



**HAL**  
open science

# Pollen risk levels prediction from multi-source historical data

Esso-Ridah Bleza, Valérie Monbet, Pierre-François Marteau

► **To cite this version:**

Esso-Ridah Bleza, Valérie Monbet, Pierre-François Marteau. Pollen risk levels prediction from multi-source historical data: Retraction of Vol 142, art no 102146, 2022. *Data and Knowledge Engineering*, 2023, 144, pp.article n°102146. 10.1016/j.datak.2023.102146 . hal-04051618

**HAL Id: hal-04051618**

**<https://hal.science/hal-04051618>**

Submitted on 30 Mar 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

## Highlights

### **Pollen risk levels prediction from multi-source historical data**

Esso-Ridah Bleza, Valérie Monbet, Pierre-François Marteau

- Pollen risk levels prediction from minimal meteorological information
- Aggregation of binary logistic regression models
- Learning with multi-source spatial and temporal data

# Pollen risk levels prediction from multi-source historical data

Esso-Ridah Bleza<sup>a,b,c</sup> (PHD student), Valérie Monbet<sup>b</sup> (Professor Researcher) and Pierre-François Marteau<sup>c</sup> (Professor Researcher)

<sup>a</sup>Lify Air, 1 avenue de champs de mars, Orléans, 45100, France

<sup>b</sup>University of Rennes 1, building 22-23, IRMAR (CNRS UMR 6625), Beaulieu Campus, Rennes, 35042, France

<sup>c</sup>South Brittany University, ENSIBS, IRISA (EXPRESSION), Tohannic Campus, BP 573, Vannes, 56017, France

---

## ARTICLE INFO

### Keywords:

Pollen

Aerobiology

Machine learning

Classification

Risk level prediction

## ABSTRACT

Numerous studies show that meteorological conditions have an impact on the emission, dispersion and suspension of pollens in the air. Several allergenic species permanently threaten the health of millions of people in France and that can be extrapolated that this is the case in most part of the world. Hence, preventive information on the risk of pollen exposure would become a real asset for allergy sufferers. The main objective of this article is to study, through statistical learning techniques exploiting historical data and meteorological parameters of the day ( $T$ ), the ability to predict three-day ahead ( $T + 3$ ) pollens presence risk levels in the air on a given territory (in metropolitan France). We are interested in the prediction of risk, discretized in four levels for three families of pollens which are among the most allergenic species (ragweed, cupressaceae and grasses). Combining binary logistic regression models for each risk level using a set of ranking rules or a random forest classifier is proposed in this study. The pollen risk level prediction performances reach 70% to more than 90% of auc, precision and recall on the majority of 68 considered sites and especially with a similar prediction capacity on sites with no previous pollen data. The comparative study with some more classical models of the literature shows that the proposed model have a slight performance advantage.


---

## 1. Introduction

Recent studies show that populations, particularly in France, are increasingly suffering from allergies to one or more pollens. It is estimated today that 30 % of the French population is concerned, against only 3 % in 1970, and that the allergic population will pass the 50 % mark in 2050 (Bettayeb, Cayrol and Girard, 2018). The only options available to allergic people today are either a recurrent treatment with its share of inconveniences (side effects, cost, dependence, etc.), or a treatment as soon as the symptoms appear with the known resulting consequences (illness, absenteeism, etc.). To date, measurements of pollen concentration in the air are reported to the public with a significant time lag. Indeed, the National Aerobiological Surveillance Network (NASN, in french RNSA<sup>1</sup>), a reference organization in France, disseminates information from HIRST sensors whose standard method, which is seventy years old, is used in many national networks for the measurement of pollen concentration (Hirst, 1952). HIRST sensors accounted for 70% of the sensors used in the world, and 98.5% in Europe in 2016 (Thibaudon, Oliver and Besancenot, 2019). The HIRST sensor (Figure 1) is made of a vacuum drum slowly rotating through a clock, equipped with adhesive blades so that the particles in the air come to stick on these blades (Thibaudon et al., 2019). It takes seven days from the start of operation of a vacuum drum to its reading, and at least two more days for it to be analyzed by an human operator before the pollen information is released (Cassagne, 2009). This technique, which is based on the analysis and identification of pollens under the microscope, causes a significant delay in the dissemination of information to the public, underlining the need to develop automatised short-term predictive models of pollen risk to complement the information disseminated by the RNSA and improve allergy prevention.

Many studies have shown that meteorological factors, such as air temperature (Howard and Levetin, 2014; Nowosad, Stach, Kasprzyk, Chłopek, Dąbrowska-Zapart, Grewling, Latałowa, Pędziszewska, Majkowska-Wojciechowska, Myszkowska et al., 2018; Ščevková, Dušička, Mičieta and Somorčík, 2015; Tseng, Kawashima, Kobayashi, Takeuchi

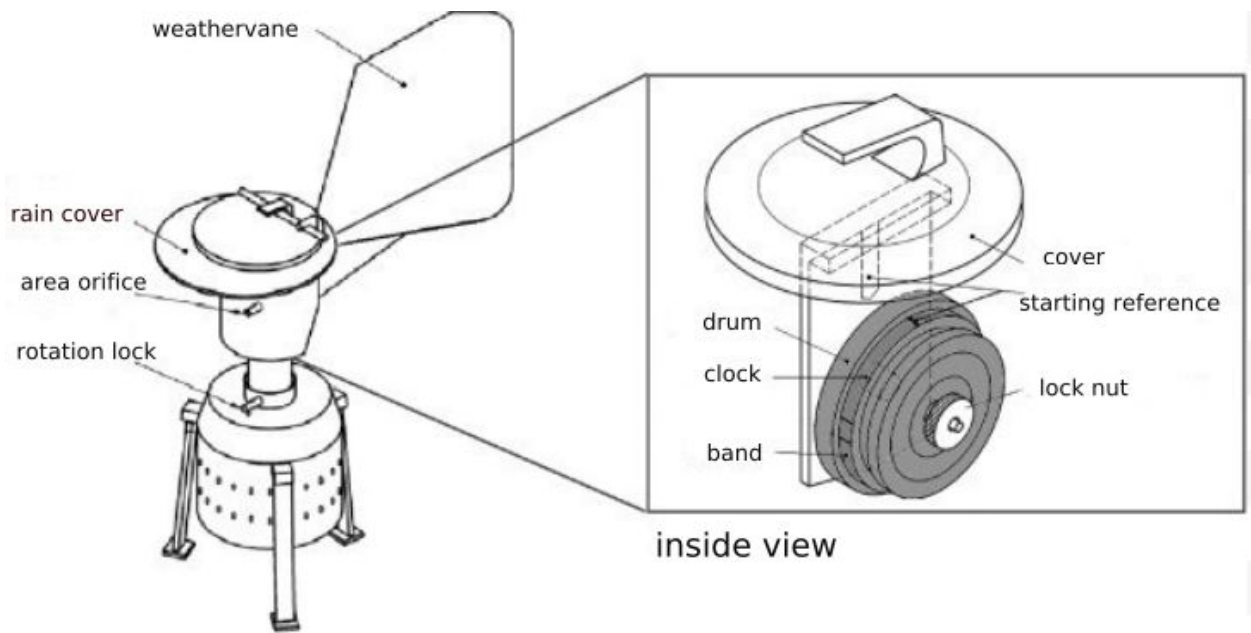
---

 [esso-ridah@bleza.fr](mailto:esso-ridah@bleza.fr) (E. Bleza); [valerie.monbet@univ-rennes1.fr](mailto:valerie.monbet@univ-rennes1.fr) (V. Monbet); [pierre-francois.marteau@univ-ubs.fr](mailto:pierre-francois.marteau@univ-ubs.fr) (P. Marteau)

 [www.bleza.fr](http://www.bleza.fr) (E. Bleza); [www.lifyair.com](http://www.lifyair.com) (E. Bleza); <https://perso.univ-rennes1.fr/valerie.monbet/> (V. Monbet); <http://people.irisa.fr/Pierre-Francois.Marteau/> (P. Marteau)

ORCID(s): 0000-0002-5998-2753 (E. Bleza)

<sup>1</sup><https://pollens.fr/>



**Figure 1:** Hirst sensor principle (Thibaudon et al., 2019)

and Nakamura, 2020), solar radiation (Iglesias-Otero, Fernández-González, Rodríguez-Caride, Astray, Mejuto and Rodríguez-Rajo, 2015; Nowosad et al., 2018; Tseng et al., 2020), sunshine duration (Myszkowska and Majewska, 2014; Rodríguez-Rajo, Valencia-Barrera, Vega-Maray, Suárez, Fernández-González and Jato, 2006), humidity (Ščevková et al., 2015; Makra, Matyasovszky, Thibaudon and Bonini, 2011; Tseng et al., 2020), and precipitation (Piotrowska, 2012; Rodríguez-Rajo et al., 2006) impact airborne pollen concentrations. Some papers show that wind also has a significant influence, in particular authors established that the wind is responsible for creating complex mixtures of pollen types that make the individual pollen detection quite difficult (Bohlmann, Shang, Giannakaki, Filioglou, Saarto, Romakkaniemi and Komppula, 2019). Hence, according to the previous studies, meteorological data play an important role in the development of predictive models, with cumulative temperature and precipitations generally proving to be highly significant variables. In addition to aforementioned meteorological variables, supervised approaches include phenological parameters, site characteristic data and pollen concentration history (Rodríguez-Rajo, Astray, Ferreiro-Lage, Aira, Jato-Rodríguez and Mejuto, 2010; Valencia, Astray, Fernández-González and al., 2019b; Zewdie, Lary, Levetin and Garuma, 2019a). Furthermore, studies have been carried out to evaluate the impact of meteorological conditions on the diffusion of fine particles in the air, i.e. particles smaller than  $1\mu\text{m}$  or  $2.5\mu\text{m}$  ( $PM_{1}$ ,  $PM_{2.5}$ ).

Our study differs from this state of the art mainly in three aspects: firstly, our study on the prediction of pollen risk level uses fewer meteorological covariates to deal with the availability of data necessary to cover the spatio-temporal domain that our study encompasses. We only consider temperature, humidity and precipitation whereas in the state of the art, atmospheric pressure, sunshine duration, diffuse radiation and wind speed are also integrated as covariates.

secondly, we are interested in the study of the French territory as a whole, by considering a total of 68 sites, whereas until now the studies concerned isolated localities on the scale of a city or a metropolis.

Finally, we study with the same methodology and the same meteorological data several of the most allergenic pollens. 21 species are considered in total, three of which are presented in this article to highlight our main results.

The article is organized as follows. The specificity of the data used to predict pollen episodes and the context of our study are introduced in Section 2. In Section 3, the predictive algorithms that we have specifically designed are detailed. Section 4 is dedicated to the assessment study through numerical experiments. The ability of the proposed algorithms to generalize is evaluated both in time and space, i.e. we explore the prediction of emissions at horizon  $T+3$  and at different geographical locations for which no training data was used. Finally, in Section 5, some concluding remarks and perspectives are given.

## 2. Nature of the data at hand

Our goal is to build an algorithm that predicts the allergic risk of various pollens on day  $T + 3$  given historical meteorological data up to the current day  $T$ . To this end, two sources of data are exploited: the first one reports pollens concentrations and the second one weather data. A detailed description of each of these data sets is given hereinafter.

### 2.1. Pollen data

Pollen emission data<sup>2</sup> are produced by RNSA which exploits the HIRST sensors, currently considered as the gold-standard technique to assess pollen emissions at a day to day basis. In practice, the emitted pollen grains are stuck on an adhesive tape which is afterwards collected and finally analyzed manually. As described in Sec. 1 the tapes of the HIRST sensors are collected, stained and analyzed with an optical microscope. Pollens and spores of different taxa are determined with their number per unit area using standardized procedures (Thibaudon et al., 2019; Buters, Antunes and Galveias, 2018). Daily pollen concentration data are freely available for twenty allergenic species and 74 sites during a period covering the years 2000 to 2017 (see map Fig. 9). After 2017, HIRST data is available on demand for research purpose. Since the sensors were not all installed at the same time, the available historical data vary according to the recording stations. The most recently equipped sites produce data from year 2012.

The distribution of the pollen concentration is asymmetric as shown in Fig.2. It is also characterized by a large number of zeros (more than 43% of the values in the first band  $[0, 1[$  as depicted in Fig. 2, bottom right) corresponding to days with no emission (zero value) or very low emissions (beginning and end of the pollen season). It presents also an heavy tail associated to some days of the year with very high emissions which can reach 3400 grains/m<sup>3</sup>/d for some species (see on Fig. 2, bottom left). Note also that emissions are less important from November to January. Moreover the pollen concentrations present a strong annual variability. As we can see on Fig. 3, the annual cumulative concentrations are fluctuating, with an inter-annual variation that can be quite marked. For instance, in 2012, the global emissions are low compared to years 2011 and 2013. 2012 has been a cold year with low yearly maximum temperature.

We detail below a descriptive analysis taking into account all 21 pollen families and all sites concerned by our study. This analysis is performed using standard unsupervised machine learning algorithms. The objective is to extract some general patterns to characterize pollens and territories. More precisely, the dimension reduction algorithms allow to extract typical situations that can highlight dependencies between pollens on the one hand and between sites on the other hand.

Consider  $X(p, s, t)$ , the standardized (to unit variance) concentration of pollen  $p \in P$  at location  $s \in S$  at time  $t \in T$ .  $P$  (resp.  $S, T$ ) is the set of all pollens (resp. sites, day indices). Two ways of representing those data are used as follows.

$$\log(X(p, s, t) + 1) = \sum_k C_k(p, t) Q_k(s) \quad (1)$$

and

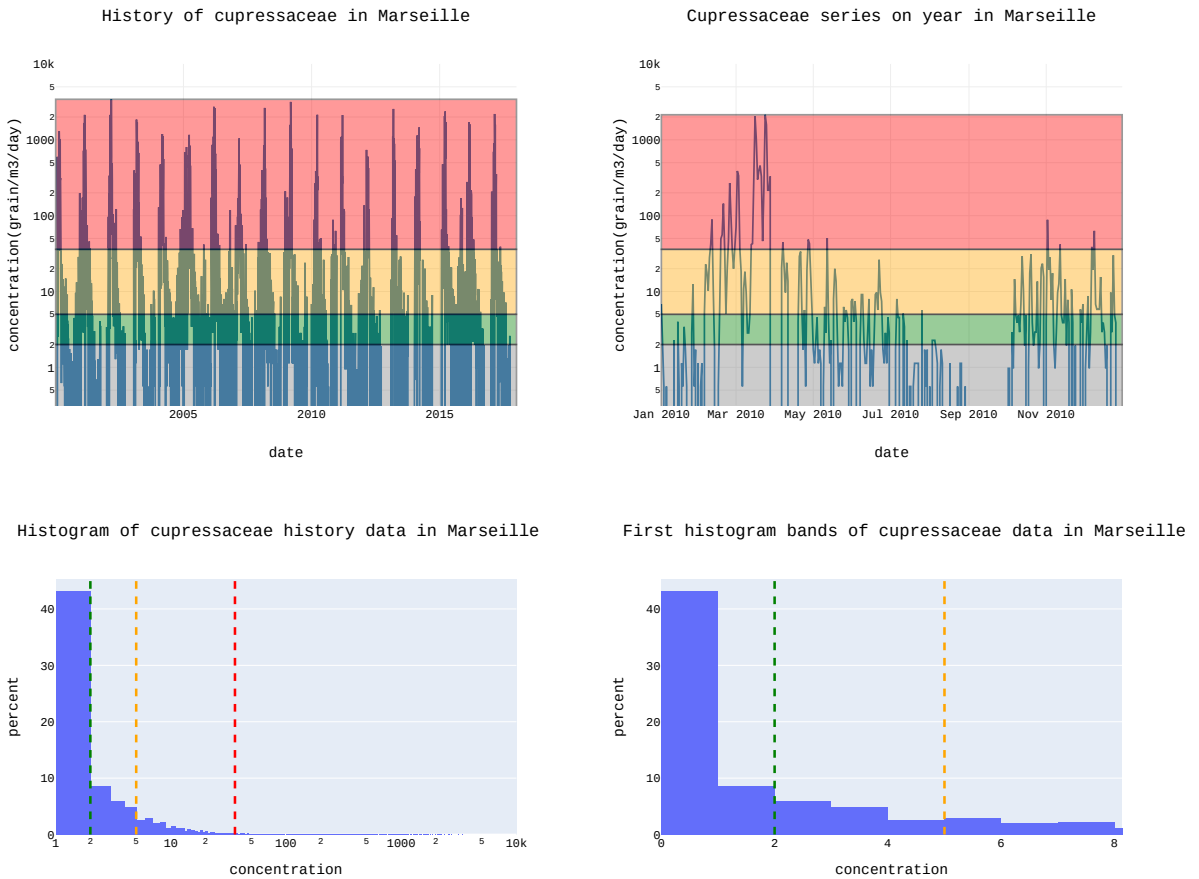
$$\log(X(p, s, t) + 1) = \sum_k \tilde{C}_k(s, t) \tilde{Q}_k(p) \quad (2)$$

under the assumptions  $\tilde{C}_k(s, t) \geq 0$  and  $\tilde{Q}_k(p) \geq 0$ .

$C_k$  (resp.  $\tilde{C}_k$ ) are coordinates in the space spanned by the eigen vectors  $Q_k$  (resp.  $\tilde{Q}_k$ ). They are obtained by minimizing the distance between the original data and the  $(C, Q)$  representation under some positivity constraints like in classical matrix decomposition methods. On Fig. 4 (resp. 5), the first eigen vectors  $Q_k$  (resp.  $\tilde{Q}_k$ ) are plotted. The first eigen vector is nothing but the mean of the total (standardized) concentrations for each site (resp. for each pollen). Pollen emissions are more important far from the coastline. Then, the next eigen vectors highlight typical situations. For instance,  $Q_2$  correspond to days where there is a lot of pollen in the air in the South-East part of France (see right top bottom panel of Fig. 4). Beside,  $\tilde{Q}_2$  correspond to days where poaceae emission are strong together with urticaceae, plantain and mugwort.

The decomposition of Eq. (2) leads to pollen typologies. Fig. 5 shows the first four eigen vectors  $\tilde{Q}_k$ . The first one mostly corresponds to days where trees are emitting pollens. The third one gathers grasses. While in the fourth one, one retrieves mainly herbs. This distribution is, mostly, explained by the seasonality of the different pollens.

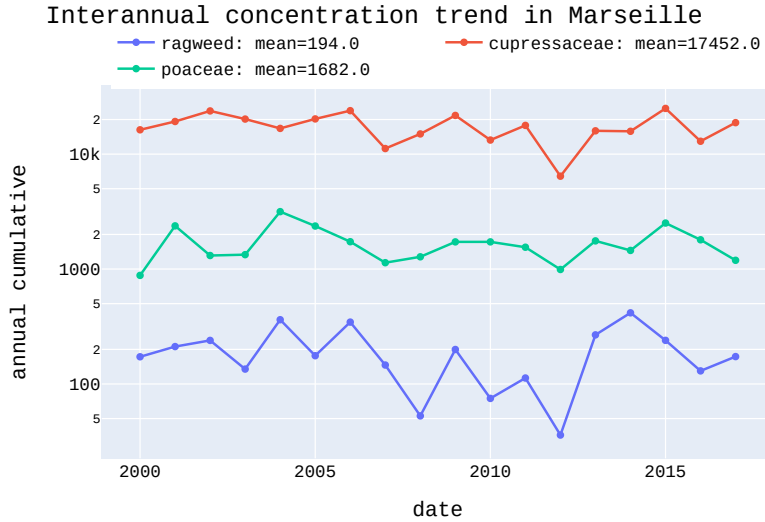
<sup>2</sup><https://www.pollens.fr/reports/database>



**Figure 2:** Example of a cupressaceae historical concentration data from 2000 to 2017 (top left) and in 2010 (top right) for Marseille site, with "null", "low", "medium" and "high" risk levels, respectively materialized by the blue, green, orange and red bands. Histogram of the same data history (in abscissa log concentration bands and in ordinate percent) is given (bottom left) and with a zoom on first bands (bottom right); the red dotted vertical lines materialize the "low", "medium" and "high" risk thresholds (from left to right)

Pollens and sites were classified by a hierarchical cluster analysis on the data previously decomposed into non-negative matrices. Manhattan distance and ward linkage have been used. Fig. 6 shows the different groups of pollen species and sites. The classification typically groups sites located in the same region, showing the importance of the geographical (spatial) dimension in pollen emission.

We focus our study on three species: ragweed (or ambrosia) , poaceae (grasses) and cupressaceae. The choice of these species is motivated by our clustering analysis 2 which shows that these three species belong to different groups in addition to our expertise concerning pollen. First of all, cupressaceae belong to the large family of "trees"; while poaceae are in the family of "grasses", ambrosia is in the family of "herbaceous plants" and then these species have very different allergenic powers. This is why the risk thresholds are different for each specie, as detailed in the table 1 which illustrates their differences in pollen emission quantities. Cupressaceae is one of the species that emit the highest quantities of pollen among all available species. Ragweed always emits very few pollen grains but is very allergenic. And poaceae is in between. This is illustrated on Fig. 3. Annual total concentrations are plotted for the three species from South-East area, especially in Marseille. The trend observed for this site is very common to what can be observed in other locations in general. In addition, these pollens have different seasonalities, see for instance the occurrence frequencies given on Fig. 7. The monthly frequency of pollen occurrence is calculated as the ratio of the number of days where the concentration is higher than a given threshold.



**Figure 3:** Inter-annual variability of pollen concentration (represented in log scale on the ordinate) from 2000 to 2017 (on the abscissa).

The goal here is to estimate for each month the probability of having a concentration for a given pollen above the threshold, based on all available historical data. The value of the threshold, if it is not chosen by an expert, is determined from the quantile  $q$  of the whole of the sample of concentrations of the days of the month concerned. Cupressaceae have a high probability of occurrence during the winter period. In Mars/April, when the amount of cupressaceae pollens starts to decrease, the poaceae occurrence probability becomes higher. The presence of the latter is decreasing from mid-July when the presence of ragweed becomes stronger until early September.

pollen	low	medium	high
ragweed	2	3	11
cupressaceae	70	142	284
poaceae	2	5	36

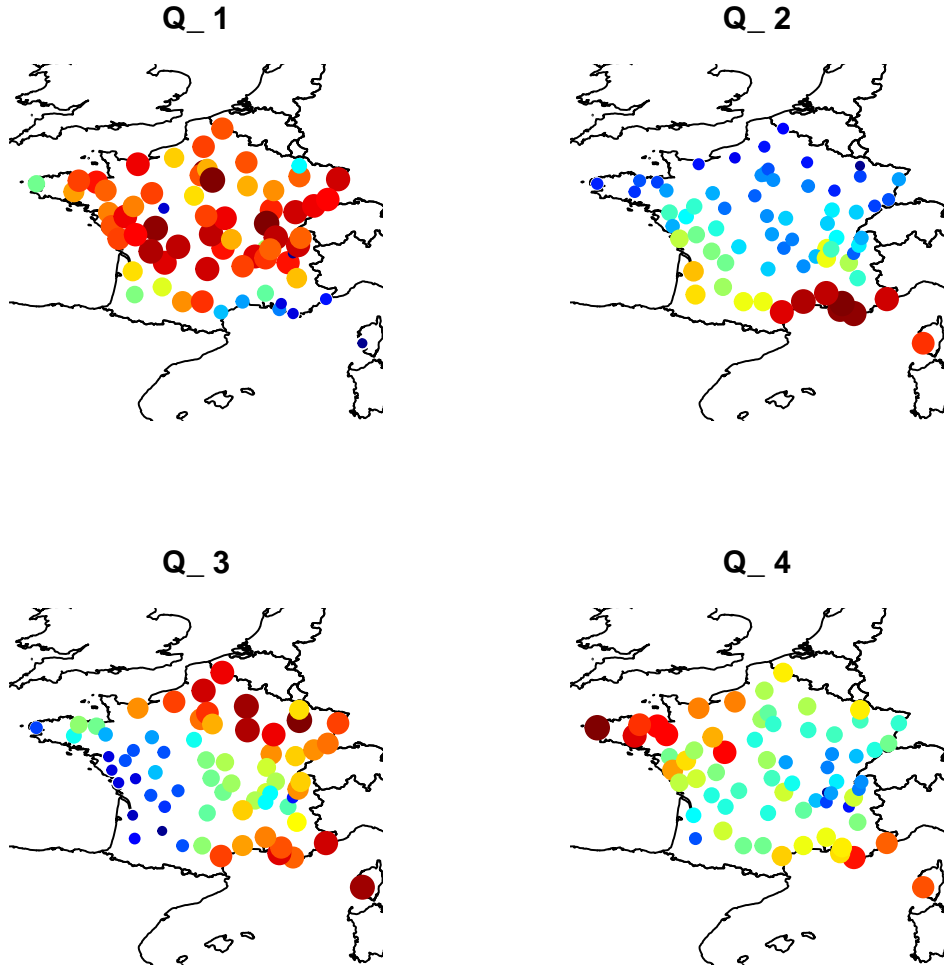
**Table 1**

Exposure risk thresholds for ragweed, cupressaceae, and poaceae, in grains per  $m^3$ /day. Thresholds are defined in (Thibaudon, 2003). They are based on pollen data, clinical and meteorological information as explained in (Thibaudon, 2003).

So far, we have presented in a general way what characterizes the nature of our data from simple observations made on few pollen species and a single site.

## 2.2. Meteorological data

The meteorological data have been downloaded from the European Climate Assessment Dataset project (<https://www.ecad.eu/>). These data are recorded by ground sensors located near airports (68 sites). They are referred to as METeorological Aerodrome Reports (METAR) data, dedicated to aviation and validated by the World Meteorological Organization. The historical data produced by ECAD were homogenized and supplemented to smooth inter-sensor variations and impute some missing data. Therefore, the weather observations are reliable and homogeneous. Some of the HIRST sensors are far from the weather station so that the very local characteristics of the meteorology may not be always totally relevant. Nevertheless, we decided to work with these observational data because they are easily accessible in real time. Moreover, on a daily scale, the fields of meteorological variables such as temperature are quite smooth and the variability over a few tens of kilometers is low. In the sequel only a part of the meteorological variables



**Figure 4:** Four first eigen vectors (Eq. 1) highlight typical situations in space. The bigger and redder the dots, the higher the total concentration of pollen in the air.

provided by ECA are used. The selected weather variables are temperature at 2 meters, minimum and maximum temperature, cumulative precipitation and humidity ; all of them are given at a daily scale. Although solar radiation is known to be a key parameter for flowering and pollen emission it is not used because it is not available for all weather stations.

Note that alternatives would be to use satellite data or reanalysis data, but the access would be either too expensive or too long.

Finally, the meteorological variables used as input in the prediction models are :

- the temperature of the day  $t$  and its differences between two successive days for days  $t$  and  $t - 1$  ( $\Delta T_t = T_t - T_{t-1}$  and  $\Delta T_{t-1} = T_{t-1} - T_{t-2}$ );
- the differences between the maximum and minimum temperatures on days  $t$ ,  $t - 1$  and  $t - 2$  ;
- degrees day  $D_t = \sum_{j=t-30}^t \frac{T_{max_j} - T_{min_j}}{2} \mathbf{1}_{T_{min_j} > 10}$  where  $\mathbf{1}$  is the indicator function and  $T_{min}$  and  $T_{max}$  are the daily min and max of temperature;
- the humidity of day  $t$  and its differences between two successive days for days  $t$  and  $t - 1$  ;
- the precipitation of day  $t$  .



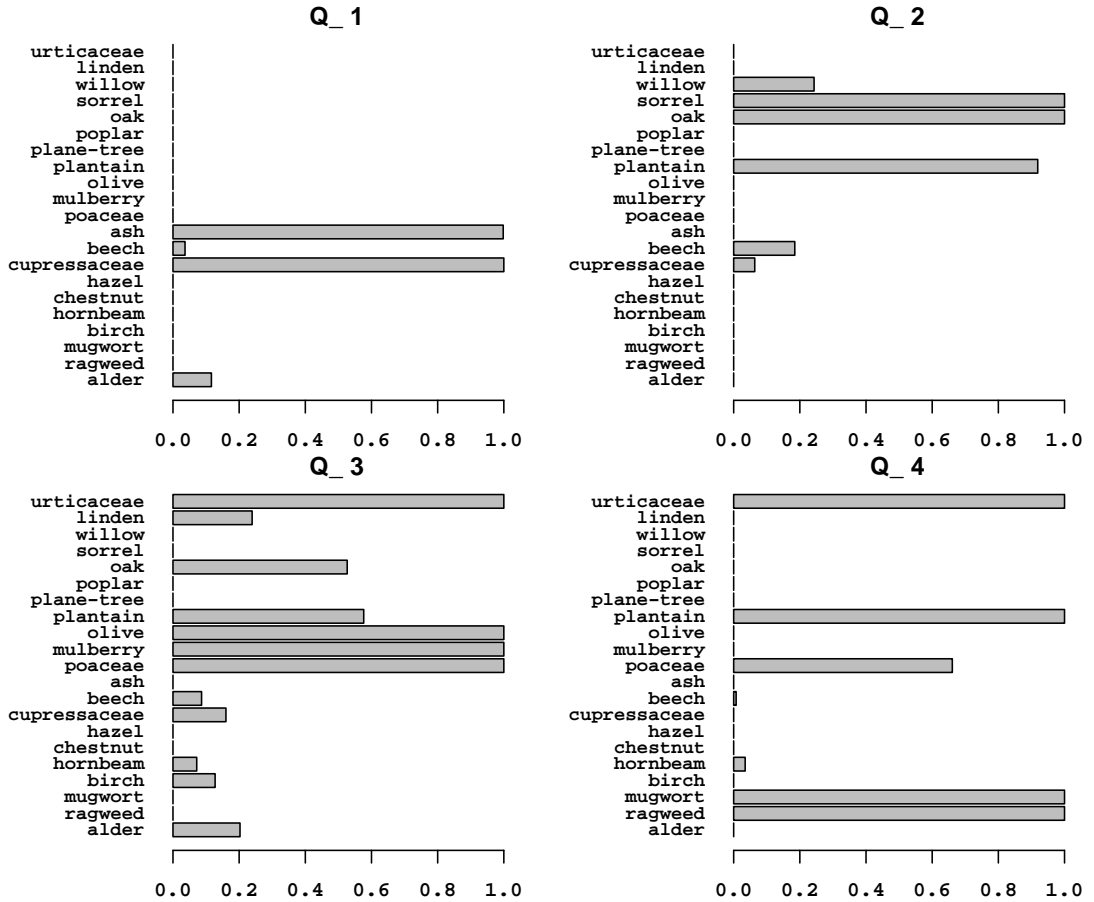


Figure 5: Four first eigen vectors (Eq. 2) highlight pollen typologies. The bar represent an amount of pollen with 0 corresponding the the mean.

Furthermore, the annual seasonality is taken into account by introducing a qualitative variable which is the number of the week and a quantitative variable  $\cos(\frac{2\pi t}{365})$ . And the spatial information is introduced via the latitude and longitude of the HIRST sensors as well as a qualitative region variable that takes the values Northeast, Northwest, Southeast and Southwest.

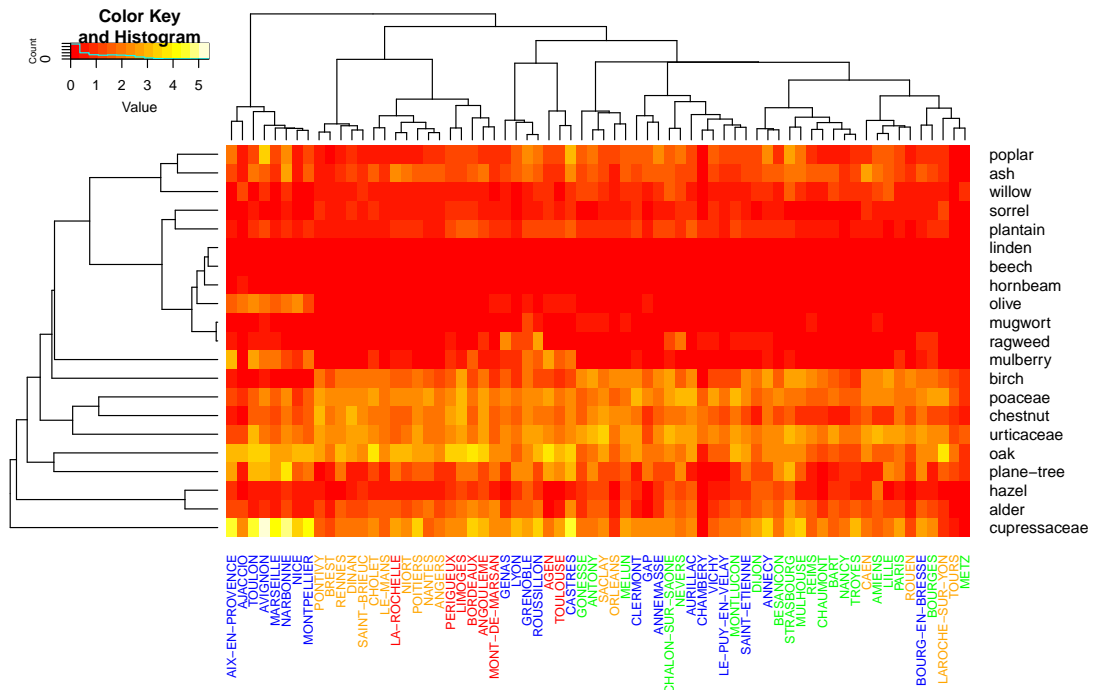
### 3. Methodology

#### 3.1. Related Works

In the field of aerobiology, as already mentioned in the introduction, machine learning algorithms have been widely proposed since the nineties. Different studies have been interested in developing models using jointly meteorological, phenological, environmental and also historical concentration data for objectives such as forecasting the presence or absence of a pollen species in a given location, estimating the onset of the pollen season of a species (Andersen, 1991; Cassagne, 2009) or predicting the inter-annual variation of pollen seasons (Spieksma, Emberlin, Hjelmroos, Jäger and Leuschner, 1995), or predict the level of pollen risk or concentration (Cordero, Rojo, Gutiérrez-Bustillo, Narros and Borge, 2021; Castellano-Méndez, Aira, Iglesias, Jato and González-Manteiga, 2005; Sánchez-Mesa, Galán, Martínez-Heras and Hervás-Martínez, 2002; Hidalgo, Mangin, Galán, Hembise, Vázquez and Sanchez, 2002; Ranzi, Lauriola, Marletto and Zinoni, 2003; Iglesias-Otero et al., 2015; Muzalyova, Brunner, Traidl-Hoffmann and Damialis, 2021)

Among others, support vector machines (SVM) have been proposed in (Zewdie, Liu, Wu, Lary and Levetin, 2019b), random forests (RF) in (Zewdie et al., 2019b; Nowosad et al., 2018), artificial neural networks (ANN) in (Cordero et al., 2021; Puc, 2012; Iglesias-Otero et al., 2015; Valencia, Astray, Fernández-González, Aira and Rodríguez-Rajo,

## Pollen risk levels prediction in France

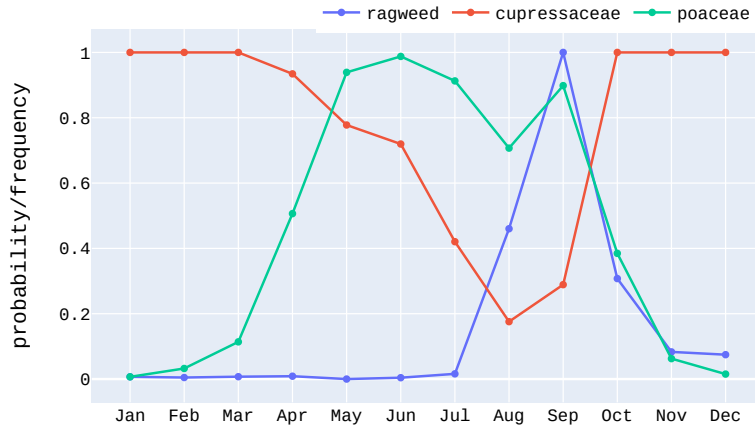


**Figure 6:** A typological analysis in space and pollen by hierarchical classification (Manhattan distance and Ward linkage after decomposition into non-negative matrix of the data) allows to establish groupings of sites and pollens. The colors of the sites are defined *a priori* according to their geographical positions. "Orange" sites are from north west, "red" sites from south west, "green" sites from north east and "blue" sites from south east.

2019a), regression models in (Box, Jenkins, Reinsel and Ljung, 2015) or the gradient boosting models (Cordero et al., 2021). These frequently used algorithms have led to satisfactory results which are detailed below. In (Cordero et al., 2021), the authors are interested in the prediction of the daily concentration of the olive tree in Madrid (Spain). They implemented models based on Light Gradient Boosting Machine (LightGBM) combined with a Generalized Additive Model (GAM) and Artificial Neural Networks (ANN) to predict the day of the year when the peak of the pollen season occurs. The average peak date is  $149.1 \pm 9.3$  and  $150.1 \pm 10.8$  days for LightGBM and ANN, respectively, which is close to the observed value ( $148.8 \pm 9.8$ ). The daily concentration is predicted with a coefficient of determination of 0.75. The authors used meteorological variables, phenological parameters, site characteristic data and historical pollen concentrations. Other studies show that the use of neural networks allows to reach satisfactory results to predict a risk of exposure to Birch. In (Castellano-Méndez et al., 2005) the authors reduced the problem to a binary classification for each level (4 concentration thresholds) using the concentration of Birch and the meteorological data history from 1993 to 2001 observed at Santiago de Compostela (Spain). Their main objective was to predict the days with high allergic risk during Birch pollination. They used the average concentration ( $\text{grains}/\text{m}^3$ ), precipitation and average daily temperature. For the construction of the artificial neural network model, they used as independent variables precipitation, average temperature and pollen concentration of the previous day to predict the risk level of the current day. By training the models on the data from 1993 to 1999 and testing on the years 2000 and 2001, the performances are: (probability of a good classification (accuracy) of a high pollen day is 83% to 100% and the probability of a good classification (accuracy) of a low pollen day is 92% to 97%).

In the study (Muzalyova et al., 2021) conducted in Augsburg (Germany), the aim was to develop a 3-hour predictive model to forecast the concentration of Birch and Poaceae pollens based on automatic pollen measurements in near-real time. Dynamic ARIMA regression models were used, as well as machine learning techniques, namely neural network auto-regression models on a history of pollen data collected at 3-hour intervals from 2016 to 2019. Air temperature, relative humidity, precipitation, atmospheric pressure, sunshine duration, diffuse radiation and wind speed were the parameters used to build the models. The authors showed that temperature and precipitation were the most significant variables. The prediction performance of the Birch model was higher with an  $R^2$  of 0.62, compared to 0.55 for the

Estimated frequency of occurrence by month in Marseille



**Figure 7:** Probability of pollen occurrence estimated from concentration history ranging from 2000 to 2017

Poaceae model. Neural autoregression led to the strongest results in predicting Birch pollen concentrations, while for Poaceae, ARIMA performed best. The authors also showed that extreme weather events are still an obstacle to good near-real-time forecasts, despite the advanced techniques available in the machine learning field. In the paper (Zewdie et al., 2019b), RF (composed with 200 decision trees), ANN (MLP with sigmoid activation function) and SVM (with Gaussian kernel function) are tested to predict daily concentrations of ragweed. The authors used in their studies, a 20-year history (1994-2014) of environmental and soil data from the NEXt generation RADar<sup>3</sup>

Table 4 in the Appendix provides a summary of several state-of-the-art studies to which we can compare our study. Most used algorithms in the literature are ANNs, linear regressions or RF and SVM.

The authors generally use additional meteorological variables (such as radiation and sunshine), in addition to environmental variables with the particularity of generally targeting a single site, unlike our study which only uses temperature, humidity and precipitation on 68 distinct sites covering the French metropolitan territory.

Our study being positioned in an industrial context, several choices and motivations are guided by the objectives of the operating company which targets the most parsimonious and explanatory models as possible, and furthermore using data of free and easy access such as the meteorological data. We thus decompose the risk prediction problem into a set of binary problems, each targeting a given level of discretized risk, associated with a final problem of aggregation of these binary decision models.

### 3.2. Prediction model by combining binary logistic regressions

The distribution of the emissions is asymmetric with really heavy tails. It is not straightforward to find a transformation to correct these features. Furthermore, the targeted application has to deliver a simple message to the allergic patients. In this study, we propose an approach in which the pollen concentration is discretized into four level of risk (classes) as in RNSA reports. More precisely, the target variable  $Y_t^{(p)}$  represents a pollen-related allergic risk which is constructed as follows.

$$\begin{aligned}
 &\text{if } C_t^{(p)} < s_{\text{low}}, & Y_t^{(p)} &= \text{"null"} \\
 &\text{if } s_{\text{low}} \leq C_t^{(p)} < s_{\text{medium}}, & Y_t^{(p)} &= \text{"low"} \\
 &\text{if } s_{\text{medium}} \leq C_t^{(p)} < s_{\text{high}}, & Y_t^{(p)} &= \text{"medium"} \\
 &\text{if } C_t^{(p)} \geq s_{\text{high}}, & Y_t^{(p)} &= \text{"high"}
 \end{aligned} \tag{3}$$

<sup>3</sup><https://www.ncei.noaa.gov/products/radar/next-generation-weather-radar>

where  $C_t^{(p)}$  is the concentration of pollen  $p$  at time  $t$ . The thresholds  $s_{low}$ ,  $s_{medium}$ ,  $s_{high}$  are defined in (Thibaudon, 2003) and depend on the pollen species (see Table 1). Thus,  $Y_t^{(p)}$  is an ordinal variable. The null class which is over represented compared to the other classes because of the seasonality of pollen occurrence.

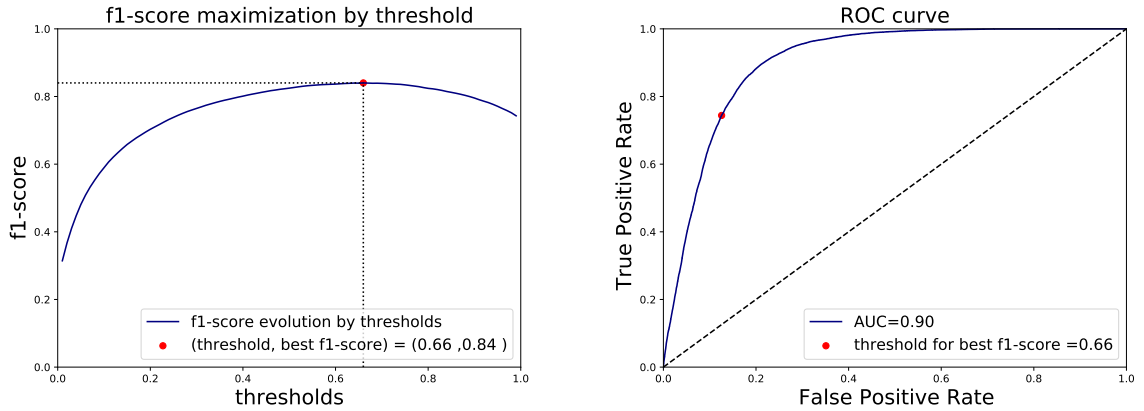
From this definition of the ordinal variable  $Y_t^{(p)}$  to be predicted, different models can be considered. The proposed algorithm is based on simple or auto-regressive binomial models for which the parameters are estimated by maximizing the likelihood. These models were considered because they have good generalization properties and their parameters are easy to estimate and to interpret.

Binomial models are used for two classes decision problems. Here, we propose an aggregation algorithm which consists in combining the predictions of three binary logistic regressions for the "low", "medium" and "high" risk level thresholds according to two approaches. In the first one, we consider the following ranking rule.

If  $\hat{R}_t(\text{medium}) == \text{True}$   
 if  $\hat{R}_t(\text{high}) == \text{True}$  then  $\widehat{Y}_t^{(p)} = \text{"high"}$   
 otherwise  $\widehat{Y}_t^{(p)} = \text{"medium"}$   
 otherwise  
 if  $\hat{R}_t(\text{low}) == \text{True}$  then  $\widehat{Y}_t^{(p)} = \text{"low"}$  otherwise  $\widehat{Y}_t^{(p)} = \text{"null"}$ ,

with  $\hat{R}_t$  the risks predicted by the binomial models and  $\widehat{Y}_t^{(p)}$  the prediction of the aggregation algorithm.

In practice, the four risk classes are unbalanced and the predictions are made by comparing the predicted probability to a threshold. For each binary logistic regression model the threshold is determined in order to maximize the F1-score on the training data set. As a reminder we have  $F1\text{-score} = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ . On Figure 8 (left panel), an example of this optimization of the F1-score using 100 values of the threshold gives the best F1-score = 0.84 for the probability threshold equal to 0.65. The associated ROC (see right panel of Fig. 8) shows an  $AUC = 0.90$ .



**Figure 8:** Example of best threshold determination using F1-score maximization on training data

In a second version of the aggregation algorithm, the probabilities  $R_t^{R(r)}$  of the binomial models associated with the risk  $R_t^r$ ,  $r \in \text{low, medium, high}$  are used as input features to a Random Forest (RF) algorithm to obtain a final ranking with four risk modalities: "null", "low", "medium" or "high".

One way to improve our prediction models is to build auto-regressive models by feeding back the forecasts at  $t + 1$  (current day+1) into the  $t + 2$  (current day+2) model and  $t + 1$ ,  $t + 2$  into the  $t + 3$  (current day+3) model. In this implementation we consider the real risks (determined using the observed pollen concentrations) at  $t + 1$  and  $t + 2$  during the training phase, whereas for the validation task, predicted risks at  $t + 1$  and  $t + 2$  are used.

	Ragweed			Cupressaceae			Poaceae		
	acc	pre	rec	acc	pre	rec	acc	pre	rec
Historical	<b>0.97</b>	0.93	<b>0.97</b>	0.83	0.74	0.83	0.76	0.70	0.76
DT	0.87	<b>0.97</b>	0.87	0.55	0.68	0.55	0.63	0.78	0.63
RF	0.92	<b>0.97</b>	0.92	0.64	0.76	0.64	0.70	0.78	0.70
MNL	0.82	0.96	0.82	0.63	0.71	0.63	0.70	0.75	0.70
ORDINAL	0.84	<b>0.97</b>	0.84	0.60	0.76	0.60	0.70	<b>0.80</b>	0.70
MM-AH	0.88	0.95	0.88	0.57	0.79	0.57	0.62	0.70	0.62
MM-RF	0.91	<b>0.97</b>	0.91	0.64	0.77	0.64	0.69	0.79	0.69
MM-AR-AH	0.96	0.96	0.90	<b>0.86</b>	<b>0.85</b>	<b>0.86</b>	<b>0.78</b>	0.78	<b>0.78</b>
MM-AR-RF	0.89	<b>0.97</b>	0.89	0.74	0.78	0.74	0.61	0.70	0.61

**Table 2**

Scores of the validation set. The accuracy (acc), precision (pre) and recall (rec) are computed for the following algorithms : decision tree (DT), random forest (RF), multinomial model (MNL), ordinal model (OL), *ad hoc* combination of binary logistic models (MM-AH), random forest combination of binary logistic models (MM-RF), *ad hoc* combination of binary logistic models with auto-regressive part (MM-AR-AH) and random forest combination of binary logistic models with auto-regressive part (MM-AR-RF). The scores of the most frequent historical risk are reported in the line Historical. All the reported validation results are mean computed over the 62 stations included in the learning set. The best performances are highlighted using a bold police.

#### 4. Experiments and Results

In this section, the prediction performances of the models proposed in Sec. 3 are studied and compared to those of other methods of the literature. The baseline algorithms considered for comparison are decision trees (DT), Random Forest (RF), multinomial classification (MNL), ordinal classification (OL). They have been selected for the following reasons. RF is used to combine the binary logistic regressions (see Section 3). So it is fair to look at its performances in a more direct approach. The DT is a particular case of Random Forest and is easy to interpret. MNL and OL are component variants involved in the proposed models. To assess the contribution of meteorological variables, the predictions of the different models are also compared to the most likely daily risk estimated from the pollen emission history (Historical).

The models are learned on 62 stations over 68. They are then used to predict the risk at horizon  $T+3$  given the weather data until time  $T$  for all the 68 stations. The 62 stations with the longest history are selected for learning. Then, 80% of the data of these 62 stations are used for the learning task and the 20% remaining data for the validation. The validation scores are computed for the 20% most recent data as well as the 6 remaining stations.

The prediction algorithms are compared thanks to the following scores: accuracy, precision and recall. All of them are computed by weighted average with weights proportional to the observed frequency of each class. Results are reported in Table 2 for the 62 stations of the learning set. The performances are good in general. The ragweed is better predicted with scores 0.82 for the worse model to 0.97 for the better. Ragweed only generate a few emissions such that the associated risk is often null. The weighted scores tends to favour the null emission class. For ragweed, one also sees that the historical prediction is one of the best prediction. It means that all the models mainly predict the periodicity of the ragweed emission. For the other pollens, cupressaceae and poaceae, the combination models have performances that are slightly better than the one of the classical models (second horizontal band of Table 2). The combination models with an auto regressive component generally lead to the best scores. The best combination is the *ad hoc* one (MM-AR-AH). The associated error is about 15% for the cupressaceae and 20% for the poaceae. Recall and precision are of the same order which means that the two type of errors (false positive and false negative) are balanced.

In Table 3, the performances obtained on the validation stations (data from the sites that were not included into the training set) are compared for all the models. The performances are of the same order as the one obtained for the stations included in the learning set. These conclusion is particularly important since it means that the models can be deployed on localities with no pollen data history.

	Ragweed			Cupressaceae			Poaceae		
	acc	pre	rec	acc	pre	rec	acc	pre	rec
Historical	-	-	-	-	-	-	-	-	-
DT	0.87	0.97	0.87	0.55	0.68	0.55	0.63	0.78	0.63
RF	0.92	0.97	0.92	0.64	0.76	0.64	0.70	0.78	0.70
MNL	0.82	0.96	0.82	0.63	0.71	0.63	0.70	0.75	0.70
ORDINAL	0.84	0.97	0.84	0.60	0.76	0.60	0.70	0.80	0.70
MM-AH	0.81	0.98	0.81	0.62	0.86	0.62	0.68	0.76	0.68
MM-RF	0.91	0.97	0.91	0.64	0.77	0.64	0.69	0.79	0.69
MM-AR-AH	0.97	0.98	0.97	<b>0.89</b>	<b>0.88</b>	<b>0.89</b>	<b>0.84</b>	<b>0.83</b>	<b>0.84</b>
MM-AR-RF	0.91	<b>0.99</b>	0.91	0.77	0.80	0.77	0.66	0.75	0.66

**Table 3**

Scores of the validation set. The accuracy (acc), precision (pre) and recall (rec) are computed for the following algorithms : decision tree (DT), random forest (RF), multinomial model (MNL), ordinal model (OL), *ad hoc* combination of binary logistic models (MM-AH), random forest combination of binary logistic models (MM-RF), ad hoc combination of binary logistic models with auto-regressive part (MM-AR-AH) and random forest combination of binary logistic models with auto-regressive part (MM-AR-RF). The scores of the most frequent historical risk are reported in the line Historical. All the reported validation results are mean computed over the 6 stations which were not included into the learning set.

Now, it is interesting to check if the performances of the risk prediction algorithms have any spatial structure. On Fig. 9, the area under the Receiver Operating Characteristic (AUC) are plotted for each station of the data set. The AUC are computed on the validation set for cupressaceae. The color scale depend on the value of the AUC whose larger dots are spatial generalization sites. The navy small dots represent the weather station. Similar figures are also available for other pollens (see in Appendix, Fig. 11 and 12 for ragweed and poaceae respectively). For the cupressaceae (Figure 9), the AUC are greater than 0.7 in majority cases except for the sites of Mediterranean littoral. The weather, especially the temperature is particularly hot and the early start of the season in this area is less well captured by the model. There is an evident spatial structure because this pollen is well distributed over the country. At the opposite for ragweed pollen, because of the quasi absence of this species on the northern part of the country. Generally the concentration rarely exceeds the medium risk threshold. For poaceae, the AUC has an almost homogeneous distribution on the whole of France. Those are one of the most popular species in geographical cover and corresponds to the distribution of these plants on the territory. About the station that were not included in the learning set (correspond to the largest dots), the AUC are similar to the one of the neighbouring stations except for Le-Puy-En-Velay (city surrounded by natural parks, close to Saint-Etienne and Valence with a specific climate) and Saclay (close to Paris) where this result is difficult to explain. It may be due to a specificity of the sensor which is located on the roof of the observatory at 15 meters above the ground without any neighbouring vegetation Sarda Estève, Baisnée, Guinot, Petit, Sodeau, O'connor, Besancenot, Thibaudon and Gros (2018).

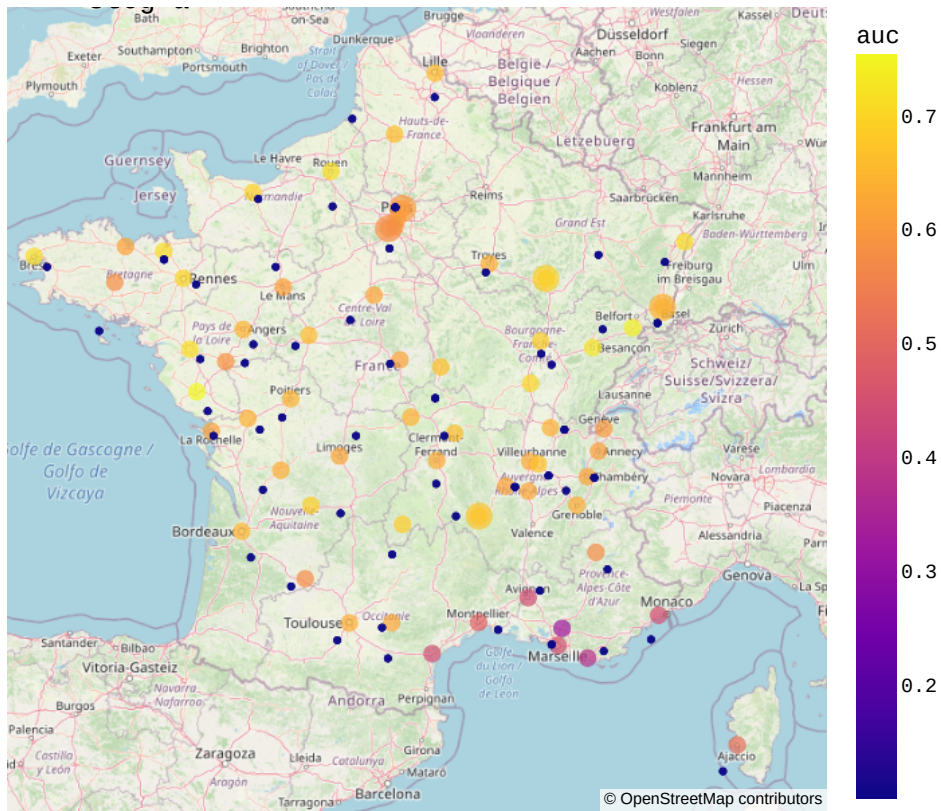
Finally, Fig. 10 gives an example of predicted risk time series. The plot shows the cupressaceae concentration in Marseille for the years of the validation data set. The background vertical green lines correspond to observed null risk days and the red ones to high risk days. The blue line exhibits the predicted risk with 4 levels from the lowest (null risk) to the highest (high risk). The blue line shows that the model is efficient to detect the beginning the pollen season. It is of utmost importance because it allows to alert allergic patients at least three days in advance. Remind that the prediction is for horizon  $T + 3$ . However, the risk is often over-estimated. Similar observation was made for other pollens (not shown).

The performance evaluation criteria used to compare the models are the area under the ROC curve but also the precision, the recall and the f1-score. The approach that consists in combining *ad hoc* (MM-AH) and random forest (MM-RF) binary regressions take as input the probability scores of the binary regressions related to the risk thresholds in Table 1.

## 5. Conclusion and perspectives

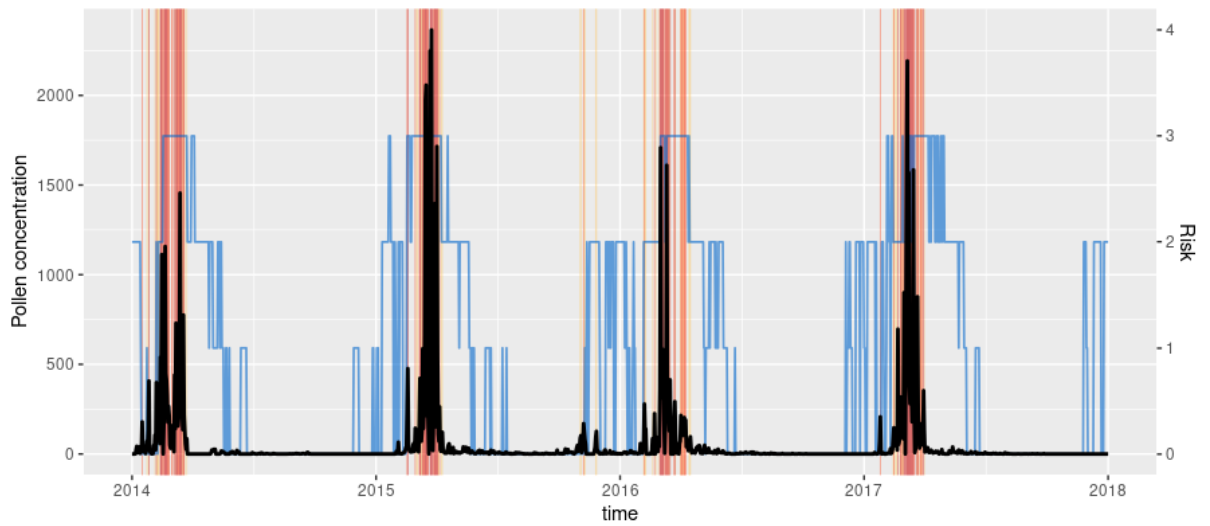
We have presented a feasibility study on three-day ahead prediction of pollen emission risk levels from historical aerobiological, meteorological, and geographic data. The originality of the approach lies in the joint consideration

## Pollen risk levels prediction in France



**Figure 9:** AUC map of cupressaceae computed with the MM-AR-AH for the prediction of risk medium or high versus low or null. The large dots highlight the HIRST locations that are not included in the learning set. The blue dots materialize the position of the weather stations.

of spatial and temporal information. Our study shows that it is generally possible to predict pollen episodes with acceptable accuracy by considering discrete risks including the presence and absence of pollen emissions. However, we observed that the risk levels tends to be overestimated. According to our study, we finally conclude that models that take into account weather conditions do only marginally better than predictive models based only on history. Models integrating meteorological data are especially interesting for sites for which no history is available. In the future, we expect to combine some of these models based on meteorological data with observation of efficient sensors to forecast the pollen concentration or emission risk levels more precisely.



**Figure 10:** Cupressaceae concentration in Marseille (black line) with risks predicted by MM-AR-AH model (blue line). The background red vertical lines materialize the observed risks (the darker the higher). Years 2014 to 2017 are in the validation data set.



## References

- Andersen, T.B., 1991. A model to predict the beginning of the pollen season. *Grana* 30, 269–275.
- Bettayeb, K., Cayrol, C., Girard, J.P., 2018. Allergies: towards new therapeutic options. *Journal du CNRS, CNRS News on-line, on-line*. URL: <https://news.cnrs.fr/articles/allergies-towards-new-therapeutic-options>.
- Bohlmann, S., Shang, X., Giannakaki, E., Filioglou, M., Saarto, A., Romakkaniemi, S., Komppala, M., 2019. Detection and characterization of birch pollen in the atmosphere using a multiwavelength raman polarization lidar and hirst-type pollen sampler in finland. *Atmospheric Chemistry and Physics* 19, 14559–14569. URL: <https://acp.copernicus.org/articles/19/14559/2019/>, doi:10.5194/acp-19-14559-2019.
- Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.
- Buters, J., Antunes, C., Galveias, A.e.a., 2018. Pollen and spore monitoring in the world. *Clinical and Translational Allergy* 8. URL: <https://doi.org/10.1186/s13601-018-0197-8>, doi:10.1186/s13601-018-0197-8.
- Cassagne, E., 2009. *Revue bibliographique des principaux seuils de détermination et méthodes de prévision de la date de début de pollinisation (ddp)*. *Revue Française d'Allergologie* 49, 571–576.
- Castellano-Méndez, M., Aira, M., Iglesias, I., Jato, V., González-Manteiga, W., 2005. Artificial neural networks as a useful tool to predict the risk level of betula pollen in the air. *International Journal of Biometeorology* 49, 310–316.
- Cordero, J.M., Rojo, J., Gutiérrez-Bustillo, A.M., Narros, A., Borge, R., 2021. Predicting the olea pollen concentration with a machine learning algorithm ensemble. *International Journal of Biometeorology* 65, 541–554.
- Hidalgo, P.J., Mangin, A., Galán, C., Hembise, O., Vázquez, L.M., Sanchez, O., 2002. An automated system for surveying and forecasting olea pollen dispersion. *Aerobiologia* 18, 23–31.
- Hirst, J.M., 1952. An automatic volumetric spore trap. *Annals of Applied Biology* 39, 257–265. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-7348.1952.tb00904.x>, doi:<https://doi.org/10.1111/j.1744-7348.1952.tb00904.x>, arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-7348.1952.tb00904.x>.
- Howard, L.E., Levetin, E., 2014. Ambrosia pollen in tula, oklahoma: aerobiology, trends, and forecasting model development. *Annals of Allergy, Asthma & Immunology* 113, 641–646.
- Iglesias-Otero, M., Fernández-González, M., Rodríguez-Caride, D., Astray, G., Mejuto, J., Rodríguez-Rajo, F., 2015. A model to forecast the risk periods of plantago pollen allergy by using the ann methodology. *Aerobiologia* 31, 201–211.
- Makra, L., Matyasovszky, I., Thibaudon, M., Bonini, M., 2011. Forecasting ragweed pollen characteristics with nonparametric regression methods over the most polluted areas in europe. *International journal of biometeorology* 55, 361–371.
- Muzalyova, A., Brunner, J.O., Traidl-Hoffmann, C., Damialis, A., 2021. Forecasting betula and poaceae airborne pollen concentrations on a 3-hourly resolution in augsburg, germany: toward automatically generated, real-time predictions. *Aerobiologia*, 1–22.
- Myszkowska, D., Majewska, R., 2014. Pollen grains as allergenic environmental factors: new approach to the forecasting of the pollen concentration during the season. *Annals of Agricultural and Environmental Medicine* 21.
- Nowosad, J., Stach, A., Kasprzyk, I., Chłopek, K., Dąbrowska-Zapart, K., Grewling, Ł., Latałowa, M., Pędziszewska, A., Majkowska-Wojciechowska, B., Myszkowska, D., et al., 2018. Statistical techniques for modeling of corylus, alnus, and betula pollen concentration in the air. *Aerobiologia* 34, 301–313.
- Piotrowska, K., 2012. Forecasting the poaceae pollen season in eastern poland. *Grana* 51, 263–269.
- Puc, M., 2012. Artificial neural network model of the relationship between betula pollen and meteorological factors in szczecin (poland). *International journal of biometeorology* 56, 395–401.
- Ranzi, A., Lauriola, P., Marletto, V., Zinoni, F., 2003. Forecasting airborne pollen concentrations: Development of local models. *Aerobiologia* 19, 39–45.
- Rodríguez-Rajo, F., Astray, G., Ferreira-Lage, J., Aira, M., Jato-Rodríguez, M., Mejuto, J., 2010. Evaluation of atmospheric poaceae pollen concentration using a neural network applied to a coastal atlantic climate region. *Neural Networks* 23, 419–425. URL: <https://www.sciencedirect.com/science/article/pii/S0893608009001087>, doi:<https://doi.org/10.1016/j.neunet.2009.06.006>.
- Rodríguez-Rajo, F.J., Valencia-Barrera, R.M., Vega-Maray, A.M., Suárez, F.J., Fernández-González, D., Jato, V., 2006. Prediction of airborne alnus pollen concentration by using arima models. *Annals of Agricultural and Environmental Medicine* 13, 25.
- Sánchez-Mesa, J., Galán, C., Martínez-Heras, J., Hervás-Martínez, C., 2002. The use of a neural network to forecast daily grass pollen concentration in a mediterranean region: the southern part of the iberian peninsula. *Clinical & Experimental Allergy* 32, 1606–1612.
- Sarda Estève, R., Baisnée, D., Guinot, B., Petit, J.E., Sodeau, J., O'connor, D., Besancenot, J.P., Thibaudon, M., Gros, V., 2018. Temporal variability and geographical origins of airborne pollen grains concentrations from 2015 to 2018 at saclay, france. *Remote Sensing* 10, 1932.
- Ščevková, J., Dušička, J., Mičičeta, K., Somorčík, J., 2015. Diurnal variation in airborne pollen concentration of six allergenic tree taxa and its relationship with meteorological parameters. *Aerobiologia* 31, 457–468.
- Spieksma, F.T.M., Emberlin, J., Hjelmroos, M., Jäger, S., Leuschner, R., 1995. Atmospheric birch (betula) pollen in europe: Trends and fluctuations in annual quantities and the starting dates of the seasons. *Grana* 34, 51–57.
- Thibaudon, M., 2003. The pollen-associated allergic risk in france. *European annals of allergy and clinical immunology* 35, 170–172.
- Thibaudon, M., Oliver, G., Besancenot, J.P., 2019. Des capteurs pas comme les autres ! trente-cinq ans de recueil du pollen en france. *Revue Française d'Allergologie* 59, 576–583. URL: <https://www.sciencedirect.com/science/article/pii/S1877032019303860>, doi:<https://doi.org/10.1016/j.reval.2019.08.003>.
- Tseng, Y.T., Kawashima, S., Kobayashi, S., Takeuchi, S., Nakamura, K., 2020. Forecasting the seasonal pollen index by using a hidden markov model combining meteorological and biological factors. *Science of the Total Environment* 698, 134246.
- Valencia, J., Astray, G., Fernández-González, M., Aira, M., Rodríguez-Rajo, F., 2019a. Assessment of neural networks and time series analysis to forecast airborne parietaria pollen presence in the atlantic coastal regions. *International journal of biometeorology* 63, 735–745.
- Valencia, J., Astray, G., Fernández-González, M., al., 2019b. Assessment of neural networks and time series analysis to forecast airborne parietaria pollen presence in the atlantic coastal regions. *Int J Biometeorol* 63, 735–745.

- Zewdie, G.K., Lary, D.J., Levetin, E., Garuma, G.F., 2019a. Applying deep neural networks and ensemble machine learning methods to forecast airborne ambrosia pollen. *International Journal of Environmental Research and Public Health* 16. URL: <https://www.mdpi.com/1660-4601/16/11/1992>, doi:10.3390/ijerph16111992.
- Zewdie, G.K., Liu, X., Wu, D., Lary, D.J., Levetin, E., 2019b. Applying machine learning to forecast daily ambrosia pollen using environmental and nexrad parameters. *Environmental monitoring and assessment* 191, 1–11.

Table 4: summary of several related works in the literature

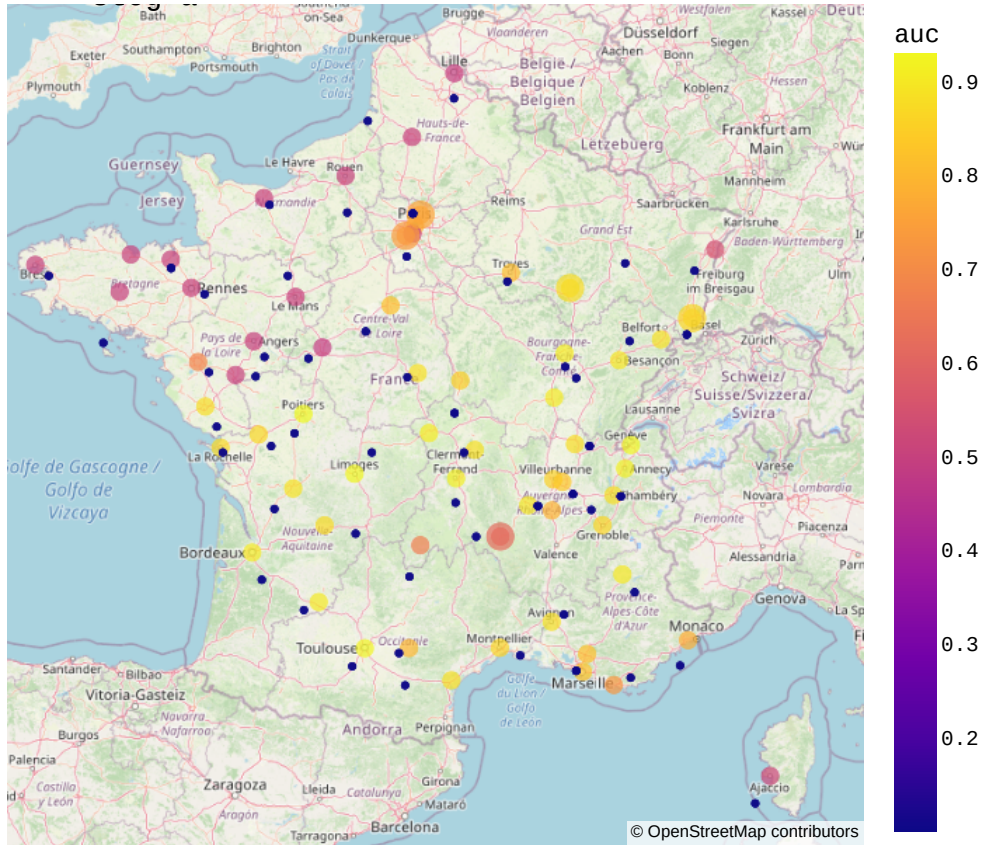
Articles	Objective	Types data	Geographical coverage	Pollen types	Models	Performance
(Andersen, 1991)	Predicting season start dates	Weather	Frederiksberg (Denmark)	Alder, ulmus, birch	CU and GDH	3-5 days prediction error
(Cassagne, 2009)	Review	Weather	Chalon sur Saône (France)	Cupressaceae, ash, birch, poaceae	Lejoly-Gabriel, GDD, Chilling Hours, Q10, Multiple regression	Linear regression is the most complete and better predictions
(Castellano-Méndez et al., 2005)	binary classification for prediction of high risk days	Weather, historical pollen count	Santiago de Compostela (Spain)	Birch	ANN	Probability of correct of correct classification between 83% and 100%
(Cordero et al., 2021)	Predicting daily concentration and season peak days	Weather, phenological, site features, historical pollen	Madrid (Spain)	Olivier	LightGBM, ANN	Estimated mean peak date(days): LightGBM = 149.1 $\pm 9.3$ ANN = 150.1 $\pm 10.8$ observed = 148.8 $\pm 9.8$ Concentration prediction: $R^2 = 0.75$
(Muzalyova et al., 2021)	Predicting pollen concentration over 3h	Weather, sunshine, duration, diffuse radiation and wind speed	Augsburg (Germany)	Birch, Poaceae	ARIMA and dynamics regression, ANN and autoregressive ANN	$R^2 = 0.62$ for birch, $R^2 = 0.55$ for Poaceae
(Sánchez-Mesa et al., 2002)	Predicting daily concentrations	Weather, historical pollen counts	southern part of the Iberian Peninsula (Spain)	Poaceae	Linear regression co-evolutionary ANN	90% of good classification

Continued on next page

Table 4: summary of several related works in the literature (Continued)

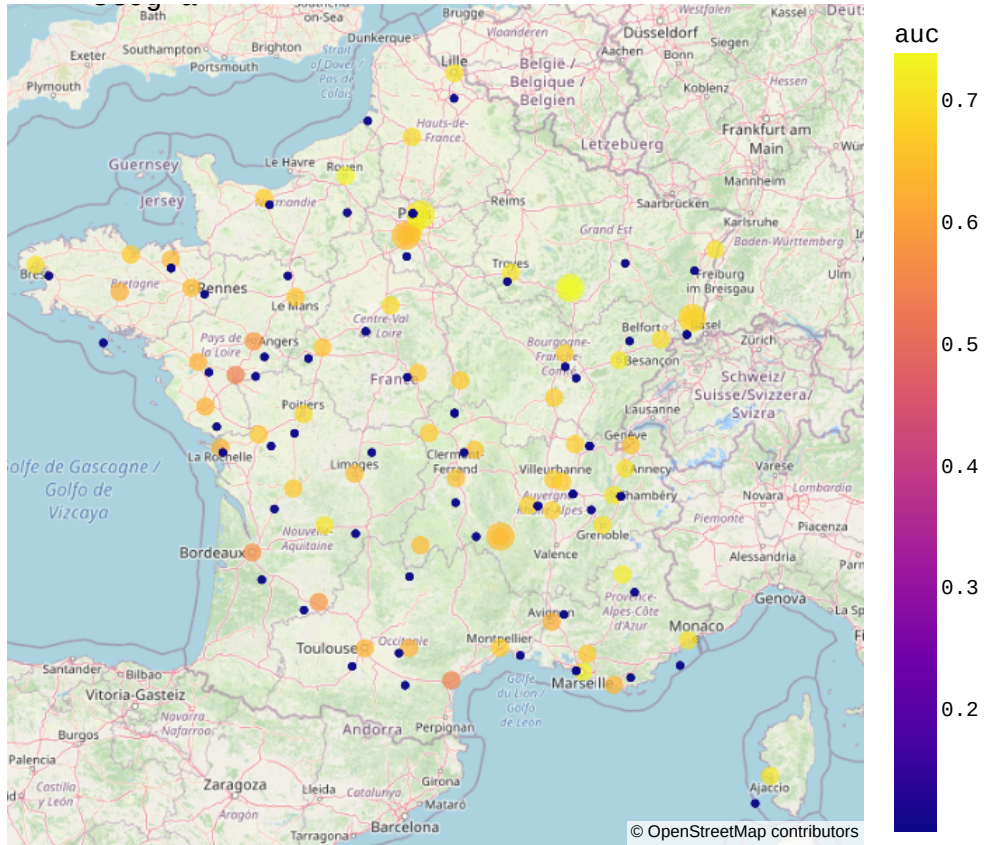
Articles	Objective	Types data	Geographical coverage	Pollen types	Models	Performance
(Iglesias-Otero et al., 2015)	Predicting concentration at d+1, d+2 and d+3	Weather	Ourense (Spain)	Plane tree	ANN	$R^2_{d+1} \in [0.449, 0.559]$ , $R^2_{d+2} \in [0.319, 0.37]$
(Hidalgo et al., 2002)	Season start date and pollen concentration	Weather, topographical, environmental (flowering)	Cordora (Spain)	Olive	Cumulative method,	Best results for ANN
(Ranzi et al., 2003)	Prediction of daily concentration, detection of pollen anomalies	Weather, phenological, auto-regressive data	Modena, Bologna (Italy)	Poaceae	ANN	Prediction of days of threshold exceedance: average error in delay or anticipation $\leq 2$ days
(Zewdie et al., 2019b)	Predicting daily concentration	Weather, land surface parameters, RADar (NEXRAD) measurements	Tulsa in Oklahoma state (USA)	Ragweed	RF, ANN, SVM	RF (R=0.61, $R^2=0.37$ ), SVM (R=0.51, $R^2=0.26$ ), ANN (R=0.46 $R^2=0.21$ )
(Valencia et al., 2019a)	Predicting concentration at 1,2,3 days in advance	Weather	Northwestern of Spain	Parietary	ANN	Prediction at 1 day ahead: $R^2 \in [0.618, 0.652]$

## Pollen risk levels prediction in France



**Figure 11:** AUC map of ragweed computed with the MM-AR-AH for the prediction of risk medium or high versus low or null. The large dots highlight the HIRST locations that are not included in the learning set. The blue dots materialize the position of the weather stations.

## Pollen risk levels prediction in France



**Figure 12:** AUC map of poaceae computed with the MM-AR-AH for the prediction of risk medium or high versus low or null. The large dots highlight the HIRST locations that are not included in the learning set. The blue dots materialize the position of the weather stations.

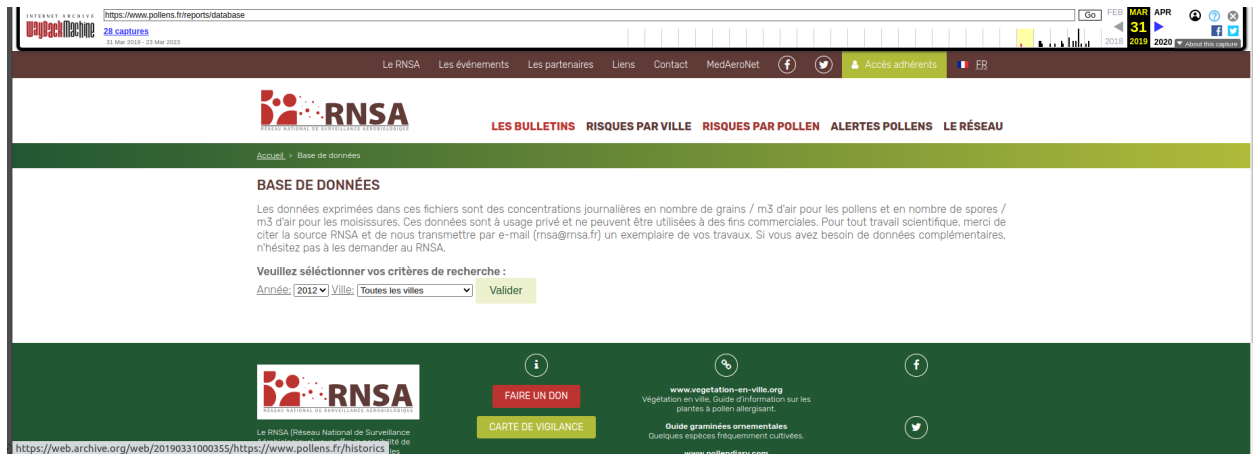


Figure 13: screen shot taken on the Wayback website archive

## 6. WARNING

This article has been accepted in *Elsvier Data & Knowledge Engineering Journal* Volume 142, November 2022, 102096. It has been since retracted due to a complaint by the owner of the dataset on which our study is based.

The substance of the complaint was the following.

1. The data set is not public and is the sole property of the National Aerobiological Survey Network (RNSA) and the French Atomic Energy Center (CEA) for the Saclay site.
2. The data set was obtained unlawfully and was not the property of the above listed authors to use in a publication. The listed authors did not formally, or informally, request the data from these organisations and the data was never communicated to the authors for their use.

Regarding the first point, we were fully aware of the proprietary nature of the data and do not contest it at all. We have dutifully mentioned the source of the data in the article, giving evidence that we know and accept the proprietary nature of the data.

Regarding the second point, we contest firmly the claim that we consider totally unfounded and libelous. At the time we get the data, in 2020 til October 2021 the web site of the RNSA institution offered a freely access to the pollens dataset for scientific use, as shown in the following screen shot taken on the Wayback website archive.

A copy of this RNSA web archive at the time of our study is accessible through the url: <https://web.archive.org/web/20190331000355/https://www.pollens.fr/reports/database>

The web site stated at that time: "Les données exprimées dans ces fichiers sont des concentrations journalières en nombre de grains / m<sup>3</sup> d'air pour les pollens et en nombre de spores / m<sup>3</sup> d'air pour les moisissures. Ces données sont à usage privé et ne peuvent être utilisées à des fins commerciales. Pour tout travail scientifique, merci de citer la source RNSA et de nous transmettre par e-mail ([rnsa@rnsa.fr](mailto:rnsa@rnsa.fr)) un exemplaire de vos travaux. Si vous avez besoin de données complémentaires, n'hésitez pas à les demander au RNSA."

which translate into:

"The data expressed in these files are daily concentrations in number of grains/m<sup>3</sup> of air for pollens and in number of spores/m<sup>3</sup> of air for moulds. These data are for private use and cannot be used for commercial purposes. For any

scientific work, please cite the RNSA source and send us a copy of your work by e-mail (rnsa@rnsa.fr). If you need additional data, do not hesitate to ask the RNSA."

We strongly confirm that we used the data only for scientific purposes, and even more precisely for academic research purposes. The three authors are a PhD student (Eso-Ridah Bléza), and two academics (Valérie Monbet professor at university of Rennes 1 and Pierre-François Marteau University Bretagne Sud) that are managing this PhD student. A paper submitted at EGC2022 conference has been sent for review the 15th October 2021, accepted for publication in January 2022 and extended during the spring of 2022 for a submission to Data & Knowledge Engineering journal.

We assure that we did not sell the RNSA data nor did we use it for any commercial activity. We exclusively use the data for academic research (even if we had a research partnership with the LifyAir company at that time, which deals with the exploration of new datasets generated with an innovative real time pollen sensor). The work presented in the paper is a research study that basically provides a baseline in the scope of forecasting pollen episodes at the scale of the French mainland territory when using meteorological and historical emission data only. Furthermore the dataset that we collected from the RNSA web site does not contain any data dated after year 2017. As such they are not at all up to date for any online / commercial application.

Finally, we send our research academic paper to the RNSA the 21st of 2022 to inform them of the use of their data in conformance with the expectation mentioned on their web site.