International Journal of
Environmental Research and Public Health

Article

Handling Complex Missing Data Using Random Forest Approach for an Air Quality Monitoring Dataset: A Case Study of Kuwait Environmental Data (2012 to 2018)

Ahmad R. Alsaber [1,*,†], Jiazhu Pan [1] and Adeeba Al-Hurban [2]

[1] Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, UK; jiazhu.pan@strath.ac.uk
[2] Department of Earth and Environmental Sciences, Faculty of Science, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait; Q8geo@hotmail.com
* Correspondence: a.alsaber@strath.ac.uk
† Current address: Livingstone Tower (Level 9), 26 Richmond Street, Glasgow G1 1XH, UK.

Abstract: In environmental research, missing data are often a challenge for statistical modeling. This paper addressed some advanced techniques to deal with missing values in a data set measuring air quality using a multiple imputation (MI) approach. MCAR, MAR, and NMAR missing data techniques are applied to the data set. Five missing data levels are considered: 5%, 10%, 20%, 30%, and 40%. The imputation method used in this paper is an iterative imputation method, missForest, which is related to the random forest approach. Air quality data sets were gathered from five monitoring stations in Kuwait, aggregated to a daily basis. Logarithm transformation was carried out for all pollutant data, in order to normalize their distributions and to minimize skewness. We found high levels of missing values for NO_2 (18.4%), CO (18.5%), PM_{10} (57.4%), SO_2 (19.0%), and O_3 (18.2%) data. Climatological data (i.e., air temperature, relative humidity, wind direction, and wind speed) were used as control variables for better estimation. The results show that the MAR technique had the lowest RMSE and MAE. We conclude that MI using the missForest approach has a high level of accuracy in estimating missing values. MissForest had the lowest imputation error (RMSE and MAE) among the other imputation methods and, thus, can be considered to be appropriate for analyzing air quality data.

Keywords: missing imputation; random forest; high dimensional data; missing data mechanism; air quality

Citation: Alsaber, A.; 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. https://doi.org/10.3390/ijerph18031333

Received: 5 December 2020
Accepted: 27 January 2021
Published: 2 February 2021

1. Introduction

Air quality monitoring is conducted with the aim of protecting public health. Numerous air contaminants have been found to have harmful effects on human health. The air quality in cities varies, due to concentrations of particulate matter 10 micrometers (PM_{10}), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), and sulfur dioxide (SO_2), from emission sources including vehicle exhaust, manufacturing operations, and chemical facilities, among other sources.

A major challenge in air quality data management is determining how to deal with missing data values. Missing information in data sets occurs for multiple reasons, such as impaired equipment, insufficient sampling frequency, hardware problems, and human error [1]. Incomplete data sets affect the applicability of specific analyses, such as receptor modeling, which generally requires a complete data matrix [2]. The occurrence of missing data, no matter how infrequent, can bias findings on the relationships between air contaminants and health outcomes [3]. Incomplete data matrices may provide outcomes that vary significantly, compared to the results from complete data sets [4].

To gain a more complete data set, researchers must decide whether to discard or impute (i.e., substitute for) missing data. Ignoring missing values is typically not warranted, as valuable information is lost, which may compromise inferential power [5]. Therefore, the most appropriate option is to impute the missing data. Yet, the systematic differences between real and substituted data can also lead to unwanted bias. Therefore, it is vital to determine an optimal approach for estimating missing values. Several problems have been linked with missing data [6]. These challenges include statistical power reduction, bias as a result of inconsistent data, difficulties in managing the data during statistical analyses, and low efficiency. The criteria implemented for measures to deal with missing data in time-series analysis rely on the missing data replacement mechanism and missing data pattern [7]. Such challenges are especially problematic when the missing data exceed 60 percent, where existing methods have significant difficulty in addressing such situations [8].

This study focuses on a case study of missing data related to air quality monitoring. The Kuwait environmental public authority (KEPA) is mandated with the responsibility for measuring air quality. A data set collected from five fixed monitoring stations was associated with missing data, likely caused by multiple reasons. One is that there were a large number of routine maintenance changes in the monitoring sites. Second, simple human error occurred. Third, there were some tagging problems that necessitated the exclusion of some data.

The main purpose of this paper was to find the best imputation method to estimate the missing values for the measured pollutants (SO_2, NO_2, CO, O_3, and PM_{10}) in the KEPA data sets. The imputation methods used in this paper are: multivariate imputation by chained equations using random forest (RF), k-nearest neighbor (kNN), Bayesian principal component analysis (BPCA), multiple imputation using expectation maximization with bootstrapping (EM with Bootstrapping), predictive mean matching (PMM), and the proposed iterative imputation method (missForest) based on a random forest. Two tests, root mean square error (RMSE) and mean absolute error (MAE), are used to compare the performances of the imputation methods. For the error indicators (RMSE or MAE), the larger the value, the greater the error. The end product is an outline of the best approaches for managing missing data in a data set that is critical for public health in Kuwait.

It is important to describe the factors that may lead to missing data in statistical analyses. The first instance of missing data is missing completely at random (MCAR), whereby the missing data result from either the observer not collecting the necessary information or the reporting of incomplete or false information. The second instance of missing data is missing at random (MAR), whereby the extent of data missing depends on the type of data under observation. MAR is recommended when the missing data can be partially retrieved, depending on the existence of information related to the variables in the same data set. The third instance is missing not at random (MNAR), whereby the missing data are dependent on the actual values absent for statistical analysis. Among the three types of missing data in statistical analysis, MAR and MNAR are the most common [9]. When the type of missing data tends towards MAR, multiple imputation techniques are more suitable than other techniques, such as listwise deletion [10].

1.1. Missing Completely at Random (MCAR)

For MCAR, the chance of missing data values is the same across all instances. It can be interpreted as the cause of missing data values not being related to the data collected. For instance, a random sample of a population, whereby each individual from the population has an equal chance of being selected for the sample. This would mean that not all members of the population were present among the selected sample. Therefore, the data and values of the members not selected would be missing from the statistical analysis. The following example describes an instance in which MCAR occurs in statistical analysis:

Suppose that Y is an $n \times p$ matrix which includes all p variables with n cases in the sample. Let the observed values be denoted as (Y_{obs}), while the missing values are denoted as (Y_{mis}). The matrix R spots the missing values locations in Y. The observations of R and Y are denoted as $\mathrm{r_{ij}}$ and $\mathrm{y_{ij}}$, respectively. Thus, $\mathrm{r_{ij}} = 1$ when $\mathrm{y_{ij}}$ is observed, while $\mathrm{r_{ij}} = 0$

when $\mathrm{y_{ij}}$ is missing. Then, the distribution of R depends upon $Y = (Y_{\mathrm{obs}}, Y_{\mathrm{mis}})$. We can write $\mathrm{Pr}(\mathrm{R}|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}, \Psi)$ when the data are said to be assumed as MCAR, if:

$$\mathrm{Pr}(\mathrm{R} = 0|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}, \Psi) = \mathrm{Pr}(\mathrm{R} = 0|\Psi), \tag{1}$$

where ψ consists of the parameters of the missing data in the model. This means that the probability of missing a data value depends only on the estimated parameters in the model.

1.2. Missing at Random (MAR)

For MAR, the chance of data values missing is equal across all categories. MAR is, therefore, a more diverse instance, compared to MCAR; for instance, when selecting a sample from a population based on certain characteristics, the resulting missing data can be categorized as MAR. Statistical software for multiple imputations usually assumes that the data are MAR [11]. Therefore, the probability of data missing is dependent on the data under observation:

$$\mathrm{Pr}\big(\mathrm{R} = 0|Y_{\mathrm{obs},}Y_{\mathrm{mis}}, \Psi\big) = \mathrm{Pr}(\mathrm{R} = 0|Y_{\mathrm{obs}}, \psi). \tag{2}$$

The KEPA data are best classified as MAR.

1.3. Missing Not at Random (MNAR)

For MNAR, the chance of data not being available is dependent on reasons unknown to the researcher. For instance, when conducting research, some respondents may decide to withhold information for reasons unknown to the researcher. Due to the nature of MNAR, it is often regarded as a more complex case in statistical analysis. It can be addressed by targeting some of the reasons respondents would choose to with hold information, Y_{mis}, itself. It is represented:

$$\mathrm{Pr}(\mathrm{R} = 0|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}, \Psi). \tag{3}$$

The data set extracted from KEPA has extensive missing values. The missing data could have been due to routine maintenance, changes in the siting of monitors, human error, or tagging problems.

1.4. Ignoring the Missing Data Mechanism

One of the major issues that arise when performing imputations is whether the missing data come from the same distribution as the observed data (Y_{obs}). As mentioned above, the observed data are made up of Y_{obs} and R with the joint density function $f(Y_{\mathrm{obs}}, R|\theta, \psi)$, which depends on the model estimated parameters θ for Y.

We can estimate θ without knowing ψ by defining the probability density function of the joint distribution of Y_{obs} and Y_{mis} as $f(Y|\theta) \equiv f(Y_{\mathrm{obs}}, Y_{\mathrm{mis}}|\theta)$. Therefore, in order to compute the marginal probability density of Y_{obs}, we integrate the missing data as:

$$f(Y_{\mathrm{obs}}|\theta) = \int f(Y_{\mathrm{obs}}, Y_{\mathrm{mis}}|\theta) dY_{\mathrm{mis}}, \tag{4}$$

where the likelihood function of θ, according to Y_{obs} while ignoring the missing data, can be defined as:

$$L_{ign}(\theta|Y_{\mathrm{obs}}) \propto f(Y_{\mathrm{obs}}|\theta). \tag{5}$$

Obtaining maximum likelihood (ML) estimates of θ can be done by maximizing the provided θ.

To build a more general model, we include R and specify the joint density distribution of Y and R as:

$$f(Y, R|\theta, \psi) = f(Y|\theta) f(R|Y, \psi). \tag{6}$$

We can find the distribution of the observed data by integrating Y_{mis} from the joint density using θ and ψ, defined as:

$$f(Y_{\mathrm{obs}}, R|\theta, \psi) = \int f(Y_{\mathrm{obs}}, Y_{\mathrm{mis}}|\theta) f(R|Y_{\mathrm{obs}}, Y_{\mathrm{mis}}, \psi) dY_{\mathrm{mis}}. \tag{7}$$

Now, we can rewrite Equation (7) as:

$$f(Y_{\text{obs}}, R|\theta, \psi) = f(R|Y_{\text{obs}}, \psi) \int f(Y_{\text{obs}}, Y_{\text{mis}}|\theta) dY_{\text{mis}} = f(R|Y_{\text{obs}}, \psi) f(Y_{\text{obs}}|\theta). \tag{8}$$

The missing data mechanism is ignorable for likelihood inference if

1. MAR: when the missing data pattern is missing at random; and
2. Distinctness: when the joint parameter space of (θ, ψ) is equal to the product of the parameter space of θ and ψ [12].

1.5. Multiple Imputation (MI)

Studies have shown that MI is unbiased if the missing rate for a variable exceeds 50% of the total missing values [13–15]. Researchers have debated the role of listwise deletion when solving for such missing data. Most research studies have concluded that, although the listwise deletion technique is not commonly used, it is applicable in some instances [16,17]. According to Marshall et al. [15], multiple imputation is favorable for computing missing data and especially applicable when the missing data rate is above 10% [18]. For instance, in a regression model, including the number of variables with a low rate of missing data. In such an instance, this may result in a rate of missing data that is higher in the full regression model, when compared to the outcomes of simple bivariant regressions. Therefore, it is critical for analysts to evaluate the total missing rate, as well as the partial missing one.

One limitation of applying a single imputation approach is that formulas of standard variance applied to filled-in data tend to underestimate the variance of the estimates; therefore, multiple imputation methods have been proposed [11]. The first step in such a method is specifying the single encompassing multivariate approach for all data sets. There are four types of multivariate models of data completion to consider [12]: (i) standard models, which impute under multivariate normal distributions; (ii) log-linear models, that have been used traditionally by social scientists in describing the associations among cross-classified data variables; (iii) general location models, which combine the log-linear approach for the variables that are definite with the multivariate model of standard regression for the continuous variables; and (iv) a two-level model of linear regression, which is mostly applied to multi-level data. The imputation model should be able to match the subsequent analysis and should be able to preserve the interactions of variables, which relates to the central point of the investigation discussed later in this paper.

A multiple imputation method balances ease of application and the quality of obtained results. The various imputations identify random errors that are appropriate to the process of imputation, making it possible to obtain unbiased estimates in all parameters. No deterministic method of imputation can achieve the same result. The technique also allows for departure from normality assumptions, while providing results that are adequate with low sample sizes or when significant amounts of data are missing.

Some requirements are necessary, in order to attain the desired results of multiple imputation [19]. First, there should be random data missing (MAR), which means that there is a dependence on observed variables and not missing observations. Second, the method of generating the values imputed should suit the analysis that subsequently follows. This maintains the associations between variables, which is a focus in the analysis shown later in this paper. Third, the model for imputation should coincide and agree with that of the investigation. Rubin has given a thorough description of these conditions. A remaining question, however, relates to adopting the most suitable practices for performing the imputations [20]. It is essential to have an awareness of the possible prediction problems, in order to reduce or minimize systematic error.

There have been many applications of multiple imputation in health, environmental [21,22], and industrial [23,24] data bases, as well as for survey data [25,26] and data mining approaches, which extract patterns from large data sets through a combination of artificial intelligence and statistical methods, that can be used for database management [23].

2. Materials and Methods

2.1. Multiple Imputation Using Random Forest Method

Let us assume that $\mathbf{X} = \left(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p\right)$ is a $n \times p$-dimensional data matrix. We propose the use of the random forest technique for imputing missing observations. The random forest algorithm has a built-in routine to handle the values that are missing by weighing the frequency of values with the proximity of a random forest after the training of an initially imputed mean data set [27]. This approach requires a response variable that is complete and useful for forest training. Instead, we estimate the values of all the missing values directly, by use of a random forest that is trained on the observed data set, where X is the matrix of the complete data. $\mathbf{X}_s$ contains all missing values at entries $\mathbf{i}_{\text{mis}}^{(s)} \subseteq \{1, \ldots, n\}$. The data set can be separated into four parts:

1. $\mathbf{y}_{\text{obs}}^{(s)}$: the observed values of $\mathbf{X}_s$.
2. $\mathbf{y}_{\text{mis}}^{(s)}$: the missing values of $\mathbf{X}_s$.
3. $\mathbf{x}_{\text{obs}}^{(s)}$: the observations, $\mathbf{i}_{\text{obs}}^{(s)} = \{1, \ldots, n\} \backslash \mathbf{i}_{\text{mis}}^{(s)}$, that belong in the other variables $\mathbf{X}_s$.
4. $\mathbf{x}_{\text{mis}}^{(s)}$: the observations, $\mathbf{i}_{\text{mis}}^{(s)}$, that belong in the other variables $\mathbf{X}_s$.

Note that $\mathbf{X}_{\text{obs}}^{(s)}$ and $\mathbf{X}_{\text{mis}}^{(s)}$ are not completely observed, as the index $\mathbf{i}_{\text{obs}}^{(s)}$ corresponds to the observed values of the variable $\mathbf{X}_s$.

According to [28], the process starts with an initial guess for the missing values in $\mathbf{X}$ using a mean imputation approach or any other imputation method, depending on the data. Then, we sort the predictors $\mathbf{X}_s, s = 1, \ldots, p$, ascending or descending, $\mathbf{X}_s, s = 1, \ldots, p$, according to the number of missing values. Then, for each variable $\mathbf{X}_s$, the missing values are imputed by random forest (i.e., the first fitting) with response $\mathbf{y}_{\text{obs}}^{(s)}$ and predictors $\mathbf{X}_{\text{obs}}^{(s)}$. Next, the missing values $\mathbf{y}_{\text{mis}}^{(s)}$ are estimated by applying the trained random forest to $\mathbf{X}_{\text{mis}}^{(s)}$. The imputation approach should be repeated until a stopping criterion is reached. Pseudo Algorithm 1 shows a representation of the missForest method (see Algorithm 1).

The stopping criterion (γ) is met when the difference between the last imputed data matrix and the previous one increases for the first time, with respect to both variable types. Here, the difference for the set of continuous variables $\mathbf{N}$ is defined as:

$$\Delta_N = \frac{\sum_{j \in \mathbf{N}} \left(\mathbf{X}_{new}^{imp} - \mathbf{X}_{old}^{imp}\right)^2}{\sum_{j \in \mathbf{N}} \left(\mathbf{X}_{new}^{imp}\right)^2}, \tag{9}$$

and that for the set of categorical variables $\mathbf{F}$ as:

$$\Delta_F = \frac{\sum_{j \in \mathbf{F}} \sum_{i=1}^{n} \mathbf{I}_{\mathbf{X}_{new}^{imp} \neq \mathbf{X}_{old}^{imp}}}{\#NA}. \tag{10}$$

Let $\mathbf{X}$ be an $n \times p$ matrix; set the stopping criterion (γ); set the initial guess for missing values. $\mathbf{k} \leftarrow$ vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing amount of missing values. $\mathbf{X}_{old}^{imp} \leftarrow$ stores the previously imputed matrix. Fit a random forest: $\mathbf{y}_{obs}^{(s)} \sim \mathbf{x}_{obs}^{(s)}$. Predict $\mathbf{y}_{mis}^{(s)}$ using $\mathbf{x}_{mis}^{(s)}$; $\mathbf{X}_{new}^{imp} \leftarrow$ update the imputed matrix using the predicted $\mathbf{y}_{mis}^{(s)}$. Update γ and the imputed matrix $\mathbf{X}^{imp}$. Where #NA is the number of missing values in the categorical variables $\mathbf{F}$.

After imputing the missing values, the performance is assessed using the normalized root mean squared error [29] for the continuous variables, defined by:

$$\text{NRMSE} = \sqrt{\frac{\text{mean}\left(\left(\mathbf{X}^{\text{true}} - \mathbf{X}^{\text{imp}}\right)^2\right)}{\text{var}\left(\mathbf{X}^{\text{true}}\right)}}, \tag{11}$$

where $\mathbf{X}^{\text{true}}$ and $\mathbf{X}^{\text{imp}}$ are the complete data matrix and the imputed data matrix, respectively. In this study, all predictors are classified as continuous observations. The mean and variance are used as a short notation for empirical mean and variance computed over the missing values only.

When an RF is fit to the part that is observed on a variable, we use the out-of-bag (OOB) estimate of an error for the variable. When we meet the stopping criterion (γ), we average it over the variable set of that type, in order to obtain an approximation of the actual errors of imputation. We assess the performance of this estimate by comparing the absolute difference between the OOB imputation error estimate in all simulation runs and the true imputation error.

2.2. Process of Multiple Imputations (MI) Using Rubin's Rules

For our data sets, we followed Rubin's rules [11] for handling missing data. The process of multiple imputations (MIs) was conducted separately for each monitoring station (see Figure 1). The first step in multiple imputation is to create values ("imputes" or "m_i"), with 10 iterations for each "m_i" to be substituted for the missing data. In order to create imputed values, we need to identify a model (say, a linear regression) that allows us to create imputes based on other variables in the data set (predictor variables). As we need to do this multiple times, in order to produce multiple-imputed data sets, we identify a set of regression lines which are similar to each other.

Figure 1 shows the process for the KEPA data sets, to process and estimate missing values using imputation methods. There were five data sets (1–5), relating to FAH, JAH, MAN, RUM, and ASA, respectively. Each data set should contain 2192 daily observations for each variable; however, due to missing values, they were all less than 2192.

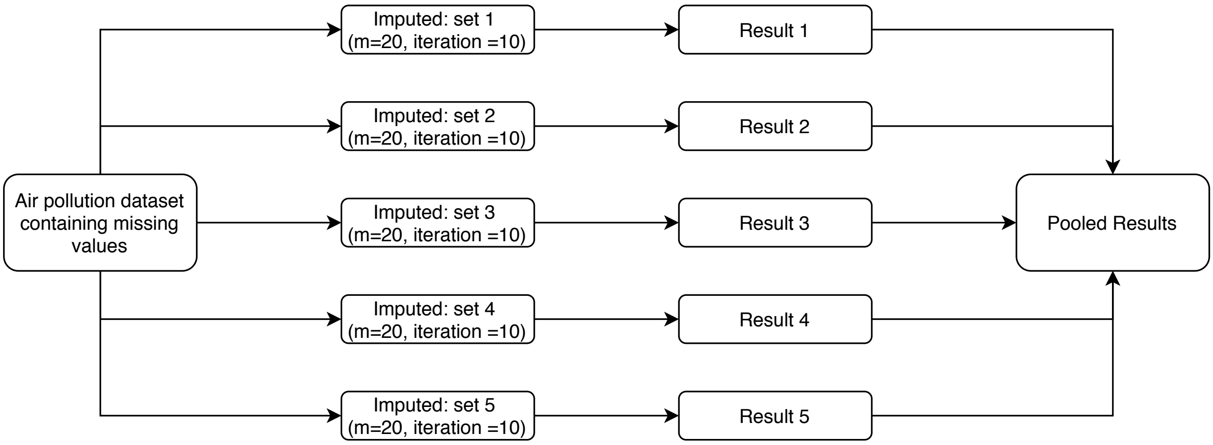

Air pollution dataset containing missing values
Imputed: set 1 (m=20, iteration =10)
Imputed: set 2 (m=20, iteration =10)
Imputed: set 3 (m=20, iteration =10)
Imputed: set 4 (m=20, iteration =10)
Imputed: set 5 (m=20, iteration =10)
Result 1
Result 2
Result 3
Result 4
Result 5
Pooled Results

Figure 1. The steps of implementing multiple imputations for PM_{10}, SO_2, O_3, CO, and NO_2 during 2012 to 2017, according to site location, in the State of Kuwait.

The power of MI lies in its multiple imputations being able to be performed for each variable in the data set. While every single imputation is ambiguous or imprecise, the combination of the computed imputations takes the uncertainty of each imputation into consideration. According to [17,18], MAR or MCAR pooled estimated parameters are less biased and the associated standard errors are corrected appropriately.

The implementation of an MI technique requires three steps: First, it imputes several values for the same observation, using at least two methods ($m \geq 2$). Then, the second step takes each individual method, m, and analyzes it using standard complete data. Finally, m (the completed data sets) is pooled by integrating the m analyses, in order to generate overall estimates and standard errors. This can be done by calculating the mean over the m repeated analyses. Pooling data from several m allows multiple imputations to ensure higher accuracy [30]. Figure 1 shows how we treated the KEPA data sets with multiple imputation, where $m = 20$.

2.3. Data Sets

We utilized a real-time air quality monitoring data set collected for 5 locations in Kuwait from the Kuwait Environmental Public Authority (KEPA), in order to evaluate and assess the performance of various imputation methods to estimate missing values in the data set. The data set contained air quality, time, and meteorological data.

1. Air quality data: The air pollutant variables in the air quality data were NO_2, CO, PM_{10}, SO_2, and O_3;
2. Meteorological data: The meteorological parameters included temperature, humidity, wind direction, and wind speed.

All these variables for the past 24 h are collected on hourly basis and features extracted from the collected data set were used for evaluation of the models, for predictions of the concentration of missing values for NO_2, CO, PM_{10}, SO_2, and O_3. Concentrations of all the pollutants are reported in $\mu g/m^3$.

We compiled pollutant data from the Environmental Public Authority of Kuwait (KEPA). The data were gathered from five environmental monitoring stations from 1 January 2013 to 31 December 2017. We used the following pollutants: Particulate matter 10 micrometers (PM_{10}), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), and sulfur dioxide (SO_2). We estimated a concentration time of 24 h (daily observation) for SO_2, NO_2, and PM_{10} at each station and 8 h for CO and O_3. We assumed 75% of the collected values as reliable averages [31]. We used the Air Quality Index (AQI), as generated by [32].

The AQI was developed, for Kuwait, based on the United States Environmental Protection Agency (USEPA) recommendations. The AQI is defined with consideration of characteristics of the air, in relation to the environmental needs of humans [32]. The AQI is an index for reporting the day-to-day air quality, providing details about the cleanliness of ambient air [33]. The following equation was used to convert between pollutant concentration to AQI:

$$I_p = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C_p - C_{low}) + I_{low}, \tag{12}$$

where I_p is the AQI for the given pollutant, C_p is the pollutant concentration, C_{low} is the concentration breakpoint that is $\leq C_p$, C_{high} is the concentration breakpoint that is $\geq C_p$, I_{low} is the index breakpoint corresponding to C_{low}, and I_{high} is the index breakpoint corresponding to C_{high} [34] (see Table 1).

Table 1. Kuwait Air Quality Index.

Categories	AQI Sub-Index	O_3 (ppm) 8-h	PM_{10} ($\mu g/m^3$) 24-h	CO (ppm) 24-h	SO_2 (ppm) 24-h	NO_2 (ppm) 24-h
	I_{low}–I_{high}	I_{low}–I_{high}	I_{low}–I_{high}	I_{low}–I_{high}	I_{low}–I_{high}	I_{low}–I_{high}
Good	0–50	0.0–0.03	0.0–90	0.0–4.0	0.0–0.03	0.0–0.03
Moderate	51–100	0.031–0.06	90.1–350.0	4.1–8.0	0.031–0.06	0.04–0.05
Unhealthy (1)	101–150	0.061–0.092	350.1–431.1	8.1–11.7	0.061–0.182	0.06–0.30
Unhealthy (2)	151–200	0.093–0.124	431.4–512.5	11.8–15.4	0.183–0.304	0.31–0.55
Very Unhealthy	201–300	0.125–0.374	512.6–675.0	15.5–30.4	0.305–0.604	0.56–1.04
Hazardous	301–500	0.375–0.504	675.1–1000	30.5–50.4	0.605–1.004	1.05–2.04

Using the data obtained from KEPA, we conducted an in-depth comparative analysis of the different imputation methods. Missing data were entered into each data set, assuming a general missing data pattern and three mechanisms of missing data: MCAR, MAR, and NMAR. Under the MCAR assumption, missing values were randomly applied to each data set. Under the MAR assumption, the probability of information being missing depended on class attribute. Under the NMAR assumption, the largest or smallest values of X_s were removed. The objective of the study was to derive a comparison of six different imputation methods for NMAR, MAR, and MCAR, concerning missing data. We simulated the rates of missing data by varying the value proportions by 5%, 10%, 20%, 30%, and 40%.

2.4. Evaluation Criteria

To determine the best imputation method, three model performance tests were considered [35]: root mean square error (RMSE), mean absolute error (MAE), and correlation coefficient (R), which are calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{13}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{14}$$

where y_i and $\hat{y}_i$ are the ith observations for the reconstructed and the comparison data sets, respectively. The error was measured based on the difference between the estimated value and the observed values. For RMSE and MAE tests, if the value obtained is small, then the estimation method is better.

2.5. R Packages Used for Imputation Process

Five well-known imputation packages accessible in R were applied. The first R package used here was VIM (https://cran.r-project.org/web/packages/VIM/VIM.pdf), which is associated with kNN imputation methods and robust model-based imputation for numerical, semi-continuous, categorical, or ordered variables [36]. The second R package was MICE (https://cran.r-project.org/web/packages/mice/mice.pdf) which stands for Multivariate Imputation via Chained Equations [37]. MICE is specialized to deal with missing values of MAR or MNAR types [38]. MICE can deal with different types of variables using different imputation methods, such as predictive mean matching for numeric variables, logistic regression for binary variables, Bayesian polytomous regression for factor variables, and a proportional odds model for ordered variables [38,39]. The third package was missForest (https://cran.r-project.org/web/packages/missForest/missForest.pdf). MissForest deals with non-parametric imputation [28]. MissForest enables the imputation of the predictors by using regression trees of resampling under the prediction classification of missing values [40]. MissForest has good computational efficiency and can work well with high-dimensional data [28]. The fourth package was Amelia (https://cran.r-project.org/web/packages/Amelia/Amelia.pdf), which enables imputation by maximizing the level of expectation with a bootstrapping algorithm. The Amelia package has also been recommended under a larger number of variables with high-dimensional data. The package also provides improved imputation models by adding Bayesian priors on individual cell values [41]. The final package used was missCompare (https://cran.r-project.org/web/packages/missCompare/missCompare.pdf). The missCompare package provides several diagnostic measurements to compare between all imputation methods, using RMSE, MAE, and other imputation performance criteria.

3. Statistical Results

Based on results for the real-time ambient air quality and meteorological data from the monitoring stations in KEPA, we inferred real-time and fine-grained ambient air quality information using means and standard deviations. The distribution analysis was conducted using the skewness and kurtosis with information of the quartiles (e.g., 25th and 75th quartiles, median, and &IQR&), where the correlation between the predictors was assessed by the Pearson correlation coefficient. The rate of missing values is presented for each monitoring station using the percentage of total number of missing values among the predictors.

Table 2 shows the average air pollutant concentrations. The overall mean and SD for PM_{10}, CO, NO_2, O_3, and SO_2 were 0.23 ± 1.07, 0.91 ± 0.90, 0.04 ± 0.02, 0.02 ± 0.01, and 0.01 ± 0.01, respectively. The missing value rates were 52.16%, 19.37%, 22.35%, 22.40%, and 22.93% from all (N = 9,006), respectively. Figures A1 and A2 from Appendix A show the missing data distribution based on year and monitoring site.

Table 2. Distribution of Kuwait ambient air pollution exposure during 2012–2017. The total daily observations for ASA are N = 1779; for FAH, N = 1820; for JAH, N = 1819; for MAN, N = 1777; and, for RUM, N = 1811.

Air Pollutant	ASA	FAH	JAH	MAN	RUM	All (N = 9006)
PM_{10}						
min	0.017	0.004	0.005	0.008	0.019	0.004
25th	0.099	0.076	0.073	0.099	0.121	0.088
median	0.154	0.109	0.107	0.142	0.211	0.140
75th	0.262	0.163	0.180	0.218	0.273	0.232
max	3.248	5.500	1.714	7.216	2.538	7.216
mean (sd)	0.26 ± 0.32	0.17 ± 0.28	0.17 ± 0.20	0.32 ± 2.38	0.25 ± 0.23	0.23 ± 1.07
%Missing	%53.16	%50.43	%53.12	%53.62	%50.48	%52.16
CO						
min	0.050	0.078	0.015	0.048	0.015	0.015
25th	0.597	0.981	0.107	0.719	0.743	0.562
median	0.720	1.265	0.235	0.922	0.971	0.860
75th	0.945	1.567	0.471	1.172	1.241	1.198
max	2.661	3.789	5.956	4.483	68.980	68.980
mean (sd)	0.80 ± 0.32	1.30 ± 0.47	0.36 ± 0.41	0.98 ± 0.41	1.08 ± 1.68	0.91 ± 0.90
%Missing	%21.57	%17.30	%20.57	%19.57	%17.84	%19.37
NO_2						
min	0.001	0.005	0.004	0.001	0.000	0.000
25th	0.028	0.032	0.014	0.018	0.018	0.020
median	0.038	0.045	0.019	0.029	0.026	0.030
75th	0.052	0.066	0.026	0.046	0.039	0.046
max	0.361	0.182	0.095	0.194	0.183	0.361
mean (sd)	0.04 ± 0.02	0.05 ± 0.03	0.02 ± 0.01	0.03 ± 0.02	0.03 ± 0.02	0.04 ± 0.02
%Missing	%20.89	%17.48	%20.89	%34.87	%17.61	%22.35
O_3						
min	0.001	0.002	0.001	0.003	0.001	0.001
25th	0.014	0.012	0.019	0.017	0.015	0.015
median	0.021	0.018	0.025	0.022	0.023	0.022
75th	0.029	0.024	0.033	0.029	0.031	0.029
max	0.073	0.076	0.062	0.065	0.075	0.076
mean (sd)	0.02 ± 0.01	0.02 ± 0.01	0.03 ± 0.01	0.02 ± 0.01	0.02 ± 0.01	0.02 ± 0.01
%Missing	%20.35	%18.48	%20.98	%34.55	%17.66	%22.40
SO_2						
min	0.000	0.000	0.000	0.001	0.001	0.000
25th	0.006	0.005	0.002	0.003	0.005	0.004
median	0.008	0.009	0.003	0.004	0.007	0.006
75th	0.011	0.019	0.005	0.005	0.011	0.010
max	0.038	0.152	0.049	0.058	0.056	0.152
mean (sd)	0.01 ± 0.00	0.02 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.01
%Missing	%20.53	%17.39	%22.80	%36.19	%17.75	%22.93

All pollutant distributions were positively skewed and we corrected the skewness by applying log transformations [31]. Figure A4 in the Appendix A shows the distribution performance after we applied logarithmic transformations to PM_{10}, SO_2, O_3, CO, and NO_2.

Table 3 shows the Pearson correlation analysis of various air pollutants and meteorological parameters. The strongest positive correlation was found between NO_2 and SO_2. This was expected, due to their common emission sources (e.g., road traffic). NO_2 had a weak association with PM_{10}, whereas O_3 had a highly negative association with NO_2. All meteorological parameters (temperature, humidity, wind speed, and wind direction) showed a negative association with NO_2.

We performed time series plot for each pollutant for each monitoring station to better understand the patterns of the missing data among all observations (see Figures 2–4). We concluded that the missing data pattern can be classified as missing at random (MAR) or missing not at random (MNAR), especially for the large missing gaps (see Appendix A, Figures A1–A3). Figure A3 from Appendix A shows missing observation ratios for each pollutant. From Figure A3, we can conclude that PM10 has the highest missing observation rate among the pollutants (see Appendix A Figure A3-left panel). The right side of the Figure A3 from Appendix A shows the missing value pattern for each pollutant. The vertical connected blocks present the non-randomness for missing data during the monitoring.

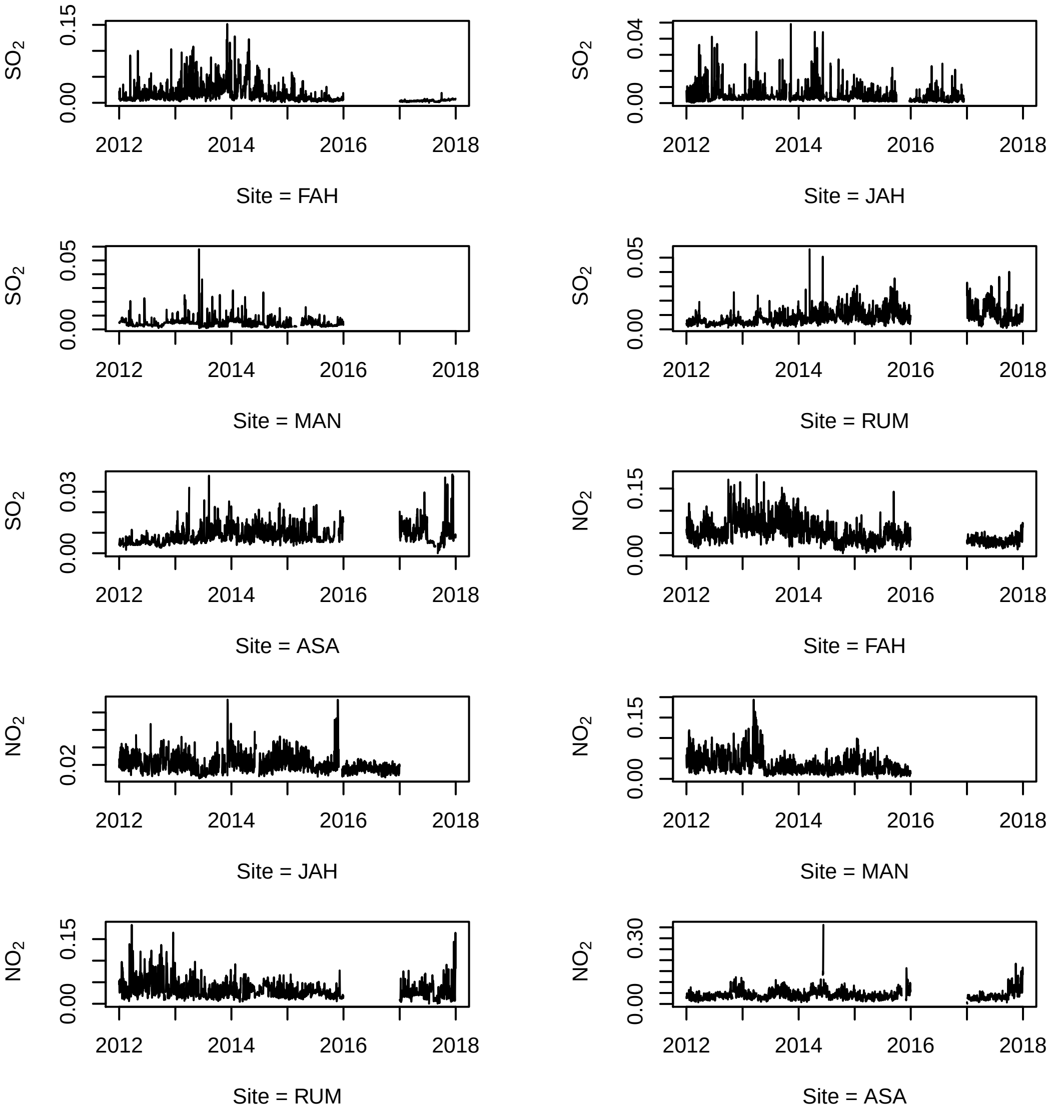


Figure 2. Time-series of air quality monitoring for SO_2 and NO_2 from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait.

Table 3. Correlation analysis between weather climatology and air-pollution components SO_2, NO_2, O_3, CO, and PM_{10}.

	NO_2	O_3	SO_2	CO	PM_{10}	Temp.	Hum.	Wind Speed
NO_2								
O_3	−0.35 ***							
SO_2	0.40 ***	−0.09 ***						
CO	0.35 ***	−0.26 ***	0.22 ***					
PM_{10}	−0.06 ***	0.05 **	−0.03 *	−0.03				
Temp.	−0.09 ***	0.45 ***	−0.06 ***	−0.14 ***	0.05 **			
Hum	−0.02	−0.25 ***	−0.08 ***	0.29 ***	−0.03	−0.61 ***		
Wind Speed	−0.20 ***	0.30 ***	0.13 ***	−0.22 ***	0.10 ***	0.24 ***	−0.32 ***	
Wind Direction	−0.25 ***	0.13 ***	−0.15***	−0.27 ***	0.06 ***	0.14 ***	−0.28 ***	0.31 ***

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$; **** $p < 0.001$.

Table 4 shows a comparison of missing rates for each monitored pollutant between monitoring stations. There were significant differences among the stations in producing missing values, where all *p*-values were less than 0.05, except for that of PM_{10}. PM_{10} was excluded from all imputation calculations, due to a missing rate level that exceeded 50% [42,43].

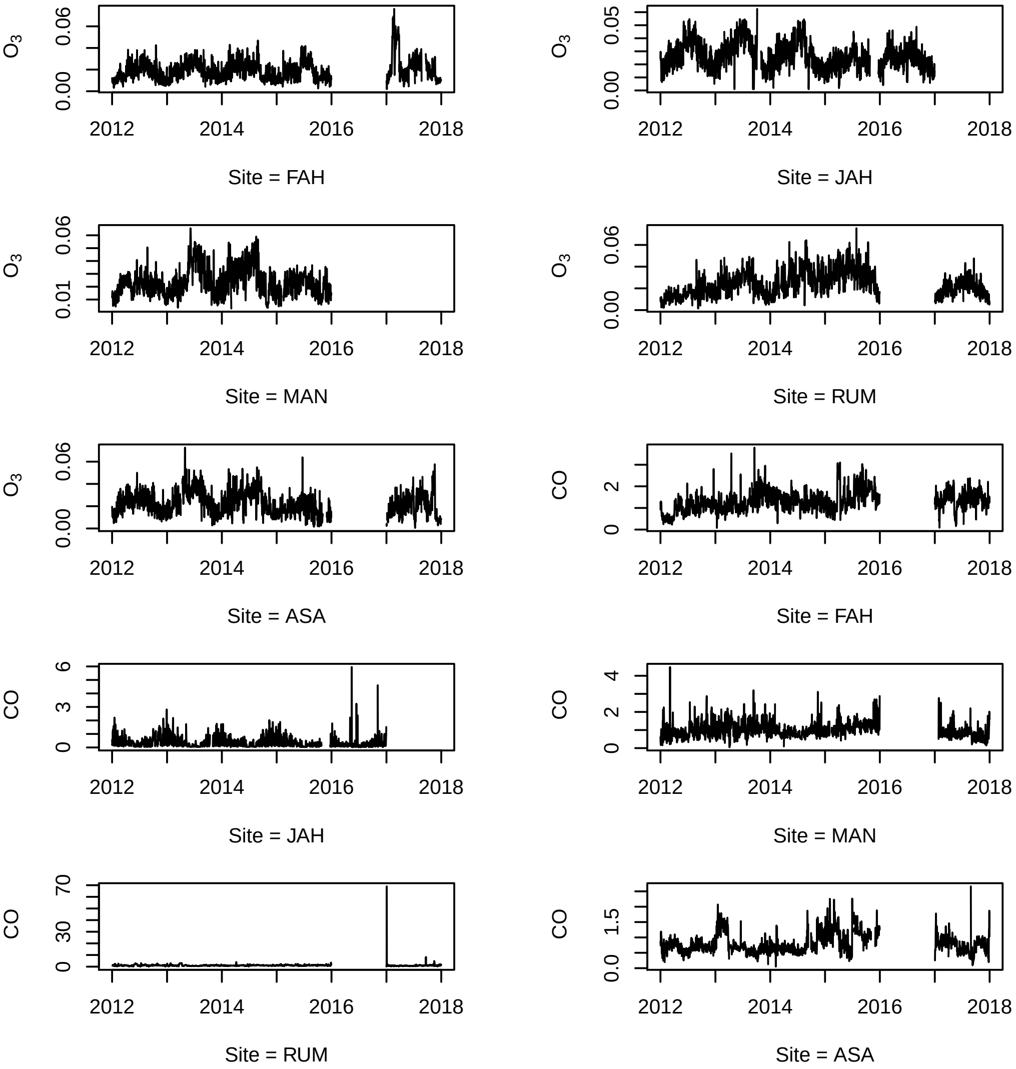


Figure 3. Time-series of air quality monitoring for O_2 and CO from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait.

Table 4. Missing data by site. From the results we conclude that all monitoring fixed stations are different in missing values amount for each pollutant.

	ASA **N = 2192**	**FAH** **N = 2192**	**JAH** **N = 2192**	**MAN** **N = 2192**	**RUM** **N = 2192**	***p*-Value**
NO_2	454 (20.7%)	379 (17.3%)	454 (20.7%)	761 (34.7%)	382 (17.4%)	<0.001
O_3	442 (20.2%)	401 (18.3%)	456 (20.8%)	754 (34.4%)	383 (17.5%)	<0.001
SO_2	446 (20.3%)	377 (17.2%)	496 (22.6%)	790 (36.0%)	385 (17.6%)	<0.001
CO	469 (21.4%)	375 (17.1%)	447 (20.4%)	425 (19.4%)	387 (17.7%)	0.001
PM_{10}	1163 (53.1%)	1103 (50.3%)	1162 (53.0%)	1173 (53.5%)	1104 (50.4%)	0.069

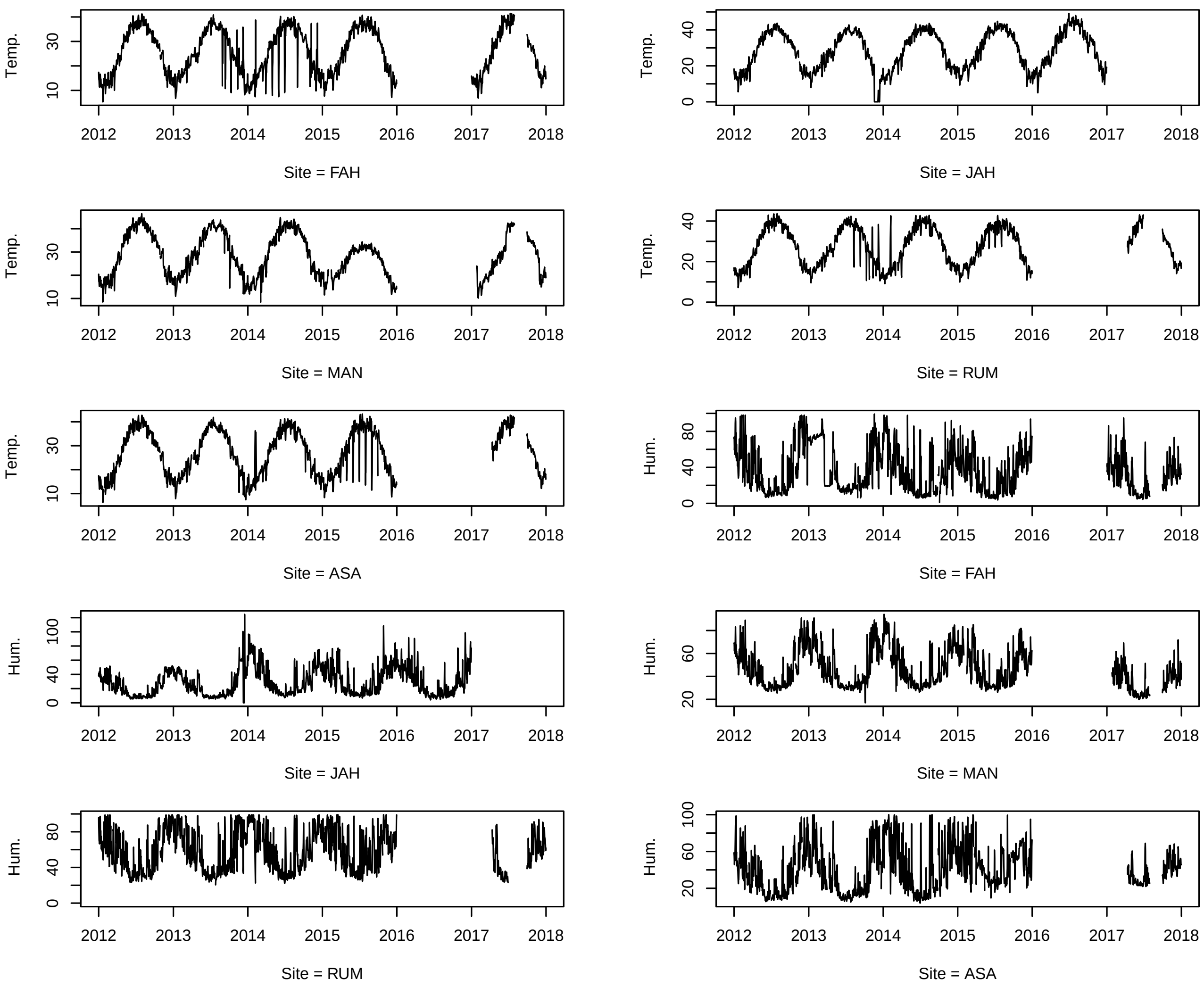


Figure 4. Time-series of weather climatology (temperature and relative humidity) from 2012 to 2017, with missing values from five different locations (stations) in the State of Kuwait.

3.1. Missing Data Patterns

As shown in Table 5 and Figure A5 from Appendix A, the RMSE ranged between 1.029 to 2.110 for MCAR, 1.028 to 1.431 for MAR, and 1.255 to 2.060 for MNAR; thus, MAR had the lowest rate of RMSE among the other missing data approaches. For MAR, the RMSE ranged between 0.821 to 1.145 for MCAR, 0.820 to 1.140 for MAR, and 1.019 to 1.478 for MNAR. This suggests that MAR had the lowest rate of MAR amongst the other missing pattern approaches. This result was consistent with previous studies [44,45]. As seen in Table 5 and appendix Figure A5, the best imputation method for estimating the simulated missing data was the missForest method. The missForest method had the smallest values of MAE and RMSE for all parameters and percentages of simulated missing data rates, this finding was consistent with the study of [1], where MTB was the best imputation method for filling the missing data, as it was able to obtain the smallest error for all percentages of missing data, in agreement with [28,44,46–49]. The second-best imputation method for estimating the simulated missing data was the k-nearest neighbor (kNN) method. This method performed better than the multiple imputation (MI) method for almost all parameters and proportions of missing data. This finding was consistent with the study reported by [42]. The worst-performing methods were multiple imputation using additive regression, bootstrapping, and predictive mean matching (PMM) methods. This was also consistent with the study reported by [42].

From Table 4, we can conclude that the missing rates are different among the selected air monitoring stations for each pollutant except PM_{10} that shows similarities in missing rates among the monitoring stations. In addition, we can figure out from Appendix A Figure A3 how the missing values are distributed for each pollutant.

The results of the missing imputation approach were diagnosed using convergent plots for the mean and standard deviation of the multiple imputation data sets using missForest (see Appendix A Figures A6 and A7). For convergence, the different streams should not show any definite trends; we did not observe any obvious trends in these data. In addition, Figure A8 shows Kernel density estimates for the marginal distributions of the observed data (blue line) and the $m = 20$ densities per variable calculated from the imputed data (red lines). This indicates stability after 10 iterations.

We imputed the missing information into the original data sets to assess if the imputed data are consistent with the existing data. Figures 5 and 6 showed how imputed datasets fit with the actual information in each station. We can see from the figures that large gaps of missing data are filled in the same pattern of the historical values for all pollutants and meteorological parameters which gives a good indication of using missForest to estimate missing air pollutants.

Table 5. RMSE comparison between the indexed original values and the imputed values using missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR) missingness patterns. From the results, it is very obvious that the MAR technique has the lowest RMSE scores among the other techniques. We can also see that missForest had the lowest RMSE and MAE, among the other imputation methods, for all missing rate criteria.

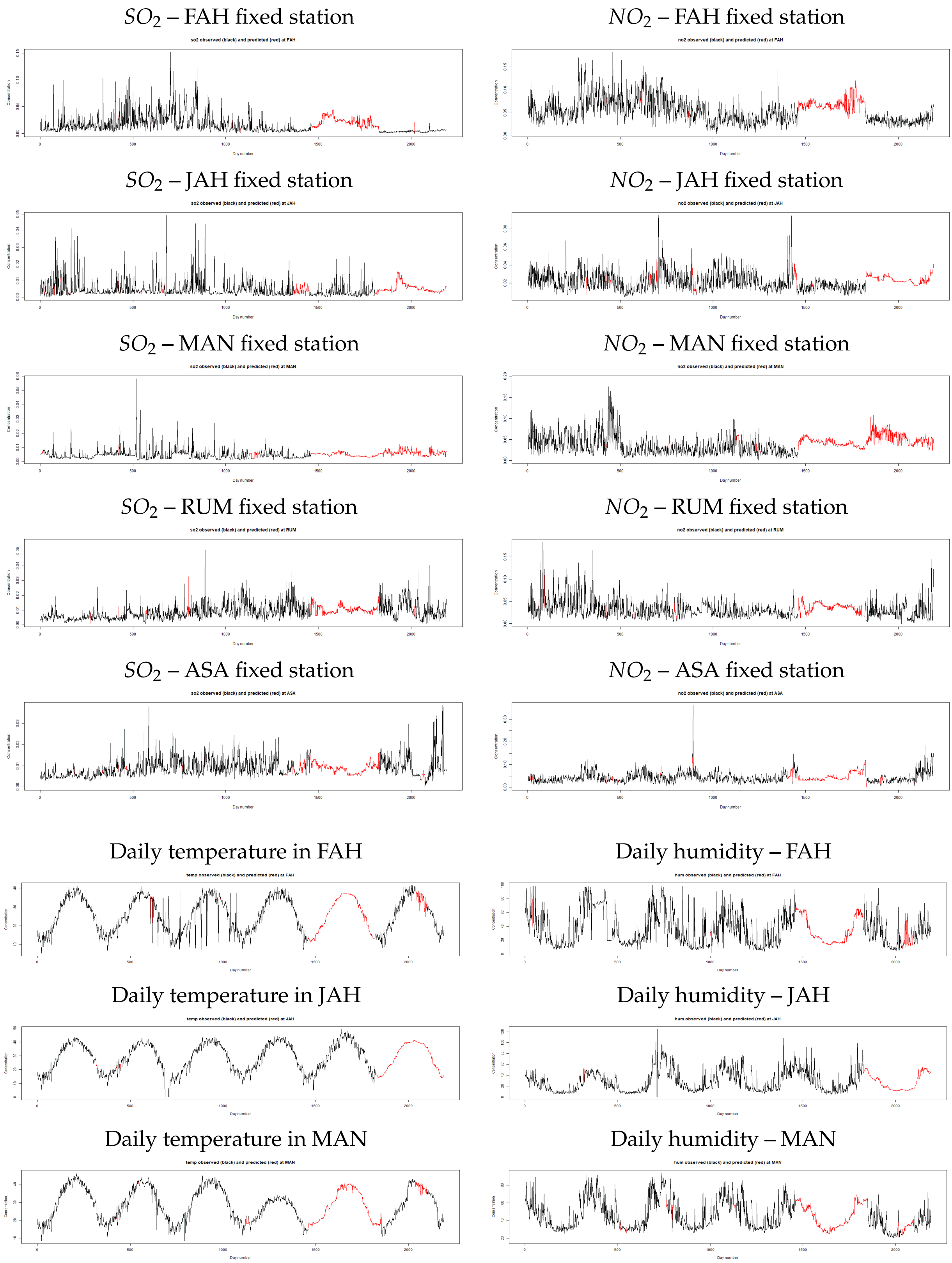
Method	5% Missingness Rate					
	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.430	1.405	1.536	1.145	1.120	1.238
PMM	1.408	1.430	1.529	1.129	1.140	1.225
RF	1.413	1.412	1.547	1.128	1.126	1.242
missForest	1.031	1.035	1.270	0.821	0.823	1.036
BPCA	2.110	1.199	1.568	1.686	0.953	1.251
kNN	1.064	1.065	1.288	0.850	0.846	1.047

Method	10% missingness rate					
	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.408	1.431	1.517	1.125	1.140	1.218
PMM	1.414	1.415	1.527	1.125	1.131	1.229
RF	1.414	1.416	1.529	1.129	1.133	1.231
missForest	1.035	1.028	1.260	0.829	0.820	1.025
BPCA	1.816	1.792	1.813	1.456	1.431	1.449
kNN	1.063	1.064	1.282	0.853	0.846	1.041

Method	20% missingness rate					
	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.415	1.410	1.523	1.129	1.124	1.225
PMM	1.418	1.417	1.528	1.129	1.131	1.226
RF	1.413	1.408	1.532	1.128	1.124	1.228
missForest	1.029	1.038	1.253	0.819	0.827	1.019
BPCA	1.653	1.548	1.856	1.319	1.233	1.478
kNN	1.062	1.065	1.270	0.847	0.850	1.032

Method	30% missingness rate					
	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.405	1.410	1.531	1.124	1.127	1.232
PMM	1.418	1.419	1.527	1.131	1.132	1.229
RF	1.419	1.419	1.521	1.136	1.134	1.224
missForest	1.034	1.033	1.255	0.825	0.823	1.023
BPCA	1.891	1.622	2.060	1.506	1.293	1.645
kNN	1.065	1.064	1.276	0.850	0.848	1.036

Method	40% missingness rate					
	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
EM	1.401	1.411	1.518	1.119	1.127	1.222
PMM	1.411	1.399	1.520	1.126	1.116	1.222
RF	1.412	1.419	1.534	1.124	1.133	1.234
missForest	1.032	1.035	1.259	0.823	0.827	1.027
BPCA	1.564	1.264	1.789	1.250	1.007	1.428
kNN	1.062	1.067	1.279	0.847	0.852	1.042


SO_2 – FAH fixed station

NO_2 – FAH fixed station

SO_2 – JAH fixed station

NO_2 – JAH fixed station

SO_2 – MAN fixed station

NO_2 – MAN fixed station

SO_2 – RUM fixed station

NO_2 – RUM fixed station

SO_2 – ASA fixed station

NO_2 – ASA fixed station

Daily temperature in FAH

Daily humidity – FAH

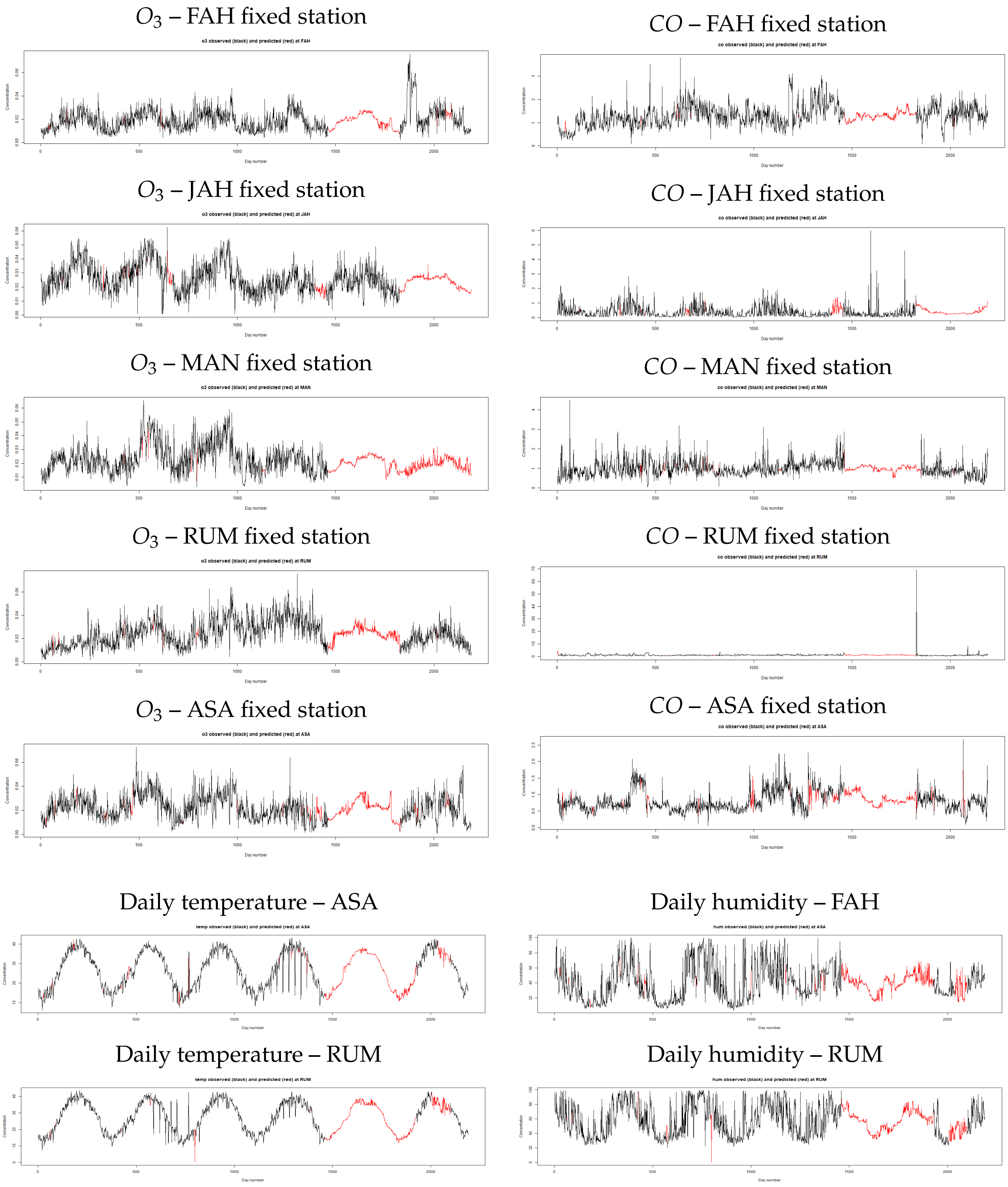
Daily temperature in JAH

Daily humidity – JAH

Daily temperature in MAN

Daily humidity – MAN

Figure 5. Daily concentrations of SO_2, NO_2, temperature, and relative humidity after estimating missing values using the missForest approach (from 2012–2017).


O_3 – FAH fixed station

CO – FAH fixed station

O_3 – JAH fixed station

CO – JAH fixed station

O_3 – MAN fixed station

CO – MAN fixed station

O_3 – RUM fixed station

CO – RUM fixed station

O_3 – ASA fixed station

CO – ASA fixed station

Daily temperature – ASA

Daily humidity – FAH

Daily temperature – RUM

Daily humidity – RUM

Figure 6. Daily concentrations of O_3, CO, temperature, and relative humidity after estimating missing values using the missForest approach (from 2012–2017).

4. Discussion

In Kuwait, the Environmental Public Authority (KEPA) is responsible for monitoring the air quality status. The data of air quality obtained from the five stations used in this study usually contain missing data, which can cause bias due to systematic errors between the observed and unobserved values [31]. Therefore, it is vital to determine the optimal approach for estimating the missing values, in order to guarantee that the analyzed data are of high quality. Incomplete data matrices may provide outcomes that vary significantly, compared to the results expected from a data set that is complete [4]. The primary purpose of any data analysis is to make valid and reasonable inferences on a particular population under study. A researcher is expected to respond to the missing data problem in a way that aligns with the population of interest.

There have been many contributions to this field, such as in environmental [1,7,50,51], statistical [52,53], and medical studies [54,55]. In the environmental field, imputation is the statistical procedure of assigning inferential values to recover all missing data using prior knowledge from other predictors.

The existence of efficient imputation algorithms has led to the extensive usage of elaborate imputation methods across the world. As more people become knowledgeable about imputation algorithms, inquisitiveness regarding the methodology increases, leading to the invention of more sophisticated imputation methods. However, the main challenge concerning imputed values is whether to consider them as actual measurements or to be handled with caution. In the field of research, it is preferable to handle assigned figures with great discretion. This is because the use of imputed figures as actual data may lead to a misguided impression, which may potentially falsify the final results. Therefore, the imputed values should be given low priority.

It is, therefore, vital for a researcher to impute missing data and assess how robust the associated data estimation is. Environmental information that relies on technological processing and simulation poses a challenge. Missing data ascription is one approach: A substantial quality of ascription methods is that they are reliable and limited to one type of variable. This variable may be considered as persistent or unmitigated. If the data type is blended, the method must deal with the different types of data separately. In conclusion, these techniques ignore the potential associations between different factor types. For the situation here, before conducting any statistical modeling or performing time-series analysis, it is better to treat the missing values and to try to estimate them using other information from other predictors. This may help to avoid any bias circumstance and to enhance model performance for better estimation.

The main contribution of this paper was to find the most appropriate method to fill in missing observations in an air pollution data set from Kuwait. Single and multiple imputation methods were adopted and their performances were compared using using the RMSE and MAE metrics. To estimate missing data for SO_2, NO_2, PM_{10}, CO, and O_3 in the KEPA database, we applied artificially introduced missing values ranging from 10% to 40%. We showed that missForest could successfully handle the missing values, particularly in data sets including different types of environmental variables.

However, this computation method also had limitations. It requires proficiency in R programming, being demanding in comparison to the kNN or PMM methods. There is also a possible connection between the pollutant values and the missing variables. Therefore, these results are not applicable in cases where the missing data are due to non-random reasons. It is evident that some of the observed air pollutant records contained erroneous information. When we ignore this factor during the examination, the results obtained tend to be misleading.

Our findings revealed that missForest was the only imputation method with a consistent and comparatively lower imputation error (of 0.82). The approach had a root mean square error of 1.04. missForest also exhibited the smallest prediction deviation in the imputed values of pollutants. Furthermore, missForest simulation provides the most readily available imputation of missing values, as its freeware R package is freely available.

While compiling the report of the study, we assumed the missing at random (MAR) tool. This premise is essential for the development of a prototype of the observation for the imputation of missing data. There was a possibility of the missing data system being not missing at random (NMAR). In such a case, the missing variables are directly related to their causes. It may be challenging to determine the actual missing data mechanism, in such a case. Therefore, distinguishing between NMAR and MAR would involve a thorough investigation of the data capturing process. Other assumptions include Gaussian-distributed data, which may have been erroneous for some variables. Using the appropriate distribution for each variable can help to reduce this error. This might increase the reliability of the MICE imputation results, which determine the mechanism for each variable.

5. Conclusions

Missing data are always lost, in their entirety and forever, but a proper imputation scheme can help to remedy the situation as much as possible. The method that performs best in each situation, in terms of the assessments, is made in this work. For this study, missForest gives the most accurate results in estimating the missing values through the multi-dimensional dataset (the datasets that came from five fixed monitoring stations). The missForest method enables imputation on virtually any kind of data. In particular, it can deal with multivariate information comprised of continuous and categorical factors at the same time. This method does not require parameter tuning, nor does it require assumptions about the distribution of the information. Finally, missForest had the least imputation error for both continuous and categorical variables at each frequency of missingness rates (5%, 10%, 20%, 30%, and 40%), and it had the smallest prediction error difference when models used imputed values.

Author Contributions: A.R.A. developed the research methodology, analyzed the data using STATA R coding, and finished writing the manuscript; J.P. contributed for review and supervision, analyzed the data, and made all figures, graphs, and tables; A.A.-H. defined air pollutants, performed the AQI calculations, and carried out the environmental literature works with provision of the KEPA data. All authors have read and agreed to the published version of the manuscript.

Funding: We declare that this research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to express thanks for the support given by the Data Management Department in Kuwait Environmental Protection Authority (EPA) for providing pollutants data through Environmental Monitoring Information System of Kuwait (eMISK). Without their contribution and support, this research would not have been possible.

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

Appendix A. Figures

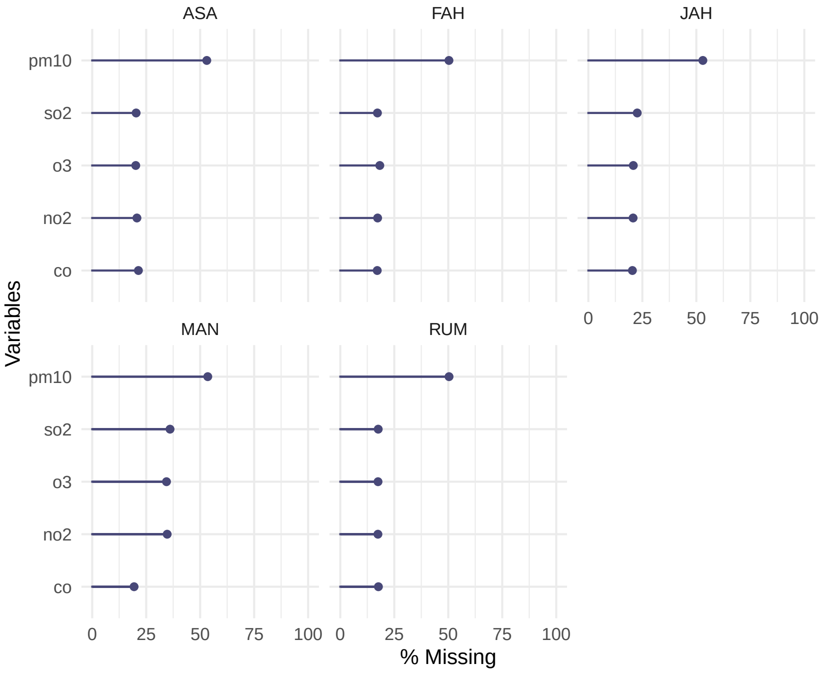


Figure A1. Missing values for air quality pollutants from 2012 to 2017 per fixed station.

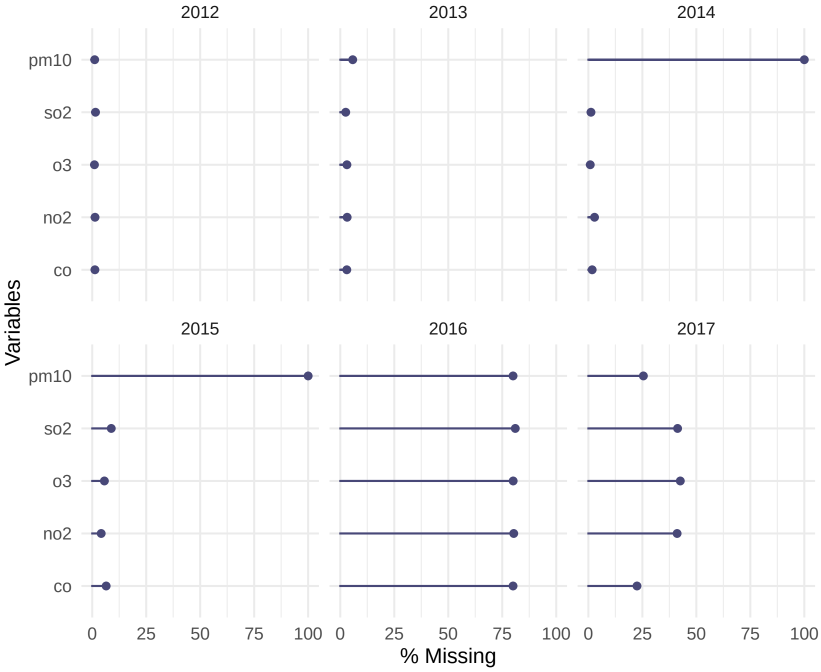


Figure A2. Missing values for air quality pollutants from 2012 to 2017 per year.

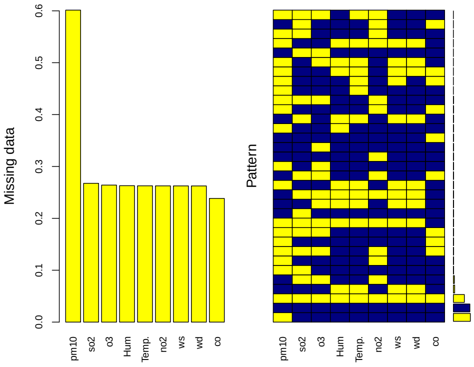


Figure A3. Missing value patterns for air quality measurements from 2012 to 2017. **Left**: Frequency of missingness in each variable. **Right**: Observed missingness patterns in the data set. The least frequent occurring patterns are located at the top of the plot, with gradually increasing frequency towards the bottom. Blue: observed, Yellow: missing.

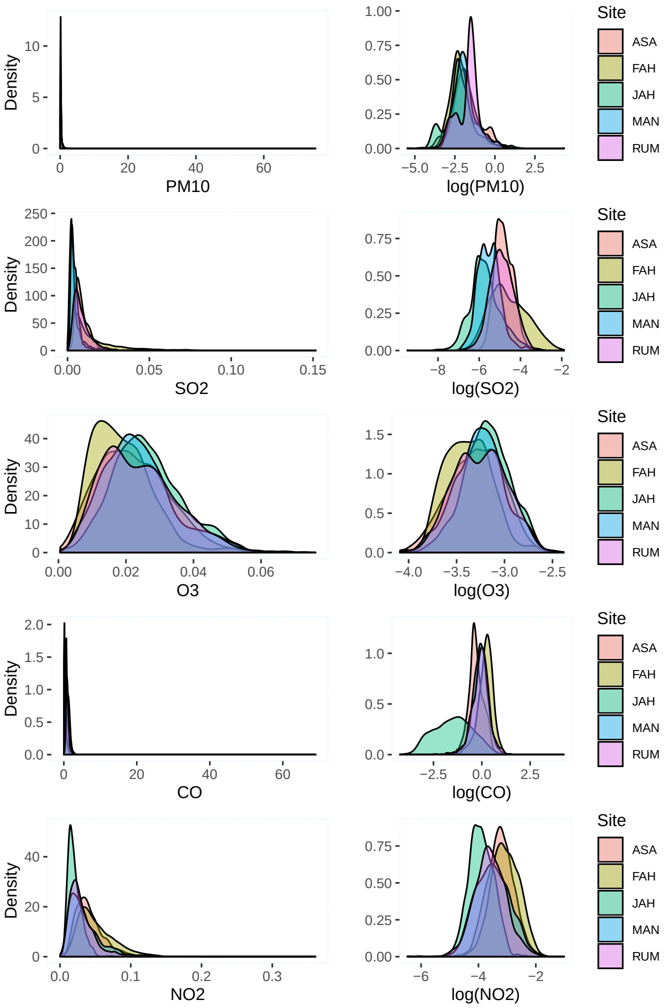
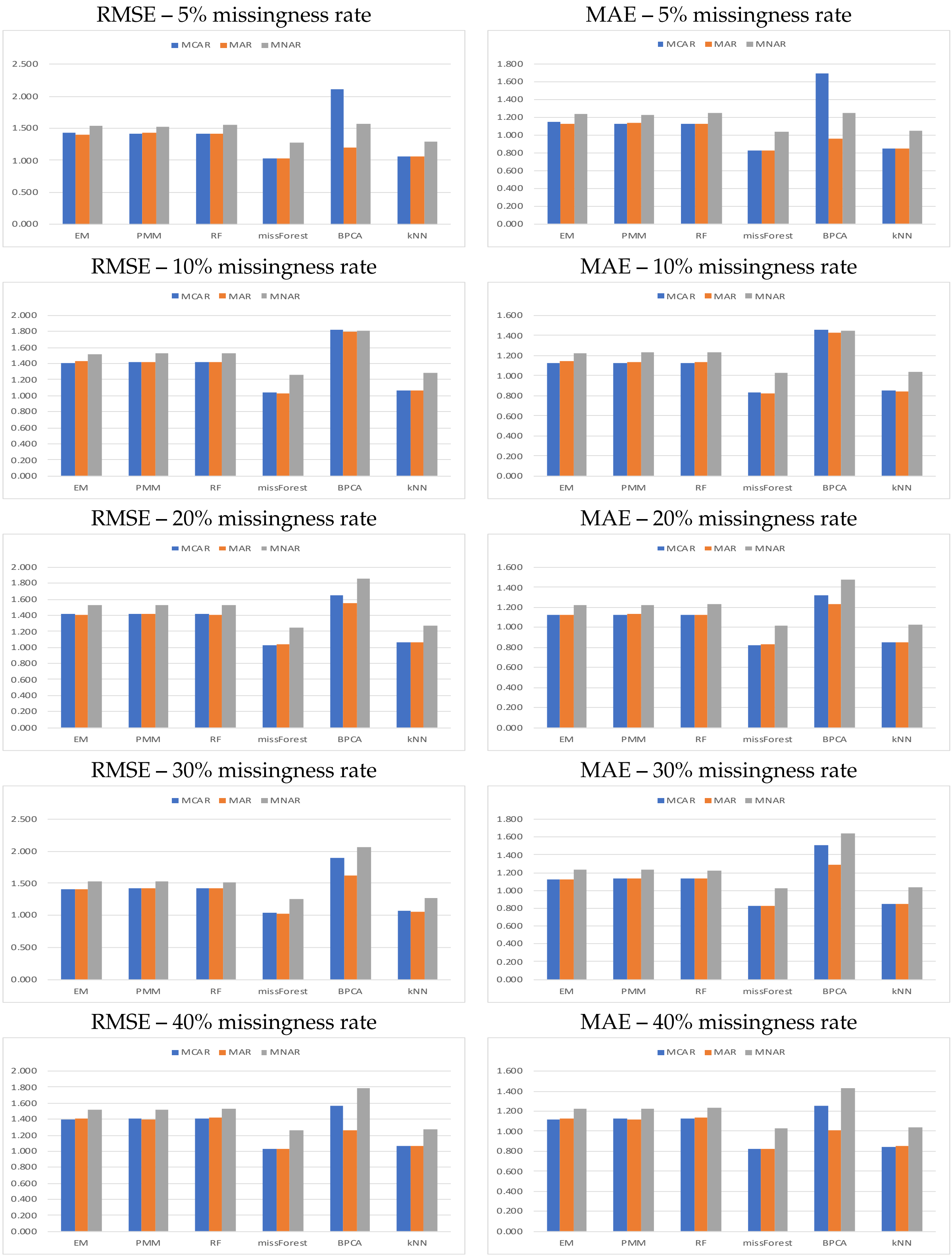


Figure A4. Distribution analysis for PM_{10}, SO_2, O_3, CO, and NO_2 during 2012 to 2017, according to site location in the State of Kuwait. It is very obvious that log transformation fixes the distribution shape for all pollutants. This step is very important—that is, normalizing the skewed data, such that they approximately conform to normality—in order to use them in the imputational calculation for more accurate results [56].



RMSE – 5% missingness rate

MAE – 5% missingness rate

RMSE – 10% missingness rate

MAE – 10% missingness rate

RMSE – 20% missingness rate

MAE – 20% missingness rate

RMSE – 30% missingness rate

MAE – 30% missingness rate

RMSE – 40% missingness rate

MAE – 40% missingness rate

Figure A5. Mean RMSE and MAE results for the Kuwait Environmental Public Authority (KEPA) data, in order to estimate missing values for SO_2, NO_2, CO, and O_3 after eliminating PM_{10} due to a high level of missing values. Results are shown for MCAR (**left**), MAR (**middle**), and MNAR (**right**) data.

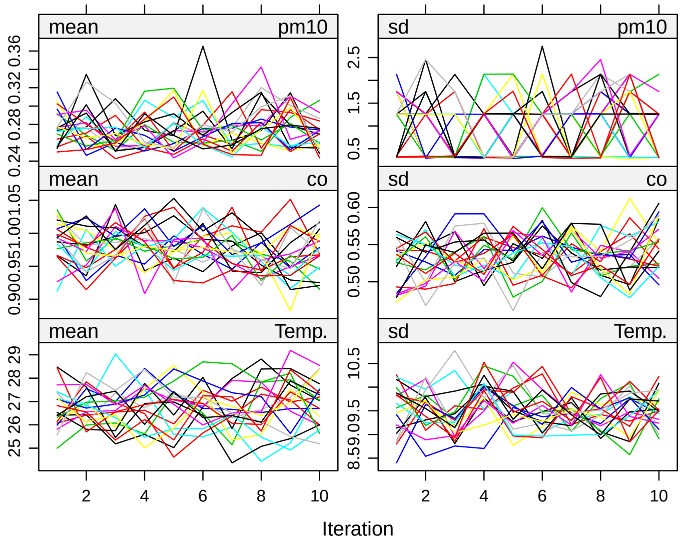


Figure A6. Inspecting the trace line convergence levels using an iterative Markov Chain–Monte Carlo type of algorithm with respect to the imputed means and standard deviations for PM_{10}, CO, and temperature. These trace plots show the imputed value summaries for all imputed data sets with $m = 20$ after applying 10 iterations, in order to reach to the convergence level of stability.

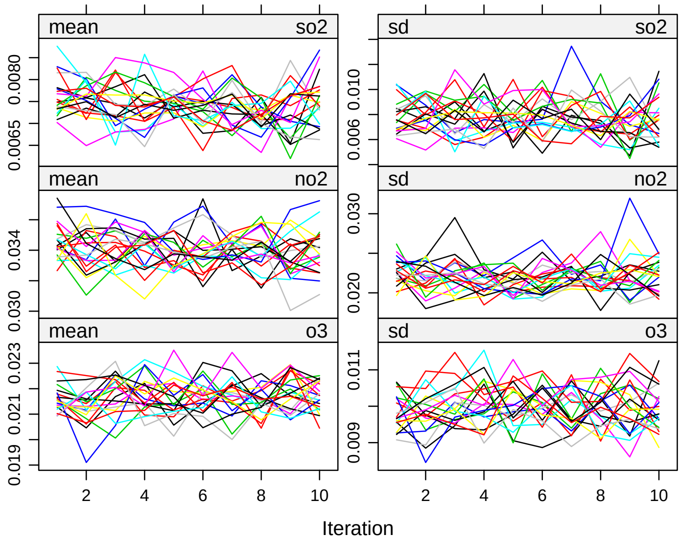


Figure A7. Inspecting the trace line convergence levels using an iterative Markov Chain–Monte Carlo type of algorithm with respect to the imputed means and standard deviations for SO_2, NO_2, and O_3. These trace plots show the imputed value summaries for all imputed data sets with $m = 20$ after applying 10 iterations, in order to reach to the convergence level of stability. Each color in the graph represents an imputed data set, where the x-axis represents the number of iterations implemented during the imputational calculation and the y-axis represents the mean (**left**-side) and standard deviation (**right**-side) of the imputed values only.

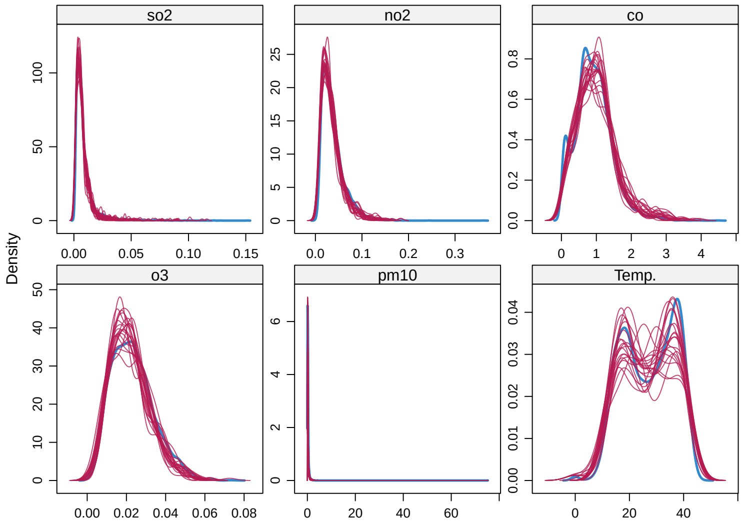


Figure A8. Density plots with multiple imputations for SO_2, NO_2, PM_{10}, CO, and O_3 data. The blue line represents the observed data and the red lines are the density plots of the 20 imputed data sets. As we can see, in all density plots, the red lines almost match the blue line (the observed data), which is an indication of matching between the observed and imputed values.

Appendix B. Algorithms-MissForest

Algorithm A1: Impute missing values with random forest [28].

```
Require: $\mathbf{X}$ is an $n \times p$ matrix, setup stopping criterion ($\gamma$)
  setup initial guess for missing values;
  $\mathbf{k} \leftarrow$ vector of sorted indices of columns in $\mathbf{X}$ w.r.t. increasing amount of missing values;
  while not $\gamma$ do
    $\mathbf{X}^{imp}_{old} \leftarrow$ store previously imputed matrix;
    for $s$ in $\mathbf{k}$ do
      Fit a random forest: $\mathbf{y}^{(s)}_{\text{obs}} \sim \mathbf{x}^{(s)}_{\text{obs}}$;
      Predict $\mathbf{y}^{(s)}_{\text{mis}}$ using $\mathbf{x}^{(s)}_{\text{mis}}$;
      $\mathbf{X}^{imp}_{new} \leftarrow$ update imputed matrix, using predicted $\mathbf{y}^{(s)}_{\text{mis}}$;
    end for
    update $\gamma$
  end while
  return the imputed matrix $\mathbf{X}^{imp}$
```

References

1. Norazian, M.N.; Shukri, Y.A.; Azam, R.N.; Al Bakri, A.M.M. Estimation of missing values in air pollution data using single imputation techniques. *Sci. Asia* **2008**, *34*, 341–345. [CrossRef]
2. Norris, G.; Duvall, R.; Brown, S.; Bai, S. *Epa Positive Matrix Factorization (pmf) 5.0 Fundamentals and User Guide Prepared for the Us Environmental Protection Agency Office of Research and Development*; Petaluma Inc.: Washington, DC, USA, 2014.
3. Junger, W.; de Leon, A.P. Missing data imputation in time series of air pollution. *Epidemiology* **2009**, *20*, S87. [CrossRef]
4. Forbes, D.; Hawthorne, G.; Elliott, P.; McHugh, T.; Biddle, D.; Creamer, M.; Novaco, R.W. A concise measure of anger in combat-related posttraumatic stress disorder. *J. Trauma. Stress Off. Publ. Int. Soc. Trauma. Stress Stud.* **2004**, *17*, 249–256. [CrossRef] [PubMed]
5. Jadhav, A.; Pramod, D.; Ramanathan, K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Appl. Artif. Intell.* **2019**, *33*, 913–933. [CrossRef]
6. Hawthorne, G.; Hawthorne, G.; Elliott, P. Imputing cross-sectional missing data: Comparison of common techniques. *Aust. N. Z. J. Psychiatry* **2005**, *39*, 583–590. [CrossRef]

7. Plaia, A.; Bondi, A. Single imputation method of missing values in environmental pollution data sets. *Atmos. Environ.* **2006**, *40*, 7316–7330. [CrossRef]
8. Farhangfar, A.; Kurgan, L.; Dy, J. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognit.* **2008**, *41*, 3692–3705. [CrossRef]
9. Graham, J.W. Missing data analysis: Making it work in the real world. *Annu. Rev. Psychol.* **2009**, *60*, 549–576. [CrossRef]
10. Rubin, D.B. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **1996**, *91*, 473–489. [CrossRef]
11. Little, R.J.; Rubin, D.B. *Statistical Analysis with Missing Data*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, **2019**; Volume 793.
12. Schafer, J.L.; Olsen, M.K. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivar. Behav. Res.* **1998**, *33*, 545–571. [CrossRef]
13. Haji-Maghsoudi, S.; Haghdoost, A.A.; Rastegari, A.; Baneshi, M.R. Influence of pattern of missing data on performance of imputation methods: An example using national data on drug injection in prisons. *Int. J. Health Policy Manag.* **2013**, *1*, 69. [CrossRef] [PubMed]
14. Lee, K.J.; Carlin, J.B. Recovery of information from multiple imputation: A simulation study. *Emerg. Themes Epidemiol.* **2012**, *9*, 3. [CrossRef] [PubMed]
15. Marshall, A.; Altman, D.G.; Holder, R.L. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: A resampling study. *BMC Med. Res. Methodol.* **2010**, *10*, 112. [CrossRef] [PubMed]
16. Heitjan, D.F.; Rubin, D.B. Inference from coarse data via multiple imputation with application to age heaping. *J. Am. Stat. Assoc.* **1990**, *85*, 304–314. [CrossRef]
17. King, G.; Honaker, J.; Joseph, A.; Scheve, K. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am. Political Sci. Rev.* **2001**, *95*, 49–69. [CrossRef]
18. Newman, D.A. Missing data: Five practical guidelines. *Organ. Res. Methods* **2014**, *17*, 372–411. [CrossRef]
19. Allison, P.D. Multiple imputation for missing data: A cautionary tale. *Sociol. Methods Res.* **2000**, *28*, 301–309. [CrossRef]
20. White, I.R.; Daniel, R.; Royston, P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Comput. Stat. Data Anal.* **2010**, *54*, 2267–2275. [CrossRef]
21. Allen, R.J.; DeGaetano, A.T. Estimating missing daily temperature extremes using an optimized regression approach. *Int. J. Climatol. J. R. Meteorol. Soc.* **2001**, *21*, 1305–1319. [CrossRef]
22. Kotsiantis, S.; Kostoulas, A.; Lykoudis, S.; Argiriou, A.; Menagias, K. Filling missing temperature values in weather data banks. In Proceedings of the 2006 2nd IET International Conference on Intelligent Environments, Athens, Greece, 5–6 July 2006; Volume 1, pp. 327–334.
23. Jagannathan, G.; Wright, R.N. Privacy-preserving imputation of missing data. *Data Knowl. Eng.* **2008**, *65*, 40–56. [CrossRef]
24. Lakshminarayan, K.; Harp, S.A.; Samad, T. Imputation of missing data in industrial databases. *Appl. Intell.* **1999**, *11*, 259–275. [CrossRef]
25. Van Ginkel, J.R.; Van der Ark, L.A.; Sijtsma, K.; Vermunt, J.K. Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Comput. Stat. Data Anal.* **2007**, *51*, 4013–4027. [CrossRef]
26. Schenker, N.; Taylor, J.M. Partially parametric techniques for multiple imputation. *Comput. Stat. Data Anal.* **1996**, *22*, 425–446. [CrossRef]
27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
28. Stekhoven, D.J.; Bühlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
29. Oba, S.; Sato, M.a.; Takemasa, I.; Monden, M.; Matsubara, K.I.; Ishii, S. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **2003**, *19*, 2088–2096. [CrossRef]
30. Norazian, M.N.; Shukri, A.; Yahaya, P.; Azam, N.; Ramli, P.; Fitri, N.F.; Yusof, M.; Mohd Mustafa Al Bakri, A. Roles of imputation methods for filling the missing values: A review. *Adv. Environ. Biol.* **2013**, 7, 3861–3869.
31. Alsaber, A.; Pan, J.; Al-Herz, A.; Alkandary, D.S.; Al-Hurban, A.; Setiya, P.; KRRD Group. Influence of ambient air pollution on rheumatoid arthritis disease activity score Index. *Int. J. Environ. Res. Public Health* **2020**, *17*, 416. [CrossRef]
32. Al-Shayji, K.; Lababidi, H.; Al-Rushoud, D.; Al-Adwani, H. Development of a fuzzy air quality performance indicator. *Kuwait J. Sci. Eng.* **2008**, *35*, 101–126.
33. Johnson, M.; Isakov, V.; Touma, J.; Mukerjee, S.; Özkaynak, H. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. *Atmos. Environ.* **2010**, *44*, 3660–3668. [CrossRef]
34. Fitz-Simons, T. *Guideline for Reporting of Daily Air Quality: Air Quality Index (AQI)*; Technical Report; Environmental Protection Agency, Office of Air Quality Planning and Standards: Research Triangle Park, NC, USA, 1999.
35. Bennett, N.; Croke, B.; Guariso, G.; Guillaume, J.A.; Hamilton, S.H.; Jakeman, A.J.; Marsili-Libelli, S.; Newham, L.T.H.; Norton, J.P.; Perrin, C.; et al. Characterising performance of environmental models. *Environ. Modell. Softw.* **2013**, *40*, 1–20. [CrossRef]
36. Kowarik, A.; Templ, M. Imputation with the R Package VIM. *J. Stat. Softw.* **2016**, *74*, 1–16. [CrossRef]
37. Royston, P. Multiple imputation of missing values. *Stata J.* **2004**, *4*, 227–241. [CrossRef]
38. Buuren, S.V.; Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **2010**, *85*, 1–68. [CrossRef]
39. Horton, N.J.; Lipsitz, S.R. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *Am. Stat.* **2001**, *55*, 244–254. [CrossRef]

40. Liao, S.G.; Lin, Y.; Kang, D.D.; Chandra, D.; Bon, J.; Kaminski, N.; Sciurba, F.C.; Tseng, G.C. Missing value imputation in high-dimensional phenomic data: Imputable or not, and how? *BMC Bioinform.* **2014**, *15*, 346. [CrossRef]
41. Honaker, J.; King, G.; Blackwell, M. Amelia II: A program for missing data. *J. Stat. Softw.* **2011**, *45*, 1–47. [CrossRef]
42. Zakaria, N.A.; Noor, N.M. Imputation methods for filling missing data in urban air pollution data formalaysia. *Urban. Arhit. Constr.* **2018**, *9*, 159.
43. Bertsimas, D.; Pawlowski, C.; Zhuo, Y.D. From predictive methods to missing data imputation: An optimization approach. *J. Mach. Learn. Res.* **2017**, *18*, 7133–7171.
44. Valdiviezo, H.C.; Van Aelst, S. Tree-based prediction on incomplete data using imputation or surrogate decisions. *Inform. Sci.* **2015**, *311*, 163–181. [CrossRef]
45. Junger, W.; De Leon, A.P. Imputation of missing data in time series for air pollutants. *Atmos. Environ.* **2015**, *102*, 96–104. [CrossRef]
46. Kokla, M.; Virtanen, J.; Kolehmainen, M.; Paananen, J.; Hanhineva, K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: A comparative study. *BMC Bioinform.* **2019**, *20*, 1–11. [CrossRef] [PubMed]
47. Tang, F.; Ishwaran, H. Random forest missing data algorithms. *Stat. Anal. Data Min. ASA Data Sci. J.* **2017**, *10*, 363–377. [CrossRef] [PubMed]
48. Ishak, A.B.; Daoud, M.B.; Trabelsi, A. Ozone concentration forecasting using statistical learning approaches. *J. Mater. Environ. Sci.* **2017**, *8*, 4532–4543.
49. Shah, A.D.; Bartlett, J.W.; Carpenter, J.; Nicholas, O.; Hemingway, H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am. J. Epidemiol.* **2014**, *179*, 764–774. [CrossRef]
50. 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]
51. Kabir, G.; Tesfamariam, S.; Hemsing, J.; Sadiq, R. Handling incomplete and missing data in water network database using imputation methods. *Sustain. Resilient Infrastruct.* **2019**, *5*, 1–13. [CrossRef]
52. Di Zio, M.; Guarnera, U.; Luzi, O. Imputation through finite Gaussian mixture models. *Comput. Stat. Data Anal.* **2007**, *51*, 5305–5316. [CrossRef]
53. Huisman, M. Imputation of missing network data: Some simple procedures. *J. Soc. Struct.* **2020**, *10*, 1–29.
54. Sartori, N.; Salvan, A.; Thomaseth, K. Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. *Comput. Stat. Data Anal.* **2005**, *49*, 937–953. [CrossRef]
55. Branden, K.V.; Verboven, S. Robust data imputation. *Comput. Biol. Chem.* **2009**, *33*, 7–13. [CrossRef] [PubMed]
56. Changyong, F.; Hongyue, W.; Naiji, L.; Tian, C.; Hua, H.; Ying, L.; Xin, M. Log-transformation and its implications for data analysis. *Shanghai Arch. Psychiatry* **2014**, *26*, 105.