applied sciences

MDPI

Article

A Method of Pruning and Random Replacing of Known Values for Comparing Missing Data Imputation Models for Incomplete Air Quality Time Series

Luis Alfonso Menéndez García [1], Marta Menéndez Fernández [1,*], Violetta Sokoła-Szewioła [2], Laura Álvarez de Prado [1], Almudena Ortiz Marqués [1], David Fernández López [3] and Antonio Bernardo Sánchez [1]

[1] Department of Mining Technology, Topography and Structures, University of León, 24071 León, Spain; lmeneg00@estudiantes.unileon.es (L.A.M.G.); laura.alvarez@unileon.es (L.Á.d.P.); almudena.ortiz@unileon.es (A.O.M.); antonio.bernardo@unileon.es (A.B.S.)

[2] Faculty of Mining, Safety Engineering and Industrial Automation, Silesian University of Technology, Akademicka 2, 44-100 Gliwice, Poland; violetta.sokola-szewiola@polsl.pl

[3] Ingeniera de Recursos Mineroindustriales S.L., C/Ortega y Gasset, 4–Bajo, 24195 Villaobispo de las Regueras, León, Spain; davidfernandez@inremin.es

[*] Correspondence: marta.menendez@unileon.es

Abstract: The data obtained from air quality monitoring stations, which are used to carry out studies using data mining techniques, present the problem of missing values. This paper describes a research work on missing data imputation. Among the most common methods, the method that best imputes values to the available data set is analysed. It uses an algorithm that randomly replaces all known values in a dataset once with imputed values and compares them with the actual known values, forming several subsets. Data from seven stations in the Silesian region (Poland) were analyzed for hourly concentrations of four pollutants: nitrogen dioxide (NO_2), nitrogen oxides (NO_x), particles of 10 µm or less (PM_{10}) and sulphur dioxide (SO_2) for five years. Imputations were performed using linear imputation (LI), predictive mean matching (PMM), random forest (RF), k-nearest neighbours (k-NN) and imputation by Kalman smoothing on structural time series (Kalman) methods and performance evaluations were performed. Once the comparison method was validated, it was determine that, in general, Kalman structural smoothing and the linear imputation methods best fitted the imputed values to the data pattern. It was observed that each imputation method behaves in an analogous way for the different stations The variables with the best results are NO_2 and SO_2. The UMI method is the worst imputer for missing values in the data sets.

Keywords: imputation; linear imputation; predictive mean matching; random forest; k-nearest neighbours; Kalman smoothing; air quality; air pollution

check for updates

Citation: Menéndez García, L.A.; Menéndez Fernández, M.; Sokoła-Szewioła, V.; Álvarez de Prado, L.; Ortiz Marqués, A.; Fernández López, D.; Bernardo Sánchez, A. A Method of Pruning and Random Replacing of Known Values for Comparing Missing Data Imputation Models for Incomplete Air Quality Time Series. *Appl. Sci.* **2022**, *12*, 6465. https://doi.org/10.3390/app12136465

Received: 3 June 2022
Accepted: 23 June 2022
Published: 25 June 2022

1. Introduction

Air pollution is one of the problems affecting cities and industrialized areas [1,2] and causes the deaths of approximately 8.5 million people each year worldwide [3].

According to European legislation [4], administrations are obliged to make information on the state of air quality in their member states available to the public.

The use of data mining (KDD) techniques is very useful for data characterization and decision making [5–7], since air quality monitoring stations that measure pollutants generate a huge amount of data. This use has become widespread in recent years for air pollution problems, and one of the most important problems encountered is that of missing data or missing values [8]. Missing data are observations that are not available in a study's dataset, either because they have not been captured or because they were removed. This is a problem that arises very frequently [9,10]. It occurs most frequently in air pollutant research studies because the data are measured by air quality monitoring stations at regular time intervals and there may be reading or recording failures. These failures may be due to maintenance shutdowns, filter clogging, periodic calibrations, power failures, etc., resulting in the absence of measurements at certain time intervals [11–13].

This absence of data creates an added difficulty in scientific research [14], firstly, because of the absence of the data itself, which impoverishes the data as a whole [15,16] and, secondly, because most of the existing data analysis procedures are not designed or adapted for the absence of observations [17].

Improper handling of missing data can lead to erroneous subsequent statistical analysis and cause the conclusions drawn to be erroneous [18,19].

There are numerous studies in the literature on the prediction of missing values in air pollution data. Each of these studies employs different techniques to determine which one best fits their data, without being able to establish a single method that is best suited to solve the problem of missing value imputation [20]. Therefore, there is no golden rule that establishes the steps to determine the best imputation method. Sometimes very sophisticated methods produce worse approximations than simpler ones [21].

In a study [22] using a cellular neural network model, the model was compared with a multiple linear regression algorithm for the prediction of the missing data corresponding to PM_{10} and SO_2 in various regions of Turkey. Ref. [23] used IDW (inverse distance squared weighting) and mean value imputation to determine missing values in the prediction of BTEX (benzene, toluene, ethylbenzene and xylene) in two areas in Ontario, Canada. In paper [24], statistical imputation methods such as the generation of a Weibull distribution were applied to substitute for missing values in gaseous pollutant data sets. In another study, [25], five different methods–mean imputation, k-NN, conditional mean imputation, multiple imputation and Bayesian principal component analysis imputation–were used to reconstruct air quality data sets in Temuco, Chile. In Ref. [26], the method for the imputation of missing values of a variable was proposed as the mean of all observations taken at the same time during a year at the same station (hour mean method), the mean of all values taken at the same time at different stations (row mean method) and the mean between the previous and next known values (last and next method). For the imputation of missing data from an air quality study in London, ref. [27] used and compared the methods of linear interpolation, cubic spline interpolation, EWMA (exponentially weighted moving average), multivariate imputation from the mean and MICE (multiple imputation by chained equation). In [28], the authors used a new method to impute missing values in two sets of air quality data. They compared MTCAN (multi-directional temporal convolutional artificial neural network) with other known methods: SVR (regression support vector machine), recurrent neural networks (RNN) and convolutional neural networks (CNN). In ref. [29], a new imputation method for a large number of consecutive missing values of air pollutant measurements for the case of $PM_{2.5}$ (particles of 2.5 µm or less) in New York was proposed.

Sometimes researchers eliminate observations with missing values or a known imputation method is chosen without the certainty that the chosen method achieves a better, equal or worse approximation than other known methods in relation to the data pattern inherent in the set. There is no perfect method and it generally depends on the preferences of the researchers [20].

The aim of this study is to investigate which of the common missing data imputation methods best fits the available data pattern. For this purpose, missing data imputation simulations were performed with different methods as an application to a real case which is used as a first step for a further study on pollution prediction in the region of Silesia, Poland.

This study provides a relatively fast and efficient computational method to choose the imputation method among the best known ones or those to be tested in order to perform the imputations of missing values of data sets prior to other subsequent studies that require complete data sets, as is the case for time series. For this purpose, the method was applied to data from seven pollution stations to observe the performance of the method for comparison.

This methodology consists of taking a data set and eliminating the observations that contain at least one missing value in any of the variables so that a data set with all known observations is obtained. For each of the variables, known values are replaced by missing values in k groups of equal size. These replacements are random and in such a way that each observation of the variable is replaced only once by a missing value and assigned to one of the k groups. All observations end up being replaced only once. In each variable the replacement is performed independently of the rest; therefore, in the same observation index (row of a table of values) there may be variables that contain a missing value and others containing a known value. The missing values are imputed with each method in each round. Thus, the performance can be compared when imputing the same missing values by different methods. Another advantage is that, since the algorithm has replaced, imputed and compared all observations only once, the entire set of values is trained and validated so that the imputation methods demonstrate their ability to represent the data set and its inherent distribution under equal conditions for all imputation methods.

Studies related to the absence of missing data in air quality research usually consider the data to be Missing at Random (MAR) [23,30,31].

For each of the seven available stations, the following imputation methods were applied: linear imputation (LI), predictive mean matching (PMM), random forest (RF), unconditional mean (UMI), k-nearest neighbours (k-NN) and imputation by Kalman smoothing on structural time series (Kalman).

To evaluate the performance of each of the above methods and for each station, RMSE was used as a metric.

2. Materials and Methods

2.1. The Database

Observations of four air pollutants were used for this study: NO_2, NO_X, PM_{10} and SO_2.

The selected air pollutant measurement stations are located in the region of Silesia (Poland). The map in Figure 1 shows their geographical distribution in the region and Table 1 shows the geographical coordinates and altitude of each station.

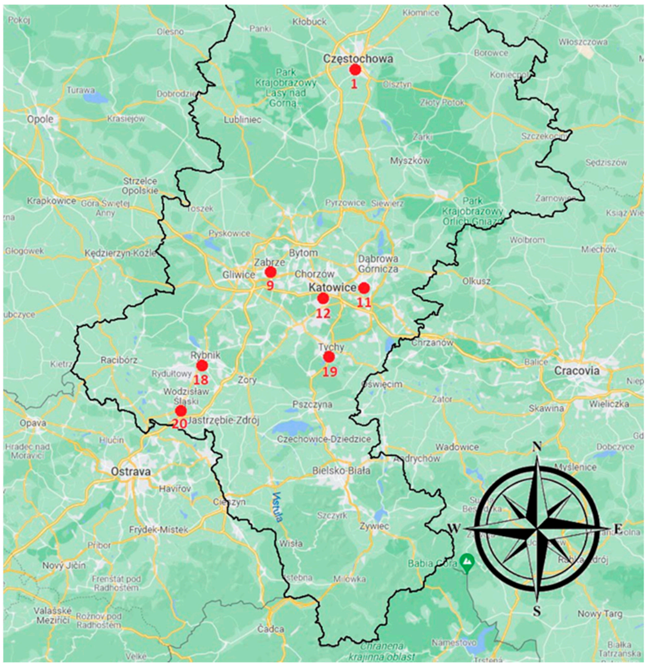

Figure 1. Location of stations in Silesia (Poland) map.

Table 1. Stations.

Station	Code	Site	Latitude (Degree)	Longitude (Degree)	Elevation (m)
ST 1	PL0184A	Czestochowa	50.836389	19.130111	265
ST 9	PL0242A	Zabrze	50.316500	18.772375	255
ST 11	PL0529A	Sosnowiec	50.285956	19.184399	250
ST 12	PL0008A	Katowice	50.264611	18.975028	273
ST 18	PL0239A	Rybnik	50.111181	18.516139	245
ST 19	PL0240A	Tychy	50.099903	18.990236	252
ST 20	PL0241A	Wodzislaw Slaski	50.007629	18.455548	271

The pollutant records belonging to these stations span from 1 January 2016 to 31 December 2020, i.e., a period of 5 full years, with the frequency of observations being hourly. Table 2 shows the total number of observations in the original dataset for each station and the number of missing observations for each pollutant and station.

Table 2. Total observations and missing values.

		ST 1	ST 9	ST 11	ST 12	ST 18	ST 19	ST 20
Total observations		43,494	43,784	43,370	43,641	43,681	43,582	43,768
Missing values	NO_2	357	343	116	311	392	62	222
	NO_X	358	344	115	312	396	62	235
	PM_{10}	174	605	173	867	238	106	168
	SO_2	419	310	255	281	203	91	150

2.2. Methodology

In order to obtain the most appropriate imputation method for the available data, the following procedure was followed, which can be generalized for j variables and k validation set. In this paper, the study was carried out for 4 pollutants (X_j), $j = 1,\ 2,\ 3,\ 4$, from 7 air quality measurement stations $\left(X_j\right)^s$ $s = 1,\ 2, \ldots,\ 7$ and 5 validation sets per pollutant (X_k^j), $k = 1,\ 2, \ldots,\ 5$.

The study was repeated in a similar way for each of the s stations and for each imputation method to be compared.

2.2.1. Pruning of Observations with Missing Values

For each of the stations belonging to the study, all the observations in which a missing value is found for any of the four pollutants were eliminated, i.e., if a value was missing for a given measurement of a pollutant on a given day, the record for that measurement of the four pollutants at that station was eliminated. This results in a data set that consists of 100% of real, observed values and has no empty observations. Table 3 shows the resulting number of observations for each station after pruning, and Figure 2 clarifies the procedure followed in a schematic way.

Table 3. Total observations after missing values purge.

ST 1	ST 9	ST 11	ST 12	ST 18	ST 19	ST 20
42,718	42,765	43,313	42,255	42,977	43,372	43,296

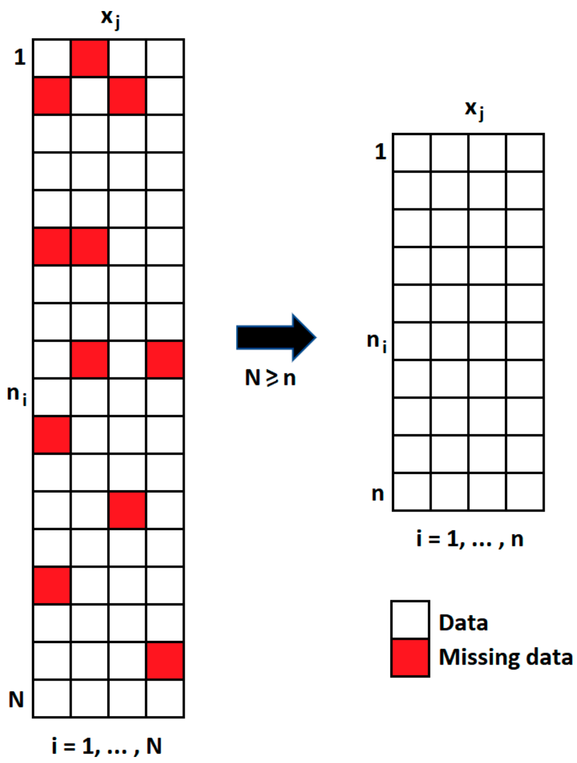

Figure 2. Purging missing values.

2.2.2. Formation of Training and Validation Sets

For each station and for each of the variables X_j, $j = 1, 2, 3, 4$ random extractions of values are performed to form training and validation sets to perform a k training and validation sets to perform a k-fold cross validation.

Extractions are made without replacement of $\frac{n}{k}$ observations of the variable X_j variable in such a way that the i indices of the observations extracted from one variable need not be equal to those of another variable for the same index k, as shown in Figure 3. In this way, known values are replaced by missing values, such that each known value is extracted once and only once among the sets and, after the extraction process, each known value is extracted once and only once among the K sets and, after the process of extractions, no value is left without being extracted and replaced by a missing value.

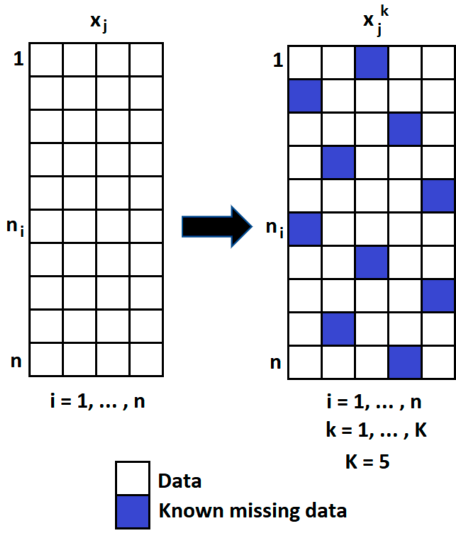

Figure 3. Extraction of known values and substitution for missing values.

The union of the values extracted in each round for each variable equals the pruned set of observations.

2.2.3. Imputation

Imputation is performed for each of the imputation methods studied on the set of values resulting from extraction k, with a training data set consisting of 80% of the data for each variable $\left(n - \frac{n}{K}\right)$. See Figure 4.

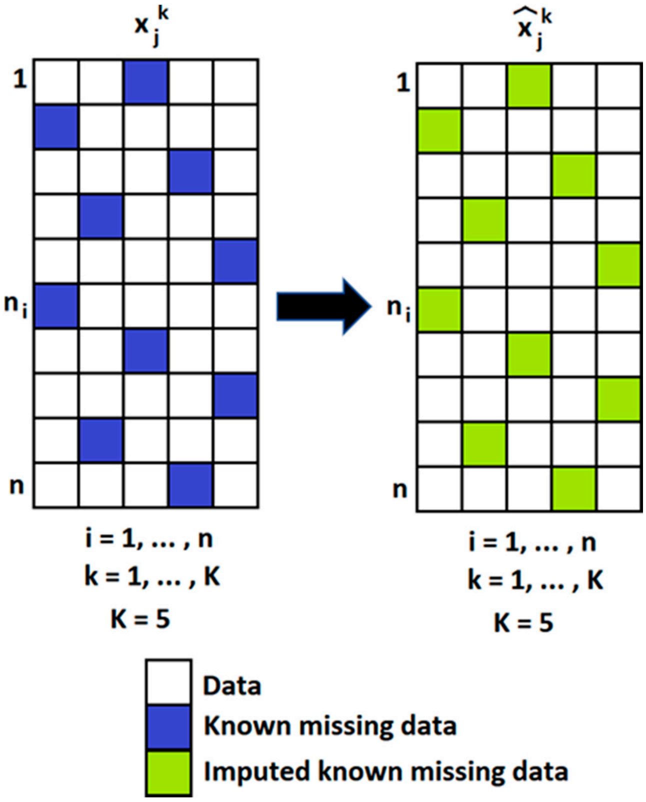

Figure 4. Imputation of known missing values.

In this way, all known values can be replaced by missing values and the imputation of all known values divided into *K* groups can be carried out and compared with 100% of the known values and determine a performance for each imputation model, variable and station. The missing values that are imputed in each round are the same for each imputation method. In this way, the performance of the methods can be compared with each other.

2.2.4. Measurement of Model Performance (Error Metrics)

To determine the validity of the imputation method, a comparison is made between the imputed missing value and the known missing value that was extracted. The performance of each imputation model and for each pair of sets (training/validation) is measured by RMSE error. To represent the performance of each model, the average RMSE value obtained from the average of each variable for each of the imputations is taken as the result. See Figure 5.

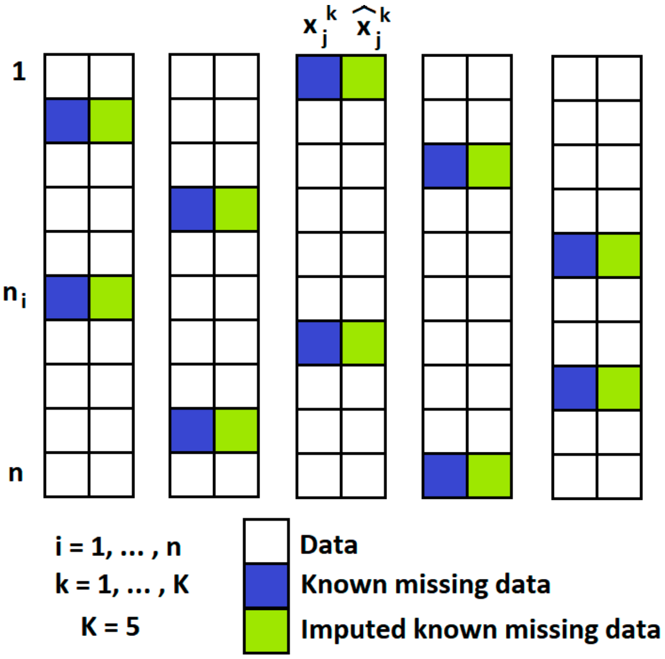

Figure 5. Comparison and measurement of performance.

Since the values in the study correspond to measurements of air pollutants and these must be greater than or equal to zero, imputation models that obtained any negative values were rejected.

In relation to possible outliers, a previous analysis was carried out detecting extreme values; however, these observations are validated data and therefore they are real and considered correct. Non-validated data were eliminated as they are not considered correct and were taken as missing values.

2.3. Imputation Models

2.3.1. Imputation by Linear Interpolation (LI)

The R package used to perform imputation by linear interpolation is imputeTS [32] and the na_interpolation function. The imputation of missing values by interpolation is a simple method and relatively good performance is obtained when the imputation is of isolated missing values of time series. The following paper, [33], gives a general overview of the package and the functions it provides and refs. [34,35] explain the theory and computation of the interpolation function.

2.3.2. Random Forest Imputation (RF)

The random forest algorithm for imputation used in this study is the one incorporated in the R package MICE [36], which is based on Breiman's algorithm and is detailed in [37,38]. This non-parametric method [39] is based on establishing a model for each of the features using the instances of the rest of the attributes and makes a prediction. In this way, the missing values for that feature are obtained. The process is repeated until the stop criterion is reached and this is performed in the same way for the rest of the attributes in an iterative way [40].

In this study the value of parameter ntree chosen was 10.

2.3.3. K-Nearest Neighbour Imputation (k-NN)

The k-NN algorithm used to perform the imputation is the one developed by [41] from R's VIM package and explained in [42]. This method identifies the k-nearest neighbour observations that have non-missing values for each feature, taking as distance measure a variation of the Gower distance [43]. Finally, it performs the imputation with the weighted average of their *k* neighbours [44].

The value of k chosen for the feature imputations is the same for each station and for each of the stations it is the root of n [9,15] where n is the number of complete observations of the station after pruning.

2.3.4. Unconditional Mean Imputation (UMI)

The method used in this study to perform the unconditional imputation of the mean is "mean" from the MICE package of R [45,46], whereby in the instances that are unknown, the algorithm completes them by introducing into all of them a single value obtained by calculating the mean of the other samples that are known [47]. In this way, the missing instances are imputed by values that lie in the centre of the distribution. This type of imputation usually has some undesirable problems such as an inadequate estimation of the variance and a deviation from the correlation between the features [48].

2.3.5. Predictive Mean Matching (PMM)

The MICE package was used to implement the predictive mean matching (PMM) method.

The general idea of this semi-parametric method is that a random instance is chosen from all complete cases that have a predictive value close to the missing case [49]. The method takes values from the data set, so they are reasonable and there are no meaningless imputed values [45,50]. In [51], the idea described above and computational details are expanded.

PMM is a method that is particularly suitable for quantitative variables that do not have a normal distribution [52].

For both random forest, unconditional mean and PMM methods, the number of multiple imputations m and maxit iterations is 5, in both cases.

2.3.6. Imputation by Kalman Smoothing (Kalman)

The Kalman imputation of the R package imputeTS, explained in [32,33] was used. This method is based on the Kalman filter [53], which consists of a recursive data assimilation system. This method was developed in [54,55]. In this paper, the imputation by Kalman smoothing on structural time series models was chosen, and to obtain a more accurate estimation, a Kalman smoothing algorithm which presents a backward recursion was used [56,57].

2.3.7. Performance Measurements

In this study, the performance of the imputation models was evaluated using the root-mean-squared error (RMSE) [58], establishing as the best model the one with the lowest RMSE value.

The k-fold cross-validation method was used to use part of the data for missing value imputations and another part to check the predicted values against the available values, being $k = 5$ parts. More information on this method can be found in [59].

3. Results and Discussion

The mean RMSE values of the $k = 5$ imputations for each variable and each imputation method at each of the station 1 are shown in Table 4.

Each of the cells of the indicated tables represents the average RMSE value of the RMSE values obtained in each of the *k* imputations with 20% of the missing values carried out for the corresponding method and variable.

Table 4. $\overline{\text{RMSE}}$ STATION ST 1.

	NO_2	NO_X	PM_{10}	SO_2
LI	2.0139	6.3158	3.4445	1.2752
PMM	4.2221	12.5949	9.2199	4.1339
RF	3.9682	12.0137	9.4291	4.1745
UMI	6.2312	15.9041	12.7478	4.8826
KNN	2.9648	10.9477	8.1196	3.5951
KALMAN	2.0322	6.3102	3.3954	1.2494

That is, each of the cells in Table 4 was obtained as follows:

As an example, the results for station ST 1 for the linear imputation method are shown in Table 5. For each variable and for each imputation method, 5 RMSE values were determined, one for each k experiment with 20% missing values (cross validation with $k = 5$) in such a way that 100% of the observations were once and only once a missing value in one of the subsets elaborated for each k, $k = 1, 2, \ldots, 5$. Thus, all observations were compared between their true value and their imputed value for each variable and for each imputation method. Once the RMSE values were determined for each of the k subsets that have 20% of missing values, the mean value of the method is calculated for each variable and then the mean RMSE value that determines the performance of the method at the station is found. Therefore, the mean RMSE values for each variable NO_2, NO_X, PM_{10} and SO_2 at station ST 1 for the linear imputation method are, respectively, 2.0138, 6.3158, 3.4444 and 1.2751. These are incorporated in Table 5 of station ST 1 for the linear imputation method in order to be able to compare these results with the rest of the method at the same station. The mean RMSE value of all the variables obtained from the mean RMSE values of each variable is the index responsible for representing the performance of the imputation method at that station and is included in the graph in Figure A, in order to compare the imputation methods at each of the stations.

Table 5. RMSE Linear imputation–STATION ST 1.

k	NO_2	NO_X	PM_{10}	SO_2
1	1.9895	6.6772	3.3359	1.6199
2	1.9484	6.5174	3.6770	1.2227
3	1.9662	6.3106	3.3713	1.1784
4	2.0774	5.8708	3.6115	1.2146
5	2.0876	6.2029	3.2261	1.1401
Mean	2.0138	6.3158	3.4444	1.2751
$\overline{\text{RMSE}}$	3.2623			

In this way, by comparing the results obtained for each method and at each station, it is possible to determine which method achieves lower RMSE results at each station and, therefore, is the most suitable method for imputation of missing data for subsequent studies. By comparing the different stations with each other, it allows us to determine whether the same imputation method is suitable for all stations or whether at each station a different method should be used to better represent the missing data.

The mean RMSE values of the $k = 5$ imputations for each variable and each imputation method at other stations are shown in Tables 6–11.

Table 6. $\overline{\text{RMSE}}$ STATION ST 9.

	NO_2	NO_X	PM_{10}	SO_2
LI	2.1045	5.9004	5.1703	1.9180
PMM	5.1576	13.6044	15.5133	5.9317
RF	4.6459	13.4225	15.5702	5.8177
UMI	7.0666	19.8137	23.4081	6.9183
KNN	3.4742	12.4591	14.0098	4.7293
KALMAN	2.0923	5.7102	5.1888	1.9426

Table 7. $\overline{\text{RMSE}}$ STATION ST 11.

	NO_2	NO_X	PM_{10}	SO_2
LI	1.9863	4.4286	3.8917	1.5538
PMM	4.0866	9.6485	11.9426	5.1804
RF	4.1306	9.6750	11.5941	5.0176
UMI	6.5735	13.4647	15.2220	5.9825
KNN	3.0239	7.6957	9.1227	3.8284
KALMAN	1.9719	4.3355	3.7875	1.5584

Table 8. $\overline{\text{RMSE}}$ STATION ST 12.

	NO_2	NO_X	PM_{10}	SO_2
LI	2.5307	9.1267	3.6796	1.3225
PMM	6.0611	23.8287	12.1254	4.7245
RF	5.8515	22.6786	11.7949	4.5565
UMI	8.1936	27.8836	14.8984	4.6735
KNN	4.3469	19.0391	9.4373	3.4929
KALMAN	2.4808	9.6010	3.6274	1.3312

Table 9. $\overline{\text{RMSE}}$ STATION ST 18.

	NO_2	NO_X	PM_{10}	SO_2
LI	2.0815	5.2487	8.5045	2.0955
PMM	4.5285	11.2938	20.5134	5.6155
RF	3.9959	11.0202	20.2237	5.3490
UMI	6.2930	15.8865	31.1047	7.3355
KNN	2.9743	9.5360	18.5270	4.6613
KALMAN	2.0940	5.1613	8.3576	2.0418

Table 10. $\overline{\text{RMSE}}$ STATION ST 19.

	NO_2	NO_X	PM_{10}	SO_2
LI	2.1384	7.3054	3.6977	1.5708
PMM	4.6330	17.7836	13.8532	5.1149
RF	4.4213	17.1405	13.2735	4.9543
UMI	6.3013	19.9789	17.8213	6.0724
KNN	3.2546	13.9541	10.9494	3.9191
KALMAN	2.1361	7.2839	3.8053	1.5541

Table 11. $\overline{\text{RMSE}}$ STATION ST 20.

	NO_2	**NO_X**	**PM_{10}**	**SO_2**
LI	1.8332	3.9101	6.6309	1.9148
PMM	3.6602	7.7101	14.8029	4.9421
RF	3.4418	7.6308	14.7711	4.8554
UMI	6.0967	11.6046	22.8886	6.6373
KNN	2.5108	6.4942	12.7487	3.8834
KALMAN	1.8604	3.8728	6.5369	1.8960

In addition to the imputation methods indicated above, other imputation methods from the MICE package of the R software such as random Bayesian linear regression (norm function), bootstrap linear regression (norm.boot function) and linear regression predicted values (norm.predict function) were also tested but were discarded because they impute negative values in data sets which, by their nature as records of air pollutant concentration measurements, contain all non-negative values.

Figure 6 shows the total $\overline{\text{RMSE}}$ value assigned to each station as the average RMSE error value of the 4 variables at each station indicated as a point for each of the imputation methods. The imputation methods are represented by different colours. The points between stations for each method were joined in the plot to emphasize how the $\overline{\text{RMSE}}$ error increases or decreases analogously at each station according to the imputation method used. Thus, a different colour curve is displayed for each imputation method.

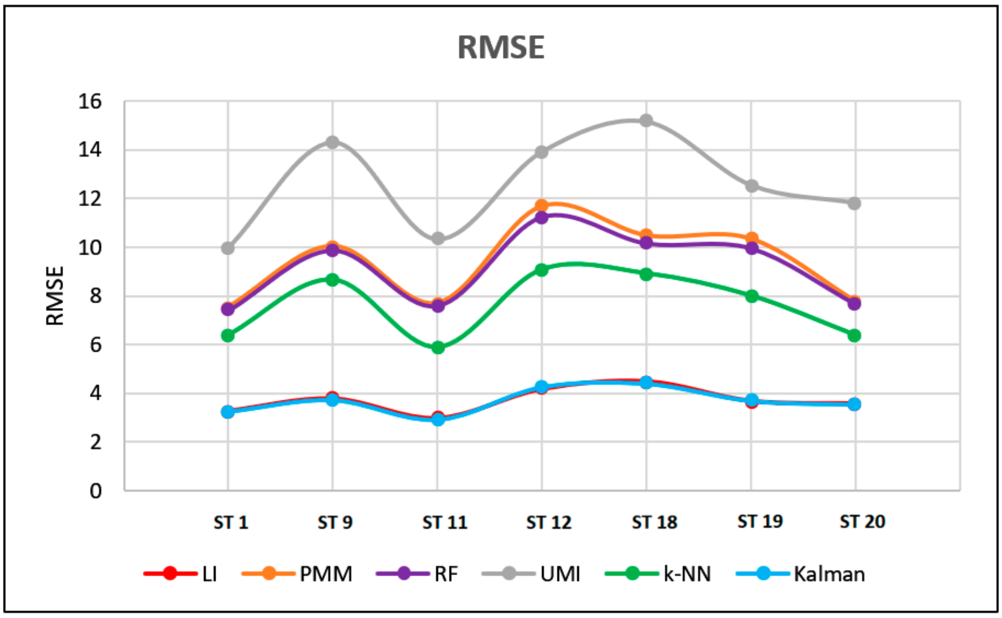

Figure 6. RMSE error for all models and stations.

As shown in Figure 6, the average RMSE for each of the methods at each station behaves similarly at all stations. It can be seen that all the imputation methods have a lower RMSE at stations ST 1, ST 11 and ST 20. The linear and Kalman imputation methods have practically identical results and their curves in the graph in Figure 6 are the ones with the smallest variations between the maximum and minimum RMSE between stations in relation to the rest of the stations, i.e., they are the ones with the flattest curves.

Figures 7–10 show the RMSE values for each pollutant: NO_2, NO_x, PM_{10} and SO_2. In each of the figures, for each method, the RMSE values obtained are shown in bar graphs. These bars are grouped by each of the imputation methods and each one represents the value obtained at each station, differentiated by colours, as shown in the legend of the graph. The following sections discuss the results obtained for each of the variables.

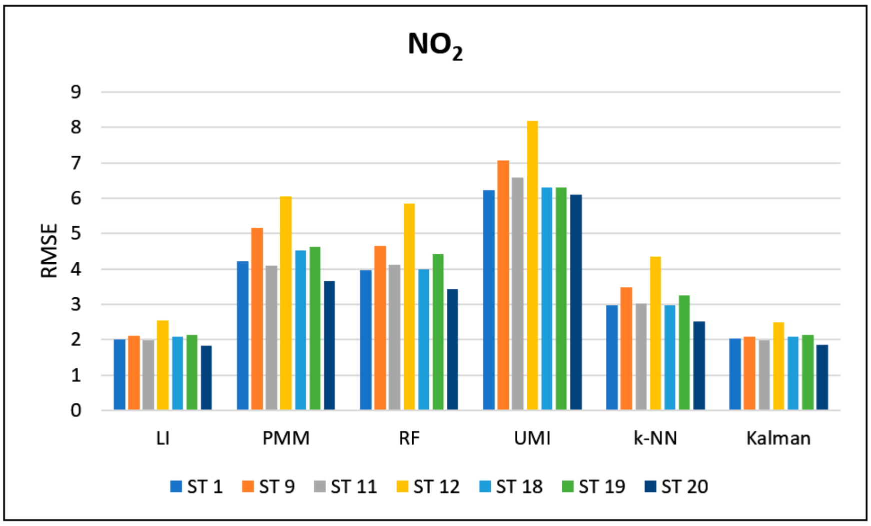

Figure 7. RMSE error for NO_2.

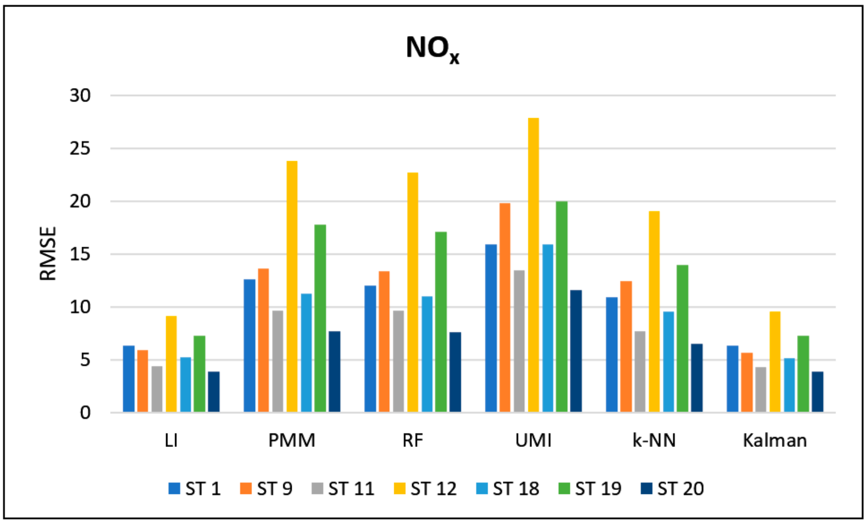

Figure 8. RMSE error for NO_X.

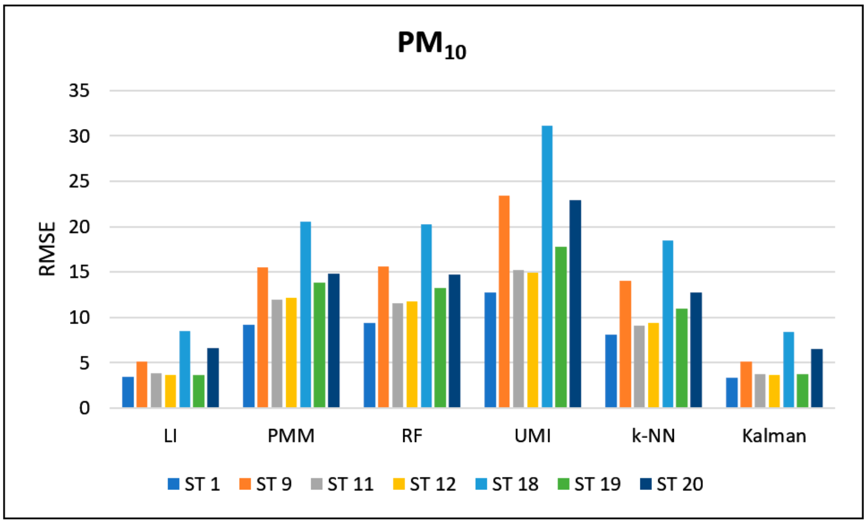

Figure 9. RMSE error for PM_{10}.

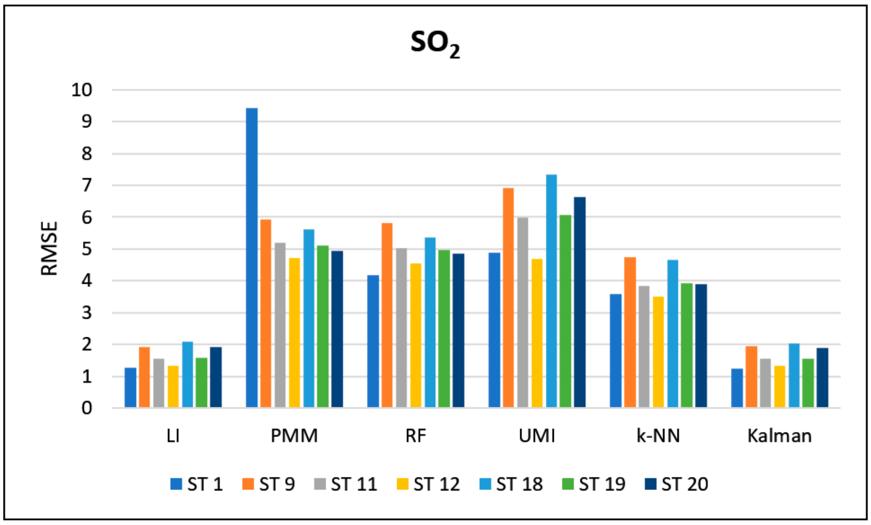

Figure 10. RMSE error for SO_2.

3.1. NO_2

Studying the performance of each variable individually, Figure 7 shows the results obtained for the variable NO_2.

It is clear that the linear and Kalman imputation methods are the ones that obtain the lowest RMSE values, being between 1.83–2.53 for LI and between 1.86–2.48 for Kalman. The values obtained are quite consistent for each method for all stations. Slightly higher values were obtained at station ST 12 (2.53 and 2.48, respectively) and the lowest RMSE values were obtained at station ST 20, with an error of 1.83 for LI and 1.86 for Kalman. ST 12 is the station that obtained the highest RMSE in all the methods, being that the UMI method is the one that generates a much higher RMSE error in all the stations.

3.2. NO_X

Figure 8 shows the mean RMSE of each imputation method used in each of the stations selected in the study for the variable NO_X. It can be seen that the profile of mean RMSE values is not as regular as in the case of the variable NO_2. The linear imputation and Kalman smoothing methods are the ones that obtain lower values, with station ST 20 followed by ST 11 having the lowest mean RMSE values. At station ST 20, the RMSE error is 3.91 for LI and 3.97 for Kalman, while at ST 11 the RMSE error is 4.42 for LI and 4.33 for Kalman. Again, ST 12 is the station with the highest mean RMSE values and the worst performing imputation method is unconditional mean imputation, with an error value of 27.88.

3.3. PM_{10}

In contrast, in Figure 9 it can be seen that the worst performance for the PM_{10} variable corresponds to station ST 18 (light blue-coloured bars) for all imputation methods. The lowest RMSE value is 3.39 and it was obtained for Kalman at ST 1, followed by 3.44 for LI, also at ST 1. However, at station ST 12, the mean values of RMSE obtained for all imputation methods remain approximately at the mean level of the rest of the stations. This station had the highest RMSE values for the variables NO_2 and NO_X, as discussed above. Again, the best performing imputation methods for all stations are the linear imputation method and Kalman smoothing. The station with the lowest mean RMSE for all the imputation methods studied is station ST 1. The RMSE value obtained for the unconditional mean imputation at station ST 18 is 31.10, which is very high with respect to the rest.

3.4. SO_2

The results for the SO_2 variable can be seen in Figure 10, with the range of mean RMSE, in relation to the stations, for each method, being lower than in the case of the NO_X and PM_{10} variables. As in the case of the other variables, the linear imputation and Kalman smoothing models are the ones that obtained the lowest mean RMSE values for all the stations, and the unconditional mean imputation method is the one with the worst performance. The lowest RMSE value is 1.24 and it was obtained for the Kalman method at station ST 1, followed by 1.27 for linear imputation, also at ST 1. However, for this variable, the highest mean RMSE value is not achieved with the unconditional mean imputation method but occurs for the PMM method at station ST 1, with a value of 9.42, which is strangely high. The rest of the stations achieved similar mean RMSE values for this method. ST 1 is the station that achieved the lowest mean RMSE value for the rest of the methods.

As we have seen for each of the variables, the worst performing imputation method is UMI. This result corresponds to [60], which indicated that, in general, the UMI method is not recommended.

In the existing literature, it has been observed that authors use different methods to create the training and validation sets as well as the way of performing the imputations. In a study conducted on the air quality monitoring dataset [61], it was determined that the method that achieved the best results in MAE and RMSE metrics was random forest imputation, with missing value rates of 5, 10, 20, 20, 20, 30 and 40%. In our study, the best results were obtained for the Kalman and linear imputation methods. The number of packets to be used depends on the methods selected for comparison. For example, ref. [61] used five R packages (MICE, VIM, AMELIA, missForest and missCompare) compared to the three R packages (MICE, VIM, imputeTS) used in this study, resulting in a lower computational cost.

Some authors, such as those in ref. [62], divide the training and validation sets according to a 70:30 ratio and according to the assumptions they use, randomly but manually removing 50% every 21 days in each quarter to simulate the MCAR and using the RMSE as a metric. In [63], they constructed sets with missing data percentages of 10% and 25%, varying the gap lengths up to 50 h. Another study, [64], proposes a time series imputation method using deep learning, with good performance, but with the drawback that the model presents difficulties for series with seasonality, using the MAPE as a performance metric.

Numerous studies, such as [44,62,65,66], use various types of neural networks to perform missing value imputations. Although they can obtain high performances, they have a very high computational cost, especially for large data sets, and require powerful computational equipment. For example, a computationally powerful computer such as Intel® Xeon® E5-2650v4 equipped with 128 GB of RAM and NDIVIA TitanX GPU acceleration was used for the calculations performed in [64], while in our study, we used an HP Pavilion laptop Intel® Core™ i7 CPU@1.30GHz equipped with 16 GB of RAM.

In the studies, the total rate of missing data in the original data set can be very different from the rate of data presented by each variable [61]. In our study the rate of data that was removed is always the same for each variable and equal to the total once the initial purging is performed, but with the consideration that, with the same number of missing values in each variable, these correspond to different observations in each variable for each of the rounds.

In [22], it is stated that linear imputation methods are not suitable because meteorology may affect the missing values. In our study, we found that this method obtains good results by replacing actual observations with missing data and imputing the same values with each imputation method at several stations and calculating the RMSE error. This may be true, but with hourly records, isolated missing values do not seem to penalize this method.

In this study we established a methodology for comparing imputation methods to be applied in subsequent studies. During the validation of the method with seven different stations and five imputation methods, it was seen that sometimes more complex methods do not perform better than simpler ones. As indicated by [67], the methods that best impute missing data from a station may change depending on the type of pollutant, the type of station or the frequency with which the data are recorded, among other factors.

4. Conclusions

Missing value imputation is a preliminary step to be performed in many research studies when incomplete data sets are obtained.

This article develops a methodology to compare various imputation methods, studying which method best fits the missing data for four pollutants recorded at seven different air quality measurement stations in Silesia, Poland as a pre-analysis to another study. The imputation methods used are linear imputation, PMM, RF, UMI, k-NN and Kalman structural smoothing.

All methods were evaluated by RMSE and validation was carried out using a cross-validation method.

The novelty of this study is the way in which the substitutions of missing values from the fully known data set are made. To perform the comparison, a pruning of missing values from the observed data set was performed beforehand and missing values were introduced to the data set with all known values so that, throughout the method, all known observations were replaced once and only once by a missing value and compared with the known value. In this way, it can be analyzed which imputation method performs best with the data pattern of the study and choose it for further investigation. This is achieved at relatively low computational cost.

The lowest RMSE error values were obtained for the NO_2 and SO_2 variables relative to NO_x and PM_{10}. However, this was not the case for all stations, indicating that not all variables have the same performance when imputations are performed on their missing data. The lowest RMSE values for each variable occurred only at two stations out of the total seven studied, stations ST 1 and ST 20. Both are located 100 km away in a straight line. At station ST 20, the pollutant NO_2 obtained an RMSE of 1.83 with the linear imputation method, and NO_x obtained an RMSE of 3.87 with Kalman imputation. At station ST 1, the pollutant PM_{10} using Kalman imputation achieved an RMSE of 3.39, and SO_2 achieved an RMSE of 1.27 using linear imputation.

The results show that the Kalman structural smoothing and linear imputation methods obtained the best results with a mean RMSE value very close to each other at all stations. The PMM and RF methods also obtained very similar results, although with higher RMSE values than the Kalman and linear imputation methods. The worst performing method in this study was UMI.

Author Contributions: Conceptualization, L.A.M.G. and A.B.S.; methodology, L.A.M.G., M.M.F., L.Á.d.P. and A.B.S.; software L.A.M.G., D.F.L. and L.Á.d.P.; validation, M.M.F., V.S.-S., L.Á.d.P., A.O.M., D.F.L. and A.B.S.; formal analysis L.A.M.G., M.M.F., A.O.M. and A.B.S.; investigation, M.M.F., V.S.-S., L.Á.d.P., A.O.M. and D.F.L.; resources, V.S.-S., D.F.L. and A.B.S.; data curation, L.A.M.G. and V.S.-S.; writing—original draft preparation, L.A.M.G., L.Á.d.P. and A.B.S. writing—review and editing, L.A.M.G., L.Á.d.P. and A.B.S.; visualization, M.M.F., V.S.-S., A.O.M. and D.F.L.; supervision, M.M.F., V.S.-S. and A.B.S.; project administration, M.M.F., A.O.M. and A.B.S.; funding acquisition, M.M.F., L.Á.d.P., A.O.M., D.F.L. and A.B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Extraordinary RENNOVA Chair in Renewable Energies and Environmental Engineering.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References and Notes

1. Kiely, G. *Environmental Engineering*; Irwin/McGraw-Hill: Boston, MA, USA, 1998.
2. Mage, D.; Ozolins, G.; Peterson, P.; Webster, A.; Orthofer, R.; Vandeweerd, V.; Gwynne, M. Urban Air Pollution in Megacities of the World. *Atmos. Environ.* **1996**, *30*, 681–686. [CrossRef]
3. Orach, J.; Rider, C.F.; Carlsten, C. Concentration-Dependent Health Effects of Air Pollution in Controlled Human Exposures. *Environ. Int.* **2021**, *150*, 106424. [CrossRef] [PubMed]
4. The European Parliament and the Council Parliament of the European Union. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. *Official Journal of the European Union*, 11 June 2008; pp. 1–44.
5. Luo, Z.; Huang, J.; Hu, K.; Li, X.; Zhang, P. AccuAir: Winning Solution to Air Quality Prediction for KDD Cup 2018. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; ACM: Anchorage, AK, USA, 2019; pp. 1842–1850. [CrossRef]
6. Li, S.-T.; Shue, L.-Y. Data Mining to Aid Policy Making in Air Pollution Management. *Expert Syst. Appl.* **2004**, *27*, 331–340. [CrossRef]
7. Menéndez García, L.A.; Sánchez Lasheras, F.; García Nieto, P.J.; Álvarez de Prado, L.; Bernardo Sánchez, A. Predicting Benzene Concentration Using Machine Learning and Time Series Algorithms. *Mathematics* **2020**, *8*, 2205. [CrossRef]
8. Zhou, Y.; De, S.; Ewa, G.; Perera, C.; Moessner, K. Data-Driven Air Quality Characterization for Urban Environments: A Case Study. *IEEE Access* **2018**, *6*, 77996–78006. [CrossRef]
9. Razavi-Far, R.; Cheng, B.; Saif, M.; Ahmadi, M. Similarity-Learning Information-Fusion Schemes for Missing Data Imputation. *Knowl.-Based Syst.* **2020**, *187*, 104805. [CrossRef]
10. Latini, G.; Passerini, G. (Eds.) Advances in management information series. In *Handling Missing Data: Applications to Environmental Analysis*; WIT Press/Computational Mechanics Inc.: Southampton, UK; Boston, MA, USA; Billerica, MA, USA, 2004.
11. Shahbazi, H.; Karimi, S.; Hosseini, V.; Yazgi, D.; Torbatian, S. A Novel Regression Imputation Framework for Tehran Air Pollution Monitoring Network Using Outputs from WRF and CAMx Models. *Atmos. Environ.* **2018**, *187*, 24–33. [CrossRef]
12. Samal, K.K.R.R.; Panda, A.K.; Babu, K.S.; Das, S.K. An Improved Pollution Forecasting Model with Meteorological Impact Using Multiple Imputation and Fine-Tuning Approach. *Sustain. Cities Soc.* **2021**, *70*, 102923. [CrossRef]
13. Liu, X.; Wang, X.; Zou, L.; Xia, J.; Pang, W. Spatial Imputation for Air Pollutants Data Sets via Low Rank Matrix Completion Algorithm. *Environ. Int.* **2020**, *139*, 105713. [CrossRef]
14. Mercer, T.G.; Frostick, L.E.; Walmsley, A.D. Recovering Incomplete Data Using Statistical Multiple Imputations (SMI): A Case Study in Environmental Chemistry. *Talanta* **2011**, *85*, 2599–2604. [CrossRef] [PubMed]
15. Hernández-Pereira, E.M.; Álvarez-Estévez, D.; Moret-Bonillo, V. Automatic Classification of Respiratory Patterns Involving Missing Data Imputation Techniques. *Biosyst. Eng.* **2015**, *138*, 65–76. [CrossRef]
16. Rubin, D.B. (Ed.) *Multiple Imputation for Nonresponse in Surveys*; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1987. [CrossRef]
17. Norris, G.; Duvall, R.; Brown, S.; Bai, S. *Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide*; EPA/600/R-14/108; EPA: Washington, DC, 2014; p. 136.
18. Junger, W.L.; de Ponce Leon, A. Imputation of Missing Data in Time Series for Air Pollutants. *Atmos. Environ.* **2015**, *102*, 96–104. [CrossRef]
19. Greenland, S.; Finkle, W.D. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *Am. J. Epidemiol.* **1995**, *142*, 1255–1264. [CrossRef] [PubMed]
20. Pollice, A.; Lasinio, G.J. Two Approaches to Imputation and Adjustment of Air Quality Data from a Composite Monitoring Network. *J. Data Sci.* **2009**, *7*, 43–59. [CrossRef]
21. Galvan, M.; Medina, F. *Imputacion de Datos: Teoria y Practica*; Estudios Estadisticos y Prospectivos; Naciones Unidas, CEPAL, Div. de Estadistica y Proyecciones Economicas: Santiago, Chile, 2007.
22. Şahin, Ü.A.; Bayat, C.; Uçan, O.N. Application of Cellular Neural Network (CNN) to the Prediction of Missing Air Pollutant Data. *Atmos. Res.* **2011**, *101*, 314–326. [CrossRef]
23. Miller, L.; Xu, X.; Wheeler, A.; Zhang, T.; Hamadani, M.; Ejaz, U. Evaluation of Missing Value Methods for Predicting Ambient BTEX Concentrations in Two Neighbouring Cities in Southwestern Ontario Canada. *Atmos. Environ.* **2018**, *181*, 126–134. [CrossRef]
24. Nosal, M.; Legge, A.H.; Krupa, S.V. Application of a Stochastic, Weibull Probability Generator for Replacing Missing Data on Ambient Concentrations of Gaseous Pollutants. *Environ. Pollut.* **2000**, *108*, 439–446. [CrossRef]
25. Quinteros, M.E.; Lu, S.; Blazquez, C.; Cárdenas-R, J.P.; Ossa, X.; Delgado-Saborit, J.-M.; Harrison, R.M.; Ruiz-Rudolph, P. Use of Data Imputation Tools to Reconstruct Incomplete Air Quality Datasets: A Case-Study in Temuco, Chile. *Atmos. Environ.* **2019**, *200*, 40–49. [CrossRef]
26. Plaia, A.; Bondi, A. Single Imputation Method of Missing Values in Environmental Pollution Data Sets. *Atmos. Environ.* **2006**, *40*, 7316–7330. [CrossRef]
27. Hajmohammadi, H.; Heydecker, B. Multivariate Time Series Modelling for Urban Air Quality. *Urban Clim.* **2021**, *37*, 100834. [CrossRef]

28. Samal, K.K.K.R.; Babu, K.S.; Das, S.K. Multi-Directional Temporal Convolutional Artificial Neural Network for PM2.5 Forecasting with Missing Values: A Deep Learning Approach. *Urban Clim.* **2021**, *36*, 100800. [CrossRef]
29. Ma, J.; Cheng, J.C.P.; Ding, Y.; Lin, C.; Jiang, F.; Wang, M.; Zhai, C. Transfer Learning for Long-Interval Consecutive Missing Values Imputation without External Features in Air Pollution Time Series. *Adv. Eng. Inform.* **2020**, *44*, 101092. [CrossRef]
30. Rubin, D.B. Inference and Missing Data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
31. Schafer, J.L.; Graham, J.W. Missing Data: Our View of the State of the Art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef]
32. Moritz, S.; Gatscha, S. *Package "ImputeTS"*, version 3.2. Time Series Missing Value Imputation. 2021.
33. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
34. Chapra, S.C.; Canale, R.P. *Numerical Methods for Engineers*, 8th ed.; McGraw-Hill Education: New York, NY, USA, 2021.
35. Davis, P.J. *Interpolation and Approximation*; Dover Publications: New York, NY, USA, 1975.
36. van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Soft.* **2011**, *45*, 1–67. [CrossRef]
37. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Doove, L.L.; Van Buuren, S.; Dusseldorp, E. Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects. *Comput. Stat. Data Anal.* **2014**, *72*, 92–104. [CrossRef]
39. Arowosegbe, O.O.; Röösli, M.; Künzli, N.; Saucy, A.; Adebayo-Ojo, T.C.; Jeebhay, M.F.; Dalvie, M.A.; de Hoogh, K. Comparing Methods to Impute Missing Daily Ground-Level PM10 Concentrations between 2010–2017 in South Africa. *Int. J. Environ. Res. Public Health* **2021**, *18*, 3374. [CrossRef]
40. Waljee, A.K.; Mukherjee, A.; Singal, A.G.; Zhang, Y.; Warren, J.; Balis, U.; Marrero, J.; Zhu, J.; Higgins, P.D. Comparison of Imputation Methods for Missing Laboratory Data in Medicine. *BMJ Open* **2013**, *3*, e002847. [CrossRef]
41. Kowarik, A.; Templ, M. Imputation with the R Package VIM. *J. Stat. Soft.* **2016**, *74*, 1–16. [CrossRef]
42. Templ, M.; Kowarik, A.; Alfons, A.; de Cillia, G.; Rannetbauer, W. *Package "VIM"*, version 6.1.1. Visualization and Imputation of Missing Values. 2021.
43. Gower, J.C. A General Coefficient of Similarity and Some of Its Properties. *Biometrics* **1971**, *27*, 857. [CrossRef]
44. Zhang, Y.; Zhou, B.; Cai, X.; Guo, W.; Ding, X.; Yuan, X. Missing Value Imputation in Multivariate Time Series with End-to-End Generative Adversarial Networks. *Inf. Sci.* **2021**, *551*, 67–82. [CrossRef]
45. Abayomi, K.; Gelman, A.; Levy, M. Diagnostics for Multivariate Imputations. *J. R. Stat. Soc. C* **2008**, *57*, 273–291. [CrossRef]
46. van Buuren, S.; Groothuis-Oudshoorn, K. *Package "Mice"*, version 3.14.0. Multivariate Imputation by Chained Equations. 2021.
47. Molenberghs, G.; Verbeke, G. *Linear Mixed Models for Longitudinal Data*; Springer Series in Statistics; Springer New York: New York, NY, USA, 2000. [CrossRef]
48. Barzi, F. Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies. *Am. J. Epidemiol.* **2004**, *160*, 34–45. [CrossRef]
49. Little, R.J.A.; Rubin, D.B. *Statistical Analysis with Missing Data*, 3rd ed.; Wiley series in probability and statistics; Wiley: Hoboken, NJ, USA, 2020.
50. Schenker, N.; Taylor, J.M.G. Partially Parametric Techniques for Multiple Imputation. *Comput. Stat. Data Anal.* **1996**, *22*, 425–446. [CrossRef]
51. van Buuren, S. *Flexible Imputation of Missing Data*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018. [CrossRef]
52. Allison, P. *Missing Data*; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 2002; Volume 136. [CrossRef]
53. Moritz, S.; Bartz-Beielstein, T. ImputeTS: Time Series Missing Value Imputation in R. *R J.* **2017**, *9*, 207. [CrossRef]
54. Kalman, R.E. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
55. Welch, G.; Bishop, G. An Introduction to the Kalman Filter. *Proc. SIGGRAPH Course* **2006**, *8*, 41.
56. Oluwasegun Agbailu, A.; Oluwafemi Clement, O.; Seno, A. Kalman Filter Algorithm versus Other Methods of Estimating Missing Values: Time Series Evidence. *Afr. J. Math. Stat. Stud.* **2021**, *4*, 1–9. [CrossRef]
57. Wijesekara, W.M.M.L.K.N.; Liyanage, L. Comparison of Imputation Methods for Missing Values in Air Pollution Data: Case Study on Sydney Air Quality Index. In *Advances in Information and Communication*; Arai, K., Kapoor, S., Bhatia, R., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2020; Volume 1130, pp. 257–269. [CrossRef]
58. Willmott, C.J.; Ackleson, S.G.; Davis, R.E.; Feddema, J.J.; Klink, K.M.; Legates, D.R.; O'Donnell, J.; Rowe, C.M. Statistics for the Evaluation and Comparison of Models. *J. Geophys. Res.* **1985**, *90*, 8995. [CrossRef]
59. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, USA, 2009. [CrossRef]
60. Little, R.J.A. Regression with Missing X's: A Review. *J. Am. Stat. Assoc.* **1992**, *87*, 1227. [CrossRef]
61. Alsaber, A.R.; Pan, J.; Al-Hurban, A. Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018). *Int. J. Environ. Res. Public Health* **2021**, *18*, 1333. [CrossRef] [PubMed]
62. Li, J.; Ren, W.; Han, M. Variational auto-encoders based on the shift correction for imputation of specific missing in multivariate time series. *Measurement* **2021**, *186*, 110055. [CrossRef]

63. Junninen, H.; Niska, H.; Tuppurainen, K.; Ruuskanen, J.; Kolehmainen, M. Methods for imputation of missing values in air quality data sets. *Atmos. Environ.* **2004**, *38*, 2895–2907. [CrossRef]
64. Kim, T.; Kim, J.; Yang, W.; Lee, H.; Choo, J. Missing Value Imputation of Time-Series Air-Quality Data via Deep Neural Networks. *Int. J. Environ. Res. Public Health* **2021**, *18*, 12213. [CrossRef]
65. Fallah, B.; Ng, K.T.W.; Vu, H.L.; Torabi, F. Application of a multi-stage neural network approach for time-series landfill gas modeling with missing data imputation. *Waste Manag.* **2020**, *116*, 66–78. [CrossRef]
66. Silva-Ramírez, E.-L.; Pino-Mejías, R.; López-Coello, M.; Cubiles-de-la-Vega, M.-D. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Netw.* **2011**, *24*, 121–129. [CrossRef]
67. Alahamade, W.; Lake, I.; Reeves, C.E.; de la Iglesia, B. A multi-variate time series clustering approach based on intermediate fusion: A case study in air pollution data imputation. *Neurocomputing* **2022**, *490*, 229–245. [CrossRef]