

RESEARCH ARTICLE

A new method based on physical patterns to impute aerobiological datasets

Sofia Tagliaferro¹, Adrián Corrochano², Pierpaolo Marchetti¹, Alessandro Marcon¹*, Soledad Le Clainche²

1 Unit of Epidemiology and Medical Statistics, Department of Diagnostics and Public Health, University of Verona, Verona, Italy, **2** School of Aerospace Engineering, Universidad Politécnica de Madrid, Madrid, Spain

* These authors contributed equally to this work.

* alessandro.marcon@univr.it



OPEN ACCESS

Citation: Tagliaferro S, Corrochano A, Marchetti P, Marcon A, Le Clainche S (2024) A new method based on physical patterns to impute aerobiological datasets. PLoS ONE 19(11): e0314005. <https://doi.org/10.1371/journal.pone.0314005>

Editor: Rajeev Singh, Satyawati College, University of Delhi, INDIA

Received: June 3, 2024

Accepted: November 4, 2024

Published: November 19, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0314005>

Copyright: © 2024 Tagliaferro et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All the materials underlying the results presented in the study are available from the Modelflows-app website (<https://modelflows.github.io/modelflowsapp/airpollution/>). These include the original pollen datasets

Abstract

Limited research has assessed the accuracy of imputation methods in aerobiological datasets. We conducted a simulation study to evaluate, for the first time, the effectiveness of Gappy Singular Value Decomposition (GSVD), a data-driven approach, comparing it with the moving mean interpolation, a statistical approach. Utilizing complete pollen data from two monitoring stations in northeastern Italy for 2022, we randomly generated missing data considering the combination of various proportions (5%, 10%, 25%) and gap lengths (3, 5, 7, 10 days). We imputed 4800 time series using the GSVD algorithm, specifically implemented for this study, and the moving mean algorithm of the “AeRobiology” R package. We assessed imputation accuracy by calculating the Root Mean Square Error and employed multiple linear regression models to identify factors independently affecting the error (e.g. pollen variability, simulation settings). The results showed that the GSVD was as good as the well-established moving mean method and demonstrated its strong generalization capabilities across different data types. However, the imputation error was primarily influenced by pollen characteristics and location, regardless of the imputation method used. High variability in pollen concentrations and the distribution of missing data negatively affected imputation accuracy. In conclusion, we introduced and tested a novel imputation method, demonstrating comparable performance to the statistical approach in aerobiological data reconstruction. These findings contribute to advancing aerobiological data analysis, highlighting the need for improving imputation methods.

Introduction

Aerobiology is a recent discipline focusing on atmospheric bioaerosols, such as pollen and spores [1, 2]. Its interdisciplinary approach allows for the examination of the impacts of climate change and the development of innovative methodologies aimed at managing allergic diseases [3]. Aerobiological data are typically measured on daily basis and are provided by local/national monitoring networks. Despite the existence of automatic sampling devices, current monitoring practices primarily rely on manual samplers, introducing the possibility of systematic errors [4–6].

downloaded from POLLnet (<https://pollnet.isprambiente.it/>), the R and Python codes to generate and input missing data, sample gappy datasets, a brief overview of the paper, and a video explanation of the methodology.

Funding: A.M. received grants to conduct the MEETOUT study from the European Union through the Italian Ministry of University and Research under the ESF REACT-EU Green and Innovation funding programme (Ministerial Decree 1061/2021) and the NextGenerationEU funding programme (Ministerial Decree 737/2021). Article processing charges were supported by the special fund at the University of Verona dedicated to Open Access publications. S.L.C. and A.C. acknowledge the grants PID2023-1477900B-I00, TED2021-129774B-C21 and PLEC2022-009235 funded by MCIN/AEI/10.13039/501100011033 and by the European Union “NextGenerationEU”/PRTR. The authors acknowledge the MODELAIR and ENCODING projects that have received funding from the European Union’s Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101072559 and 101072779, respectively. The results of this publication reflect only the authors view and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them. A.C. acknowledges the support of Universidad Politécnica de Madrid, under the program ‘Programa Propio’. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

Pollen time series are frequently incomplete due to malfunctions and maintenance of the monitoring stations [5], as well as voluntary interruptions in periods considered irrelevant for the measures. Consequently, the presence of missing data in aerobiological datasets is common, prompting the need for imputation methods. Traditional methods, as omitting to assign values to missing data, may lead to underestimation errors [1, 7, 8].

Statistical and artificial intelligence methodologies have been implemented for data imputation. Statistical approaches such as linear interpolation [9–12], cubic spline interpolation [12], the Gaussian method [13] or averaging values from other years for each day with missing data [9], are commonly used in aerobiological studies. The availability of pre-set statistical software packages facilitates the application of the most common methodologies used in data imputation [1]. Recently, the “AeRobiology” R package was developed specifically to manage and visualize aerobiological data, as well as to impute missing data [6]. In this package different interpolation methods are implemented, including linear, moving mean, spline, time series analysis, and nearby locations interpolation. The moving mean method is a statistical univariate approach. It consists of filling in missing values by averaging nearby data within a symmetrical interval that is twice the length of the gap [6].

In the last years, computational intelligence techniques have gained popularity in pollen time series analysis [5], but their application in missing data imputation is less explored. Convolutional Neural Networks [5], Denoising Convolutional Auto-encoder [13], and k-Nearest Neighbours algorithm [14] are among the approaches used. Natural systems are physical (spatio-temporal) systems characterised by dominant non-linear structures that evolve over time (such as seasonality or climate variations) that are unknown. Identifying data tendencies connected to physics enables generalization for application across various fields [15–17]. Machine learning tools could be useful to repair corrupted or incomplete datasets, using the relevant spatio-temporal information directly from the data. Of these, the Singular Value Decomposition (SVD) is a data-driven multivariate method, useful for post-processing and handling data. The SVD, based on simple linear algebra, is the primary technique behind many dimensionality reduction methods, such as the Principal Component Analysis. The SVD method is able to recognise and extract the relevant spatio-temporal information directly from the data, removing noise and filtering out spatial redundancies, thus leading to dimensionality reduction. To address missing data, the Gappy SVD (GSVD) has been implemented, utilizing SVD properties to iteratively repair and reconstruct datasets. This algorithm has already been successfully applied to reconstruct fluid flow [16, 18] or oceanographic datasets [19], but it has never been tested on aerobiological datasets.

To assess imputation accuracy, simulation studies are conducted by generating missing data scenarios in complete datasets and comparing simulated and observed values. While simulation studies on environmental datasets have been widely explored (e.g. meteorological, hydrological, and air pollution data) [20–23], the challenges of imputation in aerobiological data remain less studied. Picornell et al. tested the ability of the different interpolation methods implemented in the “AeRobiology” R package, simulating random missing data in patterns of 3, 5, 7, and 10 consecutive days in different pollen seasonal periods (pre-season, pre-peak, peak, post-peak, and post-season) [1]. Navares et al. evaluated the performance of geographical imputation via Convolutional Neural Networks, generating 10, 20, and 30% of missing values in all periods, peak and off-peak season [5].

This paper introduces a novel implementation of the Gappy Singular Value Decomposition (GSVD) algorithm, a data-driven method, specifically tailored for the application to aerobiological datasets in this study. The imputation accuracy of this method was compared to a well-known statistical method, the moving mean algorithm.

Materials and methods

Aerobiological data

POLLnet is the aerobiological monitoring network of the National System for Environmental Protection (SNPA) of Italy, which aggregates aerobiological monitoring data measured by regions and provinces into a nationwide database (<https://pollnet.isprambiente.it/>). The network's monitoring follows the European Standard UNI EN 16868 2019, using Hirst-type volumetric samplers with a calibrated pump aspirating 10 l/min of air in 24 hours. Airborne particles are captured on a rotating metallic drum with an adhesive tape. The sampling drum is extracted every seven days, and the tape is cut into fragments corresponding to each monitoring day. These fragments are then examined under a microscope at 400× magnification by a specialized technician, and daily pollen grains are counted based on their morphological characteristics. The count is recorded as the number of pollen grains per cubic meter of air (p/m^3) [24, 25].

We constructed the aerobiological datasets using RStudio version 4.2.2 [26]. For the study purposes, we selected two monitoring stations representing different environments in the Northeast of Italy (Fig 1): VI1 in Vicenza, lowlands with continental climate, and BZ2 in Bolzano, mountains with alpine climate.

We downloaded daily pollen concentrations for the period 2018–2022 using the “pollnet” R package (<https://rpubs.com/gbonafe/pollnet-data-extraction>). The dataset is available at <https://modelflows.github.io/modelflowsapp/airpollution/>. *Alnus* and Poaceae pollens were considered for the analysis due to their different seasonality, temporal distribution, and load characteristics (as shown from the 2022 time series in Fig 1).

We computed the start and end dates of the season for each pollen time series using the 95-percentage method (start: 2.5%; end: 97.5%) from the “AeRobiology” R package [6, 27]. This method was solely used to define convenient periods within the pollen seasons to generate

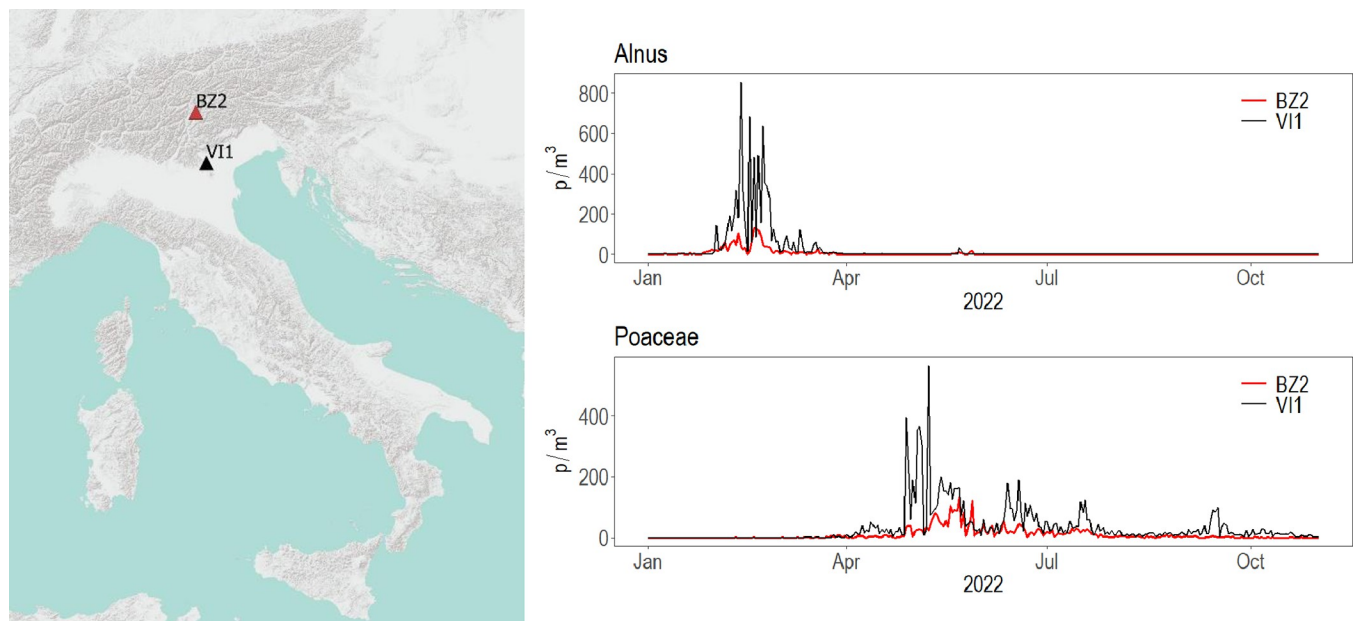


Fig 1. Location of the selected monitoring stations in the Northeast of Italy and the respective pollen time series for the year 2022. The map was produced using the QuickMapServices plugin (NextGIS, 2019) in QGIS software version 3.34.9 (QGIS Development Team. QGIS Geographic Information System. Open-Source Geospatial Foundation Project. <http://qgis.org>). The basemap used is ESRI Terrain (ESRI, Redlands, CA, USA). BZ2: Bolzano; VI1: Vicenza; p/m^3 : pollen/cubic meter.

<https://doi.org/10.1371/journal.pone.0314005.g001>

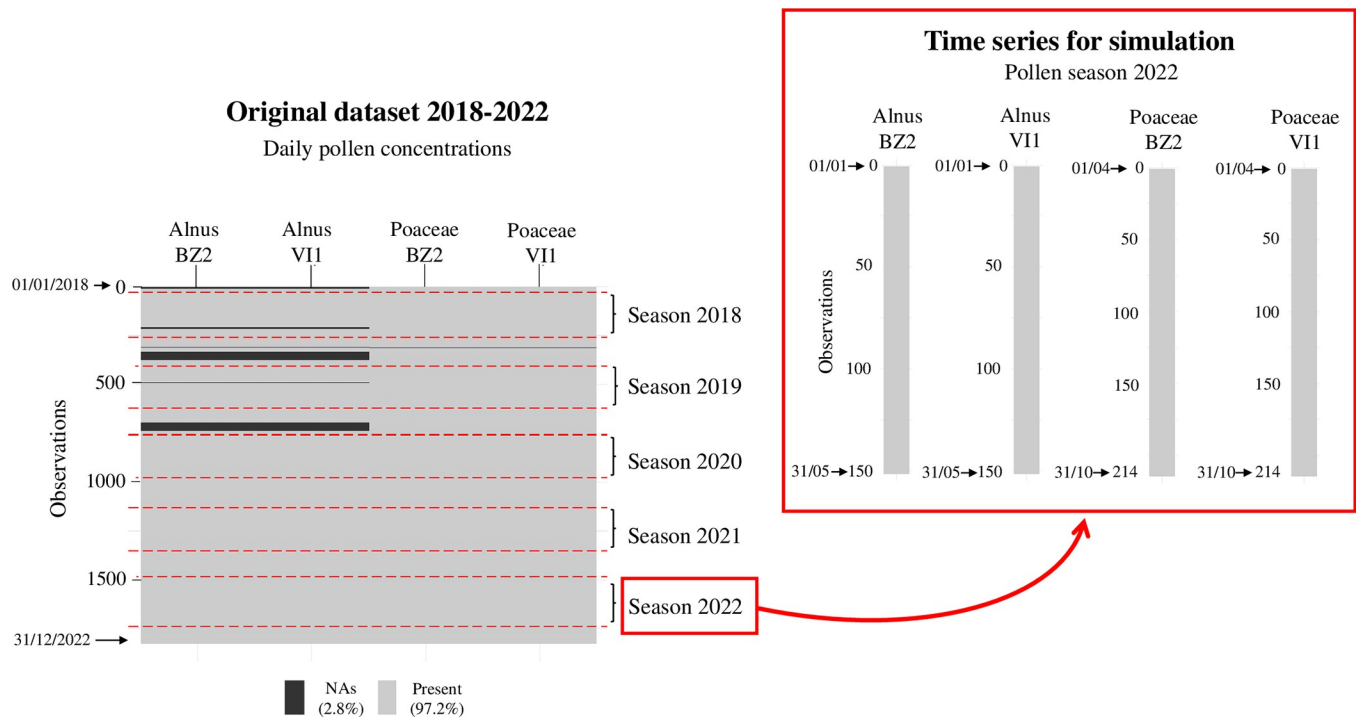


Fig 2. Scheme depicting the original dataset of daily pollen concentrations for the period 2018–2022 and time series extracted for the simulation study. BZ2: Bolzano; VII: Vicenza; NAs: missing data. Each season was obtained from the earlier start and later end day of the observed pollen seasons across the 2 monitoring stations: Season 2018 (start: 31/01/2018, *Alnus* BZ2; end: 17/09/2018, *Poaceae* VII); Season 2019 (start: 11/02/2019, *Alnus* VII; end: 17/09/2019, *Poaceae* VII); Season 2020 (start: 30/01/2020, *Alnus* BZ2; end: 05/09/2020, *Poaceae* VII); Season 2021 (start: 08/02/2021, *Alnus* VII; end: 14/09/2021, *Poaceae* VII); Season 2022 (start: 26/01/2022, *Alnus* BZ2; end: 08/10/2022, *Poaceae* VII).

<https://doi.org/10.1371/journal.pone.0314005.g002>

random missing data. Then, we examined missing data to identify the year with the most complete data coverage during the seasonal pollen period. The years from 2020 to 2022 showed no missing data at station VII, whereas station BZ2 had complete data throughout the period (2018–2022). We chose to simulate the pollen season of the year 2022 to ensure a complete data series for the preceding years, thus guaranteeing the applicability of the data-driven method. Fig 2 shows the structure of the original dataset and the time series extracted for the simulation study.

Descriptive statistics of pollen concentrations at each station in the pollen season 2022 were calculated: mean \pm Standard Deviation (SD), quartiles, coefficient of variation ($CV = (SD/\text{mean}) \times 100$) (%), and duration of the pollen season.

As the start/end dates varied depending on the monitoring station, for each pollen, we considered a common seasonal period in the year 2022 by extending the season to the first day of the month on which the minimum start date occurred and to the last day of the month on which the maximum end date occurred. As a result, the period considered for imputation was 01/01/2022 to 31/05/2022 for *Alnus* and 01/04/2022 to 31/10/2022 for *Poaceae*.

Methods of imputation investigated

We utilized the moving mean method of the “AeRobiology” R package as specifically developed for aerobiological datasets. The GSVD method used in this study was the algorithm originated from the ModelFLOWS-app (code available at <https://modelflows.github.io/modelflowsapp/airpollution/>), a novel software implementing modal decomposition

methods and hybrid machine learning tools to solve problems in complex nonlinear dynamical systems with application on patterns identification, data reconstruction, and data forecasting [16].

We initialised the GSVD algorithm assigning an initial value to the missing data. In this paper, the mean value of the time series (hereafter GSVD mean) and a linear interpolation between values of the time series (hereafter GSVD interp) were used for the initialisation. Then, SVD was applied to the initial dataset, as $X = U\Sigma V^T$, where the matrices U and V contain the modes (i.e. the spatio-temporal data decomposed by the Proper Orthogonal Decomposition mathematical approach) and the temporal coefficients, $()^T$ denotes the matrix transpose, and Σ is the diagonal matrix containing the singular values of the matrix X . The first modes contain the physical modes related to the problem, while the rest are related to noise, spatial redundancies or to fit this initial guess. Retaining the first number N of modes, which can be tuned, one can approximate the database as $X^* = U^*\Sigma^*V^{*T}$. The gaps of the original dataset were updated using the values of this approximation. Afterwards, SVD was applied again iteratively until the Mean Square Error (calculated as the ratio between the difference of the original and the reconstructed dataset and the total number of samples) of the gaps between two iterations is lower than a tolerance, set as 10^{-6} . More information about the algorithm and the implementation can be found in Díaz-Morales et al. (2024) and Hetherington et al. (2023, 2024) [15–17].

Simulation study

For each pollen type and station, we generated 12 simulation scenarios by combining 3 missing data proportions (5%, 10%, 25%) and 4 gap lengths (number of consecutive missing days: 3, 5, 7, 10 days). For each simulation scenario we obtained 100 simulated datasets. We randomly removed daily observed data from the complete pollen seasonal time series following the subsequent procedure (see Table 1):

- i. calculation of the number of days within the pollen season corresponding to the total proportions of NAs of 5%, 10%, and 25%;
- ii. calculation of the number of gaps for each gap length pattern (3, 5, 7, and 10 days) to approximate the total number of days with NAs from step i;
- iii. implementation of the algorithm to randomly remove data iteratively 100 times in RStudio, setting the number of consecutive days and the number of gaps from steps i and ii without overlapping gaps.

As a result, we obtained a total of 48 simulations (12 scenarios x 2 stations x 2 pollens), each with 100 time series for imputation. An example of the NAs generation process and resulting dataset is reported in Fig 3.

The RStudio code for the generation of missing data and examples of gappy datasets are available at <https://modelflows.github.io/modelflowsapp/airpollution/>.

Imputation and accuracy evaluation

To assess the accuracy of the imputation methods, we compared the reconstructed datasets to the observed time series, calculating the Root Mean Square Error (RMSE), i.e. the sum of the squared differences between the predicted and observed values divided by the total number of

Table 1. Settings of simulation scenarios and the resulting percentages of NAs obtained from simulations.

Pollen	Season duration (days)	NAs simulation settings				Resulting NAs	
		%	Total days	Gap length (consequent days)	Number of gaps	Total days	%
Alnus	151 (01 Jan—31 May)	5	7.55	3	3	9	6
				5	2	10	6.6
				7	1	7	4.6
				10	1	10	6.6
		10	15.1	3	5	15	9.9
				5	3	15	9.9
				7	2	14	9.3
				10	2	20	13.2
		25	37.75	3	13	39	25.8
				5	8	40	26.5
				7	5	35	23.2
				10	4	40	26.5
Poaceae	214 (01 Apr– 31 Oct)	5	10.7	3	4	12	5.6
				5	2	10	4.7
				7	2	14	6.5
				10	1	10	4.7
		10	21.4	3	7	21	9.8
				5	4	20	9.3
				7	3	21	9.8
				10	2	20	9.3
		25	53.5	3	18	54	25.2
				5	11	55	25.7
				7	8	56	26.2
				10	5	50	23.4

NAs: Missing data. The simulations are in total 12 for each pollen and station.

<https://doi.org/10.1371/journal.pone.0314005.t001>

observations (N) (1).

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Observed_i)^2}{N}} \tag{1}$$

As part of the “AeRobiology” R package, we executed the moving mean method using RStudio. We developed an algorithm to iteratively apply the “interpollen” function with “moving-mean” method to each column of individual datasets (pollen/station). After that, we merged the imputed dataset with the original corresponding pollen time series, and we implemented a function to iteratively calculate the RMSE between real data and the 100 replications of the simulated data. The final dataset contained the RMSE from all 100 simulations.

We implemented the GSVD algorithm in Python and ran it with Visual Studio Code version 1.86. As data-driven methods rely on extensive datasets to effectively capture data variability [28], we incorporated the 100 incomplete time series from each pollen, station, and simulation scenario into the original dataset including monitoring data spanning from 2018 to 2022. We studied different settings, changing the first initialisation of the values of the gaps and the number of modes, and evaluated the performance reconstruction of the gaps. Two imputation cases are shown in this paper for the sake of clarity, although other combinations

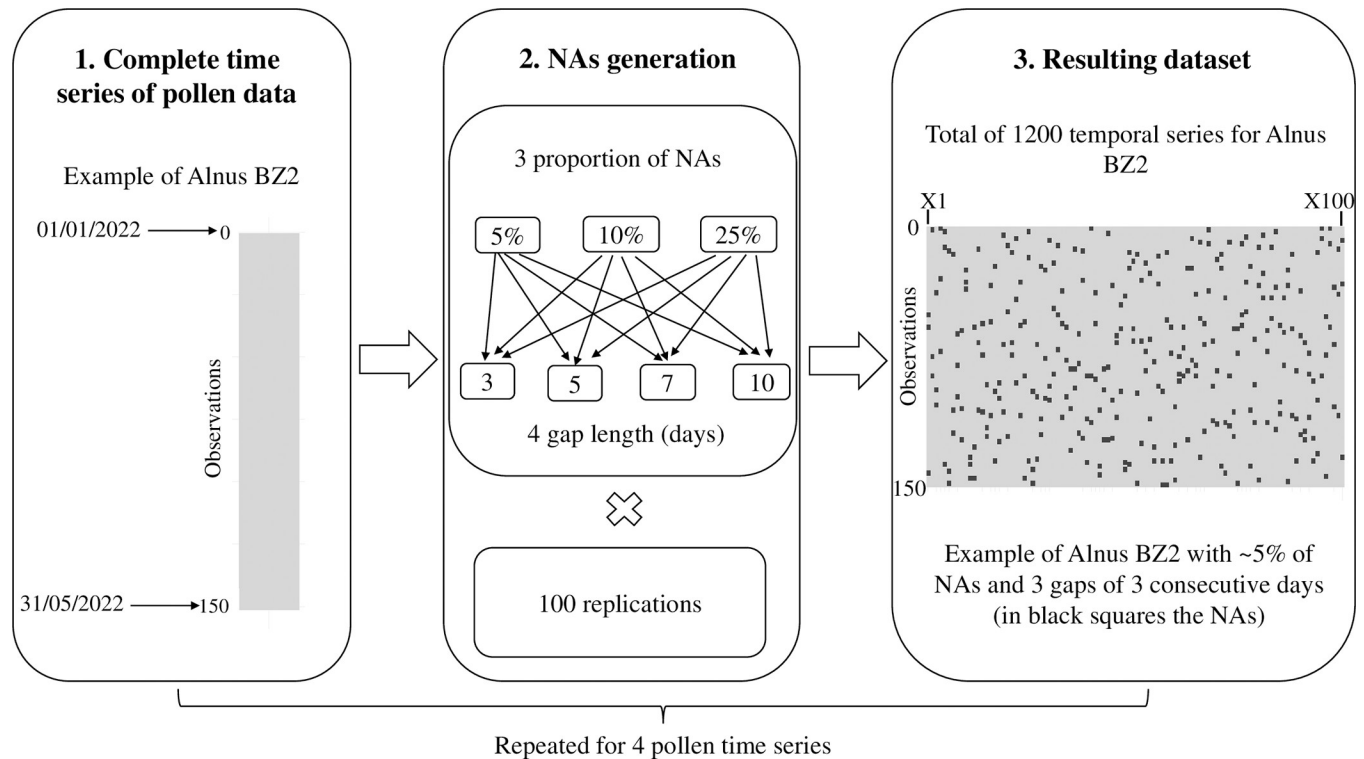


Fig 3. Example of generation of missing values (NAs) and resulting dataset for *Alnus* BZ2. BZ2: Bolzano.

<https://doi.org/10.1371/journal.pone.0314005.g003>

showed similar results: the GSVD mean 5modes and the GSVD interp 10modes. At the end of each imputation, the algorithm calculated the RMSE for each repetition and for each pollen time series and extracted the results as a dataset.

Finally, we merged the RMSE from the different imputations, and then we calculated the median RMSE for each imputation method and each combination of NAs.

Besides this, we checked if the natural variability of the pollen may affect the imputation process, as reported by Picornell et al. [1]. Indeed, pollen distribution, load, and seasonality differ according to the environment, climate, and phenology of the plant. All these factors may impact the imputation accuracy. So, for each pollen time series, we calculated the Variation index (VIn), an indicator of variability in pollen concentrations between consecutive days, based on Picornell et al (2021) [1]:

- the moving mean and SD on consecutive 2 days within the pollen season;
- the moving coefficients of variation (CV), as the ratio of the moving SD and moving mean;
- the VIn, defined as the average of the moving CV over the pollen season.

Then, we related the median RMSE and the VIn using boxplots to explore the relation between imputation accuracy and pollen variability. Moreover, we employed multiple linear regression models stratified by pollen and monitoring station (M1: *Alnus* and BZ2 station; M2: *Alnus* and VI1 station; M3: Poaceae and BZ2 station; M4: Poaceae and VI1 station) to further explore this relation. The dependent variable was the RMSE from the 100 replications by all the simulations (total of 4800 time series), which we log-transformed to satisfy the normality assumption in linear regression. In addition, we applied a robust estimator of standard errors to relax the homoskedasticity assumption. Model covariates included the imputation method,

Table 2. Descriptive statistics of pollen data in the pollen season 2022.

	Monitoring station	Mean \pm SD (p/m ³)	CV (%)	VIn (%)	1 st quartile (p/m ³)	Median (p/m ³)	3 rd quartile (p/m ³)	Maximum (p/m ³)	Duration of the season (days)
<i>Alnus</i>	BZ2	12.1 \pm 24.1	198.8	12.1	0.5	2.0	10.4	132.4	122
	VII	50.1 \pm 129.5	258.7	50.1	0.0	1.5	23.6	852.6	47
Poaceae	BZ2	14.5 \pm 22.1	152.9	14.5	1.0	4.7	20.2	135.4	150
	VII	47.2 \pm 71.3	150.9	47.2	10.3	21.4	51.8	564.4	180

SD: Standard Deviation; CV: Coefficient of Variation; VIn: Variation Index; p/m³: pollen/cubic meter; BZ2: Bolzano; VII: Vicenza.

<https://doi.org/10.1371/journal.pone.0314005.t002>

proportion of NAs, and gap length. We exponentiated the regression coefficient β ($\text{Exp}(\beta)$) to provide an estimate of the relative change in RMSE.

Results

Pollen data description

Table 2 reports descriptive statistics of pollen observations in the pollen season 2022.

For both pollen types, the mean and SD presented higher values in the VII monitoring station than in BZ2. For *Alnus*, the duration of the pollen season was shorter in VII compared to BZ2, but pollen variability was higher. Instead, Poaceae showed a shorter pollen season in BZ2 than in VII, but a higher variability in VII in terms of VIn.

Performance analysis

No specific pattern resulted in the distribution of median RMSE values for pollen and station by imputation methods (Fig 4).

The variability in the distribution of median RMSE was lower at the BZ2 monitoring station (*Alnus*: from 0.6 to 9.5 p/m³; Poaceae: from 0.9 to 8.2 p/m³) and higher at the VII monitoring station (*Alnus*: from 1.5 to 56.5 p/m³; Poaceae: from 4.1 to 27.6 p/m³).

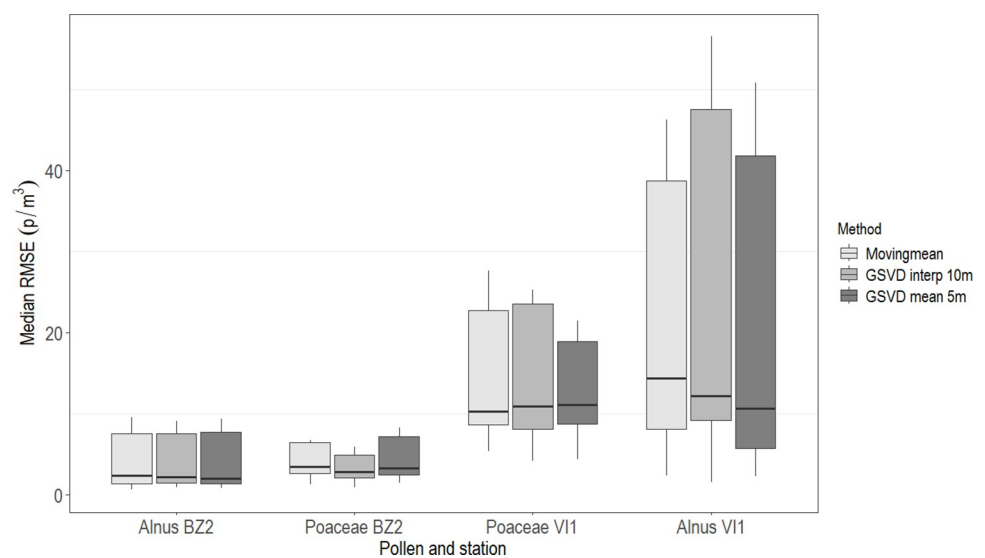


Fig 4. Distribution of the median Root Mean Square Error (RMSE) values for pollen/station by imputation method. BZ2: Bolzano; VII: Vicenza; GSVD: Gappy Singular Value Decomposition; p/m³: pollen/cubic meter. Each box represents the distribution of the median RMSE from the 12 simulations.

<https://doi.org/10.1371/journal.pone.0314005.g004>

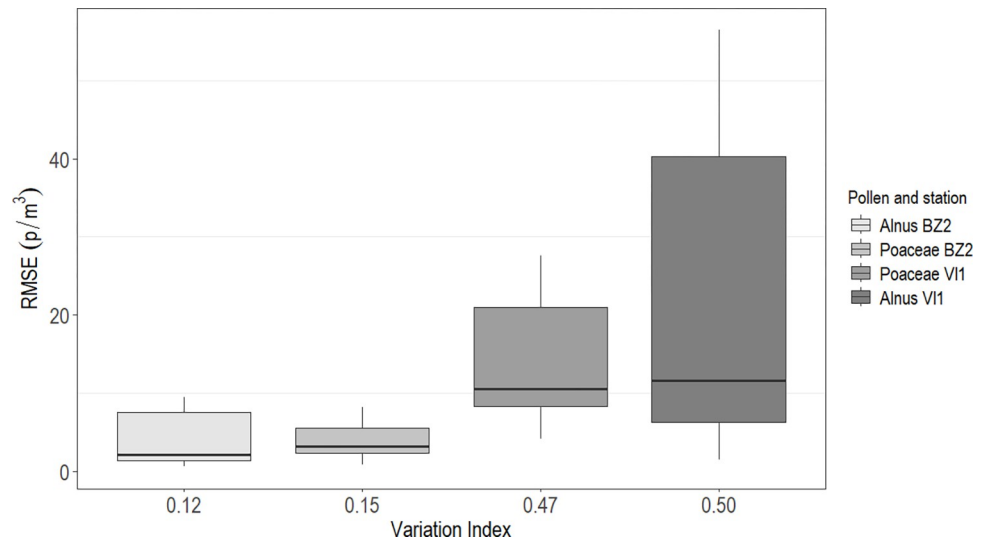


Fig 5. Distribution of median Root Mean Square Error (RMSE) for the Variation Index by pollen/station. BZ2: Bolzano; V11: Vicenza; p/m³: pollen/cubic meter. Each box represents the distribution of the median RMSE from the 12 simulations imputed with the 3 methods.

<https://doi.org/10.1371/journal.pone.0314005.g005>

When examining the relationship between the median RMSE values and VIn (Fig 5), a trend of increasing RMSE with higher VIn values emerged. Moreover, higher variability in the distribution of median RMSE values increased with higher VIn values.

Based on the results of the multiple linear regression models, there was large variability in imputation accuracy across the methods investigated, and none of them outperformed the others when adjusting for the simulation scenario (Table 3).

Moreover, no consistency was found within the GSVD imputation method, even showing contrasting results as in model M3. There was instead a consistent association between the

Table 3. Association estimates (Exp(β) representing ratios of geometric means) with 95%CI between the Root Mean Square Error (RMSE) and covariates (imputation method, % of NAs, and gap length).

	M1 (<i>Alnus</i> , BZ2)	M2 (<i>Alnus</i> , V11)	M3 (<i>Poaceae</i> , BZ2)	M4 (<i>Poaceae</i> , V11)
	Exp(β) (95%CI)	Exp(β) (95%CI)	Exp(β) (95%CI)	Exp(β) (95%CI)
Imputation method:				
Moving mean	Ref.	Ref.	Ref.	Ref.
GSVD interp 10m	1.20 (1.05–1.38)	1.00 (0.81–1.24)	0.88 (0.80–0.97)	0.94 (0.85–1.04)
GSVD mean 5m	1.03 (0.89–1.18)	1.13 (0.93–1.38)	1.10 (1.00–1.21)	0.92 (0.83–1.00)
% of NAs:				
5	Ref.	Ref.	Ref.	Ref.
10	2.12 (1.80–2.49)	3.76 (2.93–4.84)	1.94 (1.72–2.19)	1.80 (1.60–2.01)
25	5.85 (5.11–6.69)	9.86 (8.03–12.10)	3.84 (3.49–4.22)	3.74 (3.39–4.13)
Gap length (days):				
3	Ref.	Ref.	Ref.	Ref.
5	0.92 (0.79–1.08)	1.14 (0.88–1.46)	0.84 (0.75–0.94)	0.90 (0.80–1.02)
7	0.65 (0.56–0.76)	0.49 (0.39–0.63)	0.94 (0.85–1.03)	0.91 (0.82–1.01)
10	0.76 (0.66–0.88)	0.74 (0.60–0.92)	0.77 (0.69–0.85)	0.78 (0.71–0.87)

BZ2: Bolzano; V11: Vicenza; CI: Confidence Interval; GSVD: Gappy Singular Value Decomposition; NAs: missing values; Ref.: reference category.

<https://doi.org/10.1371/journal.pone.0314005.t003>

simulation settings and imputation accuracy. In fact, the RMSE increased with an increasing proportion of NAs across all models. Notably, the RMSE was 4 to 10 times higher when NAs were set to 25%, compared to the reference of 5%. On the contrary, the RMSE decreased with gap length, showing minimum values at 7 days (M1 and M2) and 10 days (M3 and M4).

Discussion and conclusions

A simulation study was conducted to compare the imputation accuracy of two methodologies, applying and evaluating for the first time the GSVD method to aerobiological datasets. Promising results emerged, demonstrating a similar performance of GSVD in comparison to the well-established moving mean method of the “AeRobiology” R package. However, it was found that both the inherent variability in observed pollen concentrations and the pattern of missing data had a more substantial impact on imputation accuracy within aerobiological datasets than the interpolation method applied. These findings contribute to filling the gap of knowledge in this field, considering the limited number of simulation studies conducted on pollen time series [1, 5, 13].

We compared univariate and multivariate methods of interpolation specifically focusing on aerobiological datasets. Previous simulation studies on other types of environmental data (e.g. hydrological, meteorological, air quality) have favoured multivariate methods, leveraging information from other temporal series, over univariate methods, which rely solely on the data series itself [20–22]. On one hand, we used the moving mean algorithm from the “AeRobiology” R package as univariate method, which was specifically designed for aerobiological datasets. This algorithm was identified within the package as the interpolation method with better performance, attributed to its reduced sensitivity to data availability, time series length, and fluctuations in pollen concentrations across consecutive days [1]. Its simplicity and increasing usage in aerobiological studies underscores its relevance and effectiveness in reconstructing time series. On the other hand, we used the GSVD algorithm as multivariate method, first evaluating its performance on aerobiological datasets. The potential of this method lies in its ability to reduce data dimensionality through data-driven decomposition, identifying the main data patterns related to physics without requiring any assumptions [15–17]. This makes it a promising tool for dataset reconstruction, as evidenced by its strong generalization capabilities across different types of data. However, more applications of the GSVD method on aerobiological data are needed to evaluate its effectiveness across diverse pollen types and environmental conditions. Indeed, the GSVD performance resulted similar to that of the statistical approach, with both methods exhibiting similarly unsatisfactory imputation accuracy in some settings. Moreover, the comparison of these two methods in our study revealed insights into the various factors influencing imputation performance. It suggested that the specific characteristics and requirements of the dataset may play a significant role in determining the most suitable interpolation approach.

The challenge of imputing missing data in aerobiological datasets is compounded by the complexity of plant phenology and pollen diffusion and advection mechanisms. Beyond their non-normal statistical distribution, each pollen type is influenced by local environmental and climatic conditions, resulting in differences in quantity, seasonality, and daily concentrations patterns [1]. Meteorological factors, particularly temperature and precipitation, are widely acknowledged to have the greatest influence on pollen variability, affecting both phenological phases and pollen behaviour in the atmosphere [29, 30]. Hence, the same pollen type may exhibit different distribution curves depending on the location characteristics [1]. Such variability has been related to decreased accuracy in imputation, as wider concentration ranges between consecutive days heighten the likelihood of errors during the imputation process [1].

This association has been observed in other environmental data as well [31]. Our findings align with Picornell et al. (2021) [1], indicating that higher variability in concentration (VIn) resulted in less accurate imputation results, both in terms of values and range of variability of the imputation error. Notably, measurements from the Vicenza station showed greater variability, likely attributable to the effect of continental climate characteristics of the lowlands on pollen, subjected to significant thermal fluctuations compared to alpine regions. Additionally, *Alnus* pollen generally displayed higher VIn values compared to Poaceae pollen, likely due to significant variability over a shorter season duration.

Besides the pollen type and location of the monitoring station, the pattern of missing data had the most substantial impact on imputation accuracy in our study. We generated missing data by introducing fixed consecutive-day gaps at various percentages in aerobiological datasets. The results showed a trend of increasing imputation error with higher percentages of NAs, regardless of the pollen/location. Our results align with the findings of Junger et al. (2015) concerning air pollution data, indicating that 5% of missing data yields satisfactory results, but accuracy decreases with more than 10% missing data [21]. In contrast, one study found opposing trends with increasing percentages of missing data for different meteorological variables [31], while another study observed no specific trend between missing data percentage and imputation error in aerobiological databases [5].

Regarding the gap length, our findings differ from those of Picornell et al. (2021) [1], as we observed that interpolation error decreases with longer gap lengths, depending on the pollen type. Specifically, the imputation error was minimum in datasets with gaps of 7 consecutive days for *Alnus*, and with gaps of 10 days for Poaceae, compared to gaps of 3 days. Despite the higher possibility of abrupt variations in longer gaps [1], the observed decrease in error with longer gaps can be attributed to the smoothing effect of interpolation. This effect leads to a reduction in the likelihood of generating peaks through interpolation, thereby minimizing errors. Notably, this effect appears to be more pronounced for pollens with wider season duration and less variability, as seen for Poaceae. The abundance of pollen-producing plants within this family, comprising over 120 genera in Italy, leads to high atmospheric pollen levels persisting over extended periods, thereby reducing day-to-day variability and smoothing peaks in pollen concentrations.

In conclusion, missing data resulting from manual measurement are common in aerobiological datasets [1, 5, 21]. Therefore, imputation remains the best solution for dealing with incomplete datasets and is useful for improving aerobiological analysis [1, 32]. In fact, even small gaps can distort estimates in environmental epidemiology or climatological studies [13]. Omitting to address missing data can result in significant errors in analysing pollen time series, which in turn can affect the definition of pollen seasonality [1, 7, 8]. We introduced and tested a novel method for missing data imputation in aerobiological research, demonstrating comparable performance to the moving mean method in data reconstruction. Both methods yielded favourable results, with the moving mean method being the simpler option. However, the imputation error remained unacceptable for certain pollen types and missing data scenarios. Additional research is required to investigate the application of the GSVD method across diverse pollen types and environmental conditions to draw a definitive conclusion. Furthermore, incorporating meteorological data into pollen datasets should be considered to improve imputation accuracy. Finally, there is a need to improve current imputation methods and develop more reliable techniques specifically tailored to pollen data, aiming to minimize the impact of temporal variability in pollen concentrations on imputation error.

Author Contributions

Conceptualization: Soledad Le Clainche.

Data curation: Sofia Tagliaferro, Adrián Corrochano.

Formal analysis: Sofia Tagliaferro, Adrián Corrochano.

Funding acquisition: Alessandro Marcon.

Investigation: Sofia Tagliaferro, Adrián Corrochano, Alessandro Marcon, Soledad Le Clainche.

Methodology: Sofia Tagliaferro, Adrián Corrochano, Pierpaolo Marchetti, Alessandro Marcon, Soledad Le Clainche.

Project administration: Alessandro Marcon.

Software: Adrián Corrochano.

Supervision: Pierpaolo Marchetti, Alessandro Marcon, Soledad Le Clainche.

Visualization: Sofia Tagliaferro, Pierpaolo Marchetti, Alessandro Marcon.

Writing – original draft: Sofia Tagliaferro.

Writing – review & editing: Sofia Tagliaferro, Adrián Corrochano, Pierpaolo Marchetti, Alessandro Marcon, Soledad Le Clainche.

References

1. Picornell A, Oteros J, Ruiz-Mata R, Recio M, Trigo MM, Martínez-Bracero M, et al. Methods for interpolating missing data in aerobiological databases. *Environmental Research*. 2021; 200: 111391. <https://doi.org/10.1016/j.envres.2021.111391> PMID: 34058184
2. Vélez-Pereira AM, De Linares C, Belmonte J. Aerobiological modeling I: A review of predictive models. *Science of The Total Environment*. 2021; 795: 148783. <https://doi.org/10.1016/j.scitotenv.2021.148783> PMID: 34243002
3. Tagliaferro S, Adani M, Pepe N, Briganti G, D'Isidoro M, Bonini M, et al. The impact of the spatial resolution of vegetation cover on the prediction of airborne pollen concentrations over northern Italy. *Agricultural and Forest Meteorology*. 2024; 355: 110153. <https://doi.org/10.1016/j.agrformet.2024.110153>
4. Matavulj P, Cristofori A, Cristofolini F, Gottardini E, Brdar S, Sikoparija B. Integration of reference data from different Rapid-E devices supports automatic pollen detection in more locations. *Science of The Total Environment*. 2022; 851: 158234. <https://doi.org/10.1016/j.scitotenv.2022.158234> PMID: 36007635
5. Navares R, Aznarte JL. Geographical Imputation of Missing Poaceae Pollen Data via Convolutional Neural Networks. *Atmosphere*. 2019; 10: 717. <https://doi.org/10.3390/atmos10110717>
6. Rojo J, Picornell A, Oteros J. AeRobiology: The computational tool for biological data in the air. Price S, editor. *Methods Ecol Evol*. 2019; 10: 1371–1376. <https://doi.org/10.1111/2041-210X.13203>
7. Smith M, Jäger S, Berger U, Šikoparija B, Hallsdottir M, Sauliene I, et al. Geographic and temporal variations in pollen exposure across Europe. *Allergy*. 2014; 69: 913–923. <https://doi.org/10.1111/all.12419> PMID: 24816084
8. Valipour Shokouhi B, De Hoogh K, Gehrig R, Eeftens M. Estimation of historical daily airborne pollen concentrations across Switzerland using a spatio-temporal random forest model. *Science of The Total Environment*. 2024; 906: 167286. <https://doi.org/10.1016/j.scitotenv.2023.167286> PMID: 37742957
9. Damialis A, Halley JM, Gioulekas D, Vokou D. Long-term trends in atmospheric pollen levels in the city of Thessaloniki, Greece. *Atmospheric Environment*. 2007; 41: 7011–7021. <https://doi.org/10.1016/j.atmosenv.2007.05.009>
10. González-Fernández E, Álvarez-López S, Garrido A, Fernández-González M, Rodríguez-Rajo FcoJ. Data mining assessment of Poaceae pollen influencing factors and its environmental implications. *Science of The Total Environment*. 2022; 815: 152874. <https://doi.org/10.1016/j.scitotenv.2021.152874> PMID: 34999063

11. Makra L, Matyasovszky I, Deák ÁJ. Trends in the characteristics of allergenic pollen circulation in central Europe based on the example of Szeged, Hungary. *Atmospheric Environment*. 2011; 45: 6010–6018. <https://doi.org/10.1016/j.atmosenv.2011.07.051>
12. Škoparija B, Marko O, Panić M, Jakovetić D, Radišić P. How to prepare a pollen calendar for forecasting daily pollen concentrations of Ambrosia, Betula and Poaceae? *Aerobiologia*. 2018; 34: 203–217. <https://doi.org/10.1007/s10453-018-9507-9>
13. Makra L, Matyasovszky I, Tusnady G, Ziska LH, Hess JJ, Nyúl LG, et al. A temporally and spatially explicit, data-driven estimation of airborne ragweed pollen concentrations across Europe. *Science of The Total Environment*. 2023; 905: 167095. <https://doi.org/10.1016/j.scitotenv.2023.167095> PMID: 37748607
14. Marchetti P, Pesce G, Villani S, Antonicelli L, Ariano R, Attena F, et al. Pollen concentrations and prevalence of asthma and allergic rhinitis in Italy: Evidence from the GEIRD study. *Science of The Total Environment*. 2017;584–585: 1093–1099. <https://doi.org/10.1016/j.scitotenv.2017.01.168> PMID: 28169023
15. Díaz-Morales P, Corrochano A, López-Martín M, Le Clainche S. Deep learning combined with singular value decomposition to reconstruct databases in fluid dynamics. *Expert Systems with Applications*. 2024; 238: 121924. <https://doi.org/10.1016/j.eswa.2023.121924>
16. Hetherington A, Corrochano A, Abadía-Heredia R, Lazpita E, Muñoz E, Díaz P, et al. ModelFLOWS-app: data-driven post-processing and reduced order modelling tools. *arXiv*; 2023. Available: <http://arxiv.org/abs/2305.17150>.
17. Hetherington A, Serfaty D, Corrochano A, Soria J, Clainche SL. Data repairing and resolution enhancement using data-driven modal decomposition and deep learning. 2024 [cited 8 Apr 2024]. <https://doi.org/10.48550/ARXIV.2401.11286>
18. Venturi D, Karniadakis GE. Gappy data and reconstruction procedures for flow past a cylinder. *J Fluid Mech*. 2004; 519: 315–336. <https://doi.org/10.1017/S0022112004001338>
19. Beckers JM, Rixen M. EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. *J Atmos Oceanic Technol*. 2003; 20: 1839–1856. [https://doi.org/10.1175/1520-0426\(2003\)020<1839:ECADFF>2.0.CO;2](https://doi.org/10.1175/1520-0426(2003)020<1839:ECADFF>2.0.CO;2)
20. Bleidorn MT, Pinto WDP, Schmidt IM, Mendonça ASF, Reis JATD. Methodological approaches for imputing missing data into monthly flows series. *Rev ambiente água*. 2022; 17: 1–27. <https://doi.org/10.4136/ambi-agua.2795>
21. Junger WL, Ponce De Leon A. Imputation of missing data in time series for air pollutants. *Atmospheric Environment*. 2015; 102: 96–104. <https://doi.org/10.1016/j.atmosenv.2014.11.049>
22. Nelsen B, Williams D, Williams G, Berrett C. An Empirical Mode-Spatial Model for Environmental Data Imputation. *Hydrology*. 2018; 5: 63. <https://doi.org/10.3390/hydrology5040063>
23. Plaia A, Bondi A. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*. 2006; 40: 7316–7330. <https://doi.org/10.1016/j.atmosenv.2006.06.040>
24. ARPAV. Il monitoraggio aerobiologico nel Veneto: i pollini allergenici. Padova: Agenzia Regionale per la Prevenzione e protezione Ambientale del Veneto; 2004. Available: <https://www.arpa.veneto.it/arpavinforma/pubblicazioni/il-monitoraggio-aerobiologico-nel-veneto-i-pollini-allergenici>.
25. Ogden EC, New York State Museum and Science Service, U.S. Atomic Energy Commission, editors. Manual for sampling airborne pollen. New York: Hafner Press; 1974.
26. RStudio Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available: <https://www.R-project.org/>.
27. Andersen TB. A model to predict the beginning of the pollen season. *Grana*. 1991; 30: 269–275. <https://doi.org/10.1080/00173139109427810>
28. Kasam AA, Lee BD, Paredis CJJ. Statistical methods for interpolating missing meteorological data for use in building simulation. *Build Simul*. 2014; 7: 455–465. <https://doi.org/10.1007/s12273-014-0174-7>
29. Blanco-Alegre C, Castro A, Calvo AI, Oduber F, Fernández-González D, Valencia-Barrera RM, et al. Towards a model of wet deposition of bioaerosols: The raindrop size role. *Science of The Total Environment*. 2021; 767: 145426. <https://doi.org/10.1016/j.scitotenv.2021.145426> PMID: 33550056
30. Schramm PJ, Brown CL, Saha S, Conlon KC, Manangan AP, Bell JE, et al. A systematic review of the effects of temperature and precipitation on pollen concentrations and season timing, and implications for human health. *Int J Biometeorol*. 2021; 65: 1615–1628. <https://doi.org/10.1007/s00484-021-02128-7> PMID: 33877430
31. Yozgatligil C, Aslan S, Iyigun C, Batmaz I. Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. *Theor Appl Climatol*. 2013; 112: 143–167. <https://doi.org/10.1007/s00704-012-0723-x>

32. Gehrig R, Clot B. 50 Years of Pollen Monitoring in Basel (Switzerland) Demonstrate the Influence of Climate Change on Airborne Pollen. *FrontAllergy*. 2021; 2: 677159. <https://doi.org/10.3389/falgy.2021.677159> PMID: [35387022](https://pubmed.ncbi.nlm.nih.gov/35387022/)