR could be expensive in order to avoid overfitting. In this point we believe RF is a more robust approach. Given the relatively high number of instances and the presence of sudden high peaks in pollen concentrations as seen in Fig. 1, RF provides stability and accuracy in presence of outliers due to the *bagging* (Breiman 2001) technique.

Several decisions need to be made in order to build a RF model and to test its predictability. In order to optimize the execution, an analysis of the parameter search space needs to be done to precisely choose the parameter set up for each predictor.

To compare the performance of the different models resulting from the parameter set up, the area under the ROC curve generated by each model (AUC) is used. An ROC curve is a two-dimensional depiction of classifier performance (Fawcett 2003). The AUC of a classifier express the probability that the classifier will rank a randomly chosen instance which is correctly classified.

To test the optimal parameter set up the system performs a grid search to identify the best set of hyperparameters for the model based on the selected metric.

One of the strengths of random forests is that they are able to provide a measure of variable importance as a by-product of the model training. Breiman (2001) and Breiman (2002) proposed the evaluation of the importance of a variable x_i by adding up the weighted Gini impurity decreases for all nodes where x_i appears, and averaging over all the trees in the forest. Every node in a decision tree is designed to split the data set into two as a condition on a single variable. The measure on which the optimal split condition is chosen is called the Gini impurity. Thus when training a tree, it can be computed how much each feature decreases the weighted impurity in a tree being the average of these decreases the rank of the feature in the forest. This gives

a view on how important each variable is, and allows for further interpretability of the results.

Experimental design

Algorithm 1 System design

Require: $X_i = \{x_{i1}, x_{i2}, \dots, x_{i71} | c_i\}$ $i \in \{1 \dots n\}$
Require: $u = \text{threshold}$

```

1: for all  $[t, y]$  in  $\{\text{horizons}, \text{years}\}$  do
2:    $AUC = 0$ 
3:    $X^S = \text{Preprocess}(X_i, u, t)$  ▷ Apply (4) (5)
4:    $X_{\text{test}}^S = X_y^S$ 
5:    $X_{\text{train}}^S = X^S - X_{\text{test}}^S$ 
6:   for  $k \in [1, 15]$  do
7:      $\text{parameter}_k = \text{Grid.Search}(\text{search\_space})$ 
8:      $\text{model}_k = \text{Random.Forest}(\text{parameter}_k, X_{\text{train}}^S)$ 
9:     if  $AUC \leq AUC(\text{model}_k)$  then
10:       $AUC = AUC(\text{model}_k)$ 
11:       $\text{best} = \text{parameter}_k$ 
12:   end if
13: end for
14:  $\text{prediction} = \text{Random.Forest}(\text{best}, X_{\text{test}}^S)$ 
15:  $E = \text{Error}(\text{prediction})$ 
16: end for

```

The aim of this work is to help allergy patients and researchers in knowing in advance the period in which pollen concentrations will reach risk levels, and to identify the most influential factors for its prediction.

Given the very different shape of pollen concentrations and of the main pollination season across the observed years, as shown in Fig. 1, the experiments were tailored to find the best model available. From sudden high peak concentration levels in short periods to prolonged moderate atmospheric concentrations, the setup of the model has to be able to capture the inner available information to successfully predict the season.

Our approach is based on the idea that the pollen concentrations can be transformed into a binary classification problem where the featured instances represent influential factors. Daily pollen concentrations are mapped to $\{0, 1\}$ depending on whether they are above the threshold (1) or not (0).

In order to avoid overfitting and to provide a more generalized overview of the performance of the model, a *leave-one-out* (LOO) cross validation approach was taken to split the data into train and test set. For each year, the observations of that year were taken out as a test set, leaving the remaining years to train the model so the final metrics consist of the average error for each iteration. By averaging the metrics from the LOO technique, the results provided are more representative than selecting, for instance, the last two

years of the period as test set which would produce results very dependent on the characteristics of the selected years for testing.

As well, to provide a wider spectrum in order to give further information both for patients and researchers, the system provides forecasts for a wide set of time horizons, ranging from 1 day to 6 months. A forecast horizon of 15 days means that with the information available up to time t , the pollen concentration at day $t + 15$ is forecast.

Given the forecast horizons, vectors are build as in Eq. 4. The LOO approach is then applied by years. At each iteration, a random search is performed on the parameters taking into account the search boundaries and comparing the results for each set up. Finally, the best candidate is validated and its forecast metrics are provided. This process is summarized in Algorithm 1.

Results

In our setup, a set of forecast horizons were tested along with a threshold of 30 grains/m³. An optimal parameterization of the RF model was done using the LOO technique for the years between 2000 and 2013, leaving the remaining year of each iteration as test set. At each iteration, several metrics are generated as an estimator of system performance for each horizon. Given the different characteristics of each year studied, this method provides generalization letting the model learn the particular characteristics of each pollination season.

The second aim of this study is to identify the best predictors of the main pollination season. It is intended to provide a robust and flexible framework to obtain a good estimation of the predictors according to different forecast horizons.

The performance of the model is tested by checking the error rate of the class when it is classified as positive, this is the daily pollen concentrations which surpass the threshold. This measure is known as sensitivity or recall, and it measures the proportion of atmospheric concentrations over the defined threshold of 30 grains/m³ that were correctly classified as such. This measure is completed by the specificity which, on the contrary, measures the proportion of pollen concentrations below the threshold correctly classified. The global precision for both classes, above and below the threshold, is measured by the accuracy.

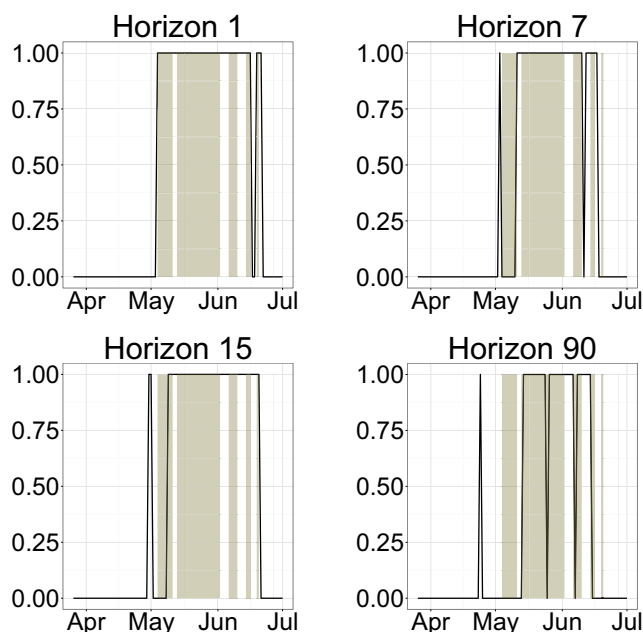
Forecast horizon

Table 4 shows the predictive metrics for each forecast horizon. Specificities and accuracies of over 90 % are achieved across the different horizons. The high values for specificity (true negative rate) indicate that the proposed model succeeds in identifying the periods of the main pollination

Table 4 Predictive Metrics. Totals based on LOO method for the study period between 2000 and 2013

Horizon	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	AUC
1	295	159	4198	36	0.891	0.964	0.958	0.972
5	282	244	4109	49	0.852	0.944	0.938	0.956
7	281	270	4081	50	0.849	0.939	0.932	0.939
15	302	333	4010	29	0.912	0.923	0.922	0.935
30	304	344	3984	27	0.918	0.921	0.918	0.935
60	308	326	3972	23	0.931	0.924	0.923	0.923
90	269	310	3972	48	0.849	0.928	0.924	0.922
120	264	318	3954	33	0.889	0.926	0.924	0.930
150	274	401	3686	22	0.926	0.902	0.904	0.924
180	262	320	3767	34	0.885	0.922	0.919	0.928

season with an acceptable rate of false negatives (predicting concentrations below the threshold inside the observed season). Figure 2 shows the prediction for 2001 with a forecast horizon of 1, 7, 15, and 90 days. Given the 30 grains/m³ threshold-based definition of the pollination season, the model manages to identify season start and end dates having a maximal error of 17 days for season start with the 90 days horizon. On the other hand, sensitivities are somehow lower, but attaining percentages over 84 % in all cases. This means that the model struggles to predict concentrations below the threshold when they appear during the main pollination season, showing a high number of false positives

**Fig. 2** Pollen observed over the threshold 30 grains/m³ (shaded) for Apr–Jul 2001 with forecast with horizon 1, 7, 15 and 90 days (solid lines)

(FP). Weather conditions during the pollination season such as heavy sudden rainfall might directly affect airborne concentrations resulting in rapid drops of pollen concentrations below the threshold. As this information is not available in the predictors, the proposal does not identify this specific conditions. We also believe this is due to the fact that the classes are unbalanced, as the pollen concentrations over the selected threshold represent only around 7 % of the total observations. Even though at each iteration of the RF double trees were built, which means bootstrap sampling from the minority class and drawing the same number of cases from the majority class to finally aggregate the predictions, there might be an improvement in this metric by penalizing misclassification of the minority class or limiting the period studied to the potential dates where high concentrations appear. This however would imply making some assumptions over the period studied which could increase the presence of missing data. For instance, missing early season start dates, i.e., end of February, if the assumption limits the study period from March to August.

It is interesting to see how the model performs for the longer forecast horizons, which in general show lower specificity and higher sensibility and, consequently, lower accuracy. This means a higher number of false positives, as illustrated in Fig 2 for the 90-days threshold. In this case, the model incorrectly predicts an early start of the season. In general, for longer horizons, there is a clear tendency of expanding the main pollination season showing a more loose decision when defining the boundary dates, and consequently increasing the number of false positives as the horizon increases.

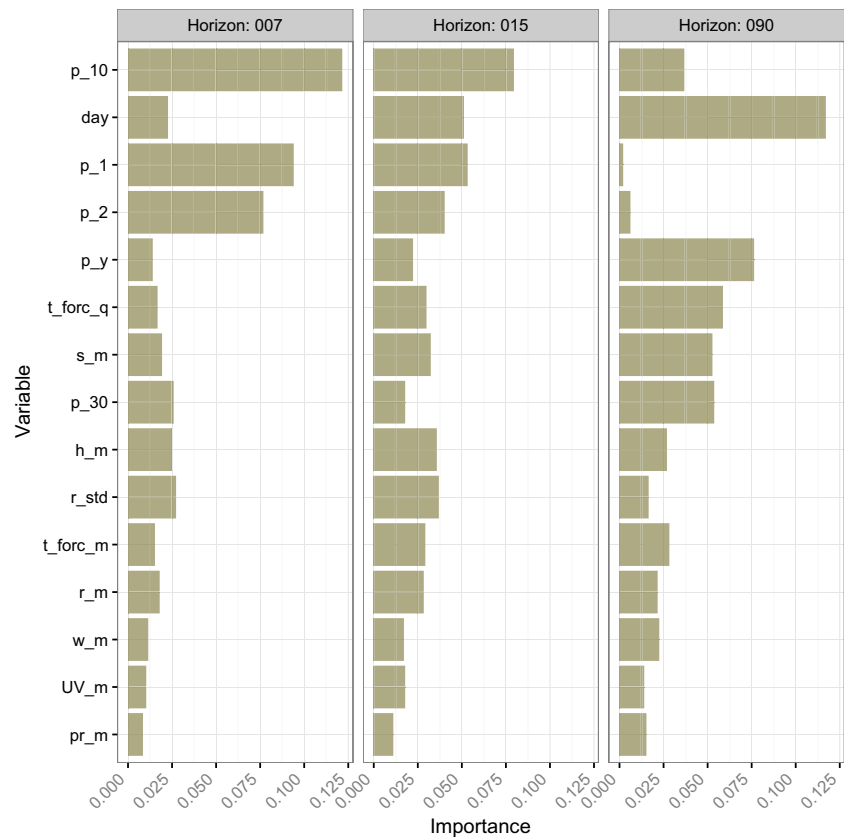
The model, on the other hand, manages to maintain a low and stable number of false negatives (FN) through the different horizons, which means that it succeeds in capturing the main periods where high concentrations appear.

It is noticeable that the decreasing accuracy pattern as the horizon increases is broken for the horizons of 60, 90, and 120 days, showing a small increase. This leads to think that the influential factors related to the previous winter period do play a key role in forecasting the start of the season.

Forecast horizon vs feature importance

In Fig. 3, the relative importance of the variables for a selected group of horizons is depicted. Each climate and pollen feature are labeled according to the method used to obtain it, as explained in Section 2. Hence, 'm' in the name of a variable denotes the accumulation of the daily featured data 30 days prior to the forecast date, 'q' represents the accumulation of daily data 90 days prior to the forecast date and 'y' the cumulative daily data from 1st January of the year in which the forecast lies. The data point of a variable x corresponding to the date $d - i$, being d the forecast date,

Fig. 3 Selection of the 15 most important variables by forecast horizon



is represented by x_i . Table 5 shows a detailed description of the most relevant features.

Clearly, for the 90 days horizon (rightmost graph), the influence of the forcing temperature is important for the prediction accounting a 6 % of the total importance compared

to the 2 % and the 2.6 % for the 7 and 15 days horizons, respectively. On the other hand, the results for short-term horizons (leftmost graph) show that the most recent pollen concentrations are the most influential factors. Previous day (p_1) and the day before the previous (p_2) pollen

Table 5 Variable Description

Variable	Description
w_m	wind speed accumulation one month prior the forecast day
UV_m	ultraviolet radiation accumulation one month prior the forecast day
t_forc_q	accumulated forcing temperature 90 days prior the forecast day
t_forc_m	accumulated forcing temperature 30 days prior the forecast day
s_m	sun hours accumulation one month prior the forecast day
r_std	standard deviation of rainfall one month prior the forecast day
r_m	rainfall accumulation one month prior the forecast day
p_y	accumulated pollen daily concentration from the first of January until the forecast date
pr_m	pressure accumulation one month prior the forecast day
p_30	daily pollen accumulation one month prior the forecast day
p_2	pollen daily concentration 2 days prior the forecast date
p_10	pollen daily concentration 10 days prior the forecast date
p_1	pollen daily concentration 1 days prior the forecast date
h_m	humidity accumulation one month prior the forecast day
day	day of the year

observations add up around 17 % of importance for the 5 days horizon while the contribution for the same features decreases to a 8.5 % and barely 1 % as the horizon increases to 15 and 90 days, respectively.

For short and medium horizons, the most influential features among the meteorological variables are the monthly cumulative humidity (h_m) and rainfall (r_m) and the 15 days standard deviation of rainfall (r_std). It is known that rainy and humid conditions wash away airborne concentrations during the pollination of the flower. Cumulative temperature features (t_forc_q) and the monthly accumulation of sun hours (s_m) are believed to boost the plant formation during the pre-flowering, thus the model weights a total of around 11 % of importance for the 90 days horizon in contrast to the 4 % achieved for 7 days. It can be clearly seen in Fig. 3 how these two variables gain importance as the horizon increases.

Discussion

As seen in Table 4, our proposal achieves accuracies which compare favorably to other studies, for example, (Brighetti et al. 2013), who obtained a value of 89.1 % for sensitivity and a value of 30.4 % for specificity for the 5 days horizon and the same threshold, whereas our model achieves a 94.4 % for sensitivity and a 85.2 % for specificity. This boost in specificity of course means that our model success in capturing the precise period when the main concentrations appear, achieving a much lower error rate outside the main pollination period which leads to an accuracy of 93.8 %. For a 1-day horizon, our model achieves an accuracy of 95.8 % compared to an average of 94.5 % of the two validation sets in Castellano-Méndez et al. (2005), being able to provide a more general approach when forecasting regardless the nature of the pollen series. Compared to the findings from Nowosad (2016), which also uses RF, our model achieves 96.4 % specificity compared to an average of 97 % for the 1 day horizon which implies a slightly lower performance when identifying low pollen concentration levels. On the other hand, our proposal achieves a 89.1 % sensitivity compared to 61, 70, and 88 % in Nowosad (2016), providing a higher hit rate when identifying high levels. This is the cause for the higher global accuracy compared to the reference techniques.

Regarding variable importance, our proposal suggests that, for horizons over 90 days, the importance of the forcing temperatures is higher compared to its role in shorter horizons, supporting the proposal of the optimal parameters in Pauling et al. (2014). Additionally, chilling temperatures are not ranked within the most influential features, confirming conclusions from Pauling et al. (2014) which hinted that chilling temperatures might lead to smaller error

reductions when forecasting. In addition, long-term horizons tend to weight more sunlight hours and rain features, which promote the formation of flowers during the pre-flowering months. Rainfall and humidity accumulations are positively related and influence the pollen release during the flowering period, in accordance to the findings of Aguilera et al. (2014). Hence, the model ranks these two features importances in accordance for short-term horizons.

Once the model is trained, producing forecasts takes less than a second on a 64-bit desktop Ubuntu machine with 6 cores and 32 GB of RAM. This of course allows the operational use of the approach.

Conclusions

The present paper introduces a new approach to forecast Poaceae pollen concentrations over different horizons making no assumptions on the phenology of the plant. It achieves consistent results in selecting the most influential factors given the forecast horizons. The selection of features from a purely data point of view is also consistent with different phenological studies while letting the model automatically select their relevance depending on the phases of the flower formation.

This study is tailored to help not only allergy patients but also research centers to prevent exposures to risk concentration levels for long-term horizons providing consistency up to 120 days prior the forecast data point. The model was tested on data from years 2000 to 2013, showing its adaptation and generalization regardless the specific characteristics of each pollen season.

The model proposed extends and supports the knowledge about the influence of meteorological factors on Poaceae pollen seasons. Although the results are promising, further efforts are required concerning the selection and generation of different features. Also, a wider experiment, using data from different sites, could shed more light into this interesting subject.

Acknowledgments This work has been partially funded by 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 his assistance in obtaining the data for this study.

The authors would also like to thank the anonymous reviewers for their useful, constructive, and valuable comments, which greatly improved the original version of the manuscript.

References

- Aguilera F, Fornaciari M, Ruíz-Valenzuela L, Galán C, Msallem M, Dhiab A, la Guardia CD, del Mar Trigo M, nd F Orlandi TB (2014)

- Phenological models to predict the main flowering phases of olive (*Olea europaea* L.) along a latitudinal and longitudinal gradient across the Mediterranean region. *Int J Biometeorology* 59:629–641
- Andersen TB (1991) A model to predict the beginning of the pollen season. *Grana* 30:269–275
- Aznarte JL, Benítez Sánchez JM, Lugilde DN, de Linares Fernández C, de la Guardia CD, Sánchez FA (2007) Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. *Expert Syst Appl* 32(4):1218–1225
- Breiman L (2001) Random forest. *Mach Learn* 45:5–32
- Breiman L (2002) Manual on setting up, using and understanding random forest. Stat Dept University of California Berkley v3.1
- Brighetti MA, Costa C, Menesatti P, Antonucci F, Tripodi S, Travaglini A (2013) Multivariate statistical forecasting modeling to predict Poaceae pollen critical concentrations by meteorological data. *Aerobiologia* 30:25–33
- 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 MJ, Iglesias I, Jato V, González-Manteiga W (2005) Artificial neural networks as a useful tool to predict the risk level of *Betula* pollen in the air. *Int J Biometeorology* 49:310–316
- Fawcett M (2003) Roc graphs: Notes and practical considerations for data mining researchers. Tech rep, HP Laboratories
- Feher Z, Jarai-Komlodi M (1997) An examination of the main characteristics of the pollen seasons in Budapest, Hungary (1991–1996). *Grana* 36:169–174
- Galán C, Emberlin J, Domínguez E, Bryant RH, Villamandos F (1995) A comparative analysis of daily variations in the Gramineae pollen counts at Cordoba, Spain and London, UK. *Grana* 34:189–198
- Green BJ, Dettman M, Yli-Panula E, Rutherford S, Simpson R (2004) Atmospheric Poaceae pollen frequencies and associations with meteorological parameters in Brisbane, Australia: a 5 year record, 1994–1999. *Int J Biometeorology* 40:172–178
- Jato V, Rodríguez-Rajo FJ, Alcázar P, Nunttiis PD, Galán C, Mandrioli P (2006) May the definition of pollen season influence aerobiological results? *Aerobiologia* 22:13–25
- Myszkowska D (2014) Predicting tree pollen season start dates using thermal conditions. *Aerobiologia* 30:307–321
- Nilsson S, Persson S (1981) Tree pollen spectra in the Stockholm region (Sweden), 1973–1980. *Grana* 20:179–182
- Nowosad J (2016) Spatiotemporal models for predicting high pollen concentration level of *Corylus*, *Alnus* and *Betula*. *Int J Biometeorology* 60:843–855
- Palacios IS, Molina RT, Rodríguez AFM (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, Srnc 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, 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
- Smith M, Emberlin J (2006) A 30-day-ahead forecast model for grass pollen in north London, UK. *Int J Biometeorology* 50:233–242
- de Weger LA, Bergmann KC, Rantio-Lehtimäki A, Dahl A, Buters J, Déchamp C, Belmonte J, Thibaudon M, Cecchi L, Besancenot JP, Galán C, Waisel Y (2013) Impact of Pollen. In: Sofiev M, Bergmann KC (eds) *Allergenic Pollen*, Springer Netherlands, pp 161–215. doi:10.1007/978-94-007-4881-1_6