# Methods for interpolating missing data in aerobiological databases

Picornell, A.[a,*]; Oteros, J.[b,c]; Ruiz-Mata, R.[a]; Recio, M.[a]; Trigo, M.M.[a]; Martínez-Bracero, M.[b,c,d]; Lara, B.[e]; Serrano-García, A [e]; Galán, C.[b,c]; García-Mozo, H.[b,c]; Alcázar, P.[b,c]; Pérez-Badia, R.[e]; Cabezudo, B.[a]; Romero-Morte, J.[e], Rojo, J.[e,f]

a. Department of Botany and Plant Physiology. University of Malaga. Campus de Teatinos s/n E-29071. Malaga (Spain).
b. Department of Botany, Ecology and Plant Physiology. Agrifood Campus of International Excellence CeiA3, University of Cordoba. Cordoba (Spain).
c. Andalusian Inter-University Institute for Earth System IISTA, University of Cordoba, Spain.
d. School of Chemical and Pharmaceutical Sciences. Technological University Dublin. Dublin (Ireland).
e. University of Castilla-La Mancha. Institute of Environmental Sciences (Botany). Toledo (Spain).
f. Department of Pharmacology, Pharmacognosy and Botany, Complutense University. Madrid (Spain).

* Corresponding author: Antonio Picornell

Department of Botany and Plant Physiology, University of Malaga.

Campus de Teatinos s/n, Malaga, E-29071, Spain.

E-mail address: picornell@uma.es

+34 952131912

## Abstract

Missing data is a common problem in scientific research. The availability of extensive environmental time series is usually laborious and difficult, and sometimes unexpected failures are not detected until samples are processed. Consequently, environmental databases frequently have some gaps with missing data in it. Applying an interpolation method before starting the data analysis can be a good solution in order to complete this missing information. Nevertheless, there are several different approaches whose accuracy should be considered and compared. In this study, data from 6 aerobiological sampling stations were used as an example of environmental data series to assess the accuracy of different interpolation methods. For that, observed daily pollen/spore concentration data series were randomly removed, interpolated by using different methods and then, compared with the observed data to measure the errors produced. Different periods, gap sizes, interpolation methods and bioaerosols were considered in order to check their influence in the interpolation accuracy. The moving mean interpolation method obtained the highest success rate as average. By using this method, a success rate of the 70% was obtained when the risk classes used in the alert systems of the pollen information platforms were taken into account. In general, errors were mostly greater when there were high oscillations in the concentrations of biotic particles during consecutive days. That is the reason why the pre-peak and peak periods showed the highest interpolation errors. The errors were also higher when gaps longer than 5 days were considered. So, for completing long periods of missing data, it would be advisable to test other methodological approaches. A new Variation Index based on the behaviour of the pollen/spore season (measurement of the variability of the concentrations every

2 consecutive days) was elaborated, which allows to estimate the potential error before the interpolation is applied.

**Keywords**

Missing data; aerobiology; time-series; modelling; interpolation; environmental sampling; bioaerosols

## 1. Introduction

Environmental time series databases require continuous and reliable monitoring systems which may be affected by technical breakdowns and human factors that can interrupt the sampling process (Oteros et al., 2013). Thus, the presence of gaps in time series data is a very widespread problem in scientific research (Junger and Ponce de Leon, 2015; Navares and Aznarte, 2019; Orlandi et al., 2014; Rubin, 1976; Schouten et al., 2018). This is why in many scientific disciplines, interpolation is commonly used to complete missing data or to increase its resolution (Lehmann et al., 1999; Luedeling et al., 2013; J. Oteros et al., 2013).

Many different methodologies have been developed for completing missing information depending on the nature of the data and the required accuracy. Some of the most extended methods are multiple imputation-based, and multiple likelihood-based estimations (Junger and Ponce de Leon, 2015). These methods are widely implemented in most statistical softwares, in particular, in several statistical R packages (e.g. “chillR”, “MICE”, “missForest”, “rrcovNA”, “mtsdi”, “mi”). Some of them use either parametric and non-parametric statistics methods (Junger and Ponce de Leon, 2015; Luedeling et al., 2013; Stekhoven and Buhlmann, 2012; Su et al., 2011; Todorov, 2020; van Buuren and Groothuis-

Oudshoorn, 2011). In general, these methods analyse the nature of the data in order to create new data that can replace the missing observations when they are randomly distributed (i.e. data are missing independently of their value or the value of other related variables) (Rubin, 1976).

Aerobiology is the scientific discipline based on the study of the atmospheric bioaerosol dynamics (pollen, spore, bacteria, virus…) (Fröhlich-Nowoisky et al., 2016). In this context, in case of gaps detection, it is not enough to create data which are statistically coherent with the rest of the database. Aerobiological data are sequential and missing data estimation must be linked with the previous and subsequent observations, given the stochastic nature of the bioaerosols in the atmosphere. In addition, aerobiological data follow a time series evolution that does not fit the stationary criterion, which increases the difficulty of predictions (Ritenberga et al., 2016).

Additionally, aerobiological samplings are complex and sometimes pollen traps are installed in non-easily accessible locations (García-Mozo et al., 2007; Oteros et al., 2019; Picornell et al., 2019c). In the case of the Hirst-type volumetric traps, the proper operation of the traps is checked once a week, but unexpected failures such as power outages or device breakdowns may happen in between, resulting in a few days period of missing data (Navares and Aznarte, 2019). Even the development of new real-time automatic sampling devices also requires interpolation methods to complete gaps during the phase of processing the database (Oteros et al., 2020). Such missing data events are produced completely at random (Missing Completely At Random; MCAR) since they are not conditioned, a priori, by any other variable or by their concentrations values (Junger and Ponce de Leon, 2015). In some cases, the gaps may may not hamper proper data analysis, but in other cases it can seriously affect the establishment of the principal dates of the main pollen season (MPS) or the main spore season (MSS) (Navares and Aznarte, 2017; Picornell et al., 2019a). Alternatively, missing concentration values might be considered as 0 pollen grains or spores/m$^3$ of air for most MPS/MSS definitions, which in many cases would produce more errors than estimated data.

The ideal method to complete missing concentration data in aerobiological databases, in terms of usability, should be independent of other variables and directly applicable. Linear interpolation is the most commonly used method to complete missing data, but its use is not so extended in Aerobiology as in other disciplines (Belmonte et al., 1999; Gabarra et al., 2002; Navares and Aznarte, 2019, 2017; Picornell et al., 2019a; Skjøth et al., 2016). Other methods are rarely applied (e.g. moving mean interpolation or interpolation by using data of nearby location) and their accuracy have never been measured (Jesús Rojo et al., 2019; Skjøth et al., 2016).

For all the aforementioned, the main aim of this study was to comparatively evaluate different methods which allow to interpolate aerobiological data, as well as to check their effectiveness and accuracy depending on the circumstances based on real life observation data.

## 2. Material and methods

To carry out this study, the databases of 6 aerobiological stations, situated in different localities of the Iberian Peninsula, have been used: Cordoba, Hornachuelos Natural Park, Malaga, Ronda, Sierra de las Nieves Natural Park, and Toledo (Fig. 1). The altitudinal range of the sampling stations varied from 58 to 1073 m a.s.l., with an average annual total precipitation between 382 mm and 996 mm, and an annual mean temperature of between 11.9 ºC and 18.4 ºC (Table 1). All the sampling stations are within the Mediterranean macroclimate (Rivas-Martínez et al., 2017).

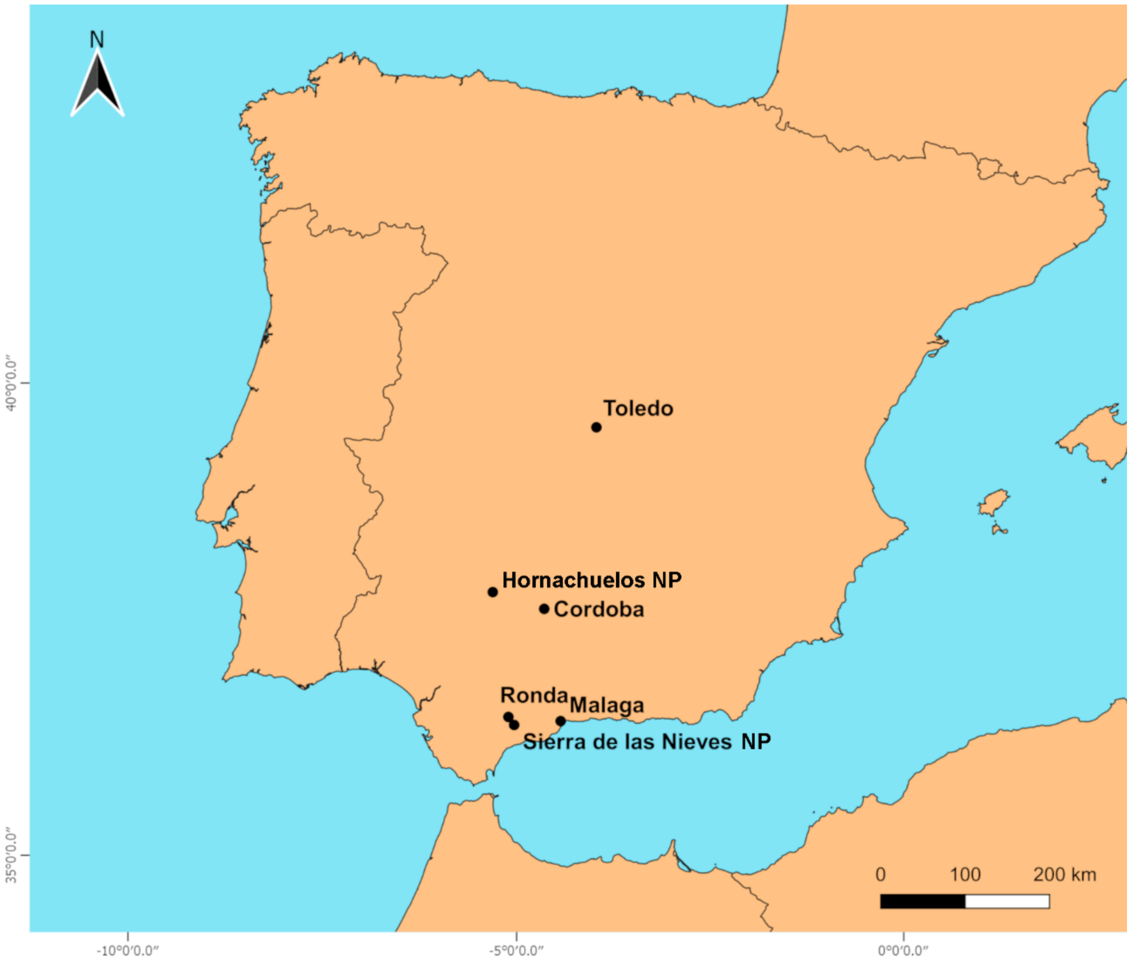


**Fig. 1.** Map of the pollen and spores sampling stations used in this study. Spatial information obtained from REDIAM (Junta de Andalucía, 2011). NP: Natural Park.

**Table 1**. Climatic parameters, sampling years and coordinates of the sampling sites included in this study. Data extracted from García-Mozo et al., 2006; Hernández-Ceballos et al., 2015; Picornell et al., 2020, 2019b.

| Location | Annual total rainfall (mm) | Annual average temperature (ºC) | Altitude (m a.s.l.) | Coordinates | Years of sampling |
|---|---|---|---|---|---|
| Cordoba | 621 | 17.8 | 138 | 37º54' N 4º43' W | 2006-2018 |
| Hornachuelos NP | 700 | 16.8 | 225 | 38°4' N 5°24' W | 1998-2019 |
| Malaga | 540 | 18.4 | 58 | 36º42' N 4º28' W | 1991-2019 |
| Ronda | 681 | 16.4 | 768 | 36º44' N 5º10' W | 2017-2019 |
| Sierra de las Nieves NP | 996 | 11.9 | 1073 | 36º39' N 5º5' W | 2018-2019 |
| Toledo | 342 | 15.8 | 450 | 39º51' N 4º2' W | 2003-2019 |

NP: Natural Park.

### 2.1. Pollen and spore data

Airborne pollen and fungal spores were collected by means of 6 Hirst-type volumetric traps, one per location (Hirst, 1952). The air flow was adjusted in all of them to 10 l/min. The aerobiological samples obtained were processed and analysed following the recommendations of both the Spanish Aerobiology Network (REA) (Galán et al., 2007) and the European Aerobiology Society (EAS) (Galán et al., 2014). More than the 10% surface of each daily sample were analysed for pollen identification and counted by light microscopy at a magnification of 400X. In the case of *Alternaria* spores, at least the 5% of each daily sample were analysed at the same magnification (Galán et al., 2021). Pollen and spore concentrations were expressed as pollen grains/m$^3$ of air and spores/m$^3$ of air, respectively, according to the international recommendations (Galán et al., 2017, 2014).

Daily pollen concentrations of Amaranthaceae, Cupressaceae, *Olea*, *Pinus*, *Plantago*, *Platanus*, Poaceae, *Quercus*, and Urticaceae were used to test the accuracy of the interpolation methods at all sampling stations, while Arecaceae and *Casuarina* pollen concentrations only were used in Malaga, Ronda, and Sierra de las Nieves, due to its scarcity in the atmosphere of the other localities. Regarding fungal spores, *Alternaria* concentrations registered in Cordoba, Malaga, Ronda, Sierra de las Nieves, and Toledo were also included in the study.

For each pollen/spore type and year, the main pollen/spore season was calculated, this being defined as the period between the first day of the year in which the 5% of the annual pollen/spore integral is reached and the first day in which the 95% annual is accumulated (Nilsson and Persson, 1981). In the case of Cupressaceae and *Alternaria* two different pollen/spore curves were detected within a year in all sampling sites. Therefore, each curve was studied separately by dividing the year in two periods: January-July (winter Cupressaceae and spring *Alternaria*) and August-December (autumn Cupressaceae and *Alternaria*). Since Urticaceae pollen type was abundantly detected during the whole year, the start and end dates of the MPS were defined by adjusting the cumulative pollen concentrations to a logistic curve and selecting the dates in which the fourth derivative of the logistic curve crossed the x-axis (Cunha et al., 2015; Ribeiro et al., 2007). The MPS/MSS were calculated with the "AeRobiology" R package (Jesús Rojo et al., 2019). The optimal definition method was applied in each case. The defined seasons, independently of the method applied, helped to categorize the time series into different periods in order to analyse any effect of the seasonal stages on the missing data estimation. Therefore, the method used to define the pollen seasons was not a crucial point in this study, and the results are not affected by them.

## 2.2. Interpolation methods tested

For this study, the different interpolation methods integrated in the “AeRobiology” R package were tested, i.e. linear interpolation, moving mean interpolation, spline interpolation, interpolation by using time series analysis, and interpolation by using data from nearby locations (Jesús Rojo et al., 2019). In all cases, calculations are based on the daily mean pollen/spore concentrations.

### 2.2.1. Linear interpolation

A linear regression is calculated by taking the first data previous and subsequent to the gap (i.e., first day before and after the gap), so, the missing data are estimated by using the regression equation (Fig. 2).

### 2.2.2. Moving mean interpolation

Each missing data is replaced by the mean value of a certain number of data placed on both sides of the gap. The number of days took for calculating the mean is the double of the gap size, and it is centred in the missing value (Fig. 2).

### 2.2.3. Spline interpolation

A spline regression is calculated by taking the first 3 data on both sides of the gap. Then, the missing data are estimated by using the regression equation (Fig. 2).

### 2.2.4. Interpolation by using temporal series analysis

For each pollen/spore type, the seasonality is calculated by taking all the daily data available for several years by performing a seasonal trend decomposition based on LOESS (Cleveland et al., 1990). Then, a linear regression is calculated between the pollen/spore curve of the year in which there are missing data and the seasonality curve in order to regulate the curve intensity based on the known data of the target pollen/spore season (Fig. 2). Missing data are estimated by using the regression equation.

#### 2.2.5. Interpolation with data from nearby locations

In this case, the database of a nearby sampling station with complete data for the missing period is used to complete the gaps. For the year in which the loss of data occurred, a linear regression is calculated between the pollen/spore curve of the nearby locality (independent variable), and the pollen/spore curve of the target locality (dependent variable). If the regression is significant (p-value $\leq 0.05$) and the regression coefficient is higher than 0.6, the data from the nearby location are transformed by applying the regression equation, and the missing data are replaced by the calculated values (Fig. 2). Regression coefficients under 0.6 have not been considered high enough to reflect a direct relationship between the concentration values of both sampling locations. In this method, it is possible to include more than one nearby location simultaneously. In such cases, the data of each nearby sampling station is included as an independent variable in a multiple linear regression and the missing data (dependant variable) is calculated by applying the regression equation.

This method was tested in Cordoba by using Hornachuelos National Park as nearby sampling station, as well as in Ronda by using Malaga, and Sierra de las Nieves databases as nearby localities, both individually and simultaneously.

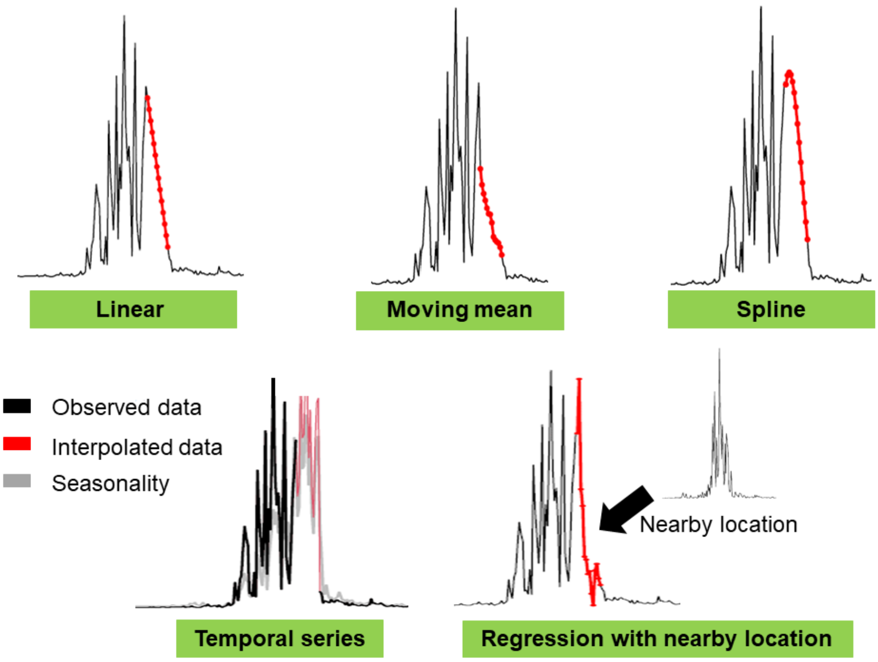


**Fig. 2.** Graphical examples to visualise the application of the different methods of data interpolation applied in this study. Example elaborated with *Olea* pollen data in Ronda during 2018.

### 2.3. Relative error calculation

To check the effectiveness of each interpolation method, some observation data were removed from the original databases of each sampling site in order to create artificial gaps. After that, the data missed were interpolated by the methods explained above and the estimated data were compared with those removed.

To avoid bias in removing the original data, an algorithm that performs random cuts in the data series was developed. This algorithm made random cuts in different periods of the pollen seasons for the different pollen/spore types, being these periods: pre-season, pre-peak, peak, post-peak and post-season. The pre- and post-season periods are those outside the MPS/MSS. The peak cut was obtained by centring the peak day in the centre of the removed data. Cuts of 3, 5, 7 and 10 consecutive days were tested.

Relative errors (RE) were calculated by means of the formula according to equation 1. In such of the mathematical formula, the error values range between 0 and 2. Cases whose observed concentrations were zero pollen grains or spores/m$^3$ caused mathematical indeterminacy when the estimated value was zero too (0/0), and relative errors of 2 when the estimated value was non-zero. Therefore, they were excluded since these concentrations were not frequent and they have scarce relevance.

$$Relative\ error = \frac{|e-o|}{\frac{|e|+|o|}{2}}\ if\ o \neq 0 \tag{1}$$

where *e* is the estimated pollen/spore concentration, and *o* the observed concentration.

Besides this, estimated and observed pollen/spore concentrations were classified into the Spanish Aerobiology Network pollen classes (nil, low, moderate and high) (Galán et al., 2007), but nil class was modified to concentrations ≤ 1 pollen grain or spore/m$^3$. Due to there are no stablished classes for *Alternaria* spores, the thresholds for the moderate and high categories were set in 30 and 50 spores/m$^3$ respectively, according to the most frequent concentrations detected in the sampling sites. After classifying the observed and estimated pollen/spore concentrations, the percentage of correct classification, i.e. the success rate (observed category = estimated category), was calculated by means of equation

2, in which, observed concentrations of 0 pollen grains or spores/m$^3$ were not excluded since they do not induce mathematical artefacts in the formula.

$$Success\ rate = \frac{Nº\ correct\ classifications}{Nº\ total\ classifications} * 100 \tag{2}$$

Differences in the relative errors and in the success rates were tested with pairwise Mann-Whitney-Wilcoxon tests with Bonferroni post-hoc corrections since data did not fit a normal distribution according to Kolmogorov-Smirnov tests with Lilliefors corrections (α=0.05).

## 2.4. Variation Index of each pollen/spore type

The same pollen/spore type can show different curve profiles in different locations depending on the abundance of the emission sources, wind dynamics, climate, meteorological conditions, and phenology of the species present in the territory (Grinn-Gofroń and Rapiejko, 2009; Picornell et al., 2019b; Velasco-Jiménez et al., 2013). Therefore, results of a certain pollen/spore type may not be directly comparable among sampling locations. Moreover, interpolation success may be strongly related with the curve shape and the variation coefficients in the concentrations of consecutive days (noise). In order to characterise the daily variations of a given pollen/spore type in the different sampling stations, we developed the so defined “Variation Index” (VIn). It consisted on calculating the average of the variation coefficients (CV, equation 3) of every two consecutive days for the main pollen/spore season (equation 4). The average VIn is then calculated for the years included in the study. This index measures the average variations during consecutive days that a certain pollen/spore type shows at a certain locality (i.e. the more variations in the concentrations among consecutive days, the highest VIn is obtained).

(3)

$$CV = \frac{\sigma}{X}$$

where *CV* is the coefficient of variation, *σ* the standard deviation and *X* the average.

(4)

$$Variation\ Index = \frac{\sum_{i=1}^{n} CV_i}{n}$$

where *CV* is the coefficient of variation, and *n* the number of days within the main pollen/spore season.

To check if there is any relationship between the VIn of a certain pollen/spore type and the errors obtained when interpolating, a linear regression was calculated between these two variables.

# 3. Results and discussion

## 3.1. Relative errors and Variation Index

Once gaps of 3, 5, 7 and 10 days were artificially created in the data series of different pollen/spore types, years and localities, and the gaps filled with the estimated data, the results obtained were analised by comparing them with the observed data. In general, regarding the different interpolation methods, the one that obtained the lowest relative error was the moving mean (0.77 as average), followed by the linear interpolation (0.80 as average) (Fig. 3A). Their relative errors showed significant differences between them as well as with the other methods. These results are due to, despite its mathematical simplicity, the moving mean takes into account the curve trend and the pollen/spore concentrations immediately before and after the gaps, what provides a more accurate adjustment to the pollen/spore curve. Although abrupt changes may happen during the missing period, this interpolation method follows the general trend of the serie and, in general, it is less likely that predictions contain major errors. The linear interpolation generally takes into account the curve trend too, but in some cases it is oversimplified and the new data obtained may be affected by punctual concentrations that do not fit the general trend.

The regressions with nearby locations, the spline regresions, and the regressions with nearby locations obtained significant higher relative errors according to Mann-Whitney-Wilcoxon tests (Fig. 3A).

The interpolation with nearby location obtained very different relative errors depending on the nearby sampling station considered. In the case of Ronda, the regressions with Sierra de las Nieves sampling station (14 km away) obtained lower relative errors (RE=0.89 as average) than the regressions with Malaga sampling station (RE=1.03), situated 62 km away (Fig. 3A). However, when Sierra de las Nieves and Malaga databases were simultaneously taken into account, the error rates were estatistically similar to those obtained when considering only Sierra de las Nieves. These last errors were also similar to the ones obtained in the case of Cordoba when using Hornachuelos Natural Park sampling station as neighbour location, which is 64 km apart, with a RE of 0.94, as average. This interpolation method is the only one of all tested that, indirectly, takes into account variables such as meteorological conditions, the effect of the vegetation or land use. The effects of these variables are also reflected in the pollen/spore daily concentrations of the nearby location, so they are indirectly integrated in the regression with the target location. Therefore, more accurate interpolations could be expected if the nearby locations had similar climatic conditions, vegetation and land use, which would be also reflected in the pollen/spore timing (phenology) and in the airborne pollen/spore load (intensity) (El-Moslimany, 2019; García-Mozo, 2017; Ruiz-Valenzuela and Aguilera, 2018). For these reasons, closer localities are generaly more likely to have similar concentration curves and, therefore, regressions analysis resulting more accurate than when using further away locations, as suggested in previous studies (Hjort et al., 2016; Lara et al., 2020; Navares and Aznarte, 2019). Nevertheless, it is possible that further sampling sites with similar conditions to the target station or with similar ornamental taxa in the vicinity of the pollen trap obtain lower errors than geographically closer sampling sites. Consequently, the errors obtained for the method of the nearby locations should be cautelously considered because the accuracy of the interpolation depends on the factors aforementioned. When applying this method, it would be interesting to select the nearby location by studing its similarity to the target location as proposed by Oteros et al. (2019).

Usually, new aerobiological sampling sites are selected in order to cover areas with different environmental conditions than other previously settled stations, including meteorological conditions, land use and vegetal coverage. Consequently, the results of the interpolation with nearby locations, as observed in the results, are expected to produce high relative errors as average, given that the different sampling stations are installed to cover the geographical heterogeneity of a territory.

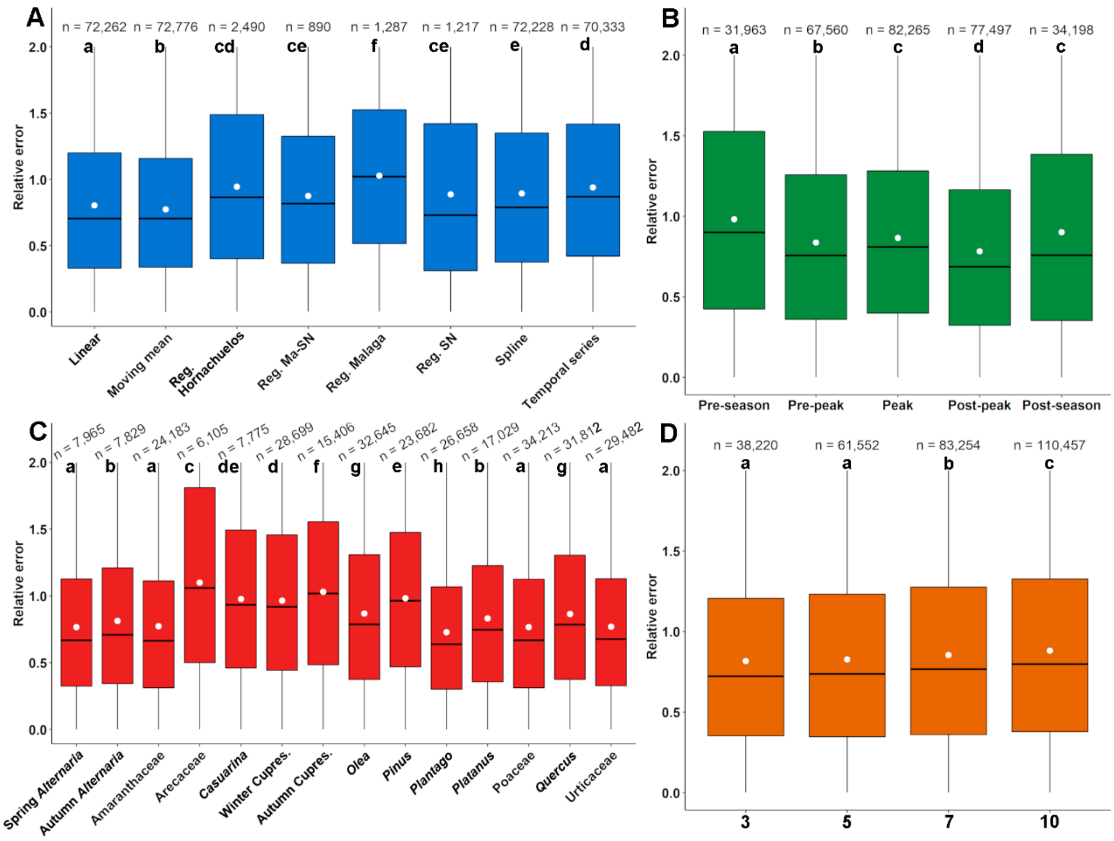


**Fig. 3.** Relative errors obtained by the different interpolation methods (A), periods of the year (B), pollen/spore types (C), and gap sizes (days; D) with all the available data. n: number of observations. Each box includes the interquartile range (Q1-Q3), bold lines indicate the median and white dots indicate the mean. Groups which share the same letter above have not any significant differences (α=0.05) between them according to Mann-Whitney-Wilcoxon tests with Bonferroni *post-hoc* corrections. Reg.: regression with; Ma-SN: Malaga and Sierra de las Nieves; SN: Sierra de las Nieves.

The spline interpolation, as well as the moving mean and the linear interpolations, also takes into account the curve trend, but they produce more pronounced curve trends than the other methods as a consequence of spline approximation. This can lead to very accurate fits or to big errors in the predictions, which, in general, gives higher errors (0.89) than the linear and moving mean interpolations (Fig. 3A).

The temporal series analysis is highly dependent on the extension of the historical database (number of years in this case). Therefore, if the number of years were not enough (as occurred in Ronda or in Sierra de las Nieves), the obtained seasonality curve might not be representative of the regular behaviour of a given pollen/spore type. Also, data series of uncommon years (according to phenology and flowering intensity) deviated from the standard behaviour may result in non-accurate interpolations when using this method (Belda et al., 2020). In wind pollinated trees, remarkable differences between pollen seasons of consecutive years are frequent due to mast seeding cycles (Bogdziewicz et al., 2017). Additionally, when performing the linear regression between the seasonality curve and the curve of the target year, more errors are accumulated (0.94 as average relative error).

Regarding the main pollen/spore season (Fig. 3B), in general, the period that obtained the lowest relative errors was the post-peak (RE average=0.78), followed by the pre-peak period (RE average=0.84). Post-peak periods usually present smoother curve shapes than the pre-peaks and the peaks periods due to these last ones are more conditioned by plant phenology and flowering intensity than the post-peaks, in which the plants progressively reduce the pollen emission intensity (Cunha et al., 2016; Kasprzyk and Walanus, 2014; Picornell et al., 2019a). Therefore, as abrupt changes in the data series are less likely during the post-peaks, fewer errors are also expected in the interpolation.

On the other hand, the peak-day concentration is difficult to predict, since it is usually an abrupt change caused by the interaction of both meteorological and biological parameters which are not easily predictable. Moreover, the peak-day concentration usually varies widely from one year to another, what makes more difficult to successfully apply interpolation methods (Devadas et al., 2018; García-Mozo et al., 2009; Picornell et al., 2019a; Valencia et al., 2019). Consequently, the relative errors obtained were higher than for the rest of the MPS/MSS period (RE average=0.86).

Outside the MPS/MSS, pre- and post-seasonal periods obtained significant higher relative errors (0.98 and 0.90 as average respectively) (Fig. 3B). During these periods, days with null value in pollen/spore concentrations are frequent, interspersed with small rises and falls, what makes it more difficult to predict or stimate the daily values. However, these errors might not be relevant for defining the MPS/MSS unless they are located near to the start or the end dates. Given that the concentrations outside the MPS/MSS usually are very low, such errors are less relevant for allergy alerts.

Gap sizes longer than 5 days obtained significantly higher relative errors (Fig. 3D). In these gaps, abrupt changes or trend changes are more likely to happen than in smaller gaps, what may lead to higher errors. Accordingly to this, the lowest error rate was obtained for gaps of 3 days (0.82), but with non-significant differences with the errors for gaps of 5 days (0.83). As expected, longer gaps produced higher errors since the uncertainty increases when it comes to estimating longer periods. In fact, relative errors of 0.85 and 0.88 were obtained for gaps of 7 and 10 days respectively. However, these errors are expected to induce less changes in the MPS/MSS calculation than leaving the gaps without data, so it is still recommendable to interpolate them.

Regarding the results obtained by pollen/spore types (Fig. 3C), spring *Alternaria* (0.76), Amaranthaceae (0.77), Poaceae (0.77) and Urticaceae (0.77) were the pollen and spore types that obtained the lowest relative errors, as average. In general, these pollen and spore types are integrated by several species that, usually, have wide distribution areas. These pollen/spore types are detected during a relatively long period of the year and it probably makes their seasonal trends be smoother than the other pollen types such as Arecaceae (in which the highest relative errors were obtained, RE=1.10), *Casuarina* (0.98) or *Platanus* (0.83). Other pollen types, such as Cupressaceae, *Pinus*, *Quercus*, and *Olea*, have several concentration peaks within their MPS, which may correspond to the flowering of the different species or varieties that integrate the pollen types. These peaks are difficult to predict, and it may increase the relative errors obtained. For the same reason, autumn *Alternaria* obtained higher relative errors than spring *Alternaria* because autumn MSS is usually shorter and contains more pronounced peaks.

Despite the pollen type has been considered, it is more interesting to consider the behaviour of the pollen curve in general, which have been characterized in this case by the Variation Index (VIn, see methods for the definition). The same pollen/spore type showed different relative errors at different sampling locations (data not shown). Therefore, it would be pointless to establish the average relative error by pollen/spore type if it is going to vary when considering a new sampling site. As observed in Fig. 4, the more variations during consecutive days (higher VIn), the more relative errors are obtained during interpolation, a linear and direct relationship existing between the VIn and the relative errors. So, by means of the regression equation, included in Fig. 5, it is possible to estimate the average error rate when interpolating values of a pollen/spore type by calculating the VIn. This error estimation is independent of the pollen/spore type, and only relies in the behaviour of its daily concentration curve. Furthermore, this regression equation has been elaborated with pollen and spore data of 6 sampling sites located at different environmental conditions and so, it can be used as a calibration curve for estimating the errors at new locations. Anyway, we recommend taking interpolation results with great caution when the VIn is higher than 0.75, since relative errors greater than 1 are expected, as average, above this value.

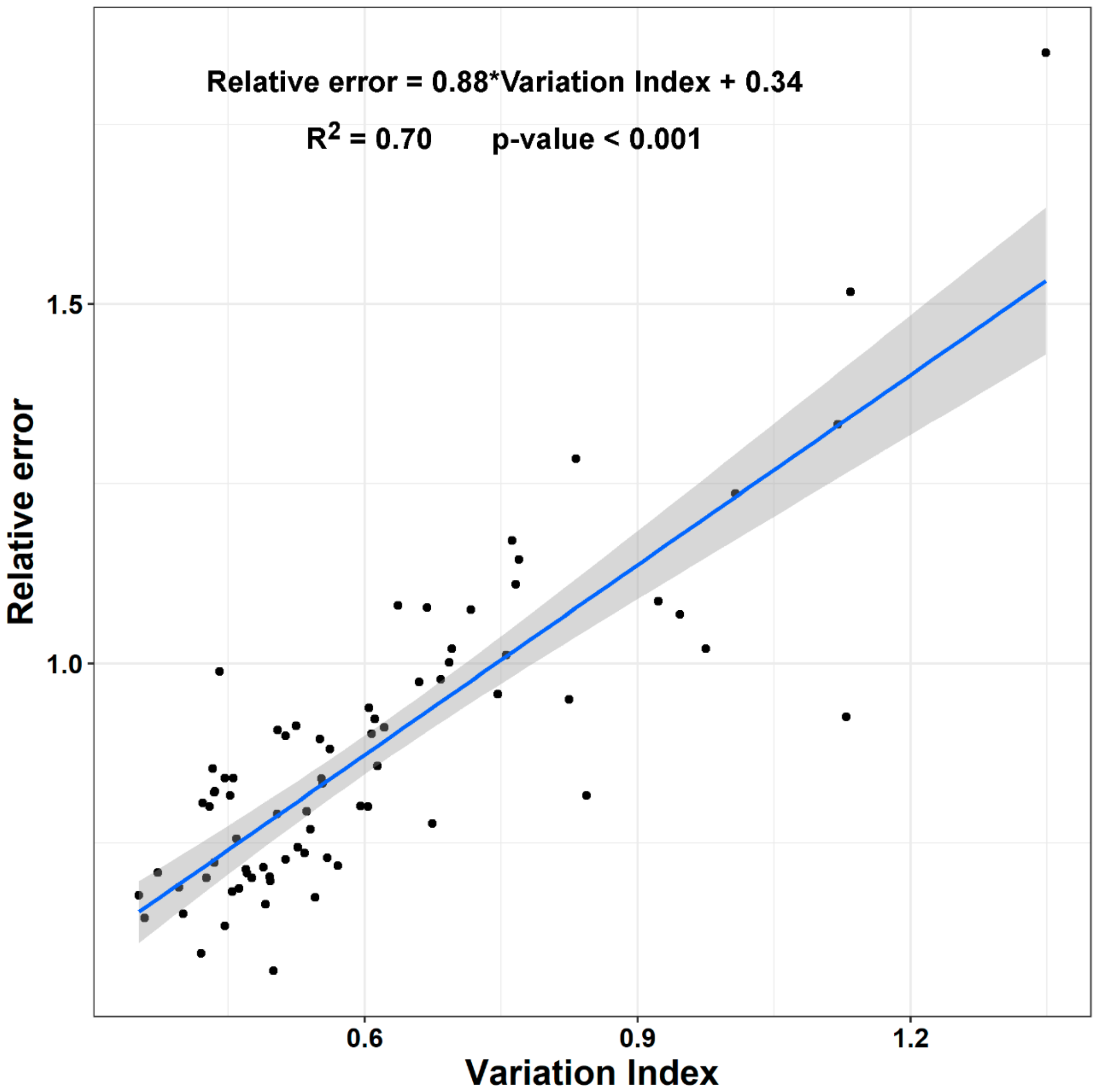

**Fig. 4.** Linear regression between the Variation Index and the relative error of the different pollen/spore type in the different sampling localities during the MPS/MSS. The grey area marks the 95% confidence interval.

Additionally, in Fig. 5 we have represented, separately, the regression lines between the Variation Indexes and the relative errors for each interpolation method. It would allow to roughly estimate the average relative error that this interpolation method would produce for a given pollen/spore type. The methods that obtained the highest coefficient of regression were the temporal series, linear and spline interpolation. The regression with nearby location is not statistically significant since the relative errors obtained depend on the similarity between sampling sites, and not on the VIn of each pollen/spore type.

According to the obtained results, some interpolation methods, such as temporal series and linear interpolations, are more sensitive to pollen/spore types with high variations in their concentrations during consecutive days than the others (Fig. 5). This can be observed in the regression equations slopes that, when significant, are higher than in the other methods. Linear interpolation may obtain lower relative errors than moving mean interpolation if the pollen/spore type had a low VIn, but the errors would be higher if the pollen type presented a higher VIn. In the case of the interpolation with data from nearby location, the points did not fit a linear regression ($p$-value$>0.05$). It can be explained, as commented above, because the errors obtained during the interpolation are related to the similarity of both sampling locations, rather than to the characteristics of the pollen/spore type.

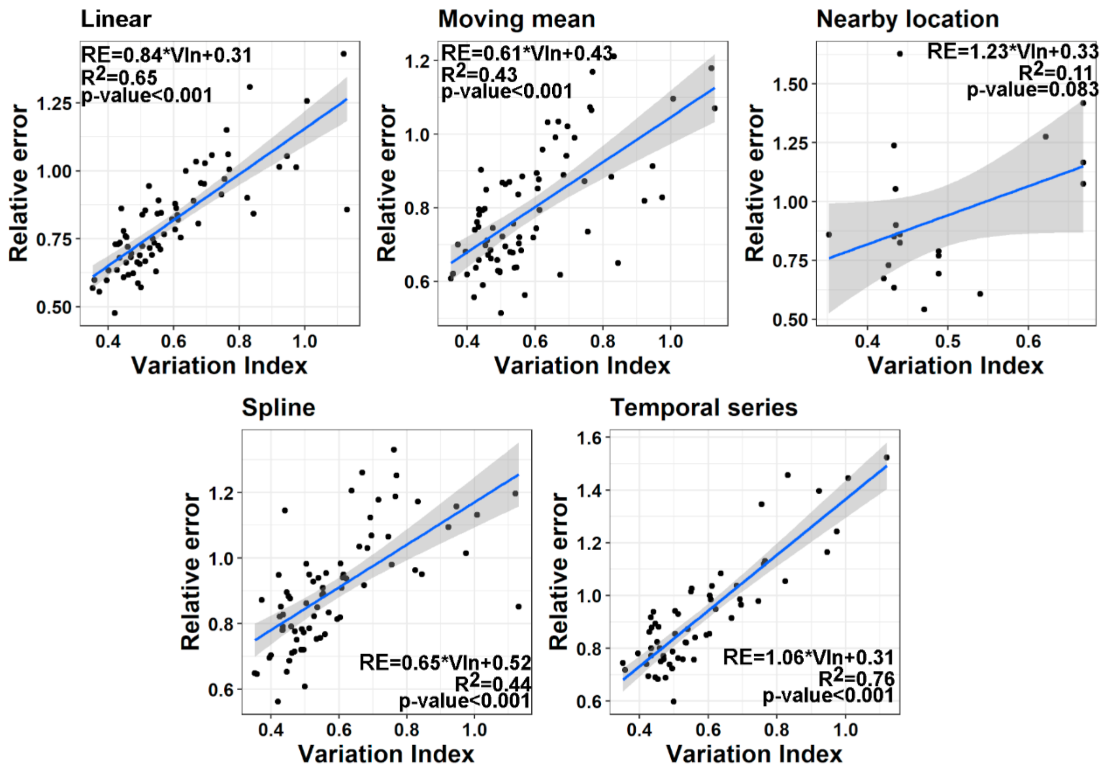


**Fig. 5.** Linear regressions between the Variation Index and the relative error of each pollen/spore type during its MPS/MSS sorted by interpolation method. The grey area marks the 95% confidence interval. RE: relative error; VIn: Variation Index.

## 3.2. Success rates

Observed and estimated pollen/spore concentrations were categorized by thresholds following the criteria of the Spanish Aerobiology Network (REA) for each pollen/spore type. Then, observed and estimated categories were compared and, in general, success rates above 60% were obtained for all the studied bioaerosols (Fig. 6). The REA and other pollen information platforms use the categories null/nil, low, moderate and high for releasing the pollen risk information to the population (Galán et al., 2007; Pérez-Badía et al., 2010).

Therefore, many of the concentrations that showed relative errors in the continuous variable, now are classified in the same risk category.

The highest average success rate (i.e. lowest errors) were obtained, once more, for the linear interpolation (71%) and the moving mean interpolation (70%), without significant differences between them, but neither with the spline interpolation (68%), or the regression with Sierra de las Nieves (62%; in the case of Malaga sampling location) (Fig. 6A). The lowest average success rates (i.e. highest errors) were detected when the interpolation was performed by using the nearby location of Malaga (56% in the case of Ronda). In general, these results were similar to the obtained by calculating the relative errors (Fig. 3A) with the exception that, in this case, using levels instead daily concentrations, the spline interpolation and the regression with Sierra de las Nieves did not show significant differences when compared to linear and moving mean methods.

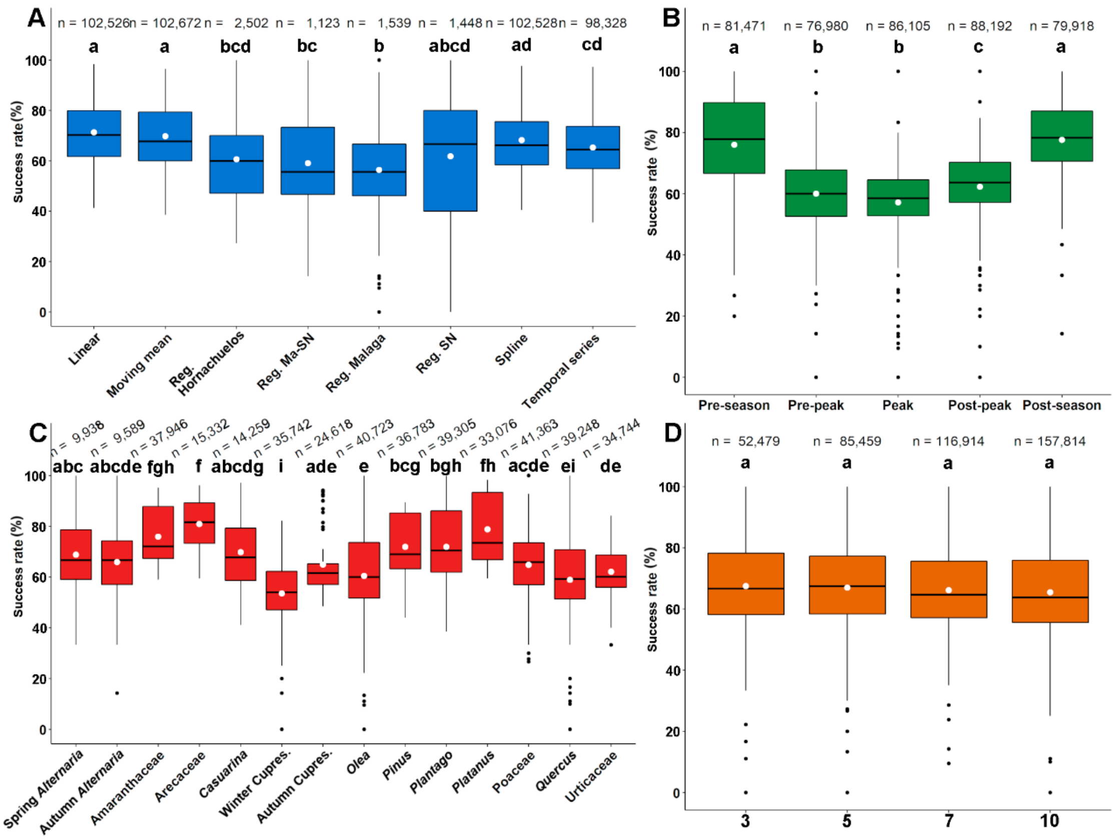

**Fig. 6.** Success rates obtained by the different interpolation methods (A), periods of the year (B), pollen/spore types (C), and gap sizes (days; D) when comparing the observed and predicted pollen/spore levels. n: number of observations. Each box includes the interquartile range (Q1-Q3), bold lines indicate the median and white dots indicate the mean. Groups which share the same letter above have not any significant differences ($\alpha$=0.05) among them according to Mann-Whitney-Wilcoxon tests with Bonferroni *post-hoc* corrections. Reg.: regression with; Ma-SN: Malaga and Sierra de las Nieves Natural Park; SN: Sierra de las Nieves Natural Park; Hornachuelos: Hornachuelos Natural Park.

Regarding the periods of the pollen season (Fig. 6B), those outside the MPS/MSS (i.e. pre-season and post-season) obtained the highest success rates (76 and 78% respectively). As previously commented, pollen and spore concentrations during these periods are generally low and, although interpolations produced high relative errors, the variations between expected and observed concentrations imply very little changes in the pollen/spore categories. Therefore, these errors are less relevant for the allergy alert systems since they do not imply high changes in the information of the atmospheric allergenic potential. The highest success rates inside the MPS/MPS were obtained for the post-peak period (62% of success), what matches the results of the relative errors, while the lowest success rates were obtained for the peak (57%), and pre-peak (60%) periods, once again the peak being the most unpredictable period.

As observed with the relative errors, there is a decrease in the average success rate (i.e. higher errors) when gaps of more than 5 days are interpolated (Fig. 6D). However, in this case, these errors did not involve any significant difference between any gap size since all average success rates were between 65 and 67%.

Finally, as can be seen in Fig. 6C, the different pollen/spore types obtained different success rates, ranged from 54 to 81%. The highest average success rate was obtained for Arecaceae (81%). This pollen type did not present significant differences with Amaranthaceae (76%) or *Platanus* (79%). Usually, Arecaceae pollen concentrations detected are low, as occurred also with *Platanus* in most sampling locations. The errors when interpolating such low concentrations may involve high relative errors but little changes in the stablished categories, a similar effect that the observed outside the MPS/MSS periods and commented above (Fig. 3B). The lowest average success rate was obtained for winter Cupressaceae (54%), followed by *Quercus* (59%), *Olea* (60%), Urticaceae (62%), and autumn Cupressaceae (65%). These pollen types are usually detected in a wider range of concentrations than in the other pollen types and so, errors during the interpolation are more likely to entail errors in the categories and, therefore, lower accuracy rates (Fig. 6C).

Although the results obtained in some cases have not been the most favourables, we consider that they have been accurate enough (relative errors are generally under 0,8) for not leaving blank the gaps in an aerobiological database, without assigning a value, due to it would lead to take these concentrations as 0 pollen grains or spores/$m^3$. This would introduce higher errors in the annual spore/pollen integral and in the MPS/MSS definition than with the interpolated data. These errors are potentially greater when a percentage definition of the MPS/MSS is applied. Moreover, when working with pollen/spore levels during these missing days, the accuracy can reach to the 70-71% of the cases (using moving mean or linear interpolation, respectively), which would allow to use these data to give pollen/spore information to the population or to make comparisons between pollen data and allergic symptoms (Karatzas, 2009).

Data quality and errors involved in the aerobiological sampling method may also play an important role in the measurement of interpolation accuracy (Oteros et al., 2015; Rojo et al., 2019). If these errors increase the variability of the bioaerosol concentrations during consecutive days, they might compromise the measurement of the interpolation accuracy (i.e., they will increase the VIn). However, this effect is not easily measurable since the data used for validating the interpolation provides from the same pollen trap and would have the same potential sampling error. In these terms, the interpolation methods that does not only rely on the data in both sides of the gap, such as the interpolation by using temporal series analysis or the interpolation with nearby locations, would be less affected by sampling errors.

Apart from the methods proposed in this study, additional methods to complete missing data may be considered, such as elaborating regional forecast models based on meteorological variables and emission maps or dispersal models (Lara et al., 2019; Verstraeten et al., 2021). Nevertheless, such models should be elaborated separately for each pollen/spore type, and it might be necessary to elaborate individual models for different climate areas (García-Mozo et al., 2008). Hence, such methods would not be easy to automatize, they require individual validation, depend on the availability of meteorological data for the target location and, in most cases, a long time series of data is required to train and validate the models.

Due to recent movility restrictions caused by the COVID-19 pandemic, many aerobiological samplings have been interrupted to a lesser or greater extend. This has caused missing data for some weeks and even months in several aerobiological sampling stations. Most of these monitoring gaps occurred during spring, which affected the MPS/MSS data collection. As observed in the results, when the gap is longer than 5 consecutive days, the error rates increase significantly. For these long gaps, most of the presented interpolation methods might not be appropriate, so, it would be interesting for further studies to test the performance of time series analysis or the regressions with nearby sampling stations when a great part of the data of the MPS/MSS is missing. Nevertheless, predicting the temporality and intensity of the MPS/MSS is not an easy task, and often requires the adjustment of the models to the local conditions (Picornell et al., 2019a; Rojo et al., 2021, 2016).

This work is an approach to perform interpolations in order to fill in the gaps that, due to different reasons, are generated in pollen/spore databases. Despite the most accurate method was generally the moving mean, for each specific case it would be necessary to select the method according to the particularities of each sampling station and pollen/spore type. Although this study was conducted with aerobiological databases, these results may be useful for interpolating missing data in other environmental databases.

## 4. Conclusions

- The moving mean interpolation is the method that generates the lowest relative errors, as average. This method is independent of the availability of additional data and of the length of the database, and it is also less sensible to variations in the pollen/spore concentrations during consecutive days than the other methods considered in this study. In addition, this method showed a success rate of the 70% when assigning the risk classes that are frequently used in the allergy alert systems.

- Periods with high variation indexes (VIn) make the pollen concentrations difficult to predict and, generally, cause high errors when interpolating data. Probably that be the reason why the pre-peak and peak periods present higher error rates.
- The Variation Index proposed, based on the pollen/spore season behaviour, is a good indicator of the success rate. Therefore, it is advisable to take this index into consideration since it allows to estimate the relative error before applying interpolation methods.
- The errors during the interpolation generally increase when gaps of more than 5 days are considered. For that reason, alternative methods should be considered for interpolating longer gaps.

**Funding:** This work was supported by the Spanish Ministry of Economy and Competitiveness [project CGL2014-54731-R]; by the Ministry of Science and Innovation [projects RTI2018-096392-B-C22]; by the Junta de Andalucía [contract 8.06/503.4764]; and by the Area of Environment and Sustainability of the Malaga City Council [contracts 8.06/5.03.4721 and 8.07/5.03.5159], and the Junta Comunidades de Castilla-La Mancha, which provides financial support for the Castilla-La Mancha Aerobiology Network (AEROCAM). Antonio Picornell was supported by a predoctoral grant financed by the Spanish Ministry of Education, Culture and Sport, in the Program for the Promotion of Talent and its Employability [grant number FPU15/01668]. The pollen trap installed in Sierra de las Nieves was funded by the Herbarium MGC of the SCAI (Central Services of Research Support) of the University of Malaga under the agreement signed between the Junta de Andalucía and the University of Malaga [contract 8.07/5.034764].

**Acknowledgments:** The authors specially want to thanks the SCAI (Central Service for Research Support) of the University of Malaga for supporting the acquisition of the pollen trap installed in Sierra de las Nieves; the Parauta City Council, the direction of Sierra de las Nieves Natural Park, Las Conejeras campsite for facilitating the installation of the pollen trap in Sierra de las Nieves; and the staff of Pérez de Guzmán High School for providing support to install and maintain the pollen trap in Ronda, and to Enresa for facilitating the installation and maintenance of the pollen trap in Hornachuelos Natural Park.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study, collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## References

Belda, S., Pipia, L., Morcillo-Pallarés, P., Rivera-Caicedo, J.P., Amin, E., De Grave, C., Verrelst, J., 2020. DATimeS: A machine learning time series GUI toolbox for gap-filling and vegetation phenology trends detection. Environ. Model. Softw. 127, 104666. https://doi.org/10.1016/j.envsoft.2020.104666

Belmonte, J., Canela, M., Guàrdia, R.A., Sbai, L., Vendrell, M., Cariñanos, P., Díaz de la Guardia, C., Dopazo, A., Fernández, D., Gutiérrez, M., Trigo, M.M., Guàrdia, R.A., Sbai, L., Vendrell, M., Cariñanos, P., Díaz de la Guardia, C., Dopazo, A., Fernández, D., Gutiérrez, M., Trigo, M.M., 1999. Aerobiological dynamics of the Cupressaceae pollen in Spain, 1992-98. Polen 10, 27–38.

Bogdziewicz, M., Szymkowiak, J., Kasprzyk, I., Grewling, Ł., Borowski, Z., Borycka, K., Kantorowicz, W., Myszkowska, D., Piotrowicz, K., Ziemianin, M., Pesendorfer, M.B., 2017. Masting in wind-pollinated trees: System-specific roles of weather and pollination dynamics in driving seed production. Ecology 98, 2615–2625. https://doi.org/10.1002/ecy.1951

Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: A Seasonal-Trend Decomposition Procedure based on Loess. J. Off. Stat. 6, 3–73.

Cunha, M., Ribeiro, H., Abreu, I., 2016. Pollen-based predictive modelling of wine production: Application to an arid region. Eur. J. Agron. 73, 42–54. https://doi.org/10.1016/j.eja.2015.10.008

Cunha, M., Ribeiro, H., Costa, P., Abreu, I., 2015. A comparative study of vineyard phenology and pollen metrics extracted from airborne pollen time series. Aerobiologia. 31, 45–56. https://doi.org/10.1007/s10453-014-9345-3

Devadas, R., Huete, A.R., Vicendese, D., Erbas, B., Beggs, P.J., Medek, D., Haberle, S.G., Newnham, R.M., Johnston, F.H., Jaggard, A.K., Campbell, B., Burton, P.K., Katelaris, C.H., Newbigin, E., Thibaudon, M., Davies, J.M., 2018. Dynamic ecological observations from satellites inform aerobiology of allergenic grass pollen. Sci. Total Environ. 633, 441–451. https://doi.org/10.1016/j.scitotenv.2018.03.191

El-Moslimany, A., 2019. Reduced Poaceae pollen under conditions of severe summer drought in the Middle East: Implications for rainfall seasonality in pollen diagrams. Rev. Palaeobot. Palynol. 271, 104068. https://doi.org/10.1016/j.revpalbo.2019.04.007

Fröhlich-Nowoisky, J., Kampf, C.J., Weber, B., Huffman, J.A., Pöhlker, C., Andreae, M.O., Lang-Yona, N., Burrows, S.M., Gunthe, S.S., Elbert, W., Su, H., Hoor, P., Thines, E., Hoffmann, T., Després, V.R., Pöschl, U., 2016. Bioaerosols in the Earth system: Climate, health, and ecosystem interactions. Atmos. Res. 182, 346–376. https://doi.org/10.1016/j.atmosres.2016.07.018

Gabarra, E., Belmonte, J., Canela, M., 2002. Aerobiological behaviour of Platanus L. pollen in Catalonia (North-East Spain). Aerobiologia. 18, 185–193. https://doi.org/10.1023/A:1021370724043

Galán, C., Ariatti, A., Bonini, M., Clot, B., Crouzy, B., Dahl, A., Fernández-González, D., Frenguelli, G., Gehrig, R., Isard, S., Levetin, E., Li, D.W., Mandrioli, P., Rogers, C.A., Thibaudon, M., Sauliene, I., Skjoth, C., Smith, M., Sofiev, M., 2017. Recommended terminology for aerobiological studies. Aerobiologia. 33, 293–295. https://doi.org/10.1007/s10453-017-9496-0

Galán, C., Cariñanos, P., Alcázar, P., Domínguez-Vilches, E., 2007. Spanish Aerobiology Network (REA): Management and Quality Manual. Servicio de Publicaciones Universidad de Córdoba, Córdoba.

Galán, C., Smith, M., Damialis, A., Frenguelli, G., Gehrig, R., Grinn-Gofroń, A., Kasprzyk, I., Magyar, D., Oteros, J., Šaulienė, I., Thibaudon, M., Sikoparija, B., 2021. Airborne fungal spore monitoring: between analyst proficiency testing. Aerobiologia. 1–11. https://doi.org/10.1007/s10453-021-09698-4

Galán, C., Smith, M., Thibaudon, M., Frenguelli, G., Oteros, J., Gehrig, R., Berger, U., Clot, B., Brandao, R., 2014. Pollen monitoring: minimum requirements and reproducibility of analysis. Aerobiologia. 30, 385–395.

https://doi.org/10.1007/s10453-014-9335-5

García-Mozo, H., 2017. Poaceae pollen as the leading aeroallergen worldwide: A review. Allergy 72, 1849–1858. https://doi.org/10.1111/all.13210

García-Mozo, H., Chuine, I., Aira, M.J., Belmonte, J., Bermejo, D., Díaz de la Guardia, C., Elvira, B., Gutiérrez, M., Rodríguez-Rajo, J., Ruiz, L., Trigo, M.M., Tormo-Molina, R., Valencia, R., Galán, C., 2008. Regional phenological models for forecasting the start and peak of the Quercus pollen season in Spain. Agric. For. Meteorol. 148, 372–380. https://doi.org/10.1016/j.agrformet.2007.09.013

García-Mozo, H., Dominguez-Vilches, E., Galan, C., 2007. Airborne allergenic pollen in natural areas: Hornachuelos Natural Park, Cordoba, Southern Spain. Ann. Agric. Environ. Med. 14, 63–69.

García-Mozo, H., Galán, C., Belmonte, J., Bermejo, D., Candau, P., Díaz de la Guardia, C., Elvira, B., Gutiérrez, M., Jato, V., Silva, I., Trigo, M.M., Valencia, R., Chuine, I., 2009. Predicting the start and peak dates of the Poaceae pollen season in Spain using process-based models. Agric. For. Meteorol. 149, 256–262. https://doi.org/10.1016/j.agrformet.2008.08.013

García-Mozo, H., Pérez-Badia, R., Fernández-González, F., Galán, C., 2006. Airborne pollen sampling in Toledo, Central Spain. Aerobiologia. 22, 55–66. https://doi.org/10.1007/s10453-005-9015-6

Grinn-Gofroń, A., Rapiejko, P., 2009. Occurrence of Cladosporium spp. and Alternaria spp. spores in Western, Northern and Central-Eastern Poland in 2004–2006 and relation to some meteorological factors. Atmos. Res. 93, 747–758. https://doi.org/10.1016/J.ATMOSRES.2009.02.014

Hernández-Ceballos, M.A., García-Mozo, H., Galán, C., 2015. Cluster analysis of intradiurnal holm oak pollen cycles at peri-urban and rural sampling sites in southwestern Spain. Int. J. Biometeorol. 59, 971–982. https://doi.org/10.1007/s00484-014-0910-9

Hirst, J.M., 1952. An automatic volumetric spore trap. Ann. Appl. Biol. 39, 257–265. https://doi.org/10.1111/j.1744-7348.1952.tb00904.x

Hjort, J., Hugg, T.T., Antikainen, H., Rusanen, J., Sofiev, M., Kukkonen, J., Jaakkola, M.S., Jaakkola, J.J.K., 2016. Fine-Scale Exposure to Allergenic Pollen in the Urban Environment: Evaluation of Land Use Regression Approach. Environ. Health Perspect. 124, 619–626. https://doi.org/10.1289/ehp.1509761

Junger, W.L., Ponce de Leon, A., 2015. Imputation of missing data in time series for air pollutants. Atmos. Environ. 102, 96–104. https://doi.org/10.1016/j.atmosenv.2014.11.049

Junta de Andalucía, 2011. Red de Información Ambiental de Andalucía (REDIAM) [WWW Document]. URL https://www.juntadeandalucia.es/medioambiente/site/rediam (accessed 2.6.20).

Karatzas, K.D., 2009. Informing the public about atmospheric quality: Air pollution and pollen. Allergo J. 18, 212–217. https://doi.org/10.1007/BF03362059

Kasprzyk, I., Walanus, A., 2014. Gamma, Gaussian and logistic distribution models for airborne pollen grains and fungal spore season dynamics. Aerobiologia. 30, 369–383. https://doi.org/10.1007/s10453-014-9332-8

Lara, B., Rojo, J., Fernández-González, F., González-García-Saavedra, A., Serrano-Bravo, M.D., Pérez-Badia, R., 2020. Impact of Plane Tree Abundance on Temporal and Spatial Variations in Pollen Concentration. Forests 11, 817. https://doi.org/10.3390/f11080817

Lara, B., Rojo, J., Fernández-González, F., Pérez-Badia, R., 2019. Prediction of airborne pollen concentrations for the plane tree as a tool for evaluating allergy risk in urban green areas. Landsc. Urban Plan. 189, 285–295. https://doi.org/10.1016/j.landurbplan.2019.05.002

Lehmann, T.M., Gönner, C., Spitzer, K., 1999. Survey: Interpolation methods in medical image processing. IEEE Trans. Med. Imaging 18, 1049–1075. https://doi.org/10.1109/42.816070

Luedeling, E., Kunz, A., Blanke, M.M., 2013. Identification of chilling and heat requirements of cherry trees-a statistical approach. Int. J. Biometeorol. 57, 679–689. https://doi.org/10.1007/s00484-012-0594-y

Navares, R., Aznarte, J.L., 2019. Geographical imputation of missing poaceae pollen data via convolutional neural networks. Atmosphere (Basel). 10, 717–727. https://doi.org/10.3390/atmos10110717

Navares, R., Aznarte, J.L., 2017. Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. Int. J. Biometeorol. 61, 647–656. https://doi.org/10.1007/s00484-016-1242-8

Nilsson, S., Persson, S., 1981. Tree pollen spectra in the Stockholm region (Sweden), 1973-1980. Grana 20, 179–182. https://doi.org/10.1080/00173138109427661

Orlandi, F., Oteros, J., Aguilera, F., Ben Dhiab, A., Msallem, M., Fornaciari, M., 2014. Design of a downscaling method to estimate continuous data from discrete pollen monitoring in Tunisia. Environ. Sci. Process. Impacts 16, 1716–1725. https://doi.org/10.1039/c4em00153b

Oteros, Jose, Galán, C., Alcázar, P., Domínguez-Vilches, E., 2013. Quality control in bio-monitoring networks, Spanish Aerobiology Network. Sci. Total Environ. 443, 559–565. https://doi.org/10.1016/J.SCITOTENV.2012.11.040

Oteros, J., García-Mozo, H., Vázquez, L., Mestre, A., Domínguez-Vilches, E., Galán, C., 2013. Modelling olive phenological response to weather and topography. Agric. Ecosyst. Environ. 179, 62–68. https://doi.org/10.1016/j.agee.2013.07.008

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M., 2015. Automatic and Online Pollen Monitoring. Int. Arch. Allergy Immunol. 167, 158–166. https://doi.org/10.1159/000436968

Oteros, J., Sofiev, M., Smith, M., Clot, B., Damialis, A., Prank, M., Werchan, M., Wachter, R., Weber, A., Kutzora, S., Heinze, S., Herr, C.E.W., Menzel, A., Bergmann, K.-C., Traidl-Hoffmann, C., Schmidt-Weber, C.B., Buters, J.T.M., 2019. Building an automatic pollen monitoring network (ePIN): Selection of optimal sites by clustering pollen stations. Sci. Total Environ. 688, 1263–1274. https://doi.org/10.1016/J.SCITOTENV.2019.06.131

Oteros, J., Weber, A., Kutzora, S., Rojo, J., Heinze, S., Herr, C., Gebauer, R., Schmidt-Weber, C.B., Buters, J.T.M., 2020. An operational robotic pollen monitoring network based on automatic image recognition. Environ. Res.

191, 110031. https://doi.org/10.1016/j.envres.2020.110031

Pérez-Badía, R., Rapp, A., Morales, C., Sardinero, S., Galán, C., García-Mozo, H., 2010. Pollen spectrum and risk of pollen allergy in central Spain. Ann. Agric. Environ. Med. 17, 139–151.

Picornell, A., Buters, J., Rojo, J., Traidl-Hoffmann, C., Damialis, A., Menzel, A., Bergmann, K.C., Werchan, M., Schmidt-Weber, C., Oteros, J., 2019a. Predicting the start, peak and end of the Betula pollen season in Bavaria, Germany. Sci. Total Environ. 690, 1299–1309. https://doi.org/10.1016/J.SCITOTENV.2019.06.485

Picornell, A., Oteros, J., Trigo, M.M., Gharbi, D., Docampo, S., Melgar, M., Toro, F.J., García-Sánchez, J., Ruiz-Mata, R., Cabezudo, B., Recio, M., 2019b. Increasing resolution of airborne pollen forecasting at a discrete sampled area in the southwest Mediterranean Basin. Chemosphere 234, 668–681. https://doi.org/10.1016/j.chemosphere.2019.06.019

Picornell, A., Recio, M., Ruiz-Mata, R., García-Sánchez, J., Cabezudo, B., Trigo, M.M., 2020. Medium- and long-range transport events of Alnus pollen in western Mediterranean. Int. J. Biometeorol. 64, 1637–1647. https://doi.org/10.1007/s00484-020-01944-7

Picornell, A., Recio, M., Trigo, M.M., Cabezudo, B., 2019c. Preliminary study of the atmospheric pollen in Sierra de las Nieves Natural Park (Southern Spain). Aerobiologia. 35, 571–576. https://doi.org/10.1007/s10453-019-09591-1

Ribeiro, H., Cunha, M., Abreu, I., 2007. Definition of main pollen season using a logistic model. Ann. Agric. Environ. Med. 14, 259–264.

Ritenberga, O., Sofiev, M., Kirillova, V., Kalnina, L., Genikhovich, E., 2016. Statistical modelling of non-stationary processes of atmospheric pollution from natural sources: Example of birch pollen. Agric. For. Meteorol. 226–227, 96–107. https://doi.org/10.1016/j.agrformet.2016.05.016

Rivas-Martínez, S., Penas, Á., del Río, S., Díaz González, T.E., Rivas-Sáenz, S., 2017. Bioclimatology of the Iberian Peninsula and the Balearic Islands, in: Loidi, J. (Ed.), The Vegetation of the Iberian Peninsula. Springer, Cham, Utrecht, Netherlands, pp. 29–80. https://doi.org/10.1007/978-3-319-54784-8_2

Rojo, J., Orlandi, F., Pérez-Badia, R., Aguilera, F., Ben Dhiab, A., Bouziane, H., Díaz de la Guardia, C., Galán, C., Gutiérrez-Bustillo, A.M., Moreno-Grau, S., Msallem, M., Trigo, M.M.M., Fornaciari, M., 2016. Modeling olive pollen intensity in the Mediterranean region through analysis of emission sources. Sci. Total Environ. 551–552, 73–82. https://doi.org/10.1016/j.scitotenv.2016.01.193

Rojo, J., Oteros, J., Pérez-Badia, R., Cervigón, P., Ferencova, Z., Gutiérrez-Bustillo, A.M., Bergmann, K.C., Oliver, G., Thibaudon, M., Albertini, R., Rodríguez-De la Cruz, D., Sánchez-Reyes, E., Sánchez-Sánchez, J., Pessi, A.M., Reiniharju, J., Saarto, A., Calderón, M.C., Guerrero, C., Berra, D., Bonini, M., Chiodini, E., Fernández-González, D., García-Sánchez, J., Trigo, M.M., Myszkowska, D., Fernández-Rodríguez, S., Tormo-Molina, R., Damialis, A., Kolek, F., Traidl-Hoffmann, C., Severova, E., Caeiro, E., Ribeiro, H., Magyar, D., Makra, L., Udvardy, O., Alcázar, P., Galán, C., Borycka, K., Kasprzyk, I., Newbigin, E., Adams-Groom, B., Apangu, G.P., Frisk, C.A., Skjøth, C.A., Radišić, P., Šikoparija, B., Celenk, S., Schmidt-

Weber, C.B., Buters, J., 2019. Near-ground effect of height on pollen exposure. Environ. Res. 160–169. https://doi.org/10.1016/j.envres.2019.04.027

Rojo, Jesús, Picornell, A., Oteros, J., 2019. AeRobiology: the computational tool for biological data in the air. Methods Ecol. Evol. 10, 1371–1376. https://doi.org/10.1111/2041-210x.13203

Rojo, J., Picornell, A., Oteros, J., Werchan, M., Werchan, B., Bergmann, K.C., Smith, M., Weichenmeier, I., Schmidt-Weber, C.B., Buters, J., 2021. Consequences of climate change on airborne pollen in Bavaria, Central Europe. Reg. Environ. Chang. 21, 9. https://doi.org/10.1007/s10113-020-01729-z

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592. https://doi.org/10.1093/biomet/63.3.581

Ruiz-Valenzuela, L., Aguilera, F., 2018. Trends in airborne pollen and pollen-season-related features of anemophilous species in Jaen (south Spain): A 23-year perspective. Atmos. Environ. 180, 234–243. https://doi.org/10.1016/j.atmosenv.2018.03.012

Schouten, R.M., Lugtig, P., Vink, G., 2018. Generating missing values for simulation purposes: a multivariate amputation procedure. J. Stat. Comput. Simul. 88, 2909–2930. https://doi.org/10.1080/00949655.2018.1491577

Skjøth, C.A., Damialis, A., Belmonte, J., De Linares, C., Fernández-Rodríguez, S., Grinn-Gofroń, A., Jędryczka, M., Kasprzyk, I., Magyar, D., Myszkowska, D., Oliver, G., Páldy, A., Pashley, C.H., Rasmussen, K., Satchwell, J., Thibaudon, M., Tormo-Molina, R., Vokou, D., Ziemianin, M., Werner, M.,

2016. Alternaria spores in the air across Europe: abundance, seasonality and relationships with climate, meteorology and local environment. Aerobiologia. 32, 3–22. https://doi.org/10.1007/s10453-016-9426-6

Stekhoven, D.J., Buhlmann, P., 2012. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597

Su, Y.-S., Gelman, A., Hill, J., Yajima, M., 2011. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. J. Stat. Softw. 45, 1–31. https://doi.org/10.7916/D8VQ3CD3

Todorov, V., 2020. rrcovNA: Scalable Robust Estimators with High Breakdown Point for Incomplete Data.

Valencia, J.A., Astray, G., Fernández-González, M., Aira, M.J., Rodríguez-Rajo, F.J., 2019. 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. https://doi.org/10.1007/s00484-019-01688-z

van Buuren, S., Groothuis-Oudshoorn, K., 2011. MICE: Multivariate imputation by chained equations in R. J. Stat. Softw. 45, 1–68. https://doi.org/10.18637/jss.v045.i03

Velasco-Jiménez, M.J., Alcázar, P., Domínguez-Vilches, E., Galán, C., 2013. Comparative study of airborne pollen counts located in different areas of the city of Córdoba (south-western Spain). Aerobiologia. 29, 113–120. https://doi.org/10.1007/s10453-012-9267-x

Verstraeten, W.W., Kouznetsov, R., Hoebeke, L., Bruffaerts, N., Sofiev, M., Delcloo, A.W., 2021. Modelling grass pollen levels in Belgium. Sci. Total Environ. 753, 141903. https://doi.org/10.1016/j.scitotenv.2020.141903