



Forecasting *Plantago* pollen: improving feature selection through random forests, clustering, and Friedman tests

Ricardo Navares¹ · José Luis Aznarte²

Received: 12 August 2018 / Accepted: 14 July 2019
© Springer-Verlag GmbH Austria, part of Springer Nature 2019

Abstract

Predicting concentrations of pollen is of great importance both for patients and for public health institutions. In this paper, we present a forecasting approach which relies on data and makes no assumptions on the underlying phenomena affecting the plants and the pollination process. Machine learning is used to build a model and to select the most important variables for prediction. Through nonparametric hypothesis testing, we show how some variables are indeed more important than others and how the careful combination of these variables can lead to more accurate and parsimonious models which avoid the huge computational times of more complex models while outperforming them in terms of the precision of the forecasts. By increasing the richness of the selected variables based on the clustered Friedman importance ranks, prediction error is reduced from 4.57 to 4.40 grains/m³ as an average, which accounts for a 3.5% average improvement across locations studied with a 50% reduction of execution times.

1 Introduction

Allergy symptoms and their severity have been increasing in Western Europe during the last decades (de Weger et al. 2013), a fact which clearly implies the usefulness of the prediction of pollen concentrations, which are used not only to limit the exposure of patients to allergens but also to prearrange resources for clinical institutions. *Plantago* is one of the most common species among the herbaceous plants. Even though its airborne atmospheric concentrations are low, positive reactions appear nearly in 50% of the sensitive patients (Subiza et al. 1995).

Up to date, several studies have been proposed to approach the pollen forecasting problem based on techniques that range from classic multivariate regression to nonlinear models (Cotos-Yáñez et al. 2004; Rodríguez-Rajo

et al. 2004; Tassan-Mazzocco et al. 2015; Tseng et al. 2018), to artificial neural networks (Astray et al. 2016; Aznarte et al. 2007; Iglesias-Otero et al. 2015). Due to the *curse of dimensionality*, in every approach, variable selection plays a crucial role. For example, complex multivariate models require a precise parameterization in order to avoid overfitting or, in the case of neural networks, the number of instances needed to train them increase with the topology of the network (which in turn makes overfitting likely).

Regardless of the chosen approach, the first aim of the practitioner is to select a set of variables which are relevant when forecasting airborne pollen concentrations. In general, it is assumed that weather conditions directly influence these concentrations. For example, heavy sudden rains during pollination wash away pollen grains from the atmosphere. Consequently, studies include meteorological variables such as previous daily precipitation (Castellano-Méndez et al. 2005; Iglesias-Otero et al. 2015) or wind speed, relative humidity, and solar radiation as a proxy of dryness, which in turn creates optimal conditions for pollen proliferation (Jones and Harrison 2004; Myszkowska 2014; Rodríguez-Rajo et al. 1983). Alternatively, other studies prefer the use of climatological indices to capture meteorological information prior the pollen release. Some such indices are cumulative weather variables (Andersen 1991; Matyasovszky et al. 2015; Myszkowska 2014; Otero et al. 2013; Pauling et al. 2014), which capture the influence of past and current weather conditions in current airborne

✉ Ricardo Navares
rnavares2@alumno.uned.es

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

concentrations (Deák et al. 2013). On the other hand, in the literature, there are also models which take a phenological point of view (Cannell and Smith 1983; Kmenta et al. 2016; Ribeiro et al. 2007; Smith and Emberlin 2006; García-Mozo et al. 2008), assuming that future airborne pollen release depends on the current and past growth state of plant buds. Finally others considered the problem as a univariate time series problem, developing predictors based exclusively on past pollen concentrations (Aznarte et al. 2007).

In any case, there is a shared necessity to reduce model complexity in order to ease its interpretability and shorten calculation times. Feature selection constitutes an important research field in Computational Intelligence and Statistics. Automatic feature selection is aimed at obtaining a subset of variables which are relevant for model construction, removing those which are redundant or irrelevant thus minimizing the loss of information.

As an alternative of automatic feature selection, some authors used statistical and probabilistic methods to reduce the dimensionality of the variables in time series forecasting. Among these techniques, factor analysis (Deák et al. 2013; Matyasovszky et al. 2015) and partial mutual information (Li et al. 2015; Tran et al. 2015) are widely used. The underlying idea is to find a new representation of the observed variables that is good at explaining the phenomenon under study at the same time avoiding collinearity among them.

The objective of this paper is to provide a framework to select the optimal combination of features to achieve a more precise forecast of airborne *Plantago* pollen concentrations while avoiding a priori assumptions about the influence of the variables, either meteorological or phenological, on future pollen releases. Through the use of random forests (RF), we perform an automatic feature selection which lets the model capture the inner information from the observed data and decide the importance of each one of the available features. Once a parsimonious model is built with the more relevant variables, it is used to forecast the series under study with good results compared with the literature. These results are in turn validated through a nonparametric ranking-based statistical test (Friedman 1937), showing the applicability of the proposal to the *Plantago* pollen concentration forecast problem in the region of Madrid.

2 Material and methods

2.1 Data description

Pollen observations correspond to daily *Plantago* concentrations registered at 7 locations distributed around the region of Madrid: Alcalá de Henares, Alcobendas, Aranjuez, Faculty of Pharmacy of Complutense University

of Madrid, Getafe, Leganés and Villalba, as shown in Fig. 1. Pollen counts were conducted following the standard methodology of the Spanish Aerobiological Network (Galán Soldevilla et al. 2007) and were provided by Red Palinológica de la Comunidad de Madrid.

Meteorological observations were obtained from sensors placed in the close surroundings of the pollen stations (each pair of stations in the same municipality separated less than 2 km), consisting of daily average temperature in Celsius degrees, solar radiation in W/m², wind speed measured in m/s, daily rainfall in mm/h, pressure in hPa, and degree of humidity in percentage. The data was provided by the Autonomous Region of Madrid¹ and was measured following the criteria from the Spanish government meteorological agency (AEMET).

Both pollen and meteorological datasets span from year 2005 to 2009 at Alcalá de Henares, Alcobendas, Aranjuez, Getafe, and Leganés. Observations at the Faculty of Pharmacy are available from 2001 to 2013 while in Villalba, only 3 years are available starting 2007 to 2009.

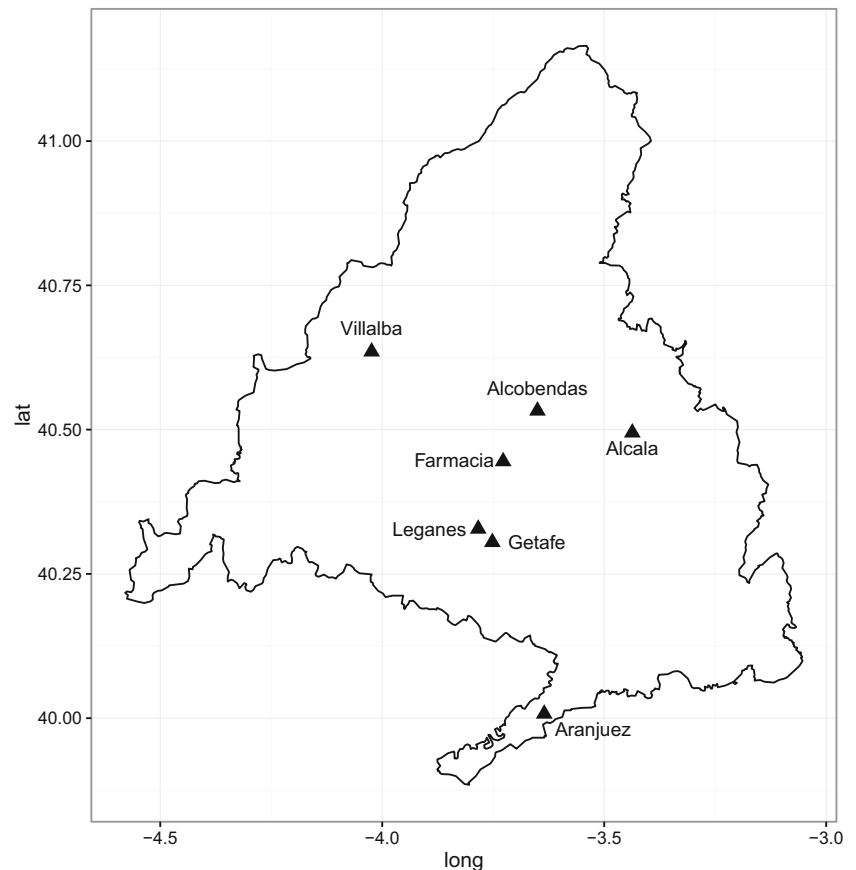
Even though the climate in the region of Madrid is continental (with a strong Mediterranean influence), its geographical characteristics provide several peculiarities at each location. Situated at Guadarrama mountains, Villalba is 903 m above the sea level with a yearly average temperature of 10–11 °C and a yearly average precipitation of 1250–1500 mm. These characteristics are related to mountain climate. Conversely, Aranjuez, which is located 495 m above the sea level, has milder yearly average temperatures (above 14 °C) with a yearly average rainfall below 400 mm. The remaining locations are metropolitan areas between 594 and 668 m above the sea level with yearly average temperatures above 15.2 °C and precipitations around 440 mm.

2.2 Features

Previous studies (Navares and Aznarte 2016b) propose a set of features based on the influence that meteorological conditions have in the development of the plant, together with several variables inspired by phenological studies along with purely analytical approaches, which extract information directly from the time series. In order to perform a deeper analysis of the optimal selection of important features for forecasting airborne concentration, this study increases the number of features from which the models are constructed.

A total of 143 features were generated as shown in Table 1. In addition, an extra variable is included which

¹https://gestion.madrid.org/azul_internet/html/web/AvisosAccion.icm?ESTADO_MENU=1

Fig. 1 Location of weather and pollen stations in the region of Madrid**Table 1** Summary of number of features generated by variable

	i	Σ_{10}	MA_i	Σ_{ytd}	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

Features include i -lagged variables, previous t days cumulative sums represented by Σ_t , moving averages (MA), and previous 15 days standard deviation (std)

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_{ytd} : year to date cumulative sum

Σ_{30} : previous month cumulative sum

Σ_{90} : previous 90-day cumulative sum

Σ_{180} : previous 180-days cumulative sum

std: previous 15 days standard deviation

represents the Julian day of the correspondent year. Thus, the model is trained using a matrix of 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, y_{t+1} represents next day pollen concentration (day-ahead forecast), and $x_{i,t}$ represents the value at time t of the variable i .

2.3 Random forests for regression

Proposed for the first time in Breiman (2001), a random forest (RF) is an ensemble approach which leverages the performance of many weak learners (trees) by combining them to form a strong learner. RF is a supervised learning procedure which generates several randomized regression trees over sample fractions of data, and combines them by averaging. Different random selections are computed by each tree improving stability and accuracy; this technique is known as bootstrap aggregating or *bagging* (Breiman 1996).

Bagging intervenes at two levels, first in data selection and second in variable selection, ensuring that each tree uses a different set of data and a different set of variables. Then, by averaging, the likelihood of *overfitting* diminishes providing a clear advantage over other computational intelligence methods such as neural networks. Also, this procedure gives robustness against the presence of outliers

which favors the selection of RF when compared with support vector regressions.

Several measures have been proposed to estimate variable importance in RF. The most advanced is the permutation accuracy importance measure. Its rationale is based on randomly permuting each predictor variable so the original association with the response is broken. Thus, when a variable is permuted, the accuracy is recorded and averaged over all trees providing the relative variable importance, which increases as the difference in accuracy becomes bigger before and after the permutation. As opposed to classification trees, the accuracy (or prediction error) in the regression problem is measured by the variance.

In our case, model performance is checked using the root mean squared error (RMSE) defined by Eq. 1, along with the coefficient of determination R^2 :

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

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 Nonparametric hypothesis testing

In order to evaluate the variable relevance obtained by the RF algorithm, we use the nonparametric Friedman test (Friedman 1937) over variable importances with a post hoc procedure as described in Navares and Aznarte (2016b). The aim is to investigate, in significance terms, which variables are considered better predictors.

2.5 Experimental design

A first experiment (which will be considered as a benchmark for the subsequent) consists in building, for each location in which we have data, a model designed to predict one day-ahead pollen concentrations using the full set of 144 variables. As a result, we obtain the model's performance along with the identification of the most important variables. Secondly, a verification step is done through the application of statistical inference. The purpose is to generalize and corroborate if the average rankings of variables per location are statistically significant. This is done through the non-parametric tests applied to the importance of variables obtained per location and per year. Lastly, we compare the results obtained from the benchmark with different models created using subsets of the most important variables ranked in the hypothesis tests.

The benchmark model approaches the one day-ahead airborne concentration forecast problem by building a RF model per location. Given the number of years available, as

a cross-validation technique, we use a *leave-one-out* (LOO) setup by years. Hence, the series are split into training and test set at each location, leaving one of the available years in each iteration as a test set and training the model with the remaining years. Through this validation technique, we aim to increase the generalization power of the test set, providing it with data points across all seasonal characteristics. Also, by averaging the outcome of the LOO, the results obtained are more independent of peculiar characteristics of each year.

The second experiment uses a different subdivision of the data to add an extra layer of generalization, which is location independent and it is not subordinated to the shape of the pollen curve of each year. For each combination of year and location, the variable importances resulting from building a RF model are distributed in a $N \times M$ matrix where N represents each pair {location, year} present in the dataset and M the available variables. The random nature of RF justifies the averaging of 50 instances of each model in order to obtain the expected variable importances as they might slightly change from one RF execution to other. A Friedman hypothesis test is then conducted to prove the existence of significant differences within the full set of variables and, if applicable, a post hoc analysis is performed.

Once the test is applied, we compare the models in terms of accuracy and time complexity with the benchmark from the first experiment, which uses the full set of variables. Thus, the results of this full-sized model are compared with those obtained by using a reduced set of the 5 and 15 most important variables out of the statistical inference process.

Considering the trade-off between accuracy and computational complexity, the last experiment is tailored to find the combinations of variables that improve current accuracies subject to minimizing the complexity of the model. As the number of combinations is large, and consequently the number of models, a restricted set of combinations is defined based on clustering the ranked importances obtained from the nonparametric Friedman test.

3 Results

3.1 Predicting with all the variables

In our first setup, RF is applied using the LOO technique across all locations with the full set of variables. The relative importance of the variables is calculated and then averaged, resulting in Fig. 2.

It can be seen that the variables Σ_{10}^p , p_1 , and Σ_{180}^h stay ranked as the top 3 most important variables across the locations except for Villalba. At this location, 90 days cumulative pressure (Σ_{90}^{pr}) and wind speed (Σ_{90}^w) increase up to 6% and 4% of the total importance, respectively.

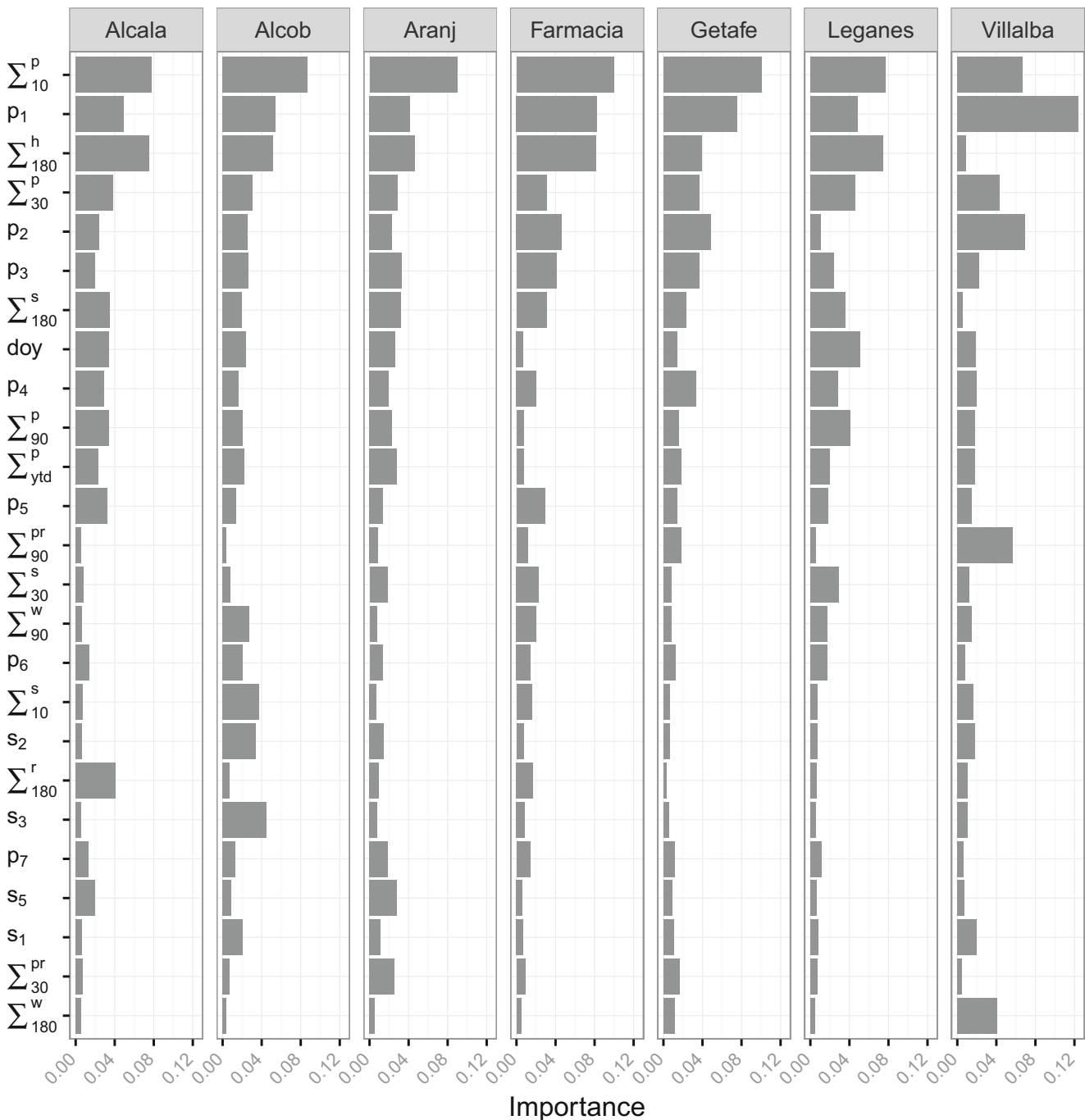


Fig. 2 Relative importance of the most important variables by location using the permutation accuracy importance measure (variables are ordered by the sum of their importance across locations)

Due to the fact that this station is located at 903 m above the sea level, its particular climate conditions, related to mountain climate, produce higher correlation between the wind variable and airborne concentrations. Also, lower atmospheric pressure due to the elevation eases plant formation and consequently pollen releases, which explains higher importance for Σ_{180}^{pr} , compared for example with Aranjuez which is located 495 m above the sea level.

Furthermore, the different urban configuration, for example, with no high buildings, as compared with urban areas such as Getafe, Leganés, Farmacia, Alcobendas, and Alcalá, causes pollen concentrations to differ considerably.

However, excluding the particularity of Villalba, Σ_{10}^p stays as the most important variable across locations, followed by p_1 and Σ_{180}^h which alternate positions depending on the location.

3.2 Nonparametric tests to compare variable importance

As we have seen in Section 3.1, there is some stability in the top-ranked variables. However, there is no clear pattern on the grouping of the variables and their interrelations. In order to investigate this, and to provide statistical evidence of the existence of variables (or groups of variables), which are more important than others, a nonparametric test was performed.

For each location and year available, a RF model was trained using each combination $\{location, year\}$ in isolation and the relative importance of each variable is computed. Out of this setup, the Friedman rank test was performed over 16 variables, which represent more than 1% of the total variance. This reduced set is considered as the pairwise computation with the full set is expensive. A Friedman statistic of $F = 165.84$, which is distributed according to chi-square with 15 degrees of freedom, obtains a p value of $1.04e-10$ with $\alpha = 0.05$, which provides strong evidence of the existence of significant differences between the variables, which are ranked as shown in Table 2.

Figure 3a shows the dendrogram based on the rankings from Friedman. It can be clearly seen that there are two main groups which consist of the top four ranked variables by importance and the remaining ones. As Friedman's null hypothesis was rejected, a post hoc procedure can be carried out in order to check the differences between pairs of variables.

The rejected hypotheses with $\alpha = 0.05$ from the pairwise comparison is shown in Table 3. As stated above, there is no significant difference between the four most important variables Σ_{10}^p , p_1 , Σ_{180}^h , and Σ_{30}^p , suggesting the first cluster of variables shown in the dendrogram from Fig. 3a. This means it is likely these variables alternate ranking positions depending on the location as seen in Fig. 2. It is noticeable that there is no evidence of the difference between the group variables $\{p_1, \Sigma_{180}^h, \Sigma_{30}^p\}$ and $\{p_2, p_3\}$

although there is significant evidence that Σ_{10}^p is different from p_2 and p_3 implying that $\{p_1, \Sigma_{180}^h, \Sigma_{30}^p\}$ has to be different from $\{p_2, p_3\}$, according to the logical relation between the combination of the hypotheses proposed by Shaffer (1986).

Regarding other groups of variables, there is a statistical evidence that p_2 is different from Σ_{30}^s and Σ_{90}^w which constitute the third cluster (Fig. 3a, top), and between p_3 and Σ_{180}^s , which also belongs to the third group. In summary, the tests prove the existence of features which are more relevant than others which are grouped according the three top levels of the dendrogram in Fig. 3a. There is no evidence that members of the same group differ from each other, which means that variables from the same cluster might alternate rank positions within the array of its constituent variables.

3.3 Predicting with a reduced set of variables

After verifying the statistical evidence of the relative importance of the variables, we repeated the experiment in Section 3.1 for two reduced set of variables selected from the findings in Section 3.2. The two sets consist of the 15 (setup B) and the 5 (setup C) most important variables from Friedman test, which are tested against the full set of 144 variables (setup A) from Section 3.1, in order to check how the setup influences model performance in terms of predictability and computational efficiency.

For each reduced set, 50 instances of a RF model were trained to forecast one day-ahead *Plantago* concentrations via the LOO cross-validation by years and observation station. Execution times for a single instance of each setup are 22.66, 17.08, and 11.86 s for setup A (144 variables), setup B (15 variables), and setup C (5 variables), respectively.

By using the setup B of 15 variables, execution time reduces by 25% per iteration while setup C (5 variables) produces a substantial reduction of almost 50%. However, the aim is to find a balanced setup which provides the shortest execution times without an excessive impact in the predictive performance of the models in terms of RMSE and R^2 . Table 4 shows the performance metrics per location and setup; it can be seen that setup A produces in average the least accurate results with an RMSE equal to 4.80 compared with 4.64 for both setups B and C, which clearly indicates the benefits of reducing the number of variables both in terms of accuracy and computational performance. Having a close look at the locations, reducing the number of variables to 5 (C) produces mixed results when compared with setup B being the performance dependent on the location and, in any case, the error increases or decreases in a relatively small amount when both setups are compared.

It is clear that there is a consistent improvement on the errors across locations when a reduced set of variables

Table 2 Rank of the 16 most important features computed by averaging the rank of each variable importance for each year and location

i	Variable	Ranking	i	Variable	Ranking
1	Σ_{10}^p	3.219	9	p_5	9.073
2	p_1	5.073	10	Σ_{90}^p	10.024
3	Σ_{30}^p	5.591	11	p_6	10.097
4	Σ_{180}^h	5.975	12	doy	10.170
5	p_2	6.975	13	Σ_{ytd}^p	10.268
6	p_3	7.878	14	Σ_{90}^w	10.804
7	Σ_{180}^s	8.487	15	Σ_{30}^s	11.487
8	p_4	8.682	16	Σ_{180}^r	11.829

Fig. 3 **a** Dendrogram based on Friedman’s ranks. **b** Combinations of 5 variables tested by clusters and groups used in Section 3.4. ${}^nC_k$ represents the combination of a subset of length k selected from a set of n variables being $k \leq n$

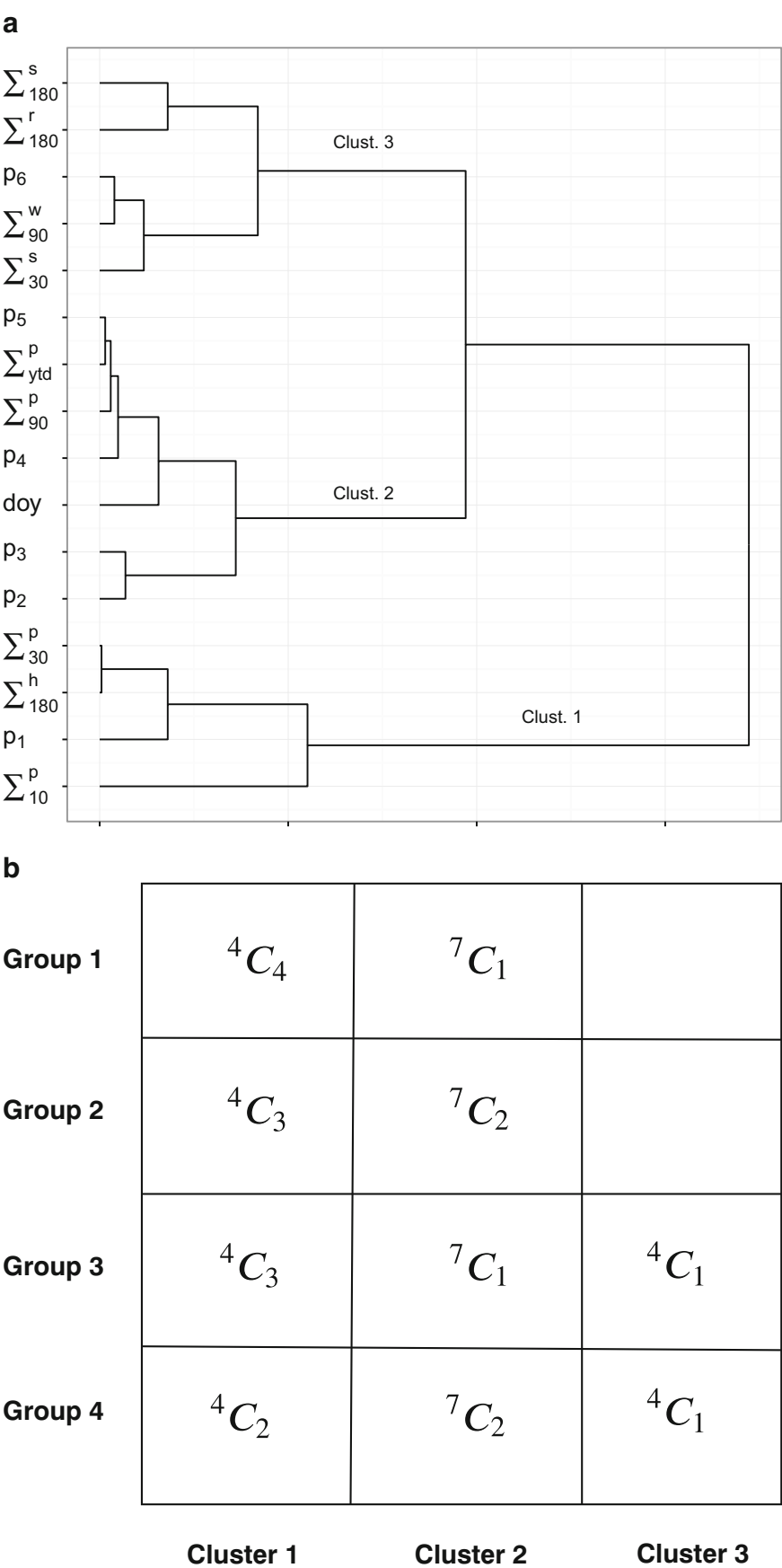


Table 3 Pairwise rejected hypothesis at $\alpha = 0.05$ with unadjusted p value and adjusted Holm and Shaffer p values

i	Hypothesis	p	p_{holm}	p_{shaff}
1	Σ_{10}^p vs Σ_{180}^r	2.66e-16	3.19e-14	3.19e-14
2	Σ_{10}^p vs Σ_{30}^s	3.74e-15	4.46e-13	3.93e-13
3	Σ_{90}^w vs Σ_{10}^p	5.44e-13	6.42e-11	5.72e-11
4	Σ_{10}^p vs Σ_{ytd}^p	2.04e-11	2.38e-09	2.14e-09
5	doy vs Σ_{10}^p	3.83e-11	4.44e-09	4.02e-09
6	Σ_{10}^p vs p_6	6.11e-11	7.02e-09	6.41e-09
7	Σ_{90}^p vs Σ_{10}^p	9.70e-11	1.11e-08	1.02e-08
8	p_1 vs Σ_{180}^r	1.32e-10	1.49e-08	1.38e-08
9	Σ_{30}^s vs p_1	1.06e-09	1.19e-07	1.11e-07
10	Σ_{30}^p vs Σ_{180}^r	2.27e-08	2.52e-06	2.38e-06
11	Σ_{10}^p vs p_5	2.59e-08	2.85e-06	2.72e-06
12	Σ_{180}^h vs Σ_{180}^r	2.59e-08	2.85e-06	2.72e-06
13	Σ_{90}^w vs p_1	5.01e-08	5.41e-06	5.26e-06
14	Σ_{30}^p vs Σ_{30}^s	1.40e-07	1.50e-05	1.47e-05
15	Σ_{30}^s vs Σ_{180}^h	1.59e-07	1.68e-05	1.67e-05
16	Σ_{10}^p vs p_4	2.04e-07	2.14e-05	2.14e-05
17	Σ_{10}^p vs Σ_{180}^s	5.44e-07	5.66e-05	5.00e-05
18	p_1 vs Σ_{ytd}^p	7.79e-07	8.02e-05	7.17e-05
19	doy vs p_1	1.25e-06	1.27e-04	1.15e-04
20	p_6 vs p_1	1.77e-06	1.79e-04	1.63e-04
21	Σ_{90}^p vs p_1	2.49e-06	2.49e-04	2.29e-04
22	p_2 vs Σ_{180}^r	3.91e-06	3.88e-04	3.60e-04
23	Σ_{90}^w vs Σ_{30}^p	3.91e-06	3.88e-04	3.60e-04
24	Σ_{90}^w vs Σ_{180}^h	4.38e-06	4.24e-04	4.03e-04
25	Σ_{10}^p vs p_{-3}	9.41e-06	9.03e-04	8.66e-04
26	p_2 vs Σ_{30}^s	1.78e-05	0.00	0.00
27	Σ_{30}^p vs Σ_{ytd}^p	4.03e-05	0.00	0.00
28	Σ_{180}^h vs Σ_{ytd}^p	4.46e-05	0.00	0.00
29	doy vs Σ_{30}^p	6.00e-05	0.01	0.01
30	doy vs Σ_{180}^h	6.62e-05	0.01	0.01
31	Σ_{30}^p vs p_6	8.04e-05	0.01	0.01
32	p_6 vs Σ_{180}^h	8.85e-05	0.01	0.01
33	Σ_{90}^p vs Σ_{30}^p	1.07e-04	0.01	0.01
34	Σ_{90}^p vs Σ_{180}^h	1.18e-04	0.01	0.01
35	p_1 vs p_5	1.42e-04	0.01	0.01
36	Σ_{180}^r vs p_{-3}	1.72e-04	0.01	0.01
37	p_2 vs Σ_{90}^w	2.71e-04	0.02	0.02
38	p_2 vs Σ_{10}^p	3.54e-04	0.03	0.03
39	p_4 vs p_1	5.97e-04	0.05	0.05

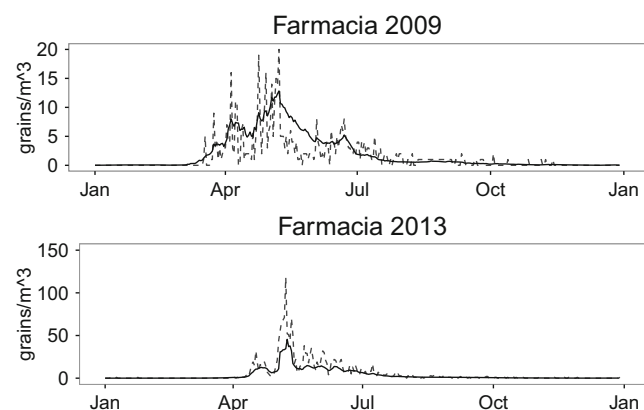
(setup B and C) are used, compared with the full set. This provides evidence about the convenience of ignoring redundant or noisy variables, which eases the predictive capability of the model. However, this gain is not shown in the R^2 measure, which obtains as an average of 60% in setup A compared with 60% in setup B and 55% in setup C. Given the similarity of results between choosing 5 or 15 variables,

Table 4 Average RMSE in grains/m³ and R^2 of the forecasted test years studied at each location

Station	RMSE _A	RMSE _B	RMSE _C	R_A^2	R_B^2	R_C^2
Alcalá	5.75	5.69	5.80	0.58	0.54	0.51
Alcobendas	6.64	6.39	6.10	0.66	0.62	0.58
Aranjuez	3.83	3.76	3.81	0.54	0.51	0.49
Farmacia	3.71	3.63	3.63	0.67	0.65	0.59
Getafe	5.58	5.50	5.68	0.65	0.62	0.59
Leganés	5.48	5.15	5.11	0.59	0.64	0.52
Villalba	2.79	2.36	2.20	0.52	0.59	0.55
Average	4.82	4.64	4.64	0.60	0.60	0.55

there is still an open question on how to improve setup C, which is objectively more convenient in terms of execution, on those locations where it is less accurate than setup B.

On the other hand, through a close examination of the results, we have identified some difficulties of the model in capturing sudden peak pollen concentrations over 100 grains/m³ as can be seen in Fig. 4. Concretely, this sudden extreme peaks increase RMSE to 10.06 grains/m³ at the Faculty of Pharmacy (Farmacia) in 2013, while year 2009 produces a RMSE = 2.27 grains/m³. The reduced amount of observed data with these characteristics explains the difficulties of the model to identify the inner behavior of the data when forecasting. However, this situation can be mitigated by limiting airborne concentration levels, imposing a threshold which substitutes the data points above it (Navares and Aznarte 2016a). This limit should be related to the concentration levels which are considered to be risky for human health; for instance, symptoms 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 observed at 25 grains/m³ (Rodríguez-Rajo et al. 1983).

**Fig. 4** Forecast (solid line) and observed (dashed) pollen concentration values for two years in the Faculty of Pharmacy (Farmacia)

3.4 Optimizing predictions with reduced sets of variables

In Section 3.3 we left an open question about how the setup C, consisting of 5 variables, can be improved, knowing it is the optimal configuration in terms of execution time and that it does not significantly differ in accuracy from setup B (15 variables). Concretely, the following question is relevant: is there a subset of 5 variables out of the 15 most important variables which outperform setup B accuracies?

Although random forests are known to be in general robust against overfitting, in the presence of correlated features, strong features can obtain low scores and the method can be indeed biased towards certain variables. For this reason, the experiment in Section 3.1 was repeated with a reduced set of 5 variables selected from the results obtained in the nonparametric test from Section 3.2. However, Fig. 3a showed that the 15 most important variables showed a cluster structure based on the results of Friedman test, which leads to the aforementioned question. The idea here is to use the cluster structure to find combinations of the 15 variables of setup B taken in groups of 5 which minimize the model's RMSE.

Of course, there are 3003 possible combinations of 15 variables taken in groups of 5, and this represents a high burden of executions of the experiment if one wants to explore them all. Thus, we propose to limit the number of combinations in such a way that the cluster structure, and the possible interactions between variables from different clusters, is taken into account. Taking as a reference setup C which includes the 5 most important variables, the idea is to step by step increase the presence of variables with lower importance by replacement in the original setup C. Hence, a first group includes the 4 most important variables from setup C while the fifth variable is replaced by one from cluster #2. This is done a number of times equal to the number of variables in cluster #2, resulting in the execution of 7 models as the second cluster contains 7 variables. The second group reduces the representation of the first cluster to 3 variables and increments the number of variables selected from cluster #2 to 2. The number of models executed is equal to the number of combinations

resulting from taking 3 variables from cluster #1 and 2 variables from cluster #2 (always keeping 5 as the total number of variables). This logic is followed for subsequent groups selecting the correspondent number of combined variables per cluster as summarized in Fig. 3b.

As shown above, based on the pairwise comparison test, 3 clusters were obtained as in Fig. 3a. Cluster #1 contains the 4 most important variables, cluster #2 the next 7 variables in importance, and cluster #3 the last 4. Notice we excluded the least important variable from cluster #3 (Σ_{180}^s) for direct comparison with setup B of 15 variables. Given the restrictions mentioned, the groups of combinations explored are those shown in Fig. 3b.

In that figure, ${}^nC_k = \frac{n!}{k!(n-k)!}$, being n the total number of variables belonging to a cluster and k the number of selected variables of the same cluster. The best results obtained are shown in Table 5 along with the reference benchmark, that is, represented by the best performing setup from Section 3.3 in terms of accuracy (RMSE).

The results obtained support the initial statement of this section about producing biased models when variables are selected regarding only importance, as this rules out the consideration that they might be related to each other. Increasing the richness of the variable selection space, in terms of including candidates from less important clusters according to Friedman test, and consequently with statistical evidence of differences, produces an increase of accuracy across all locations. Table 5 shows that including at least 2 variables from cluster #2 or cluster #3 or both increases accuracy by an average of 4%, with the largest increase of 7% in Getafe. Most of the improved results were obtained from groups 3 and 4, which include combinations of variables extracted from all the clusters considered, thus diversifying the distribution of importances and avoiding redundancies among the top-ranked variables.

4 Discussion

Automatic feature selection has been gaining relevance in computer sciences and statistics since the end of last century, when specific issues were published in specialized

Table 5 Comparison of the benchmark RMSE in grains/m³ obtained from the best setup in Section 3.3 with the best performing combination along with its constituent variables and the test group they belong

Station	Benchmark	RMSE _{best}	Group #	Cluster #1	Cluster #2	Cluster #3
Alcalá	5.69 (B)	5.67	4	{p ₁ , Σ_{180}^h }	{p ₂ , p ₅ }	Σ_{180}^r
Alcobendas	6.10 (C)	5.73	4	{ Σ_{10}^p , p ₁ }	{p ₃ , p ₄ }	p ₆
Aranjuez	3.76 (B)	3.65	4	{ Σ_{10}^p , p ₁ }	{p ₂ , p ₄ }	Σ_{30}^s
Farmacia	3.63 (C)	3.49	3	{ Σ_{10}^p , p ₁ , Σ_{180}^h }	{p ₅ }	Σ_{180}^r
Getafe	5.50 (B)	5.11	4	{ Σ_{10}^p , p ₁ }	{p ₄ , p ₅ }	Σ_{180}^r
Leganés	5.11 (C)	4.95	3	{ Σ_{10}^p , p ₁ , Σ_{180}^h }	{p ₄ }	Σ_{30}^s
Villalba	2.20 (C)	2.20	4	{ Σ_{10}^p , p ₁ }	{p ₂ , p ₅ }	p ₆

journals, including many related papers (Blum and Langley 1997; Kohavi and John 1997). Numerous approaches and extended literature exist on the topic (Bolón-Canedo et al. 2013), and there is no widespread agreement on a so-called "best method" as the solution is always domain and problem specific. In this paper, we used the Friedman nonparametric hypothesis test to support the conclusions about our proposal selected as well as to provide statistical evidence of the results.

As seen in Section 3, using the proposed approach makes it possible to identify the most important variables for one day-ahead airborne pollen concentrations forecasting. Variable ranks obtained from the model (shown in Fig. 2) are consistent with the rank-based nonparametric hypothesis test, which provides statistical evidence of the findings without any *a priori* assumption about the influence of each climate or phenological feature.

The different geographical locations considered in the study, which do not share the same climatic conditions, prove the generalization abilities of the proposed approach, which obtain stable variable ranks across the observation stations. The dendrogram in Fig. 3a shows the expected rank interchangeability among variables which belong to the same cluster, being these clusters corroborated by a pairwise comparison post hoc procedure, which provides statistical evidence of the difference between pairs of variables.

Our approach identifies the previous day pollen observation and 10- and 30-day cumulative daily concentrations as the most important features in one day-ahead pollen forecast. In phenological studies, these features have been previously proved to serve as an indicator of the state of the plant (Ribeiro et al. 2007; Smith and Emberlin 2006). From the point of view of time series analysis, previous day pollen observation is generally found to be a relevant variable (Astray et al. 2016; Iglesias-Otero et al. 2015; Levetin 2014). Notwithstanding, there is weak evidence that 10-day cumulative sum has a tendency to be the most influential as it is the only variable in the top-ranked cluster which is clearly set aside from the variables of next group (composed by two- and three-day past pollen observations).

Along with these three variables, the model identifies previous 180-day relative humidity accumulation as one of the most influential features which is also in accordance with the findings in Jones and Harrison (2004), Levetin (2014), and Rodríguez-Rajo et al. (2004), as it promotes plant growth during the development state and mitigates pollen spread during release phase of the plant. Furthermore, relative humidity has more effect on airborne pollen concentrations compared with 180-day rainfall accumulation which is shown in cluster # 3 (Fig. 3a). This result is inline with previous studies which show non significant correlation between pollen concentrations and rainfall (Bartková-Scevková 2003) while humidity shows a

significant effect. Pollen grains release allergen due to hydration which it becomes more intense before rainfalls (Grote et al. 2001).

Even though Barnes et al. (2001) show a higher influence of heavy rains on pollen concentrations when intradaily measured only daily pollen counts were available for this study, consequently, it was not possible to determine intraday strong weather conditions changes which have effects on pollen concentrations (Barnes et al. 2001). For instance, sporadic short heavy rains followed by long dry periods within the day does not considerably change daily pollen counts (Bartková-Scevková 2003).

The sensitivity analysis presented by Puc (2012) indicates that the most influential variables to predict airborne pollen concentrations are maximum daily temperature and humidity, being minimum daily temperature ranked the least influential. Since only daily average temperatures were used in this proposal, there is a mitigation effect in terms of maximum daily temperature influence. However, cumulative solar radiation appears among the top most important features computed (Table 2). A positive radiative forcing involves climate warming and, as a result, an increase of temperatures (Lean and Rind 1998).

In Section 3.3, we applied three configurations based on three sets of variables selected by importance from the rankings. The results from setups B and C outperform the initial model with 144 variables. This implies that removing redundant or irrelevant features tend to improve model performance both in accuracy and execution time. Given the parity of results of these setups and to reduce model complexity, in Section 3.4, we tested different combinations of 5 variables selected among the 15 most important (setup B). This mitigates redundancies among variables which do not significantly differ according to Friedman test and, as a result, improves both model accuracy and performance (measured in execution time).

The accuracy of the results are comparable with other studies such as Iglesias-Otero et al. (2015), who reports a R^2 of 0.66 compared to 0.60 in setup B and an average of 0.55 from the experiments in Section 3.4. We believe that optimizing the number of variables to select, instead of the arbitrary 5 or 15, would improve this metric. However, Iglesias-Otero et al. (2015) used artificial neural networks, which require substantial data preprocessing as the inputs shall be limited when the number of training instances is not large enough, increasing the number of assumptions to be taken *a priori*. Not only does RF mitigate these drawbacks, but also is considered more robust against overfitting.

When compared to similar algorithms, our proposal outperforms the most accurate regression model among those proposed by Csépe et al. (2014) which achieves a RMSE of 28.26 using an implementation of regression trees. With a RMSE of 4.17, RF reduces the estimation variance

(and consequently the prediction error) due to the *bagging* technique. Also, averaging the results through the LOO approach provides more sound and general results compared with predefined test sets, whose results rely on the specific characteristics of the years selected.

The proposal presented in this paper has immediate practical application as it is able to identify the top-ranked variables producing satisfactory results. However, there are some drawbacks which deserve further investigation: (1) model error increases with the appearance of sudden high concentrations levels, concretely those over 100 grains/m³, given the reduced number of observations of this type. However, this situation is persistent in most approaches found in the literature. This circumstance can be addressed by limiting the observations without altering the practicality of the model by setting meaningful thresholds (Navares and Aznarte 2016a, 2017). For instance, Rodríguez-Rajo et al. (1983) conclude that first allergy symptoms in Spain are reported over the risk threshold of 25 grains/m³, which could be used as threshold. (2) There is an open question about the number of variables which guarantee optimal results.

In summary, we have presented statistical evidence that the approach proposed in this paper identifies the most important variables without any a priori assumptions over their influence. Removing redundant and irrelevant features improve model accuracy, being setup C with 5 variables the best performer among the subset tested and producing satisfactory results. Including a diverse subset of variables avoid biased results due to the mitigation of redundancies, thus improving the accuracy of the model.

5 Conclusions

The present study introduces a feature selection approach for forecasting airborne pollen concentrations. Statistical evidence of the consistency in identifying the most important features is provided through nonparametric hypothesis testing, using as a case study a one day-ahead forecast of *Plantago* airborne concentrations in the region of Madrid.

The results indicate that the proposed approach is a valid and efficient way to rank independent variables in the task of airborne pollen time series prediction. Precisely, we have determined that the cumulative sums of recent pollen concentrations along with previous days pollen observations are among the most important features. This represents a data-driven confirmation of main findings from phenological studies. Also, the influence of cumulative daily relative humidity was shown as an important factor during plant formation and pollen release state, which

coincides with the conclusions of previous researches based on the influence of meteorological variables.

Furthermore, from this ranking of variables, it has been shown how it is possible to build new, more parsimonious models which produce better results than benchmark approaches: by selecting the best set of independent variables, we achieved a 3.5% average improvement across locations with an average 50% reduction in execution times. This effect, due to the elimination of redundancies and irrelevant features, is to be expected in any regression technique using the same set of variables, and, as shown, is the main strength of our proposal when compared with other techniques.

Acknowledgments 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.

References

- Andersen TB (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 JL, Benítez Sánchez JM, Lugalde 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
- Barnes C, Pacheco F, Landuyt J, Hu F, Portnoy J (2001) The effect of temperature, relative humidity and rainfall on airborne ragweed pollen concentrations. *Aerobiologia* 17(1):61–68
- Bartková-Scevková J (2003) The influence of temperature, relative humidity and rainfall on the occurrence of pollen allergens (betula, poaceae, ambrosia artemisiifolia) in the atmosphere of Bratislava (Slovakia). *Int J Biometeorol* 48(1):1–5
- Blum A, Langley P (1997) Selection of relevant features and examples in machine learning. *Artif Intell* 97:245–271
- Bolón-Canedo V, no NSM, Alonso-Betanzos A (2013) A review of feature selection methods on synthetic data. *Knowl Inform Syst* 34:483–519
- Breiman L (1996) Bagging predictions. *Mach Learn* 25:123–140
- Breiman L (2001) Random forest. *Machn 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 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
- 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 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, https://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
- Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Statist Assoc* 32:674–701
- Galán Soldevilla C, Cariñanos González P, Alcázar Teno P, Domínguez Vélchez E (2007) *Manual de Calidad y Gestión de la Red Española de Aerobiología*. Universidad de Córdoba
- García-Mozo H, Chuine I, Aira M, Belmonte J, Bermejo D, de la Guardia CD, Elvira B, Gutiérrez M, Rodríguez-Rajo J, Ruiz L, Trigo M, Tormo R, Valencia R, Galán C (2008) Regional phenological models for forecasting the start and peak of the quercus pollen season in Spain. *Agr Forest Meteorol* 148:372–380
- Grote M, Vrtala S, Niederberger V, Wiermann R, Valenta R, Reichelt R (2001) Release of allergen-bearing cytoplasm from hydrated pollen: a mechanism common to a variety of grass (poaceae) species revealed by electron microscopy. *J Allergy Clin Immunol* 108(1):109–115
- Iglesias-Otero MA, Fernández-González M, Rodríguez-Caride D, Astray G, Mejuto JC, Rodríguez-Rajo FJ (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
- Kohavi R, John G (1997) Wrappers for feature subset selection. *Artif Intell* 97:273–324
- Leanh J, Rind D (1998) Climate forcing by changing solar radiation. *J Climate* 11(12):3069–3094
- Levetin E (2014) Daily ragweed pollen forecasting. *J Allergy Clin Immunol* 133:AB17
- Li X, Maier H, AC Z (2015) Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models. *Environ Model Softw* 65:15–29
- Matyasovszky I, Makra L, Csépe Z, Sümeghy 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 (2016a) Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. *Int J Biometeorol*. <https://doi.org/10.1007/s00484-016-1242-8>
- Navares R, Aznarte J (2016b) What are the most important variables for poaceae airborne pollen forecasting? *Sci Total Environ* 579:1161–1169
- Navares R, Aznarte J (2017) Forecasting the start and end of pollen season in madrid. In: *Advances in time series analysis and forecasting*. Springer International Publishing, pp 387–399. chap 26
- 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
- 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
- Puc M (2012) Artificial neural network model of the relationship between betula pollen and meteorological factors in Szczecin (Poland). *Int J Biometeorol* 56(2):395–401
- 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
- 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
- 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
- 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
- Tran H, Muttill N, Perera B (2015) Selection of significant input variables for time series forecasting. *Environ Model Softw* 64:156–163
- Tseng Y, Kawashima S, Kobayashi S, Takeuchi S (2018) Algorithm for forecasting the total amount of airborne birch pollen from meteorological conditions of previous years. *Agr Forest Meteorol* 249:35–43

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.