UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

MASTER THESIS

Forecasting Airborne Pollen Concentrations through Random Forests

Author:
Ricardo NAVARES

Supervisor:
José Luis AZNARTE, PhD

A thesis submitted in fulfilment of the requirements
for the degree of MSc. Advanced Methods in Artificial Intelligence

in the

Dept. Artificial Intelligence

September 18, 2016

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Abstract

Faculty of Computer Science Engineering
Dept. Artificial Intelligence

MSc. Advanced Methods in Artificial Intelligence

Forecasting Airborne Pollen Concentrations through Random Forests

by Ricardo NAVARES

Poaceae is the largest family of monocotyledonous flowering plants, known as grasses and considered to be one of the most important aeroallergens in Europe. The increase of allergy cases and the severity of the reactions motivates the prediction of atmospheric concentrations in order to minimize the exposure to risky pollen levels. Also, it is of large interest for clinical institutions in order to apply preventive measures and plan in advance the implications of an increase number of allergy patients.

Phenological and meteorological parameters characterize the stages of vegetation development during the growing season. Thus, they can be potentially related to the biological definition of plant phenology. In this thesis, time series of airborne pollen concentrations and meteorological variables measured at the region of Madrid were used to predict risk pollen concentrations for patients. Detailed relationships were established between future airborne concentrations, meteorological data, and flowering states derived from the inner information of the underlying data via computational intelligence models. Therefore, these data were used to develop predictive models for a range of forecast horizons. The proposal will be beneficial to the medical fields related with allergies affections planning and treatment, demonstrating that computational intelligence holds a great potential for aerobiology.

In this research, we demonstrated several novel approaches that significantly contribute to the field of aerobiology including: (i) developing a computational intelligence-based model to predict risk concentration levels, and consequently the start and end of pollen season, for long term horizons up to 6 months, on which none of previous works succeed to obtain satisfactory results, (ii) identifying and characterizing the most influential factors which induce the presence of high airborne concentrations for a given set of horizons, using an assumption-free approach which supports biometeorological findings from other authors. The findings of the research are related to producing more accurate prediction models and providing a comprehensive analysis of the relationship of airborne concentrations with various meteorological parameters.

As a product of this study, a version of Chapter 3 of this thesis was accepted and presented at the *International Work Conference on Time Series Analysis (2016)*[1], and it has been selected to be extended and submitted as a book chapter in the Springer series *Contributions to Statistics*[2]. A version of Chapter 5 has been published as a full paper in the *International Journal of Biometeorology*[3] and Chapter 4 is in preparation to be sent to an international journal for publication.

[1] http://itise.ugr.es/

[2] http://www.springer.com/series/2912

[3] http://link.springer.com/journal/484

Acknowledgements

Me gustaría que estas líneas sirvieran para expresar mi más sincero agradecimiento al Dr. José Luis Aznarte por aceptarme para realizar este trabajo bajo su dirección, por su apoyo y confianza. Su aporte ha sido invaluable no solamente en el desarrollo del trabajo, sino también en mi formación. Las ideas propias, siempre enmarcadas en su orientación y rigurosidad, han sido claves para el desarrollo de las actividades. Le agradezco también el haberme facilitado los medios suficientes para llevar a cabo todas las actividades propuestas.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This introductory chapter gives a general description of the problems faced in forecasting pollen time series. It offers a presentation of the biological and meteorological factors involved as well as the motivation and objectives of this research. Finally there is a brief description of the structure of this document.

1.1 Presentation

Continuously increasing allergy symptoms in developed countries, and the clinic and socioeconomic relevance of this problem, has 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 [40]. During the pollination season, those affected by allergies experience symptoms such as nasal congestion, ear inflammation (*otitis* media with effusion), sinus infection (sinusitis), coughing, sneezing, itchy, watery eyes and more. These symptoms can seriously impair labor and recreational activities, and result in sleeping disorders, causing fatigue, learning impairment and irritation. In addition to the reduced quality of life, expenses related to pollen allergies are extensive.

The first advice for allergy patients is usually allergen avoidance which is impossible most of the time. Hence, medication is needed being antihistamine the most commonly used. The medication should be started before the pollen season to give maximum effect during the season. Likewise, the effect of the strongest medication, corticosteroids injections, only last for 2-3 weeks, making it necessary to receive the injection at the appropriate time. Pollen forecasting is essential for both medication and allergen avoidance. In order to enable preventive measures and reduce the exposure for patients, field experts focus on the prediction of pollen concentration levels which imply high risk for allergic population.

These predictions could help research centers and clinical institutions to plan in advance the implications of high pollen counts and their duration, as well as allergy patients to be able to limit their exposure to risky pollen levels.

The main pollination season is defined as the period where high concentrations of pollen counts are measured. In the literature, several definitions of what it is considered a pollen season have been established [23]. It is possible to classify them into two main approaches, those based on the cumulative daily pollen counts [2, 18, 26] and those based on a predefined threshold level over which the season is defined to start and end [37]. It is therefore important to determine a single criterion to delimit the main period during which, pollen is present in the air. It is needed to analyze the terms used in aerobiological literature, with a view to selecting the most appropriate of these for defining the period in question, and to examine the extent to which aerobiological results and pollen curves are influenced by the criteria used to fix the start and end of the pollen season.

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 [9, 31]. On the other hand, a dry and windy weather spreads the airborne quickly, leading to higher distributions [25, 36]. The most widely used approach to forecast the start of the pollination period is temperature sum models based on the idea that the development of plants depends on temperature sums [31]. During the most recent decades, the approach of temperature sums to forecast the state of the vegetation has been used in a number of aerobiological studies. There are still large gaps in the knowledge regarding the effects of weather parameters on physiological processes of the plants because most of them are not visible, although several models have been proposed, trying to simulate the physiological processes [9]. As opposed to biometeorological forecasts, many authors address the problem by using univariate regression models based on past information of atmospheric concentrations.

The difficulty of modelling using time series lies in the nature of the pollen series, since it consists of zero pollen counts throughout almost the entire year, interrupted by one or several random intervals, of short duration, of high values with very fast fluctuations. Depending on the attributes of the time series, some techniques are more suitable than others leading to a first decision about which model should be used.

1.2 Motivation

Historically, the problem of forecasting pollen time series has been addressed by means of time series analysis. These techniques are highly dependent on the selection and parametrization of the models. In recent years, several authors have been proposing the use of computational intelligence techniques, mainly neural networks, as a framework to produce more accurate predictions.

Both approaches rely heavily in the selection of the variables to input the model. This selection is based on findings from either meteorlogical or phenological studies which establish the relations between pollen release and different weather variables of plant states. On the one hand, time series analysis proposals are very sensitive to collinearity of the variables, conversely, neural networks are computationally expensive and their outcomes hard to explain.

This subjectivity in selecting the features limits the scalability of these approaches, making them location and pollen specie specific. Besides, as the variables were already preselected based on assumptions about their influence, very few studies provide information about the relevance of a wide range of variables during the flowering states of the plant.

Up to date, the predictions proposed estimate short-term pollen concentrations, which do not fully fill the needs of, for example, clinical institutions in planning the resources to attend an increase of allergy cases. Therefore, there is a necesity for longer term accurate forecasts to deal with this type of situations. Additionally, the results provided by the models are not interpretable from the point of view of a user with no background in the field, for instance, an allergy patient ignores the implications and the significance of having 10 grains/m^3 in the environment. This draws a line between the scientific community and the users.

Achieving methods able to deal with all the information considered from different fields of biology and meteorology, and at the same time provide easily interpretable and accurate results, even for mid and long-term horizons, would be as of great interest not only for the scientific community, but also for all type of interested users. The methods could derive important benefits in a wide range of fields as well as in human health. These are the core motivations for this research.

1.3 Objectives

The main objective of this research are:

- Select an assumption-free model which can cope with all the different requirements considered, attending to accuracy, scalability and interpretability.
- Test model ability to rank the most influential variables to widen the information provided by previous proposals and support meteorological and phenological studies.
- Provide a framework for mid and long-term accurate forecasts which will set an edge in the field.

As an additional objective, we intend to broaden the audience by providing an easy interpretation of the results.

1.4 Outline of the thesis

A more detailed description of the work presented in this thesis will be given in the following.

Chapter 1 is an introductory chapter with preliminary information and the motivation of this study. The main contributions of this work were detailed compared to previous related researches as well as the present detailed outline of the thesis

Chapter 2 gives a detailed description of the materials used in the following chapters. It drives the reader through a formalization of the main pollination season, which is the period where high atmospheric concentrations might appear, visiting different proposals among the authors. By revising the literature based on both climatological and phenological studies, it is intended to ease the system in predicting pollen concentrations by generating relevant features which might be influential.

The first paper included as Chapter 3 was presented at the *International Work Conference on Time Series Analysis*[1] in Granada, Spain the 27^{th} of June, 2016. The study has been one of the selected papers among all proposals as a book chapter in the Springer series *Contributions to Statistics*[2]. The paper addresses the problem of predicting he start and the end dates of the pollen season of grasses (the family Poaceae) in the city of Madrid. A classification-based approach was taken by discretizing atmospheric concentrations according to a range of thresholds. Several computational intelligence approaches are tested including Random Forest, Logistic Regression and Support Vector Machines accross a set of forecast horizons ranging from short to mid term. The proposal allows to select the most accurate model for prediction in order to limit risk exposure for patients, and allow preventive measures for clinical institutions.

[1] http://itise.ugr.es/

[2] http://www.springer.com/series/2912

Based on the conclusions obtained in the previous chapter, Chapter 4 consists on the second paper of this series which approaches the problem of identifying the most influential features in predicting Poaceae pollen concentrations in seven locations accross the Autonomous Community of Madrid, Spain. Following the conclusions from the previous chapter, Random Forests were used to provide a framework to detect the significance of weather and physiological factors which influence the formation of airborne concentrations. As opposed to previous works, no assumptions were taken in determining the influential variables such as spring rainfall, but letting the proposal to capture the inner information from the data available. A nonparametric Friedman test and *post hoc* procedures were applied in order to give statistical evidence of the results obtained.

The last paper of the series which constitutes Chapter 5 was accepted for publication in the *International Journal of Biometeorology*[3]. This paper wraps all conclusions from previous chapters to address the problem of forecasting the dates in which risk concentration levels are observed based of Random Forests. Unlike previous works, the proposal extends the range of forecasting horizons up to 6 months ahead. Furthermore, it allows to identify the most influential factors for each forecast horizon providing support to findings from other researchers in the fields of biology.

Finally in the last part, this work will be discussed in its entirety along with its relations to recent studies and the future works which will be implemented.

[3] http://link.springer.com/article/10.1007%2Fs00484-016-1242-8

Chapter 2

Materials and methods

2.1 Data description

Study location and pollen stations. Daily Poaceae concentrations were provided for several locations by Red Palinológica de la Comunidad de Madrid. The observations consist of 8 locations from 2000 to 2013. The data is registered in Alcalá de Henares, Alcobendas, Aranjuez, Coslada, Faculty of Pharmacy of Complutense University of Madrid, Getafe, Leganés and Villalba.

Weather data. Weather data was provided by the Autonomous Community of Madrid[1]. Weather observations consist of average daily temperature in Celsius degrees, solar radiation in W/m^2, wind speed measured in m/s, daily rainfall in mm/h, pressure in mbar and degree of humidity in percentage and Ultra Violet radiation in mW/m^2. Data sets for locations Alcalá de Henares, Alcobendas, Aranjuez, Getafe and Leganés consist of 5 years observations from 2005 to 2009. At the Faculty of Pharmacy, mapped to weather stations located in Casa de Campo, Plaza de España and Cuatro Caminos, provided this time by Ayuntamiento de Madrid, data is available from 2000 to 2013 while in Villalba only three years are available starting 2007 to 2009.

2.1.1 Pollen Data

Some cleaning is required in the pollen time series as missing values appear. A study about the nature of the missing data points was led by the verification of their distribution. It is intended to avoid a naive approach which will result in an lost of important features. 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. Figure 2.1 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 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 29^{th} 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

[1] `http://gestiona.madrid.org/azul_internet/html/web/InformExportacionAccion.icm?ESTADO_MENU=7_4`

hence the resulting matrices are weighted to give more relevance to most recent data (within each year), as in:

$$p_t = \beta \cdot \mathrm{r}_{\mathrm{row}} + (1 - \beta) \cdot \mathrm{r}_{\mathrm{col}}, \tag{2.1}$$

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

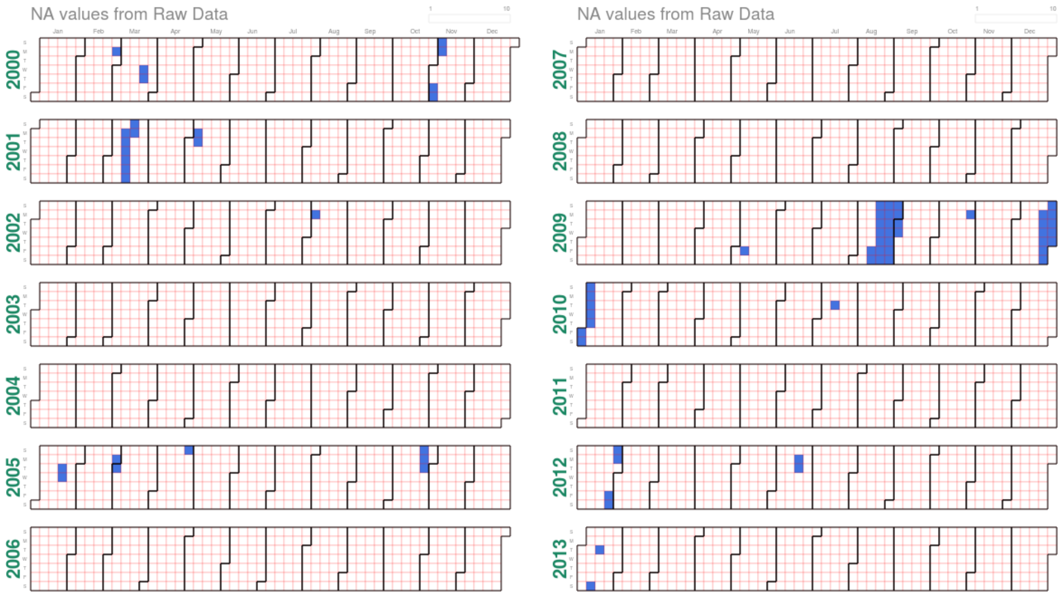


FIGURE 2.1: Discretized NA values distribution at Faculty of Pharmacy. Filled points represent missing data.

In order to have an idea of the distribution, pollen values are mapped into a $\{0, 1\}$ space where 1 represents data points with *NA* value. On Figure 2.1 can be appreciated that not only do we find missing data around estimated season start but long sequences as well. It can be seen for instance in year 2009 that it appears a three weeks period of missing data at the end of August, carrying the last data point forward can incur in misrepresentation. Also in 2001 we appreciate a one week missing data in March which lays close enough to the spring equinox [2], it is the start of the spring season, a well-known critical period for grass bud burst and pollen concentrations may reach critical values for allergic effects. Thus the approach to follow requires some extra processing to avoid delaying the season start. Conversely, this situation could lead to a season end extension.

2.1.2 Weather Data

A sample of weighted weather normalized observations for year 2008 is shown in Figure 2.2. A preliminary visual analysis tells that temperatures and sun hours are positively related to pollen concentrations in Figures 2.2 (A) and (D) respectively . The higher the temperatures and the number of sun hours, the bigger the number of grains per cubic meter . On the other hand, Figure 2.2 (B) shows a negative relation between

[2] An equinox is an astronomical event in which the plane of Earth's equator passes through the center of the Sun, which occurs twice each year, around 20 March and 23 September.

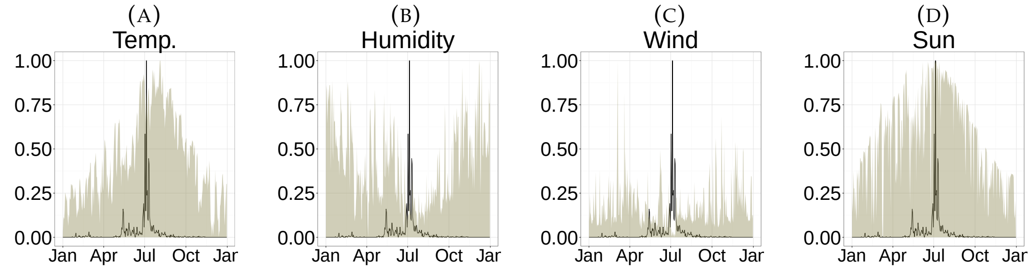


FIGURE 2.2: Normalized weather values with pollen concentrations for year 2008 at Faculty of Pharmacy.

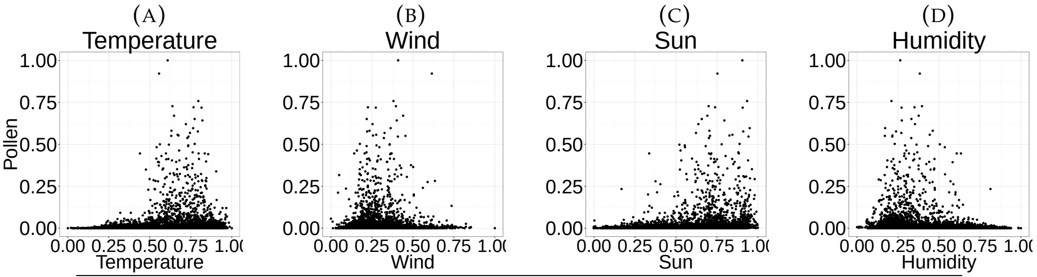


FIGURE 2.3: Normalized scatter plots with pollen concentrations at Faculty of Pharmacy.

humidity and airborne concentrations which also can be seen in Figure 2.2 (D) with wind speed.

We have just seen that there can be a visual intuition on the relation between the grain concentration and the weather data. Having said that, it makes sense a preliminary analysis on the correlation between the variables. Figure 2.3 shows the scatter plot between normalized weather variables and pollen data which contributes with not much information but the absence of collinearity. We find no significant evidence to conclude any direct relation between pollen counts and climate data.

It is needed further analysis to capture the nonlinearity. At this point we have no clear conclusion about the relation between the concentration of grains and the influence of climate data.

2.2 Season definition

In literature, the main pollination season is defined according to two different approaches [23]. The first one is based on daily cumulative counts and the second considers the season started when a pollen concentration threshold is consistently surpassed.

Among the authors who define the pollination main season according to the days in which the daily pollen counts exceed a predefined threshold, some definitions are based on exceeding certain predefined quantity, e.g. 30 grains/m^3 [37]. Others consider when the daily concentration consistently surpass certain level during a period, e.g. 3 grains/m^3 during at least 4 days in the following week [16]. On the other hand, the authors that follow the cumulative approach define the season based on percentages of the yearly total sum. Depending on the authors, the definition of the main pollination period is defined as the period between the day in which the sum of daily pollen concentrations reached 5% of the total sum and the day in which the sum reaches 95% [26], or when it reaches 2.5% and 97.5% [2], or even 1% and 99% [18].

TABLE 2.1: Sample Start and End of pollination season accordig to different definitions

Approach	Definition	Year	Start	End
Nilsson *et al.* [26]	*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 2004 2009 2012	09 Feb 26 Feb 04 Apr 17 May	03 Nov 11 Jul 09 Sep 31 Aug
Galán *et al.* [18]	*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 2004 2009 2012	20 Jan 11 Jan 08 Mar 7 Feb	27 Dec 15 Sep 30 Oct 30 Nov
Andersen *et al.* [2]	*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 2004 2009 2012	26 Jan 21 Jan 14 Mar 03 Mar	01 Dec 03 Aug 29 Sep 21 Sep
Sánchez-Mesa *et al.* [37]	*The first day in which the daily pollen concentration reaches values over (start) and below (end) 30 grains/m*^{3}	2002 2004 2009 2012	17 May 11 Jan 07 May 25 May	03 Jun 15 Sep 30 Oct 18 Jun
Feher *et al.* [16]	*The first day in which the threshold reaches values over (start) and below (end) 3 grains/m*^{3} *for 4 consecutive days*	2002 2004 2009 2012	05 Feb 11 Apr 26 Sep 22 Aug	19 Jun 03 Jul 30 Oct 12 Oct

Season dates might differ according to their definition. Table 2.1 shows the differences between approaches on selected years. Threshold-based approaches such as [16] and [37] tend to limit the season where peak concentrations appear, and this implies a high sensitivity to isolated peak counts. In contrast, cumulative approaches widen the pollination period being sensitive to early moderate concentrations, as is the case for 2002 in the table. Figure 2.4 shows how restrictive the proposal of [37] is compared to [26] and how the season period varies by reducing the threshold to 15 grains/m^3.

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 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 [32, 34], symptoms appear over 30 grains/m^3 in Finland and Croatia, while in Spain the first symptoms are observed between 25 grains/m^3 [36] and 30 grains/m^3 [37]. By far, the most common threshold level found in the literature is 30 grains/m^3 [10, 20, 37] which corresponds to the concentration at which the first allergy symptoms appear. Therefore, this level is selected as a representative in this study.

2.3 Features Definition

In the pollen forecasting problem, the independent variables should contain relevant meteorological data and the pollen levels themselves, as all of them are known to play a crucial role in the development of the pollination process. At the same time, due to the "curse of dimensionality" and to ease the computational burden, it is important to

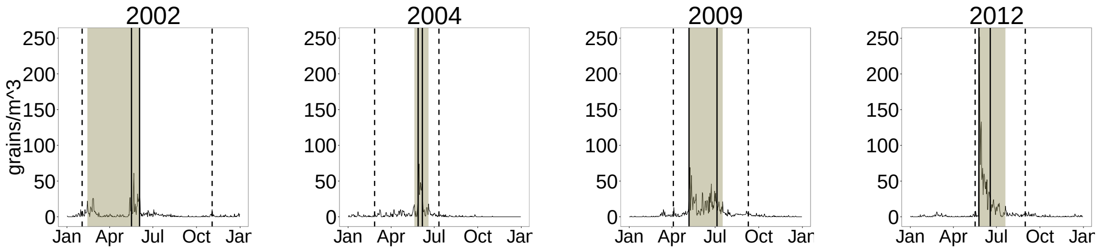


FIGURE 2.4: Pollen concentrations for years 2002, 2004, 2009 and 2012 and definition of the season according to [26] (vertical dashed line) and [37] (vertical solid line). The shaded rectangle represents the latter approach relaxing the threshold to 15 grains/m^3.

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 three 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 [35, 39]. Correspondingly, 10 and 30-days cumulative sums of daily pollen counts 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 [9]. Sum of temperatures up to some point are usually considered as a good representation of this absorbed energy [2, 9, 36]. Other authors [31] 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:

$$F_{\text{sum}}(d) = \sum_{i=d-n}^{d} R_{\text{forc}}(i), \tag{2.2}$$

where

$$R_{\text{forc}}(i) = \begin{cases} 0 & \textit{if } T(i) < T_{\text{forc}} \\ T(i) - T_{\text{forc}} & \textit{if } T(i) \geq T_{\text{forc}} \end{cases}, \tag{2.3}$$

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 (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 $\{1°\text{C}, 16°\text{C}\}$ for the forcing period and $\{-6°\text{C}, 8°\text{C}\}$ for the chilling period at thresholds of 10 grains/m^3 and 50 grains/m^3 respectively, as in [31].

Given the definition of the threshold in this study of 30 grains/m^3 the corresponding base temperatures are 8°C for the forcing and 6°C for the chilling (Fig 2.5).

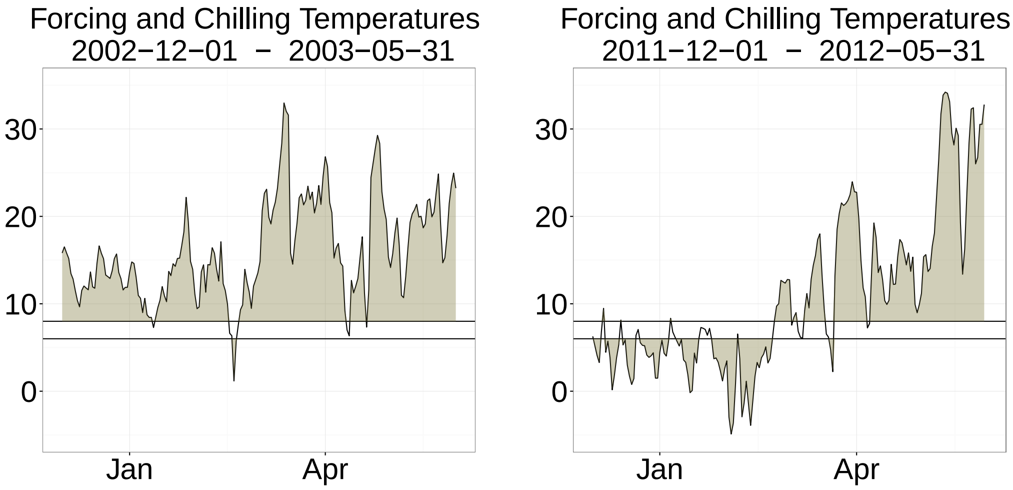


FIGURE 2.5: Sample seasons Forcing and Chilling Avg Temperatures. Sample of mild winter-spring season (left) and cold to warm winter-spring (right) with chilling and forcing base temperatures of 6°C and 8°C respectively

TABLE 2.2: 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

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 [30]. Hence a 30-days cumulative sum of wind speed features is generated. For all climate data, similar as for the pollen counts, the prior 7 daily raw data observations are also included.

This leads to the availability of 71 features as detailed in Table 2.2, which are distributed in a matrix corresponding to the desired forecast horizon.

For a classification problem approach, the class is discretized according to the threshold u,

$$\left[\begin{array}{ccc|c} x_{1,1} & \dots & x_{1,71} & p_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & \dots & x_{n,71} & p_n \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} x_{1,1} & \dots & x_{1,71} & c_{1+t} \\ \vdots & \ddots & \vdots & \vdots \\ x_{n-t,1} & \dots & x_{n-t,71} & c_n \end{array}\right] \tag{2.4}$$

$$c_i = \begin{cases} 0 & \textit{if}\, p_i < u \\ 1 & \textit{if}\, p_i \geq u \end{cases}, \tag{2.5}$$

where p_i is the daily pollen observation at time i, t the forecast horizon in number of days and u is the threshold. In order to define a regression problem the pollen observation is lagged accordingly to the forecast horizon,

$$\left[\begin{array}{ccc|c} x_{1,1} & \cdots & x_{1,71} & p_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,71} & p_n \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} x_{1,1} & \cdots & x_{1,71} & p_{1+t} \\ \vdots & \ddots & \vdots & \vdots \\ x_{n-t,1} & \cdots & x_{n-t,71} & p_n \end{array}\right] \tag{2.6}$$

Similarly, the proposal applies to the wind speed, which eases pollination for grasses. In this *a priori* scenario it makes sense to include short term wind features as well as moving averages as it might count as high relevance when combined with dryness. The assumption that dryness is represented by sunlight hours is taken as it is the observable variable provided instead of humidity. It can be also find in the literature that sunlight, which also implies temperature increase, in combination with preceding rainy periods assists on the speed of formation of the grasses. Hence, the model also includes as a feature long and short term cumulative sum of sunlight hours, applying same reasonings as above described.

Chapter 3

Comparing three algorithms to forecast the start and end of grass pollen season

In this chapter we approach the problem of predicting the start and the end dates for the pollen season of grasses (the family Poaceae) in the city of Madrid. A classification-based approach is introduced to forecast the main pollination season, and the proposed method is applied to a range of parameters such as the threshold level, which defines the pollen season, and several forecasting horizons. Different computational intelligence approaches are tested including Random Forests, Logistic Regression and Support Vector Machines. The model allows to predict if there will be a risk exposure for patients and preventive measures should be activated for clinical institutions.

3.1 Introduction

Given the fact that allergenic pollen is responsible for allergic rhinoconjunctivitis, asthma and the oral allergy-symptom in about 15 million people in Europe, aerobiological studies have a clear clinical interest. Allergies have been continuously increasing in developed countries, not only in the number of affected patients but also in the severity of allergic reactions [40].

Poaceae pollen is one of the most prevalent allergens causing allergic reactions. The establishment and the prediction of a pollen calendar is essential to reduce the exposure of allergic patients to pollen during the days of higher pollen concentration. It is also important to enable the development of other preventive measures.

There is no consensus on how to define the pollination season [23] which is the period where airborne concentrations of pollen are measured. Some authors define it based on the cumulative daily pollen counts [2, 18, 26] and other authors define it based on predefined threshold levels over which the season is considered to be started and ended [37]. This study uses the threshold-based approach in order to define the season which is going to be forecast.

Climate directly or indirectly defines the vegetation and acts on two levels: (1) during the stages prior to flowering [9, 31], and (2) during the pollen season [25, 36]. In this study, we characterize different features of the pollen season in order to determine the effect of meteorological parameters on the incidence of Poaceae pollen in Madrid, Spain. Once the features are defined, several computational intelligence techniques are applied and compared according to their performance on this problem. We cast the season predicting problem into a binary classification one, in order to obtain the most accurate estimates for the start and end of the pollination season with special attention to the threshold at which allergy reactions might appear.

3.2 Materials and methods

3.2.1 Data description

The study uses observations of Poaceae pollen from the Faculty of Pharmacy of Complutense University of Madrid, Spain ($40°26'52.1''$ N, $3°43'41.1''$ W) from 1994 to 2013, provided by Red Palinológica de la Comunidad de Madrid. Meteorological data is provided by weather stations located in Barajas, Cuatro Vientos, Getafe and Colmenar and consists of hours of sunlight per day, the speed of wind in km/h, rainfall in mm/h and daily maximum, minimum and average temperature in degrees Celsius. The stations are distributed at varying distances of the Faculty of Pharmacy, so data are interpolated based on its distance to the pollen sampler.

One of the objectives of the experiments was to evaluate, in terms of their predictive ability in the framework of forecasting the grass pollen season in Madrid, different general purpose machine learning or statistic methods based on different paradigms. Hence, in order to select the best suited model, we tried them against the data described in Section 2.1.

Although 30 grains/m^3 threshold is of the most interest for allergy patients and clinical and research institutions as explained in Section 2.2, to extend the information provided by the algorithms, the threshold space is extended. For a set of thresholds $u = \{5, 15, 30, 50\}$ and for a set of forecast horizons $h = \{1, 2, 5, 7, 10, 15\}$, we trained the three methods using the training set and checked their performance against the test data set represented by the years 2011, 2012 and 2013.

The main objective of this work is to help allergy patients in knowing in advance between which dates the pollen concentrations will be at risk levels. Given the above definitions of pollination season start and end, we aim at developing a model which forecasts these dates.

As seen in Section 2.2, there is no consensus as to which are the pollen concentrations considered as risk levels. Hence, several thresholds, ranging from 5 to 50 grains/m^3, will be used in this work in order to provide a variety of options and to compare them.

Another important element that needs to be fixed is the forecasting horizon, which corresponds to the number of days in advance pollen concentrations will be forecast. There is always a trade off between precision and anticipation, and in the literature we can find predictions of the pollen season which range from 1 to 10 days in advance. In order to test its predictive capacities, the model will produce forecasts for several forecasting horizons ranging from 1 to 15 days.

Finally, for each combination of thresholds and horizons, different derived meteorological and pollen features are computed to set up the instances on which different machine learning algorithms will be trained.

Our approach is based on the idea that one can cast the forecasting problem into a binary classification problem where the featured instances represent influential factors for the predictions. Hence, daily pollen concentrations are mapped to $\{0, 1\}$ depending on whether they are above the threshold (1) or not (0). Given the definition of season start, the first data point classified as 1 will indicate the start of the season.

The feature generation process described in 2.3 leaves us with a total of 71 features. Depending on the desired threshold and forecast horizon, the data is set up according to the parameters in order to transform it into a classification problem.

3.2.2 Computational intelligence models

Different classification approaches are trained using the training set in order to forecast the start and end of the season for test set. Concretely we compare Random Forests (RF) [7], Logistic Regression (LR) [11] and Support Vector Machines (SVM) [33].

Random Forest. Proposed in 2001 by Leo Breiman [7], 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 algorithm draws n trees bootstrap from the training data. For each bootstrap sample, grows a classification tree on where, at each node, randomly samples the predictors and choose the best split from among those variables. Finally predict new data by aggregating the predictions of the n trees using the majority votes.

Its main advantage over classification trees is its robustness against overfit by building many randomized, partial trees and vote to determine the class of the new observation. Each tree embraces a subset of the training data and captures the specific information it contains. Different random selections are computed by each tree improving the stability and accuracy, this technique is know as bootstrap aggregating or *bagging* [7].

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 [15]. 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.

Logistic Regression. Logistic regression is part of a broader family of generalized linear models where the conditional distribution of the response falls in some parametric family, and the parameters are set by a linear predictor. In binary logistic regression the response represents the absence or presence of a specific event, which is in this case whether the data point is over the predefined threshold or not.

The stability of the estimation of the parameters suffers when those covariate in a similar fashion. As several features were derived from others as they were phenologically justified, it is likely to find dependencies between them, thus it is intended to avoid the misbehavior of the maximum likelihood parameter estimation. Thus, a ridge estimator [11] was introduced to add penalty on weights learned to avoid over-fitting.

RF and LR make different assumptions about the data and has different rates of convergence. On the one hand, RF assumes that the decision boundaries are parallel to the axes based on whether a feature is $\geq, \leq, <$ or $>$ to certain value so the feature space is chopped into hyper-rectangles. On the other hand, LR finds a linear decision boundary in any direction by making assumptions on $P(C|X_n)$ applied to weighted features so non-parallel to the axes decision boundaries are picked out. This trade off motivates to take into account SVM as an alternative.

Support Vector Machines. The current SVM standard algorithm, proposed by Vladimir N. Vapnik and Corinna Cortes [13] in 1995, is a learning method used for binary classification which finds a hyper-plane which separates the d-dimensional data perfectly into its two classes. However, since sample data is often not linearly separable, SVM's introduces the notion of a *kernel induced feature space* which casts the data into a higher dimensional space where the data is separable. A good classifier is achieved when the hyperplane has maximum distance to the closest point of each class.

The radial basis function kernel (RBF) was used for the experiment in order to handle the nonlinear relations between the class and the features and to ease the numerical difficulties. To identify the optimal parameter set up a grid-search was perform over C and γ with a 10-fold cross-validation to prevent overfitting.

Feature Selection. It is known that LR is highly sensitive to collinearity in the features. In order to compare all the algorithms in equal conditions a correlation based feature selection was performed as a preprocess step. Some of the features might be redundant as they were generated from others. Given the computational cost of the algorithms, we need to reduce the number of features to those which are relevant for the class.

Hence, a filter algorithm based on [21] and on the definition of feature relevance by [24] is applied to rank subsets of features according to a correlation based evaluation function. This algorithm will select subsets that contain features highly correlated with the class and uncorrelated with each other. A feature is accepted when it predicts the class in areas of the instance space not already predicted by other features. The features are treated uniformly by discretization in a pre-processing step, and then a correlation based heuristic is repeatedly applied to test the merit of a subset, defined as

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{3.1}$$

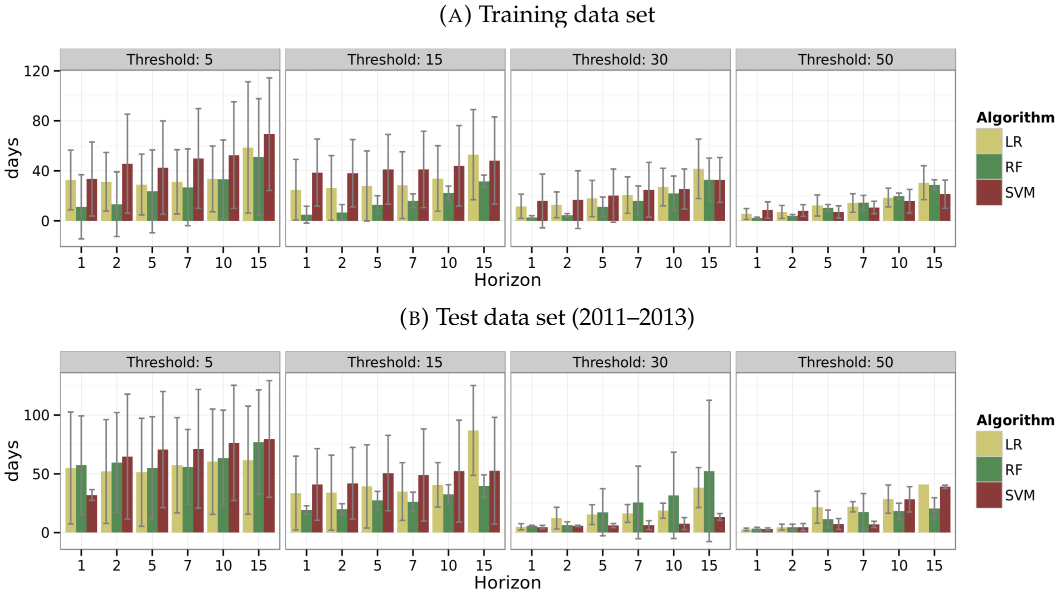
where M_s the merit of a subset S containing k features and $\overline{r_{cf}}$ is the mean feature-class correlation and $\overline{r_{ff}}$ the average feature-feature correlation.

This methodology leads to a significant reduction in the number of features to be used by the algorithms. For instance, with a threshold of $u = 30$ grains/m^3 and a forecast horizon of $t = 1$ days, the feature space consisting of 71 attributes is reduced to 6 features given the metric in (3.1). Being d the date on which the system forecasts the pollen concentration at day $d + t$, the subset of features for the previous example are: the daily pollen concentration for day d, the sum of daily pollen concentration for the period $[d - 10, d]$, the daily pollen concentration for day $[d - 30, d]$, the sum of chilling maximum temperature for the previous quarter, the sum of chilling minimum temperature for the previous semester and the sum of forcing minimum temperature for the previous semester.

In sum, the experiments are tailored to compare the models and compute their forecasts for each threshold and time horizon previously defined. Both parameters, threshold and horizon, define a set up of the data presented to the models according

Algorithm 1 Experimental design - Model Selection

```
Require: $X_i = \{x_{i1}, x_{i2} ... x_{i71} | c_i\}$    $i \in \{1...n\}$
for all $[m, u, t]$ in $\{models, thresholds, horizons\}$ do
    $[X_{train}, X_{test}] = \text{Preprocess}(X_i, u, t)$    ▷ Apply (2.4) (2.5) & Split
    $\text{Classifier} = \text{Tune}(m, X_{\text{train}})$
    $\text{Prediction} = \text{Classifier}(X_{\text{test}})$
    $E = \text{Error}(\text{Prediction})$
end for
```



(A) Training data set

(B) Test data set (2011–2013)

FIGURE 3.1: Mean forecasting errors and standard deviation for each forecasting horizon by threshold and model.

to Equations (2.4) and (2.5). Then a three step process applies consisting on feature selection and evaluation of the learning algorithm over the training set and prediction on the test set. The whole procedure is resumed in Algorithm 1.

3.3 Results and Discussion

Figure 3.1 shows the mean forecasting errors and standard deviations by method for each threshold and horizon. In the training data set, RF clearly outperforms the other two, especially for lower thresholds, being SVM the worst except in the case of $u = 50$. In this case, SVM is the best algorithm for horizons over 5 days. As could be expected, the error increases with the forecasting horizon, as the start of the season is easier to predict in the short term than in the long term. It is also interesting that all the methods obtain smaller mean errors and standard deviations with higher thresholds. The definition of a high threshold makes it easier to predict the start and end of the season, but the results are rougher and probably less informative: with higher thresholds, the start of the pollen season is sharper and more precisely defined.

The results with the training data set indicate that RF manages to capture the inner behavior of the data, but say nothing about its generalization abilities. On the other hand, concerning the testing data set, the results are mixed. LR shows slightly best

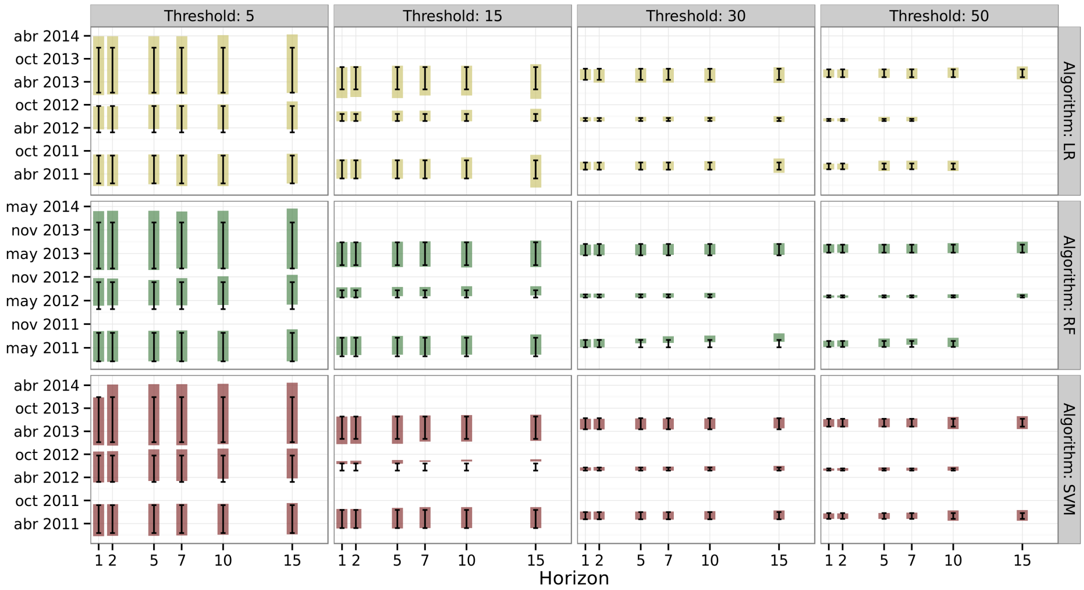


FIGURE 3.2: Predicted (colored rectangles) and observed (black vertical lines) season start and end dates for 2011, 2012 and 2013, by algorithm and threshold.

averages for thresholds fixed at $u = 5$ grains/m^3 (except for the day after, which is best predicted by SVM), while RF clearly outperforms the others for $u = 15$. SVM manages to get by far the best results in the case of $u = 30$.

The results with the test data set might derive from the fact that the models do not have enough data to properly generalize, as we only have 13 years, which means only 13 season starts and ends.

However, it is clear that high threshold levels lead to more satisfactory results, enabling the classifier to identify the patterns which influences the season start and end even for long forecasting periods as can be seen in Figure 3.2.

From a clinical point of view, predicting the moment in which most of the patients will start having symptoms is of a greater interest than predicting the moment when they will experience relief. Hence, Table 3.1 shows the predictive metrics by each model for all the horizons considered the threshold $u = 30$, following [3] (all patients experience moderate or severe symptoms).

Althought the results are mixed regarding only the season start and end date, it is clear that RF managed to outperform the other two classification algorithms when capturing the inner information of the series in this problem. Table 3.1 shows SVM with higher sensitivity than RF but, on the other hand, it achieves higher specificity which implies RF limits better the main pollination season without classifying high risk levels outside the period. In addition, RF provides a more accurate prediction. Compared to LR, RF manages to obtain a 30% better sensitivity at the cost of around 1% underperformance on specifity.

Results lead to choose RF as the best performer. Also 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. It is intended to extend the functionality of the system to cope with regression problems

TABLE 3.1: Test data set average errors for $u = 30$, of the predictions of the season.

Model	Horizon	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	AUC
LR	1	81	0	923	91	0.471	1.000	0.917	0.735
LR	2	63	0	922	109	0.366	1.000	0.900	0.683
LR	5	122	8	911	50	0.709	0.991	0.947	0.850
LR	10	97	13	901	75	0.564	0.986	0.919	0.775
LR	15	76	0	909	96	0.442	1.000	0.911	0.721
Avg LR						0.5105	0.9954	0.9189	0.7529
RF	1	137	7	916	35	0.797	0.992	0.962	0.894
RF	2	135	7	915	37	0.785	0.992	0.960	0.889
RF	5	123	6	913	49	0.715	0.993	0.950	0.854
RF	10	113	25	889	59	0.657	0.973	0.923	0.815
RF	15	110	26	883	62	0.640	0.971	0.919	0.805
Avg RF						0.7186	0.9845	0.9425	0.8515
SVM	1	143	70	853	29	0.831	0.924	0.910	0.878
SVM	2	144	82	840	28	0.837	0.911	0.899	0.874
SVM	5	122	36	883	50	0.709	0.961	0.921	0.835
SVM	10	112	82	832	60	0.651	0.910	0.869	0.781
SVM	15	117	199	710	55	0.680	0.781	0.765	0.731
Avg SVM						0.7419	0.8975	0.8729	0.8197

as well, given the relatively high number of instances and the presence of sudden high peaks in pollen concentrations as seen in Figure 2.4, RF provides stability and accuracy, due to the *bagging* [7] technique, in presence of outliers which motivates its selection against SVM.

3.4 Conclusions

This study introduces a new approach for helping allergy patients prevent the start and end of the pollination season, as well as the anticipation of resources for medical research. It is shown that tackling the problem from the data point of view obtains good results specially for thresholds higher than 20 grains/m^3. The proposal gets accurate forecasts of the pollination season even in years with particularly odd characteristics as it is 2012, which shows a specially short main pollination period with a sudden start.

We have seen RF as the most general model for prediction on this problem having a very accurate results for horizons within a week. The definition of the threshold, which dictates the start and end of the pollination season, takes an important role on the performance of the model. This study shows that levels above 20 grains/m^3 allow an accurate prediction. It is to note that the influential authors studied by Jato *et al.* [23] set the threshold at 30 grains/m^3 or above.

Chapter 4

Identifying influential factors for Poaceae pollen prediction

This chapter approaches the problem of identifying the most influential features in predicting Poaceae pollen concentrations in seven different locations distributed across the province of Madrid, Spain. As opposed to previous works, no prior assumptions were made about the significance of weather variables, instead they were estimated by using random forests, as concluded in the previous chapter. Hypothesis testing was used to provide statistical evidence of the results before applying the technique in a day ahead pollen concentration forecast. The results obtained suggest that the proposal is useful to support phenological and climatological studies in identifying the most relevant factors for forecasting. As a consequence, the results from the predictions ease allergy patients and clinical institutions in preventing the exposure to this aeroallergen.

4.1 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 classic statistical time series analysis to machine learning and computational intelligence.

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 behavior of just the pollen concentrations data [4]. 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 [9, 35, 39], or based on the relation with climate conditions [2, 31, 36], or both. Hovever, 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 [10, 29]. Others prefer the use of autoregressive indexes, as, for example, thermal indexes during plant formation season, in order to capture climatological information prior to pollen emission [2, 25, 28, 31].

On the other hand, automatic feature selection is an important research field in computational intelligence. Feature selection techniques are used, amongst 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.

The objective of this chapter 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 relevance of the variables, neither based on the phenology of the plant nor on meteorological consideration nor on derived indexes, 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 [17], 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. 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 [37]. The increase of allergy cases and the severity of the reactions [41] motivates the need for prediction of Poaceae concentrations.

4.2 Materials and methods

4.2.1 Data description

Pollen data. Daily Paoceae concentrations were provided for several locations by Red Palinológica de la Comunidad de Madrid. The observations consist of 8 locations from 2000 to 2013. The data is registered in Alcalá de Henares, Alcobendas, Aranjuez, Coslada, Faculty of Pharmacy of Complutense University of Madrid, Getafe, Leganés and Villalba.

Weather data. Weather data was provided by the Autonomous Community of Madrid Weather observations consist of average daily temperature in Celsius degrees, solar radiation in W/m^2, 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 5years 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. Figure 4.1 shows the distribution of the stations.

Very few missing data points were observed in the meteorological series so these were directly linearly interpolated. On the other hand, pollen series contain missing observations in what is believed *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. Concentration levels for the same day in different years do not imply similar levels in another given the differences in climate conditions, and thus in the phenology of the plant. Hence the regressed data is weighted to give more relevance to the data within the year.

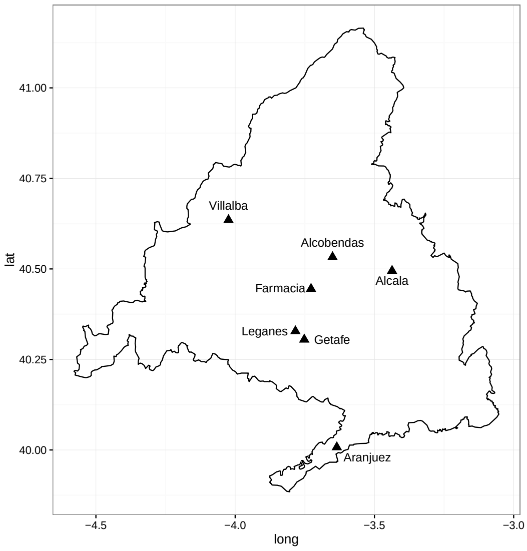


FIGURE 4.1: Location of weather and pollen stations.

4.2.2 Random forest for regression

There has been a lot of interest in ensemble learning which aggregates the results of many of the methods selected to boost their predictive performance. A well-known method is called *bagging* or bootstrap aggregating proposed by [5]. [7] proposed random forests (RF) which add 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 method such as neural networks. Given this randomness, our proposal intends to provide reproducible results and model is boosted by averaging several RF to mitigate the random effect on the results.

The procedure combines several randomized regression trees generated over sample fractions of the data, and aggregates their prediction by averaging. This average mitigates the influence of outlied data points giving RF advantage over support vector regression (SVR) which is highly sensitive in presence of outliers. As opposed to classification trees, the optimal split condition is the Variance, which at the same, is used to measure the importance of the variables. The relevance of a variable is estimated by looking how much prediction error or variance increases when data from that variable is permuted while the others are left unchanged.

In order to check the performance of the model a general purpose error metric for numerical predictions named root mean squared error (RMSE) defined by (4.1) was used along with the coefficient of determination R^2, which indicates the proportion of variance of the observed data was predicted.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.1}$$

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.

4.2.3 Nonparametric statistical test

In order to improve the evaluation process of the relevance of the features the use of statistical tests becomes necessary to confirm the conclusion obtained by the RF. Parametric tests are based on assumptions which are likely to be violated such as normality, or at least symmetry, on the distribution of the data.

Additionally, nonparametric tests can be applied to continuous data adjusting the input to test requirements via a ranking-based transformation. The Friedman test [17] it is a multiple comparisons test to detect significant differences between a set of at least two samples. In our approach the scope is to prove the existence of features which are more relevant than others regardless the year and the location observed. The first step of the procedure is converting the original variable importance for each year and location to its correspondent rank within the set to obtain the average rank $R_j = \frac{1}{n}\sum_i r_i^j$. Where j denotes the feature, i refers to each year and location and n is the total number of pairs $\{\text{location}, \text{year}\}$. Thus the null hypothesis of equality of medians is tested through the statistic,

$$F = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{4.2}$$

where k is the number of variables and $F \sim \chi^2_{k-1}$. 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 [12]. The main drawback is that these p-values are not suitable for multiple conparison as the probability error of a certain comparison, does not take into account the remaining comparisons belonging to the family.

To solve this problem, 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 adjustments is the Holm procedure [22] 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 he 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 [38] 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 relevant than v_2, it is not possible that v_1 is as relevant as v_3 and v_2 has the same relevance 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 $j \in \{1, ..., i\}$

In order to contrast the difference between the relevance of two variables we can use as an estimator the medians of the differences of each variable importance across locations and years [19]. 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,

$$m_{v_i} = \frac{\sum_{j=1}^{k} Z_{v_i,v_j}}{k} \tag{4.3}$$

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 relevance.

4.2.4 Experimental design

The first scope of this study is to provide evidence and support to phenological researches on the influential factors which might determine future atmospheric pollen concentrations. Once the influence of each weather and pollen feature studied is determined, a second step is performed to forecast the day ahead atmospheric concentrations. During the forecast, three models are proposed; (i) a forecast using all variables available in order to check whether the relevance of the variables maintains the same rank as in the theoretical results obtained in the nonparametric test, (ii) a model with only the 15 most relevant features from previous experiment, (iii) using only the 5 most relevant variables. Finally, all the results are compared and the trade-off between performance and execution time is discussed.

Given the different years observed at each location, our proposal is based on identifying, via the nonparametric test explained in Section 4.2.3, the most influential factors for a day ahead forecast regardless the location and the shape of pollen concentrations across the years. The approach consists in distributing the importances resulting from the RF in a matrix of dimension $N \times M$ where N represents each pair of $\{\text{location}, \text{year}\}$ observed and M the variables. Given the random nature of RF, an iterative approach was taken to avoid the likelihood of overfitting, the final result consists on the average of the iterations of the RF. 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

Algorithm 2 Nonparametric test

```
Require: series; iterations = $50$; $i = 0$
Require: $M \leftarrow NULL$
for all $[l, y]$ in $\{\text{location}, \text{years}\}$ do
    var.imp $\leftarrow NULL$
    for all iterations do
        var.imp = var.imp + RF.imp(series$_{l,y}$)
    end for
    var.imp = var.imp/iterations
    $M[i,]$ = var.imp
    $i{+}{+}$
end for
$F$ = friedman.test($M$)
if $F \leq \alpha$ then
    post.hoc($M$)
end if
```

Regarding the second scope of the study, our proposal tests the results in a more operational manner by conducting a day ahead prediction of airborne concentrations and comparing the importances of the variables with the results obtained in the nonparametric tests. The approach consists on using a *leave-one-out* (LOO) technique to split the series into training and test set at each location. By using this cross validation approach, the proposal provides results from a variety of pollen time series avoiding a fixed test set, which is highly dependent on the nature of the serie selected.

For each location, data is transformed as in (2.6) and the LOO is then applied by year. As implemented in the nonparametric test, iterations of RF executions are also included at each iteration of the LOO in order to mitigate both overfitting and parameter set up influence by averaging the results. This experiment is repeated three times, with all the variables, the 15 and the 5 most relevant obtained from the post-hoc procedure. The process for each model is summarized in Algorithm 3

Algorithm 3 Prediction

```
Require: series; iterations = 50
for $l$ in locations do
    $X^l$ = Preprocess($X^l_i$)                                  ▷ Apply (2.6)
    for $y$ in years(location) do
        $X^l_{\text{test}} = X^l_{\text{y}}$
        $X^l_{\text{train}} = X^l - X^l_{\text{test}}$
        for $i$ in iterations do
            $\text{RF}_i = \text{RF}(X^l_{\text{train}})$
            importance[$y$][$i$] = RF.importance($X^l_{\text{train}}$)
            prediction[$y$][$i$] = $\text{RF}_i(X^l_{\text{test}})$
        end for
    end for
    total.importance[$l$] = average(importance[$y$][$i$])
    total.prediction[$l$] = average(prediction[$y$][$i$])
end for
```

4.3 Results and discussion

4.3.1 Nonparametric test

The nonparametric test was tailored to give statistical evidence of the existence of features which are more influential than others. At each location the iterative approach of RF was taken over each year to obtain the relevance of each variable. Given the large number of variables, the pairwise computation might be expensive, thus the hypothesis test is reduced to those variables which represent more than 1% of the total variance, this leads to 16 variables. Out of this set up 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 $\chi^2_{15} = 24.99$ at $\alpha = 0.05$, leading to a computed p-value of $1.11\text{e}{-10}$, which strongly suggests the existence of significant differences among the variables. Table 4.1 shows the average Friedman ranking of the variables.

TABLE 4.1: Average ranking of the 16 most relevant features.

i	Variable	Ranking	i	Variable	Ranking
1	p_10	4.195	9	p_5	8.707
2	p_1	4.439	10	w_q	8.829
3	p_2	5.780	11	p_6	8.878
4	p_3	6.829	12	t_forc_q	10.561
5	h_Q	7.732	13	p_7	10.732
6	p_30	7.902	14	s_m	10.780
7	day	8.488	15	s_3	11.390
8	p_4	8.707	16	t_ma5	12.049

Due to the fact the null hypothesis for Friedman's test is rejected, a post-hoc test can be performed to detect the pairs which produce the difference. Table 4.2 shows the contrast estimation of medians of the relevance in percentage, it is noticeable how p_1 and p_10 obtain as an average around 5% more importance compared to other variables, supporting the rankings obtained by Friedman's test. These two variables are followed by p_2 and p_3 which outperform around 2% and 1.5% respectively when compared to the remaining set. On the other hand, t_ma5 achieves the lowest relevance across all stations and years.

Carrying out the post-hoc pairwise test will tell the evidence in the differences among pairs of variables. Table 4.3 shows the rejected hypothesis (p-value $\leq \alpha$) with a

	t_forc_q	p_10	s_m	day	p_6	w_q	p_1	h_Q	s_3	p_2	p_3	t_ma5	p_5	p_4	p_30	p_7
t_forc_q	0.00	-5.58	-0.04	-0.62	-0.73	-0.65	-6.22	-1.32	-0.05	-2.88	-1.87	0.13	-1.12	-1.07	-1.01	-0.46
p_10	5.58	0.00	5.54	4.96	4.85	4.92	-0.64	4.26	5.52	2.69	3.71	5.71	4.46	4.51	4.57	5.12
s_m	0.04	-5.54	0.00	-0.59	-0.70	-0.62	-6.19	-1.29	-0.02	-2.85	-1.83	0.17	-1.08	-1.03	-0.97	-0.42
day	0.62	-4.96	0.59	0.00	-0.11	-0.03	-5.60	-0.70	0.57	-2.26	-1.25	0.76	-0.49	-0.44	-0.39	0.16
p_6	0.73	-4.85	0.70	0.11	0.00	0.08	-5.49	-0.59	0.68	-2.15	-1.14	0.86	-0.39	-0.34	-0.28	0.27
w_q	0.65	-4.92	0.62	0.03	-0.08	0.00	-5.57	-0.67	0.60	-2.23	-1.22	0.79	-0.46	-0.41	-0.36	0.20
p_1	6.22	0.64	6.19	5.60	5.49	5.57	0.00	4.90	6.17	3.34	4.35	6.36	5.11	5.16	5.21	5.76
h_Q	1.32	-4.26	1.29	0.70	0.59	0.67	-4.90	0.00	1.27	-1.56	-0.55	1.46	0.21	0.26	0.31	0.86
s_3	0.05	-5.52	0.02	-0.57	-0.68	-0.60	-6.17	-1.27	0.00	-2.83	-1.82	0.19	-1.06	-1.01	-0.96	-0.41
p_2	2.88	-2.69	2.85	2.26	2.15	2.23	-3.34	1.56	2.83	0.00	1.01	3.02	1.77	1.82	1.87	2.42
p_3	1.87	-3.71	1.83	1.25	1.14	1.22	-4.35	0.55	1.82	-1.01	0.00	2.00	0.75	0.80	0.86	1.41
t_ma5	-0.13	-5.71	-0.17	-0.76	-0.86	-0.79	-6.36	-1.46	-0.19	-3.02	-2.00	0.00	-1.25	-1.20	-1.14	-0.59
p_5	1.12	-4.46	1.08	0.49	0.39	0.46	-5.11	-0.21	1.06	-1.77	-0.75	1.25	0.00	0.05	0.11	0.66
p_4	1.07	-4.51	1.03	0.44	0.34	0.41	-5.16	-0.26	1.01	-1.82	-0.80	1.20	-0.05	0.00	0.06	0.61
p_30	1.01	-4.57	0.97	0.39	0.28	0.36	-5.21	-0.31	0.96	-1.87	-0.86	1.14	-0.11	-0.06	0.00	0.55
p_7	0.46	-5.12	0.42	-0.16	-0.27	-0.20	-5.76	-0.86	0.41	-2.42	-1.41	0.59	-0.66	-0.61	-0.55	0.00

TABLE 4.2: Contrast estimation in %.

significance level of $\alpha = 0.05$ for each pair compared. It can be seen there is evidence that p_10 and p_1 significantly differs from most of the other variables included but there is no evidence they differ from each other. Out of the contrast estimation and the ranks from Friedman, it was shown these two variables represent the higher influence. All lead to statistical evidence that p_10 and p_1 are grouped as, in general, the most influential variables regardless the year and location.

There exists evidence of difference between p_2 and p_3 and the group composed by t_ma5, s_3, s_m, p_7 and t_forc_q which its higher rank is 10.56, separating the relevance of these two variables from the lowest ranked group in this study. Table 4.3 does not show clear distinction between {p_10, p_1} and {p_2, p_3} but referring to the logical relation between the combination of the pairwise hypotheses proposed by Shaffer [38], it can be seen an evidence of difference between p_10 and p_1 and p_4 which does not exist for p_2 and p_3, leading to conclude that p_10 and p_1 are more relevant than p_4, but there is no evidence that p_2 and p_3 are more relevant than p_4 so it can be stated that p_10 and p_1 are more influential than p_2 and p_3.

In summary, the nonparametric test evidences the existence of features which are more relevant than others. Among them, there exist groups of features which significantly differ from others but there is no statistical evidence of difference between group members. This means that, within a group, it is expected their members maintain or alternate their ranks within the bounds of the rank of the group. For instance, we have seen that p_10 and p_1 differ from the rest of the variables but do not from each other, constituting the top ranked group. As shown in Table 4.1, these variables have a Friedman's rank of 4.195 and 4.439 respectively, which is translated to position 1 and 2 in relevance. It is expected that p_10 and p_1 maintain their correspondent position or p_10 takes position 2, dragging p_1 to position 1 as they do not differ from each other.

4.3.2 Checking test results in an operational case

Firstly, in order to compare the results obtained in Section 4.3.1 in a more operational approach, the study performs RF using the LOO technique, using all variables, for the years available at each location independently, leaving the remaining year of each iteration as a test set. At each location, the relevance, averaged by iteration and test sets, of the most important variables is provided as shown in Figure 4.2. It can be seen the relation among the four most important variables p_10, p_1, p_2 and p_3 is maintained across locations as statistically evidenced in the rank test, except for Villalba. At

TABLE 4.3: 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	p_10 vs t_ma5	8.09e-14	9.71e-12	9.71e-12
2	p_1 vs t_ma5	4.59e-13	5.46e-11	4.82e-11
3	p_10 vs s_3	7.78e-12	9.18e-10	8.17e-10
4	p_1 vs s_3	3.83e-11	4.48e-09	4.02e-09
5	p_10 vs s_m	3.78e-10	4.39e-08	3.97e-08
6	p_10 vs p_7	5.09e-10	5.85e-08	5.34e-08
7	p_10 vs t_forc_q	1.41e-09	1.61e-07	1.48e-07
8	p_1 vs s_m	1.63e-09	1.84e-07	1.71e-07
9	p_1 vs p_7	2.17e-09	2.43e-07	2.28e-07
10	p_2 vs t_ma5	2.50e-09	2.78e-07	2.63e-07
11	p_1 vs t_forc_q	5.81e-09	6.40e-07	6.11e-07
12	s_3 vs p_2	9.56e-08	1.04e-05	1.00e-05
13	p_3 vs t_ma5	6.91e-07	7.47e-05	7.26e-05
14	s_m vs p_2	1.98e-06	0.00	0.00
15	p_2 vs p_7	2.49e-06	0.00	0.00
16	t_forc_q vs p_2	5.46e-06	0.00	0.00
17	p_10 vs p_6	8.45e-06	0.00	0.00
18	p_10 vs w_q	1.05e-05	0.00	0.00
19	s_3 vs p_3	1.44e-05	0.00	0.00
20	p_10 vs p_5	1.78e-05	0.00	0.00
21	p_10 vs p_4	1.78e-05	0.00	0.00
22	p_1 vs p_6	2.43e-05	0.00	0.00
23	w_q vs p_1	2.98e-05	0.00	0.00
24	h_Q vs t_ma5	4.03e-05	0.00	0.00
25	p_10 vs day	4.46e-05	0.00	0.00
26	p_1 vs p_5	4.92e-05	0.00	0.00
27	p_1 vs p_4	4.92e-05	0.00	0.00
28	t_ma5 vs p_30	8.04e-05	0.01	0.01
29	p_1 vs day	0.00	0.01	0.01
30	s_m vs p_3	0.00	0.02	0.02
31	p_3 vs p_7	0.00	0.02	0.02
32	t_forc_q vs p_3	0.00	0.03	0.03
33	p_10 vs p_30	0.00	0.04	0.03
34	h_Q vs s_3	0.00	0.04	0.04
35	day vs t_ma5	0.00	0.06	0.06

this location, the importance of daily wind speed accumulated during 90 days prior the forecast date (w_q) accounts for 15% of relevance. Being Villalba located at 903 m above the sea level while Aranjuez and Getafe have an elevation of 495 m and 622 m respectively, its meteorological conditions, related to mountain climate, and the highly correlated this variable is with the atmospheric concentrations during the study period, increases its relevance at this location. For instance, during those years w_q 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. Similar applies to pr_Q, which also gains importance due to the elevation, dropping average daily pressure and influencing, apparently, to flower formation and consequently pollen release.

The relation between the 4 most important variables stays as concluded in Section 4.3.1. Variable p_10 keeps the best rank among them except for the stations located in Farmacia, Villalba and Leganés, where its position is exchanged with the second most important variable from the test (p_1). This situation is expected as the pairwise hypothesis does not evidence difference between them. On the other hand, ranks for p_2 and p_3 are perfectly maintained across all locations as well as their relation to the top 2 ranked features. We have seen the ranks and the relations established in the nonparametric test are maintained among those locations which share similar meteoroloical conditions. The particular climate in Villalba makes some weather features

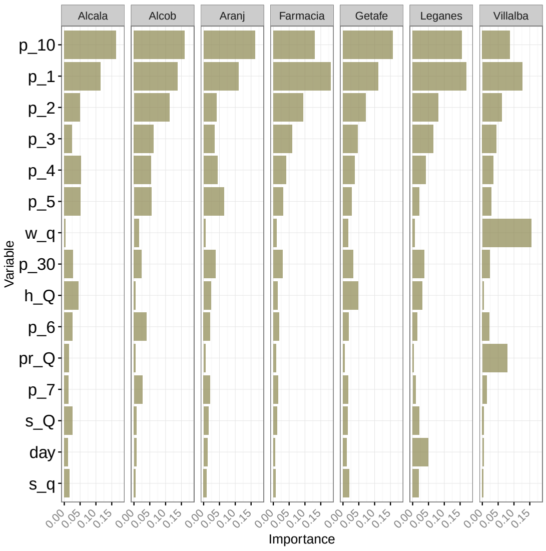


FIGURE 4.2: Selection of the 15 most important variables among all of them by station.

TABLE 4.4: Variable Description

Variable	Description
p_10	pollen daily accumulation 10 days prior the forecast date
p_i	pollen concentration the ith $\in [1, 7]$ day prior the forecast date
w_q	wind speed accumulation 90-days prior the forecast day
p_30	daily pollen accumulation one month prior the forecast day
h_Q	humidity accumulation 180 days prior the forecast day
pr_Q	pressure accumulation 180 days prior the forecast day
s_Q	sun accumulation 180 days prior the forecast day
s_q	sun accumulation 90 days prior the forecast day
day	day of the year

gain importance.

4.3.3 Day ahead forecast

For each location the LOO technique was used for the years available, RF for regression was trained using an average of 50 repetitions with different parameter set up within the parameter optimal space, leaving the remaining year of each iteration as a test set. The one day ahead predictive performance of the model is tested by the root mean square error (RMSE) using all variables, the 15 and the 5 most important as shown in Table 4.5.

TABLE 4.5: Average RMSE and R^2 of the test years studied at each location.

Station	RMSE$_{all}$	RMSE$_{15}$	RMSE$_5$	R$^2_{all}$	R$^2_{15}$	R$^2_5$
Alcalá	18.96	17.99	18.52	0.62	0.57	0.49
Alcobendas	14.98	15.35	14.94	0.69	0.68	0.60
Aranjuez	16.48	16.49	18.20	0.64	0.58	0.47
Farmacia	18.61	17.84	17.98	0.62	0.58	0.51
Getafe	16.23	16.33	17.63	0.70	0.71	0.55
Leganés	18.83	16.56	21.75	0.69	0.71	0.54
Villalba	17.47	15.97	18.80	0.62	0.69	0.55

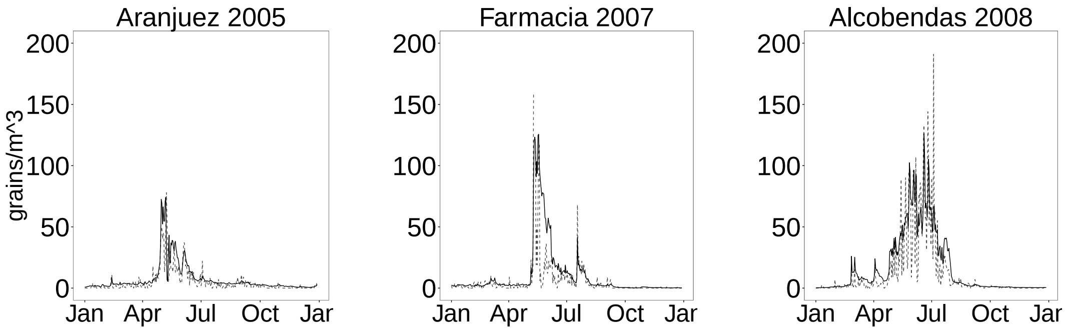


FIGURE 4.3: Sample prediction (solid line) vs observed (dashed) with minimum (left), average (middle) and maximum (right) RMSE.

The proposal using all variables, achieves an average RMSE_{all} of 17.37 grains/m^3 with a R^2_{all} of 0.65 on average. We have seen an error equal to 3 grains/m^3 as an average of the best performing test years across the locations while the worst performing years goes up to 30.32 grains/m^3. The main reason for this situation is the appearance of sudden extreme pollen concentration peaks during the pollination season. The model struggles to capture the whole path of the peak contributing with an increase of the error. It can be seen in Figure 4.3 the situation in Alcobendas in 2008 when comparing the predicted airborne concentration versus the observed. The presence of sudden high peaks, even up to 200 grains/m^3, makes the model predict with an RMSE equal to 28.64 grains/m^3, its worst performance test year for that location. The higher the atmospheric concentration peak, the bigger the error as can be seen in Farmacia the same year, where peaks up to 400 grains/m^3 can be observed during the main season and, consequently increasing the error to 36.72 grains/m^3.

On the other hand, the model performs well when there is a lack of concentration peaks over 100 grains/m^3 achieving an average error of 2 grains/m^3 in the full year not taking into account the location at Villalba, which contributes with an average of 8.42 grains/m^3 on its best performing year. This is due to the limited data available of 3 years (2007-2009), out of which two of them are used for training the algorithm. We believe more data observations would let the model to capture the inner information of different characteristics of the series, thus decreasing the mean error.

Looking at the forecasting error when reducing the number of variables, the model achieves $\text{RMSE}_{15} = 16.65$ with the same R^2 as if all variables would have been used. Although RF is robust against overfitting, reducing redundancies in the number of variables eases its avoidance, which results in this case in higher accuracy. Although, at some locations this does not apply, as for example Getafe, where the error increases in 0.10 grain/m^3 compared to the model with the full set of variables. A parallel experiment was done to investigate this issue, showing the stability of the errors when sequentially reducing the variables until around 10 features. This increase in the error is assumable as the complexity of the algorithm is reduced to almost $\frac{1}{5}$ of the original version. As opposed to reducing the number of variables to 15, using only 5 features results in an increase of the error to $\text{RMSE}_5 = 18.26$ as an average with a $\text{R}^2_5 = 0.53$. Errors across all stations increase except for Alcobendas, which allows a bigger reduction in the number of features when the sequential experiment is performed. Clearly, there is an inconvenience in limiting in excess the number of features as relevant information can be missed.

Our proposal provides statistical evidence of the existence of relevant groups of variables when forecasting day ahead pollen concentrations. The model succeeds in weighting and identifying influential features, concretely top ranking previous pollen daily observations and 10-days cumulative airborne concentrations, which ave been proved to serve as an indicator of the development of the plant according to the findings in [35, 39]. As a first step of this proposal, a Friedman test and post-hoc procedures were applied to identify the most relevant features for prediction from a purely data point of view, the results obtained support the selection of the variables using botanical criteria. As a second step, top ranked variables were used the results form the models were compared to the performance using all variables available.

Figure 4.2 shows the most influential variables differ at Villalba from the findings in the nonparametric test. This situation was led by the particular climate conditions at this location the period studied compared to other stations. As a result the model leads to weight more some weather influential variables in detriment to pollen features. On the other hand the results obtained from the tests showed in Tables 4.1 and 4.3 provides evidence that in general p_10, p_1, p_2 and p_3 are the critical features for a day ahead forecast.

Regarding the predictions, our proposal struggles to forecast airborne concentrations which exceed 150 grains/m^3 negatively impacting the error as seen in Table 4.5, specially at the Faculty of Pharmacy (Farmacia) and Alcalá where peak concentrations up to 400 grains/m^3 were observed. To our knowledge, this situation affects to all the models proposed so far. This is not a particularly worrying issue as the metric can be improved by limiting the observed atmospheric concentration to levels which are considered risk levels for human health, for instance, according to [32] and [34], symptoms appear over 30 grains/m^3 in Finland and Croatia, while in Spain the first symptoms are observed at 25 grains/m^3 [36]. Despite this fact, the results achieved by the proposal achieves accuracies which compare to other studies such as [29], who obtains an average of $R^2 = 0.6633$ compared to an average $R^2 = 0.65$ in our model with 15 variables. Our proposal provides a framework which makes no assumptions on the structure of the method used, letting the algorithm to adapt itself to the inner information of the data, having the advantage of limiting the efforts in investigating the hyperparameter of the algorithm as in neural networks. The proposal presented in this study achieves an average RMSE$= 17.43$ across all configurations, being the model with 15 variables the best configuration with a $\text{RMSE}_{15} = 16.65$, overperforming the best regression model proposed by [14] which achieves a RMSE=33.53. Additionally, we believe a more generalized results are presented due to the LOO technique and the multiple locations as opposed to a predefined test set, on which the results rely on the specific characteristics of the years selected.

Limiting the number of variables form 71 to 15 not only reduces the complexity of the algorithm but also increases the general accuracy in this study. Although there is an open debate about the optimal number of variables to reduce. The results were satisfactory but, it deserves further research on the optimization of the trade-off between number of variables and accuracy reduction.

4.4 Conclusions

This paper introduces a new approach to identify the influential factors for Poaceae pollen forecast. Through nonparametric tests the proposal provides statistical evidence of the consistency in selecting the most influential variables given the forecast horizon. Being previous pollen observations and p_10 the most relevant factors, supporting from a purely data point of view the findings from phenological studies. The forecast models proposed in the second step, reinforce the conclusions from the theoretical rank test introduced. The test was performed over yearly data available at the locations proposed, but it can be also tailored to limit the study period according to the needs of the researcher, for example, recognizing the main influential factors only during the pollination season. Besides, establishing longer forecast horizons, for instance one month, will arrange the importance of the variables and contribute in extending the knowledge about the influence of different variables from those presented in this study.

In order to provide a more practical approach, the tests were extended to forecast the day ahead airborne concentrations at each location using the LOO technique, which we believe it provides generalization. Additionally, several configurations of the model were tested based on the results from the nonparametric test, opening a debate about which is the optimal configuration of variables. The results are promising although further research is required when predicting extreme high atmospheric concentrations and selecting the optimal number of parameters. A research line was commented to address this issue by limiting up atmospheric concentrations to risk levels for allergy patients, which it would improve the forecast results and consequently would help patients and clinical centers to apply preventive measures.

Chapter 5

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

We haven concluded in previous chapters that Random Forests seems to be the best performer among the algorithms proposed. Besides, the algorithm is able to retrieve the most relevant information to forecast future airborne concentrations. In Chapter 4 we have seen that Random Forest for regression struggles in capturing the full path of main season peaks, and based on clinical studies, discretizing pollen concentrations to certain threshold provides enough useful information. In this chapter 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.

5.1 Introduction

Continuously increasing allergy symptoms in developed countries, and the clinic and socioeconomic relevance of this problem, has 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 [40]. 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 concentrations of pollen counts are measured. In the literature, several definitions of what it is considered a pollen season have been established [23]. It is possible to classify them into two main approaches, those based on the cumulative daily pollen counts [2, 18, 26] and those based on a predefined threshold level over which the season is defined to start and end [37].

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 [9, 31]. On the other hand, a dry and windy weather spreads the airborne quickly, leading to higher distributions [25, 36]. 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 the influence of meteorological conditions [2, 25, 31, 36] or previous pollen concentrations [10]. The aim of this research is to provide an assumption-free predictive model using a computational intelligence technique known as random forests (RF) [7]. 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.

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 [2, 10, 25]. 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 pollen counts 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.

5.2 Materials and methods

5.2.1 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/m^2. 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, and are measured in grains per cubic meter of air.

This study will be restricted to threshold-based season definitions from 2.2. In what follows, if u is a fixed 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 [32, 34], symptoms appear over 30 grains/m^3 in Finland and Croatia, while in Spain the first symptoms are observed between 25 grains/m^3 [36] and 30 grains/m^3 [37]. By far, the most common threshold level found in the literature is 30 grains/m^3 [10, 20, 37] which corresponds to the concentration at which the first allergy symptoms appear. Therefore, this level is selected as a representative in this study.

5.2.2 Random forest

Proposed in 2001 by Leo Breiman [7], 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 Figure 2.4, RF provides stability and accuracy in presence of outliers due to the *bagging* [7] 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 [15]. 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 [6, 7] 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.

5.2.3 Experimental design

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 Figure 2.4, the experiments were tailored to find the best model available. From sudden high peak concentration levels in short periods to prolonged moderate pollen counts, 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.

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. (2.4). 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 4.

Algorithm 4 System design

```
Require: $X_i = \{x_{i1}, x_{i2}...x_{i71} | c_i\}$      $i \in \{1...n\}$
Require: $u = threshold$
for all $[t, y]$ in $\{horizons, years\}$ do
    $AUC = 0$
    $X^S = \text{Preprocess}(X_i, u, t)$                    ▷ Apply (2.4) (2.5)
    $X^S_{\text{test}} = X^S_{\text{y}}$
    $X^S_{\text{train}} = X^S - X^S_{\text{test}}$
    for $k \in [1, 15]$ do
        $parameter_k = \text{Grid.Search}(search\_space)$
        $model_k = \text{Random.Forest}(parameter_k, X^S_{\text{train}})$
        if $AUC \leq \text{AUC}(model_k)$ then
            $AUC = \text{AUC}(model_k)$
            $best = parameter_k$
        end if
    end for
    $prediction = \text{Random.Forest}(best, X^S_{\text{test}})$
    $E = \text{Error}(prediction)$
end for
```

5.3 Results and Discussion

5.3.1 Forecast horizon

Table 5.1 shows the predictive metrics for each forecast horizon. Specificities and accuracies of over 90% are achieved across the different horizons.

TABLE 5.1: Predictive Metrics. Totals based on LOO method for the study period between 2000 and 2013

t	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

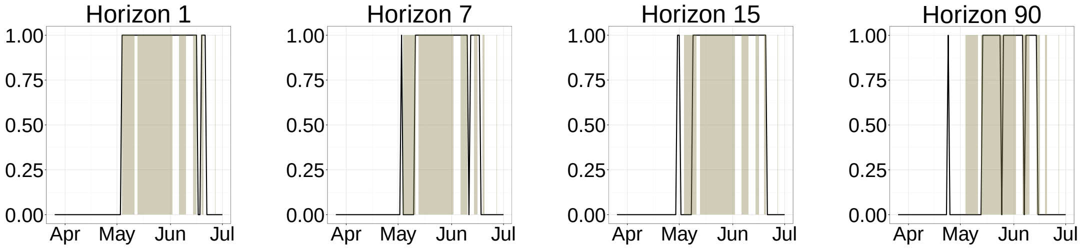


FIGURE 5.1: Pollen observed over the threshold 30 grains/m^3 (shaded) for Apr-Jul 2001 with forecast with horizon 1, 7, 15 and 90 days (solid lines).

The high values for specificity (true negative rate) indicates that the proposed model succeeds in identifying the periods of the main pollination season with an acceptable rate of false negatives (predicting concentrations below the threshold inside the observed season). Figure 5.1 shows the prediction for 2001 with a forecast horizon of 1, 7, 15 and 90 days. Given the 30 grains/m^3 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 over the threshold when they appear outside the main pollination season, showing a high number of false positives (FP). We 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 Figure 5.1 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,

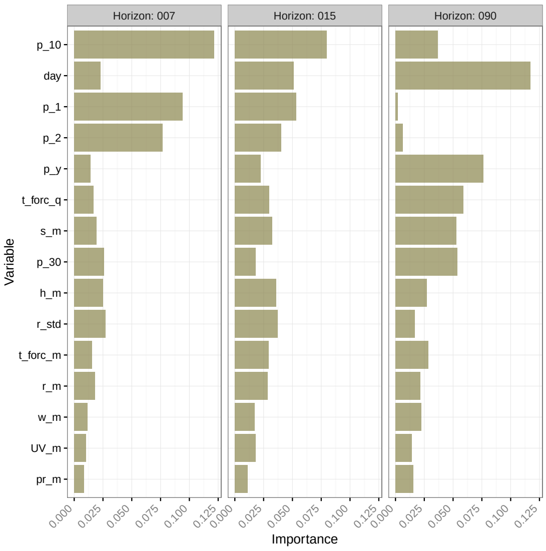


FIGURE 5.2: Selection of the 15 most important variables by forecast horizon.

TABLE 5.2: 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

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.

5.3.2 Forecast horizon vs feature importance

In Figure 5.2, the relative importance of the variables for a selected group of horizons is depicted. Each climate and pollen feature is labelled according to the method used to obtain it, as explained in Section 2.3. 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 $\mathrm{d} - i$, being d the forecast date, is represented by x_i. Table 5.2 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 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 rainfalls (r_std). It is known that rainy and humid conditions wash away pollen counts 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 Figure 5.2 how these two variables gain importance as the horizon increases.

5.3.3 Comparison with previous works

As seen in Table 5.1, our proposal achieves accuracies which compare favorably to other studies, for example [8], which 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 [10], being able to provide a more general approach when forecasting regardless the nature of the pollen series. Compared to the findings from [27], 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 [27], 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 [31]. Additionally, chilling temperatures are not ranked within the most influential features, confirming conclusions from [31] 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 [1]. Rainfall and humidity accumulations are positively related and influence the pollen release during the flowering period, in accordance to the findings of [1]. 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.

5.4 Conclusions

The present chapter 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.

Chapter 6

General Discussion

6.1 General conclusion

Concerns about the increasing number of allergy patients and symptoms have been raised during the past decades in developed countries. The environment today is very different from decades ago, providing evidence that current climate conditions, combined with other factors as pollution, exacerbate existing allergies. While it is difficult to act on some of these factors in the short and mid term, it is necessary to apply preventive measures to mitigate allergy symptoms, and consequently to reduce the socioeconomic impact.

Knowing in advance future exposure to pollen allergens enables an efficient application of correspondent preventive measures, such as scheduling intake of medication or planning the resources to respond the increasing number of patients. Several studies address the problem of forecasting airborne concentrations, but there is still a large gap between the information provided by the models and the needs of end users. To cover this gap, it is needed not only an accurate model but also a model which provides interpretable results about risk levels of pollen and their causes.

We have seen in Chapter 2 that there is no consensus on the definition of the main pollination season where the vast majority of the symptoms appear. Thus, there is a need of a fixed definition which considers the needs of the patients and clinical institutions. Based on clinical studies, the presented research establishes a threshold-based definition, being 30 grains/m^3 the level above which allergies appear in the study location.

Traditionally, the models developed in aerobiological studies used statistical tools such as time series analysis or linear parametric models. More recent studies use computational intelligence approaches such as artificial neural networks to predict risk levels in pollen concentrations. Each approach has its advantages and disadvantages and there is not a clear winner, from models which are highly dependent on the parameterization (traditional models) to those whose interpretability of the results cannot be easily deduced (artificial neural networks). As a result of the first part of this research, random forests (RF) was selected because both its performance and its robustness against variable collinearity, preventing from discarding features or expensive parameterizations as it happens with traditional models. Also, it enables a more robust approach against overfitting compared to other models.

Several studies make assumptions on the influence of meteorological parameters or flowering stages over pollen release. Consequently, the proposals and the results are highly dependent on the inputs. Instead, given the robustness of RF against variable collinearity, we use as input parameters to the system all available influential features based on proposals from bio-meteorological studies and univariate time series analysis. In Chapter 4, we found evidence about the capabilities of our proposal in identifying the most relevant features for predicting future atmospheric concentrations, which answers one of the open questions at the beginning of this research: what are the conditions which influence high pollen levels?

Most of the previous proposals forecast airborne concentrations and so we did using RF for regression. In all cases, the identification of extremely sudden high concentrations is one of the main difficulties. This has two main implications, (i) the model is influenced by the outlier values, thus biasing the forecasts, (ii) users with non-scientific background might not know the interpretation of pollen levels, for example, when concentrations are at 10 grains/m^3. To tackle this problem we proposed to cast pollen concentrations into two levels, below and above a threshold of 30 grains/m^3 which was clinically proved as a risk border in the region. Few studies approach the problem of predicting high risk atmospheric pollen concentration by discretizing pollen counts, and turn the regression problem into a classification problem. A binary classification problem eases the interpretability of results for a broader amount of users, as it simplifies the output information to risk or no-risk levels. Thus, covering another underlying question in a simply manner: is the population affected going to suffer from allergy? Is there going to be an increase of the allergy cases?

In the literature can be found forecast horizons up to 10 days. This seems insufficient to address the problems clinical institutions face in planning the resources for an increase of allergy cases. Gathering all the knowledge obtained along the research so far, we proposed in the last chapter to extend the horizons achieving exceptional results and easing the decision to be taken by providing forecasts up to 6 months. None of the previous studies provide long-term forecasts being [39] the only reference which manages to achieve satisfactory results for a 30-day-ahead forecast.

The research presented in this document deals with the prediction of future concentrations of pollen in the air. Starting from simple underlying questions on whether there is going to be risk of allergy or a preventive plan is needed, we found in the literature neither easily interpretable nor direct answers. Attending to those requirements, we investigated several computational intelligence techniques to select the most adequate for the objective. Random forests were chosen as the strongest candidate in order to satisfy both performance issues and the easy interpretability of the results. We were able to provide evidence about the ability of the model in identifying the most relevant factors for predicting future pollen concentrations according to the forecast horizon. Thus, our proposal contributes to provide information about the most influential variables, clarifying certain points which were studied in very few previous studies. Furthermore, it also serves as a support for different meteorological and phenological findings. By turning the problem into a binary classification problem, the interpretability of the results was increased, which eases the use of the output information by a broader public. The developed approach manages to outperform previous predictive proposals over the short-term, which establishes an edge in the field and opens a research path as few studies in biometeorolgy use the computational intelligence approach. Given the good results in the short term, we proposed an extension of the forecast horizon to long term, which was one of the objectives. The results were satisfactory, proving the suitability of the model, which was able to success in forecasting the main season and to identify the relevant factors according to the horizon.

6.2 Further developments

The novel methods used in this field and the deliverables produced from the research are related to producing more accurate pollen prediction models, which are adaptable and have already contributed to the knowledge of the biometeorology field. By extending the forecast horizon, and given the results obtained, the proposal is of great interest for clinical and research institutions in order to plan in advance the impact of high pollen concentrations, also patients can limit their exposure to risk levels. Aiming this research to clinical institutions, it would be interesting to match the resulting predictions from the proposal with the registered number of allergic cases in hospital patients, thus its functionality can be extended to forecast daily number of patients in nearby hospitals. In addition, by adding pollution data two further points can be explored; (i) how pollution influences pollen releases (ii) how pollution and pollen releases impact on society, by applying the analysis of the relevance of the variables presented in Chapter 4 to the number of hospitalized cases due to respiratory related symptoms, providing a clear clinical character.

A combination of thresholds and forecast horizons were provided to give the users the possibility to adapt the information to their requirements. Additionally the model extends and supports the knowledge about the influence of meteorological factors and flowering stages in the formation of Poaceae pollen releases. Although, further research is required on the generation of different features which might be relevant. Through a technique known as Grammatical Evolution (GE) a set of expressions, functions and operations can be applied to the underlying variables, and by using a genetic algorithm optimize a cost function which can be the prediction error, achieving the optimal combination of features.

Including numerical weather predictions as new factors and generating features from them will definitely increase the accuracy of the proposal, specially during the main pollination season where, as we have seen in Chapter 5, a large number of false positives appeared as a result of sudden climate conditions the date to forecast.

All this research lines constitute a promising and interesting topic which encouraged the author of this work to extend this study to a doctorate level.

This research project introduces a novel application for helping allergy patients, as well as the anticipation of resources for medical research. Also, supports and extend the knowledge of the influence of meteorological factors in pollen formation. Although the results are promising, allergic diseases have been on the increase. Hence, it is important to continue studying the changes in pollen season, with the aim of providing an always more precise picture about airborne concentrations under meteorological factors. The hope is contributing to widen the knowledge in this field, providing useful information to improve life conditions.

Appendix A

Author's proof IJB

Int J Biometeorol
DOI 10.1007/s00484-016-1242-8

ORIGINAL PAPER

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

Ricardo Navares[1] · José Luis Aznarte[2]

Received: 24 May 2016 / Revised: 24 August 2016 / Accepted: 25 August 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 performace 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/m^2. 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

	Month											
Year	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:

$$p_t = \beta \cdot \mathrm{r}_{\mathrm{row}} + (1 - \beta) \cdot \mathrm{r}_{\mathrm{col}}, \qquad (1)$$

where $\mathrm{r}_{\mathrm{row}}$ is a linear regression within the year, $\mathrm{r}_{\mathrm{col}}$ 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/m^3.

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

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

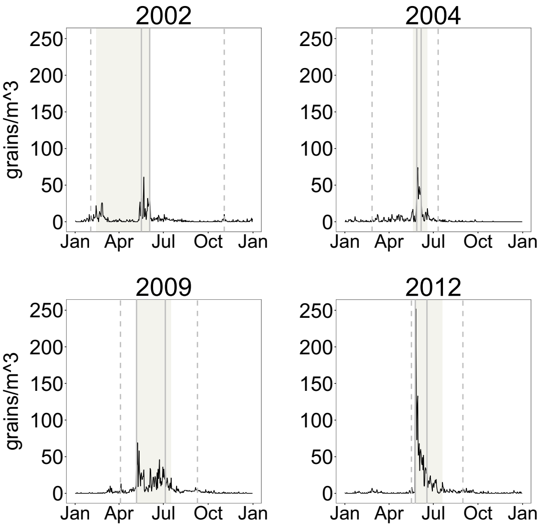


Fig. 1 Pollen concentrations for years 2002, 2004, 2009 and 2012 and definition of the season according to Nilsson and Persson (1981) (*vertical dashed line*) and Sánchez-Mesa et al. (2003) (*vertical solid line*). The shaded rectangle represents the latter approach relaxing the threshold to 15 grains/m^3

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/m^3 in Finland and Croatia, while in Spain the first symptoms are observed between 25 grains/m^3 (Rodríguez-Rajo et al. 1983) and studies such as Sánchez-Mesa et al. (2003) use 30 grains/m^3 . By far, the most common threshold level found in the literature is 30 grains/m^3 (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:

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

where

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

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/m^3 and 50 grains/m^3 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:

$$\left[\begin{array}{ccc|c} x_{1,1} & \dots & x_{1,71} & p_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & \dots & x_{n,71} & p_n \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} x_{1,1} & \dots & x_{1,71} & c_{1+t} \\ \vdots & \ddots & \vdots & \vdots \\ x_{n-t,1} & \dots & x_{n-t,71} & c_n \end{array}\right] \quad (4)$$

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

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 cummulative sum
30: previous 30-day cummulative sum
y: year to date cummulative sum
m: previous month cummulative sum
q: previous 90-day cummulative 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: $X_i = \{x_{i1}, x_{i2}...x_{i71}|c_i\}$    $i \in \{1...n\}$
Require: $u = threshold$
for all $[t, y]$ in $\{horizons, years\}$ do
    $AUC = 0$
    $X^S = \text{Preprocess}(X_i, u, t)$            ▷ Apply (4) (5)
    $X^S_{\text{test}} = X^S_{\text{y}}$
    $X^S_{\text{train}} = X^S - X^S_{\text{test}}$
    for $k \in [1, 15]$ do
        $parameter_k = \text{Grid.Search}(search\_space)$
        $model_k = \text{Random.Forest}(parameter_k, X^S_{\text{train}})$
        if $AUC \le \text{AUC}(model_k)$ then
            $AUC = \text{AUC}(model_k)$
            $best = parameter_k$
        end if
    end for
    $prediction = \text{Random.Forest}(best, X^S_{\text{test}})$
    $E = \text{Error}(prediction)$
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/m^3. 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/m^3 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/m^3 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 (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.

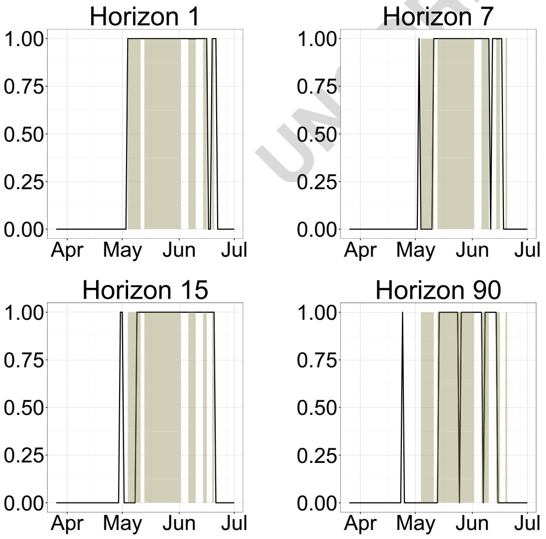


Fig. 2 Pollen observed over the threshold 30 grains/m^3 (*shaded*) for Apr-Jul 2001 with forecast with horizon 1, 7, 15 and 90 days (*solid lines*)

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,

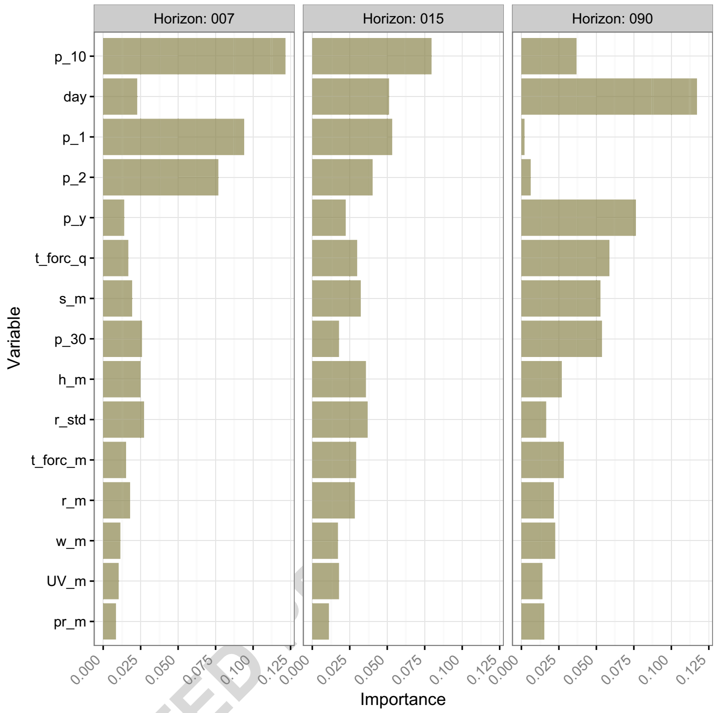


Fig. 3 Selection of the 15 most important variables by forecast horizon

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 observations add

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	accummulated forcing temperature 90 days prior the forecast day
t_forc_m	accummulated 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	accummulated 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

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 Bioeteorology 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 seeting 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 meteoclimatic 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, Nuntiis 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, Srnec 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-Lehtimaki 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

Bibliography

[1] F. Aguilera et al. "Phenological models to predict the main flowering phases of olive (Olea europaea L.) along a latitudinal and longitudinal gradient across the Mediterranean region." In: *Int. J. Bioeteorology* 59 (2014), pp. 629–641.

[2] T. B. Andersen. "A model to predict the beginning of the pollen season". In: *Grana* 30 (1991), pp. 269–275.

[3] I. Antépara et al. "Pollen allergy in the Bilbao area (European Atlantic seaboard climate): pollination forecasting methods". In: *Clinical & Experimental Allergy* 25.2 (1995), pp. 133–140.

[4] J. L. Aznarte et al. "Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models". In: *Expert Systems with Applications* 32.4 (2007), pp. 1218–1225.

[5] L. Breiman. "Bagging predictiors." In: *Machine Learning.* 25 (1996), pp. 123–140.

[6] L. Breiman. "Manual on seeting up, using and understanding random forest". In: *Stat. Dept. University of California Berkley.* v3.1 (2002).

[7] L. Breiman. "Random Forest". In: *Machine Learning* 45 (2001), pp. 5–32.

[8] M. A. Brighetti et al. "Multivariate statistical forecasting modeling to predict Poaceae pollen critical concentrations by meteoclimatic data." In: *Aerobiologia* 30 (2013), pp. 25–33.

[9] M.G.R. Cannell and R.I. Smith. "Thermal time, chill days and prediction of budburst in Picea sitchensis." In: *Journal of Applied Ecology* 20 (1983), pp. 269–275.

[10] M. Castellano-Méndez et al. "Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air." In: *Int. J. Biometeorology* 49 (2005), pp. 310–316.

[11] S. le Cessie and J.C. van Howelingen. "Ridge estimators in logistic regression." In: *Applied Statistics* 41 (1992), pp. 191–201.

[12] W. J. Conover. "Nonparametric methods". en. In: *Practical nonparametric statistics*. Ed. by Brad Wiley and Mary O'Sullivan. DOI: 10.1002/bimj.19730150311. John Wiley and Sons, 1999, pp. 233–305. ISBN: 978-0-471-16068-7.

[13] C. Cortes and V.N. Vapnik. "Support-vector networks." In: *Machine Learning* 20 (1995), pp. 273–276.

[14] Z. Csépe et al. "Predicting daily ragweed pollen concentrations using Computational Intelligence techniques over two heavily polluted areas in Europe." In: *Science of the Total Environment* 476–477 (2014), pp. 542–552.

[15] M. Fawcett. "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers." In: *Tech. rep., HP Laboratories* (2003).

[16] Z. Feher and M. Jarai-Komlodi. "An examination of the main characteristics of the pollen seasons in Budapest, Hungary (1991-1996)". In: *Grana* 36 (1997), pp. 169–174.

[17] M. Friedman. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." In: *J. of American Statistical Association* 32 (1937), pp. 674–701.

[18] C. Galán et al. "A comparative analysis of daily variations in the Gramineae pollen counts at Cordoba, Spain and London, UK". In: *Grana* 34 (1995), pp. 189–198.

[19] S. García et al. "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power." In: *Information Science* 180 (2010), pp. 2044–2064.

[20] B. James Green et al. "Atmospheric Poaceae pollen frequencies and associations with meteorological parameters in Brisbane, Australia: a 5 year record, 1994–1999." In: *Int. J. Biometeorology* 40 (2004), pp. 172–178.

[21] M. A. Hall. "Correlation-based feature selection for machine learining." In: *PhD. Thesis. University of Waikato* (1999).

[22] S. Holm. "A simple sequentially rejective multiple test procedure." In: *Scandinavian Journal of Statistics* 6 (1979), pp. 65–70.

[23] V. Jato et al. "May the definition of pollen season influence aerobiological results?" In: *Aerobiologia* 22 (2006), pp. 13–25.

[24] R. Kohavi and G.H. John. "Wrappers for feature subset selection." In: *Artificial Intelligence* 97 (1997), pp. 273–324.

[25] D. Myszkowska. "Predicting tree pollen season start dates using thermal conditions". In: *Aerobiologia* 30 (2014), pp. 307–321.

[26] S. Nilsson and S. Persson. "Tree pollen spectra in the Stockholm region (Sweden), 1973-–1980". In: *Grana* 20 (1981), pp. 179–182.

[27] J. Nowosad. "Spatiotemporal models for predicting high pollen concentration level of Corylus, Alnus and Betula." In: *Int. J. Biometeorology* 60 (2016), pp. 843–855.

[28] J. Otero et al. "Biometeorological and autoregressive indices for predicting olive pollen intensity." In: *Int. J. Biometeorology* 57 (2013), pp. 307–316.

[29] M. A. Otero et al. "A model to forecast the risk periods of Plantago pollen allergy by using ANN methodology." In: *Aerobiologia* 31 (2015), pp. 201–211.

[30] I. Silva Palacios, R. Tormo Molina, and A. F. Muñoz Rodríguez. "Influence of wind direction on pollen concentration in the atmosphere." In: *Int. J. Biometeorology* 44 (2000), pp. 128–133.

[31] A. Pauling, R. Gehrig, and B. Clot. "Toward optimized temperature sum parametrizations for forecasting the start of the pollen season". In: *Aerobiologia* 30 (2014), pp. 45–57.

[32] R. Peternel et al. "Poaceae pollen in the atmosphere of Zagreb (Croatia), 2002–2005." In: *Grana* 45 (2005), pp. 130–136.

[33] A. Rakotomamonjy. "Variable Selection Using SVM-based Criteria." In: *Journal of Machine Learning* 3 (2003), pp. 1357–1370.

[34] A. Rantio-Lehtimäki et al. "Significance of sampling height of airborne particles for aerobiological information." In: *Allergy* 46 (1991), pp. 68–76.

[35] H. Ribeiro, M. Cunha, and I. Abreu. “Definition of main pollen season using logistic model.” In: *Ann Agric Environ Med* 14 (2007), pp. 259–264.

[36] F.J. Rodríguez-Rajo, G. Frenguelli, and M.V. Jato. “Effect of air temperature on forecasting the start of the Betula pollen season at two contrasting sites in the south of Europe (1995-2001)”. In: *Int J. of Biometeorology* 47 (1983), pp. 117–125.

[37] J.A. Sánchez-Mesa et al. “Characteristics of grass pollen seasons in areas of southern Spain and the United Kingdom”. In: *Aerobiologia* 19 (2003), pp. 243–250.

[38] J. Shaffer. “Modified sequentially rejective multiple test procedures.” In: *J. of AmericanStatistical Association* 81 (1986), pp. 826–831.

[39] M. Smith and J. Emberlin. “A 30-day-ahead forecast model for grass pollen in north London, UK.” In: *Int J. Biometeorology* 50 (2006), pp. 233–242.

[40] M. Sofiev and K.C. Bergmann. “Allergenic Pollen: A review of the production, release, distribution and health impacts.” In: Springer Science and Business Media, 2012. Chap. Impact of pollen, pp. 161–215.

[41] Letty A. de Weger et al. “Impact of Pollen”. en. In: *Allergenic Pollen*. Ed. by Mikhail Sofiev and Karl-Christian Bergmann. DOI: 10.1007/978-94-007-4881-1_6. Springer Netherlands, 2013, pp. 161–215. ISBN: 978-94-007-4880-4 978-94-007-4881-1. (Visited on 08/22/2016).