Penerbit Akademia Baru

Journal of Advanced Research Design

Journal homepage:
https://akademiabaru.com/submit/index.php/ard
ISSN: 2289-7984


# Comparative Analysis of Imputation Methods for Missing Environmental Data: A Case Study on Ozone Concentrations

Nurliyana Juhan[1,*], Siti Noradiah Jamaludin[2], Yong Zulina Zubairi[3], Dg Siti Nurisya Sahirah Ag Isha[4], Nur Idayu Ah Khaliludin[5]

[1] Preparatory Center for Science and Technology, University Malaysia Sabah (UMS), Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia
[2] Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia
[3] Institute for Advanced Studies, University of Malaya, 50603 Kuala Lumpur, Malaysia
[4] Faculty of Science and Natural Resources, University Malaysia Sabah (UMS), Jalan UMS, 88400, Kota Kinabalu, Sabah, Malaysia
[5] Tahmidi Centre, Universiti Sains Islam Malaysia, Bandar Baru Nilai 71800, Nilai, Negeri Sembilan, Malaysia

**ARTICLE INFO**

***Article history:***
Received 10 January 2025
Received in revised form 17 February 2025
Accepted 2 June 2025
Available online 13 June 2025

***Keywords:***
Imputation method; missing data; mean-before-after; ozone concentrations

**ABSTRACT**

Handling missing values is crucial to environmental data analysis since missing datasets can lead to biassed results. Using Weibull distributions, this study compared six single-imputation methods (mean, median, mean-before-after (MBA), cubic interpolation, linear interpolation, last observation carried forward (LOCF)) for estimating missing ozone concentration data in Petaling Jaya, Selangor. The present study simulated data for sample sizes of 50 and 150 with varying missing value percentages (5%, 10%, 15%, 20% and 25%). The performance of each imputation method was evaluated using prediction accuracy, root mean square error (RMSE) and mean absolute error (MAE). The findings suggested that the MBA approach outperformed all examined cases, followed by linear interpolation and LOCF. Conversely, cubic interpolation, mean, and median substitution approaches performed poorly, especially as the proportion of missing data increased. This study emphasises the critical role of selecting appropriate imputation methods to enable accurate and trustworthy environmental data analysis. The findings can help researchers select efficient approaches for addressing missing values in air quality datasets, thus boosting the reliability of environmental studies.

## 1. Introduction

The notion of missing values pertains to a scenario when the dataset comprises either empty or incomplete values [1]. There is a possibility that missing values are present in a diverse range of industry and research databases, including in the environmental and air pollution datasets [2-5]. The presence of these missing values in air pollution data arises from multiple sources, including equipment malfunctions, errors during manual data entry, and inaccurate measurements themselves [2,3,6,7]. These incomplete observations can introduce bias and hinder the interpretability of results during data analysis [5]. Moreover, mishandling these incomplete observations can significantly

* *Corresponding author.*
*E-mail address: liyana87@ums.edu.my*

https://doi.org/10.37934/ard.134.1.6376

impair the effectiveness of the data, introduce bias, and ultimately, erroneous inferences being made from the research [5,7,8].

The imputation method serves as a crucial tool for managing missing values [5,6,9,10]. It addresses missing data by replacing them with plausible estimates, creating a complete dataset for subsequent analysis using standard methods and software [9,11,12]. This approach offers the advantage of retaining all available information. Additionally, in cases where the observed data offers clues into the underlying pattern of missingness, it can be leveraged to achieve more accurate predictions for the missing values [13,14]. Imputation's ability to maximize the utility of observed data positions it as a preferred strategy over other methods for handling missing data [6,15].

Missing data imputation techniques can be broadly categorized into deterministic and stochastic methods [10]. Deterministic methods consistently assign the same imputed value to units with missing data within a specific sample. Conversely, stochastic methods introduce an element of randomness into the imputation modeling process, may not always create the same values, and potentially yield different imputed values for the same missing value across replications [10]. This study focuses on two basic deterministic methods: mean imputation and median imputation. These methods were chosen due to their simplicity and ease of implementation [16,17] compared to other deterministic techniques, such as ratio imputation, logical imputation and regression imputation.

The academic literature documents the frequent use of convenient missing data techniques like pairwise deletion (available case analysis) and listwise deletion (case deletion or complete case analysis) [18-20]. Nevertheless, these methods have been criticized for their reliance on observed data only, essentially editing the dataset to achieve completeness [5,6,12,21,22]. Besides, it produces a substantial loss of information, weakens statistical power and potentially introduces significant bias [5,7,9,12,17].

Therefore, to address the challenge of missing data in environmental datasets, various imputation techniques have been proposed [5,6,9,10]. Priti *et al*., [4] evaluated the performance of various imputation methods for particulate pollutant time series data with varying missing percentages. They compared six univariate single imputation methods [median, mean, last observation carried forward, Spline, Kalman and Seadec) and four multivariate multiple imputation approaches (predictive mean matching, multiple imputation by automatic, distance-aided donor selection, random forest (RF), multiple imputed using PCA (MIPCA)]. Their findings suggested that Kalman AutoRegressive Integrated Moving Average (ARIMA) imputation performed well for long-missing gaps and most missing levels (excluding 60-80%), resulting in low errors and high R-squared values. However, for the highest missing percentage scenario, MIPCA outperformed Kalman-ARIMA across all target stations. In a related study, Middya and Roy [3] proposed a novel multi-view-based missing value imputation method (MVDI) specifically designed for air pollution time series data. Their work demonstrated that MVDI outperformed various baseline methods, including traditional statistical approaches AutoRegressive, ARIMA, RF Regressor, Artificial Neural Network (ANN), Linear Interpolation, Nearest Neighbors, Mean Imputation, Convolutional Neural Network (CNN) and Convolutional LSTM.

Furthermore, Wardana *et al*., [5] proposed a spatiotemporal convolutional autoencoder for imputing missing air pollutant data. This method outperformed traditional univariate (median, mean) and multivariate (extra-trees, decision tree, Bayesian ridge regressors, k-nearest neighbors (KNN)) imputation techniques, particularly for discontinuous and long-missing data, as evidenced by significant improvements in root mean squared error (RMSE). In a separate study, Peña *et al*., [23] investigated regularized regression methods (LASSO and Ridge) to estimate missing values in air pollutant time series. Their findings suggest that LASSO models achieved slightly lower errors and higher performance than Ridge regression. Besides, Alsaber *et al*., [6] compared various imputation

methods for air quality data with different missing data percentages. They found that the proposed missForest imputation method based on a RF algorithm achieved superior accuracy in estimating missing values compared to other techniques such as RF, KNN, Bayesian principal component analysis, expectation-maximization with bootstrapping, and predictive mean matching.

Hadeed *et al*., [7] investigated imputation approaches for short-term air pollutant monitoring data with varying missingness percentages. They compared univariate techniques (median, mean, last observation carried forward, random, Kalman filter, Markov) and multivariate approaches (row mean, predictive mean matching) to address missing values. Their findings suggested that univariate methods, particularly mean imputation, random and Markov, yielded the best results with the lowest errors and highest R-squared values across varying missingness levels; Markov obtained the best-performing method. Conversely, multivariate methods consistently performed worse. In a similar study, Rumaling *et al*., [24] compared the Nearest Neighbor Method (NNM) and Expectation Maximization (EM) for imputing missing PM10 concentration data in five air quality monitoring stations in Sabah with varying missing data percentages. Their results indicated that NNM outperformed EM in imputing data for three stations.

In addition, Shaadan and Rahim [11] investigated imputation methods for time series air quality data (PM10). They compared six techniques, including spline interpolation, linear interpolation, exponential moving average, random sample, mean before-after (MBA), and Kalman filter with ARIMA modeling. Their analysis revealed that the Kalman filter using ARIMA was the most effective method for their specific dataset. In a separate study by Xu *et al*., [12], the performance of four common imputation methods (deletion, mode imputation, hot-deck, and multiple imputation) was evaluated for mental health questionnaires within a population-based survey. Their findings indicated that multiple imputations yielded the best results, although it requires slightly more data processing expertise and programming skills. Besides, Libasin *et al*., [9] investigated the performance of single imputation techniques for addressing missing values in Malaysia's air particulate matter (PM10) data. They evaluated four methods: mean of nearby points, series mean, linear trend and linear interpolation. Their analysis revealed that linear interpolation yielded the lowest mean absolute error (MAE) and demonstrated the best overall accuracy in replacing missing PM10 data.

This research examined six single imputation approaches to identify the most effective technique for managing missing data. While these techniques are among the simplest available, a comprehensive evaluation is undertaken to identify the most effective method within the method pool of simple imputation techniques, specifically comparing just the single method. This study investigates the application of the chosen imputation method to univariate data. A set of performance indicators is employed to assess the method's effectiveness. Subsequently, a comprehensive simulation study is conducted to assess the comparative performance of the different imputation methods under consideration.

## 2. Methodology

### *2.1 Data Acquisition*

The present study utilized data obtained from the Malaysian Meteorological Department. Specifically, this study focused on hourly ozone concentration data for Petaling Jaya, Selangor, spanning five months from January to May 2020. This dataset comprises ozone concentration measurements recorded on an hourly timescale. Missing data points are defined as instances where no ozone concentration value is available.

### 2.2 Single Imputation Methods

This study addressed missing values through an imputation technique based on interpolation. Two specific interpolation methods are employed: linear and cubic. Linear interpolation, the simpler method, estimates missing values by connecting two data points with a straight line [11,14,25]. The specific formula for linear interpolation involves calculating the slope and intercept of the line connecting these two known points, which is shown below in Eqs. (1) to (4),

$$\frac{f_1(x) - f(x_0)}{x - x_0} = \frac{f(x_1) - f(x_0)}{x - x_0} \tag{1}$$

$$f_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x - x_0}(x - x_0) \tag{2}$$

Where,

$$b_0 = f(x_0) \tag{3}$$

$$b_1 = \frac{f(x_1) - f(x_0)}{x - x_0} \tag{4}$$

Thus, the interpolation function is given in Eq. (5),

$$f_1(x) = b_0 + b_1(x - x_0) \tag{5}$$

Where, $x$ is the explanatory variable, where $x_i$ $(i = 0,1,2,\dots)$ is a value of the explanatory variable and $b_i$ is coefficients in the case of $f = f_1$.

Cubic interpolation is also used as a second approach to interpolation in the present study. This method is particularly suited to scenarios where data for four known points is available [26]. The mathematical expression for cubic interpolation can be written as in Eq. (6),

$$f_1(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + b_3(x - x_0)(x - x_1)(x - x_2) \tag{6}$$

Where, the coefficients $b_0$ and $b_1$ are attained from (3) and (4) and $b_2$ and $b_3$ are given in Eq. (7) and (8),

$$b_2 = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0} \tag{7}$$

and

$$b_3 = \frac{\frac{f(x_3) - f(x_2)}{x_3 - x_2} - \frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_3 - x_0} \tag{8}$$

with $f = f_3$.

A third imputation technique, known as the mean-before-after (MBA) technique, imputes missing values by employing the average of the preceding and subsequent data points [11,27]. Let considers a time series denoted by $y_1, y_2, \ldots, y_n$ with $n$ observation of which $k$ values denoted by $y_1^*, y_2^*, y_k^*$ are missing [11]. Consequently, the observed data with missing values can be represented as in Eq. (9),

$$y_1, y_2, \ldots, y_{n_1}, y_1^*, y_{n_1+1}, y_{n_1+2}, \ldots, y_{n_2}, y_2^*, y_{n_2+1}, y_{n_2+2}, \ldots, y_k^*, y_n \tag{9}$$

Hence, the initial missing values appears after $n$ observation, with subsequent missing values occurring at intervals of $n$ observations thereafter. It is important to note that there may be instances of consecutive missing observations. Therefore, in this context, $y_1^*$ is replaced using Eq. (10),

$$\overline{y}_1 = \frac{y_{n_1} + y_{n_1+1}}{2} \tag{10}$$

and $y_2^*$ will be substituted by Eq. (11),

$$\overline{y}_2 = \frac{y_{n_2} + y_{n_2+1}}{2} \tag{11}$$

and so forth.

The fourth method used in the current study was the last observation carried forward (LOCF). It addresses missing data points in longitudinal studies with repeated measures [4,7]. LOCF imputes missing values by substituting them with each subject's most recent non-missing observation [4,7]. This approach assumes that the most recently observed value provides the most accurate prediction for subsequent missing values within the same subject.

The fifth method, mean substitution, imputes missing values by replacing them with the average of all observed data points [16]. Mathematically, this can be expressed as in Eq. (12),

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{12}$$

Where $n$ is the amount of available data, and $y_i$ is the data points.

Lastly, this study employed the median substitution method. The mean's susceptibility to outliers makes the median a more robust choice [16]. Consequently, missing values within each feature were replaced with the median value of the corresponding data set [16]. In the case of ozone concentration, the median of the ozone data set was used to fill in missing values.

### *2.3 Performance Evaluation Criteria*

This study engaged three metrics to assess the imputation method's performance: prediction accuracy (PA), root mean square error (RMSE) and mean absolute error (MAE) (Table 1) and are presented as in Eqs. (13) to (15).

**Table 1**
Performance evaluation metrics

| No. | Evaluation metrics | Function | Range | Formula | |
|---|---|---|---|---|---|
| 1. | Prediction accuracy (PA) | Accuracy of the imputation method | 0 to 1, higher value indicates a better fit | $PA = \sum_{i=1}^{N} \frac{\lvert(P_i - \overline{P})(O_i - \overline{O})\rvert}{(N-1)\sigma_P \sigma_O}$ | (13) |
| 2. | Root means square error (RMSE) | Quantifies the discrepancy between observed and imputed concentrations and provides the model's average error | Lower RMSE values indicate superior model performance | $RMSE = \left(\frac{1}{N}\sum_{i=1}^{N}[P_i - O_i]^2\right)^{\frac{1}{2}}$ | (14) |
| 3. | Mean absolute error (MAE) | Average discrepancy between predicted and observed values. | 0 to ∞, 0 indicates perfect fit | $MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert P_i - O_i\rvert$ | (15) |

**Source:** Chen *et al*., [28], Libasin *et al*., [9], Middya and Roy [13], Priti *et al*., [4], Shaadan and Rahim [11]

Where, $N$ is the total of imputations, $O_i$ is the observed data and $P_i$ is the imputed data point, $\overline{P}$ is the imputed data's average, $\overline{O}$ is the observed data's average, $\sigma_P$ is the imputed data's standard deviation and $\sigma_O$is the observed data's standard deviation.

### *2.4 Simulation Study*

To evaluate and compare the imputation methods introduced previously, a simulation study is conducted. This study assessed each method's performance using several performance metrics. Hourly ozone concentration data from Petaling Jaya, Selangor, Malaysia, served as the basis for simulating missing values. The present study adopted the Weibull distribution, known for its flexibility in extreme value analysis, to produce the simulated data. This distribution is characterized by α, the shape parameter, and β, the scale parameter [29-31]. Eq. (1) details the Weibull probability density function as in Eq. (16),

$$f(x,\alpha,\beta) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} exp\left[-\frac{x^{\alpha}}{\beta}\right], \quad x > 0, \alpha > 0, \beta > 0 \tag{16}$$

and the cumulative distribution function (cdf) takes the form as in Eq. (17),

$$F(x,\alpha,\beta) = 1 - exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \quad x > 0, \alpha > 0, \beta > 0 \tag{17}$$

### *2.5 Weibull Distribution Parameters*

This study obtained the simulated data by first fitting a Weibull distribution to the ozone dataset. The estimation process yields the shape (β) and scale (α) parameters, which characterize the distribution [29-31]. These parameters, estimated to be 3.9 for β and 0.02 for α, are then employed to randomly generate data that closely resembles the distribution of the actual air quality measurements used in the study. It is important to note that when data are missing at varying percentages, the estimated parameters deviate from the true values of the underlying distribution. The extent of these deviations is presented in Tables 2 and 3.

Across both sample sizes, the shape parameter (β) exhibits greater sensitivity to the proportion of missing values. When the percentage of incomplete observation increases from 5% to 25%, the

estimated β deviates more substantially from the true value of 3.9, ranging between 4.42 and 4.46. The scale parameter (α) demonstrates a similar trend but with generally smaller deviations from the true value of 0.02. In most cases, the estimated α remains close to 0.02.

Informed by the estimated shape (β) and scale (α) parameters, two sample sizes (n = 50 and n = 150) were chosen for the simulation. The generated data were introduced with missing values at varying percentages (5%, 10%, 15% and 25%). These missing values were assumed to follow a normal distribution. Each combination of sample size and missing value percentage underwent 5,000 replications of the simulation process implemented using R software. Figure 1 presents a flowchart that visually clarifies the systematic execution of the simulation study. This flowchart effectively summarizes the key steps involved in conducting the simulation.

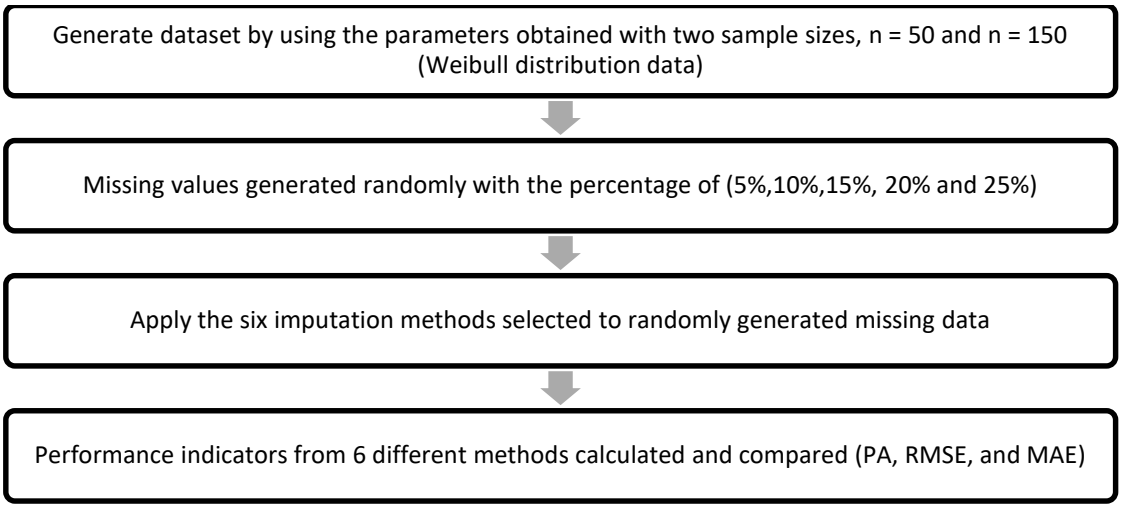


**Fig. 1.** The flow chart of the simulation study

**Table 2**
Shape and scale parameters for n = 50

| n = 50 | SHAPE (β) | | | | |
|---|---|---|---|---|---|
| METHOD | 5% | 10% | 15% | 20% | 25% |
| MBA | 4.11 | 3.67 | 3.88 | 4.03 | 4.42 |
| LINEAR | 4.19 | 4.13 | 4.10 | 3.84 | 4.09 |
| CUBIC | 4.14 | 3.92 | 2.78 | 3.47 | 3.66 |
| LOCF | 4.19 | 4.02 | 4.01 | 3.81 | 4.12 |
| MEAN | 4.25 | 4.15 | 4.25 | 4.27 | 4.46 |
| MEDIAN | 4.25 | 4.15 | 4.22 | 4.25 | 4.46 |
| n = 50 | SCALE (α) | | | | |
| METHOD | 5% | 10% | 15% | 20% | 25% |
| MBA | 0.02114 | 0.01870 | 0.01961 | 0.02063 | 0.01982 |
| LINEAR | 0.02117 | 0.01878 | 0.01959 | 0.02007 | 0.01972 |
| CUBIC | 0.02111 | 0.01912 | 0.02056 | 0.02005 | 0.01969 |
| LOCF | 0.02120 | 0.01888 | 0.01982 | 0.02026 | 0.02033 |
| MEAN | 0.02126 | 0.01851 | 0.01952 | 0.01981 | 0.01955 |
| MEDIAN | 0.02127 | 0.01851 | 0.01946 | 0.01977 | 0.01955 |

**Table 3**
Shape and scale parameters for n = 150

| n = 50 | SHAPE (β) | | | | |
|---|---|---|---|---|---|
| METHOD | 5% | 10% | 15% | 20% | 25% |
| MBA | 3.92 | 3.87 | 3.74 | 3.83 | 4.13 |
| LINEAR | 3.92 | 3.96 | 3.77 | 4.00 | 4.47 |
| CUBIC | 3.71 | 3.49 | 3.52 | 3.71 | 3.71 |
| LOCF | 3.86 | 3.95 | 3.78 | 3.80 | 4.25 |
| MEAN | 4.07 | 4.04 | 4.01 | 4.23 | 4.71 |
| MEDIAN | 4.06 | 4.04 | 4.01 | 4.26 | 4.75 |
| n = 50 | SCALE (α) | | | | |
| METHOD | 5% | 10% | 15% | 20% | 25% |
| MBA | 0.02006 | 0.01941 | 0.02049 | 0.02030 | 0.02013 |
| LINEAR | 0.02022 | 0.01974 | 0.02046 | 0.02069 | 0.02043 |
| CUBIC | 0.02031 | 0.02000 | 0.02033 | 0.02084 | 0.02060 |
| LOCF | 0.02028 | 0.01968 | 0.02065 | 0.02068 | 0.02056 |
| MEAN | 0.02026 | 0.01951 | 0.02072 | 0.02031 | 0.02021 |
| MEDIAN | 0.02025 | 0.01951 | 0.02072 | 0.02029 | 0.02028 |

## 3. Results and Discussion

Table 4 presents the results of 5,000 simulations conducted with a sample size (n) of 50. The data demonstrated a consistent trend: as the percentage of missing values increases, prediction accuracy (PA) decreases [27]. This observation aligns with the principle of goodness-of-fit, suggesting that a higher proportion of missing data leads to a model with reduced predictive power. In contrast, the mean absolute error (MAE) and root mean squared error (RMSE) values exhibit a positive relationship with the percentage of missing values [27,32]. This implies that the amount of error tends to rise as the quantity of missing information in the dataset increases.

Furthermore, the PA values for both mean and median substitution methods is consistently zero. This arises from the inherent structure of the PA equation. As shown in Eq. (13), $\overline{P}$ (imputed data's average) is identical to $P_i$ (the value of individual imputed data points). This occurs because both mean and median substitution techniques replace missing values with the same value, either the mean or median of the entire dataset. Consequently, the difference between these values ($P_i - \overline{P}$)) becomes zero, resulting in a PA value of zero for both methods when calculated using the given equation.

Moreover, among the imputation methods considered, the mean-before-after approach (MBA) yielded the most favorable results for a sample size of 50, as similarly found by Zakaria and Noor [27]. This is because, as indicated by the PA values, MBA has PA values ranging from 0.99 to 0.85 across all missing value percentages, and the closer the PA values are to one, the better the fit is than the other approaches. Furthermore, the MBA technique demonstrably produced the lowest error than the others for all missing value percentages. Linear interpolation achieved the second-highest PA, with values ranging from 0.98 to 0.81 across different missing value percentages. While it surpassed the last observation carried forward (LOCF) method in terms of PA, linear interpolation also yielded the highest MAE compared to LOCF for all missing value percentages. However, the error increase for linear interpolation was minimal when measured by RMSE.

Interestingly, the LOCF method demonstrated a competitive performance in terms of PA at lower missing value rates (5%) compared to both interpolation techniques at 5% missing values. However, this advantage diminished at higher missing value percentages (10% and 20%), where LOCF's PA surpassed only cubic interpolation. While LOCF exhibited higher MAE values compared to cubic

interpolation at 15%, 20% and 25% missing data, it produced lower RMSE for these percentages. In contrast, cubic interpolation exhibited the least favorable performance across all missing value percentages. While this method achieved relatively high PA at lower missing value rates (5% and 10%), it consistently yielded the highest values for MAE and RMSE compared to other imputation techniques. Like Morelli *et al*., [33], they found that linear interpolation has lower RMSE than cubic interpolation.

**Table 4**
Performance of methods for n = 50

| P | Technique | PA | MAE | RMSE |
|---|---|---|---|---|
| 5% | MBA | 0.99 | 0.00490 | 0.00690 |
| | LINEAR | 0.98 | 0.02300 | 0.00630 |
| | CUBIC | 0.98 | 0.04780 | 0.01592 |
| | LOCF | 0.99 | 0.02334 | 0.00710 |
| | MEAN | 0.00 | 0.00426 | 0.00470 |
| | MEDIAN | 0.00 | 0.00428 | 0.00470 |
| 10% | MBA | 0.98 | 0.00500 | 0.00530 |
| | LINEAR | 0.98 | 0.02130 | 0.00730 |
| | CUBIC | 0.89 | 0.02474 | 0.01472 |
| | LOCF | 0.96 | 0.02080 | 0.00820 |
| | MEAN | 0.00 | 0.00421 | 0.00500 |
| | MEDIAN | 0.00 | 0.00424 | 0.00500 |
| 15% | MBA | 0.96 | 0.00500 | 0.00600 |
| | LINEAR | 0.91 | 0.02080 | 0.00820 |
| | CUBIC | 0.85 | 0.01901 | 0.01424 |
| | LOCF | 0.82 | 0.01956 | 0.00900 |
| | MEAN | 0.00 | 0.00425 | 0.00510 |
| | MEDIAN | 0.00 | 0.00428 | 0.00520 |
| 20% | MBA | 0.88 | 0.00490 | 0.00600 |
| | LINEAR | 0.88 | 0.01957 | 0.00880 |
| | CUBIC | 0.80 | 0.01595 | 0.01426 |
| | LOCF | 0.81 | 0.01754 | 0.00950 |
| | MEAN | 0.00 | 0.00424 | 0.00510 |
| | MEDIAN | 0.00 | 0.00428 | 0.00520 |
| 25% | MBA | 0.85 | 0.00510 | 0.00620 |
| | LINEAR | 0.81 | 0.02078 | 0.00910 |
| | CUBIC | 0.78 | 0.01713 | 0.01433 |
| | LOCF | 0.77 | 0.01896 | 0.00980 |
| | MEAN | 0.00 | 0.00425 | 0.00520 |
| | MEDIAN | 0.00 | 0.00430 | 0.00520 |

***Note:*** P-Percentage of missing values, PA-Prediction accuracy, MAE- Mean absolute error, RMSE-root mean square error, MBA-Mean-before-after method, LOCF-Last observation carried forward

Table 5 summarizes the findings of 5,000 simulations conducted for 150 sample sizes. The MBA technique emerged as the most effective imputation strategy, consistently delivering superior results compared to other approaches, as similar found by Zakaria and Noor [27]. This method achieves a high PA range from 0.85 to 0.80 across varying percentages of missing values. Linear interpolation appeared as the second-best method. At the 10%-mark, linear interpolation surpasses the MBA method in terms of PA. Additionally, these two methods exhibited comparable PA values at 5% and 15% missing values. However, a closer examination revealed a key distinction between their performances: the amount of error produced, whereas the MBA method demonstrated a clear advantage in error minimization [27]. Linear interpolation consistently generated higher MAE and RMSE values than the MBA method. This observation underscores the superiority of the MBA

approach for this particular scenario (sample size of 150). In contrast to Shaadan and Rahim [11], the MBA method generated higher MAE and RMSE values than the linear interpolation method.

**Table 5**
Performance of methods for n = 150

| P | Technique | PA | MAE | RMSE |
|---|---|---|---|---|
| 5% | MBA | 0.85 | 0.005100 | 0.006080 |
| | LINEAR | 0.85 | 0.023800 | 0.007000 |
| | CUBIC | 0.84 | 0.038200 | 0.015709 |
| | LOCF | 0.84 | 0.023600 | 0.007800 |
| | MEAN | 0.00 | 0.005100 | 0.005071 |
| | MEDIAN | 0.00 | 0.005100 | 0.005083 |
| 10% | MBA | 0.84 | 0.004964 | 0.006095 |
| | LINEAR | 0.85 | 0.021200 | 0.007500 |
| | CUBIC | 0.78 | 0.024500 | 0.014600 |
| | LOCF | 0.84 | 0.020400 | 0.008400 |
| | MEAN | 0.00 | 0.005100 | 0.005156 |
| | MEDIAN | 0.00 | 0.005200 | 0.005200 |
| 15% | MBA | 0.84 | 0.004928 | 0.006100 |
| | LINEAR | 0.84 | 0.021200 | 0.008200 |
| | CUBIC | 0.82 | 0.019435 | 0.014645 |
| | LOCF | 0.77 | 0.019800 | 0.009000 |
| | MEAN | 0.00 | 0.005200 | 0.005138 |
| | MEDIAN | 0.00 | 0.005200 | 0.005200 |
| 20% | MBA | 0.82 | 0.004838 | 0.006111 |
| | LINEAR | 0.80 | 0.019800 | 0.008700 |
| | CUBIC | 0.77 | 0.014785 | 0.013500 |
| | LOCF | 0.75 | 0.016727 | 0.009400 |
| | MEAN | 0.00 | 0.005200 | 0.005177 |
| | MEDIAN | 0.00 | 0.005200 | 0.005200 |
| 25% | MBA | 0.80 | 0.005107 | 0.005967 |
| | LINEAR | 0.79 | 0.020800 | 0.009200 |
| | CUBIC | 0.77 | 0.016434 | 0.013686 |
| | LOCF | 0.75 | 0.018100 | 0.009800 |
| | MEAN | 0.00 | 0.005200 | 0.005191 |
| | MEDIAN | 0.00 | 0.005200 | 0.005205 |

***Note:*** P-Percentage of missing values, PA-Prediction accuracy, MAE- Mean absolute error, RMSE-root mean square error, MBA-Mean-before-after method, LOCF-Last observation carried forward

The LOCF method outperformed cubic interpolation at lower missing value rates (5% and 10%), mirroring the trend observed for a sample size of 50. This is reflected in their prediction accuracy (PA) values. However, the advantage shifts at higher missing value percentages (15%, 20% and 25%), where cubic interpolation yielded superior PA compared to LOCF. Interestingly, the error profiles for these methods remained consistent with the findings for sample size 50. LOCF continued to exhibit higher MAE but lower RMSE compared to cubic interpolation at 15%, 20% and 25% missing values. Cubic interpolation maintained its position as the method with the least favorable results across all missing value percentages, in which it obtained the highest overall error across all missing value percentages for both MAE and RMSE.

These findings align with previous research. Morelli *et al*., [33] reported that linear interpolation resulted in lower RMSE compared to cubic interpolation. Similarly, Priti *et al*., [4] observed that LOCF outperformed the mean and median for missing value levels below 20%. However, the discrepancy between the imputed and observed means increased for higher missing value percentages, leading to a significant performance decline reflected in high MAE and RMSE values. Hadeed *et al*., [7] also

found that LOCF performed well at missing value rates between 20% and 40%, but its performance declined as the duration of missingness increased.

For overall comparison, the findings presented in Tables 4 and 5 revealed that the MBA imputation method consistently outperformed other techniques for both sample sizes (50 and 150) across varying missing value percentages, as similarly found by Zakaria and Noor [27]. This method demonstrably achieved the lowest error rates while maintaining high PA values. Conversely, the mean substitution method yielded the poorest PA results, consistently producing a value of zero across all missing value percentages. Among the interpolation techniques, linear interpolation emerged as the most effective technique among the interpolation methods, particularly for lower missing value rates (5%, 10% and 15%). Cubic interpolation only demonstrates favorable performance at the 5% missing value mark. This aligned with the findings of Morelli *et al*., [33], who reported lower RMSE for linear interpolation compared to cubic interpolation.

The LOCF method outperformed cubic interpolation at lower missing value percentages (5% and 10%), but cubic interpolation surpassed LOCF at higher rates (15%, 20% and 25%). Besides, while median and mean substitution methods produce a PA value of zero, they exhibited the lowest error measures compared to other techniques. Nevertheless, studies by Priti *et al*., [4], Wardana *et al*., [5] and Hadeed *et al*., [7] discovered mean and median imputation among the methods that obtained high values in MAE and RMSE.

In essence, the MBA method stands out as the most effective strategy for predicting missing values. It consistently delivered superior performance in terms of both error minimization and maintaining high PA. Linear interpolation follows as a strong contender, mainly for handling smaller amounts of missing data. LOCF demonstrated some utility at lower missing value rates. Conversely, cubic interpolation and mean/median substitution consistently produced the least favorable results. These findings corroborate the conclusions of Wardana *et al*., [5] that the mean and median imputation often leads to the most inaccurate imputations.

### *3.1 Application of Mean-Before-After (MBA) Imputation*

The MBA imputation method, identified as the most effective through a simulation study, was employed to address missing values within an ozone dataset. This study centered on a year's worth of hourly data pertaining to ozone concentration measurements gathered in Petaling Jaya, Selangor, Malaysia. The data comprises ozone concentrations recorded on an hourly basis for a single month. There were 720 hourly ozone concentration values available for January, with 5% (34 observations) containing missing data.

Table 6 presents a comparison of descriptive statistics between the original data with missing observations and the corresponding data following imputation (obtained using the MBA method within the R package, pastecs, to address missing entries). Descriptive statistics encompass basic metrics summarizing the data: minimum value (min), maximum value (max), the sum of all non-missing values (sum), and range (max-min). Additionally, the table incorporates statistics describing the central tendency and spread of the data. It is noteworthy that the confidence interval on the mean (CI.mean) is calculated using a default probability level of p = 0.9.

Table 7 presents the estimated values and standard deviations of the shape (β) and scale (α) parameters for both the original ozone data with missing observations and the corresponding data following imputation (obtained using Mean-Before-After imputation). Statistical software is unable to directly estimate these parameters when missing values are present. Therefore, to obtain parameter estimates for the missing data, these values were excluded from the initial analysis. Notably, the parameters estimate for the shape (β) and scale (α), along with their standard

deviations, are lower in the imputed ozone data than the original data containing missing values, indicating good estimation.

**Table 6**
Comparison of descriptive statistics between incomplete ozone data and imputed ozone data

| Statistic | Ozone data with missing values | Imputed ozone data |
|---|---|---|
| Minimum value (Min) | 0.001 | 0.001 |
| Maximum value (Max) | 0.095 | 0.095 |
| Range (Max-Min) | 0.094 | 0.094 |
| Sum of all non-missing values (Sum) | 11.74 | 12.30 |
| Median | 0.0065 | 0.0070 |
| Mean | 0.017115 | 0.017084 |
| Standard error of mean (SE.mean) | 0.000780 | 0.000757 |
| Confidence interval of the mean a specified significance level (p) (CI. mean.0.95) | 0.001531 | 0.001487 |
| Variance (Var) | 0.000417 | 0.000413 |
| Standard Deviation (Std. Dev) | 0.020421 | 0.020318 |
| Coefficient of variation (coef.var) | 1.193142 | 1.189307 |

**Table 7**
Estimated parameters of original and imputed data

| Estimated parameters | Original data with missing observations | Imputed data |
|---|---|---|
| Shape, β | 0.76 | 0.77 |
| (Standard deviation) | (0.0228965515) | (0.02245908) |
| Scale, α | 0.014571 | 0.014636 |
| (Standard deviation) | (0.0007579287) | (0.0007408004) |

## 4. Conclusion

This investigation explored the efficacy of various simple imputation techniques for addressing missing data. This study evaluated the performance of six simple imputation methods for missing values. Simulated datasets with varying percentages of missing data were used to assess the effectiveness of each method. Performance was measured using a combination of metrics: prediction accuracy (PA), mean absolute error (MAE), and root mean squared error (RMSE). PA served as an indicator of the quality of the imputed values, while MAE and RMSE quantified the error introduced by each imputation method. Across all missing data scenarios, the study observed a progressive deterioration in prediction accuracy as the percentage of missing data escalated. Conversely, both MAE and RMSE values exhibited a positive correlation with the increasing percentage of missing data.

Among the evaluated imputation methods, the Mean-Before-After (MBA) technique emerged as the most effective for ozone data, achieving prediction accuracy close to one. Linear interpolation and the Last Observation Carried Forward (LOCF) method also demonstrated comparable performance. Conversely, the cubic interpolation method yielded the least favorable results. As anticipated, the performance of interpolation methods is inherently limited. As evidenced by the findings, interpolation techniques are generally only suitable for imputing short gaps. The cubic interpolation method's performance demonstrably declined with increasing missing value percentages. This is attributable to the concurrently increasing length of gaps between the data points used for interpolation, ultimately leading to a higher degree of error observed with this method.

Furthermore, the findings revealed discrepancies between the methods, particularly highlighting the unreliability of mean and median substitution techniques. This is concerning as these methods are frequently reported in the literature and serve as default options within numerous statistical

software packages. Consequently, the simulation study results within this research indicate that the most effective approach for handling missing data points in the ozone dataset is the MBA imputation method. This method replaces each missing data point with the average of the two preceding and subsequent data points.

**Acknowledgement**
This research was supported by Universiti Malaysia Sabah through UMSGreat Scheme (Project Code: GUG0642-2/2023). Also, the authors acknowledge the invaluable contribution of the Department of Environment and the Department of Meteorological Malaysia in providing the data used in this study.

## References

[1] Rochman, Eka Mala Sari, and Herry Suprajitno. "Overcoming missing values using imputation methods in the classification of tuberculosis." *Commun. Math. Biol. Neurosci.* 2022 (2022): Article-ID. https://doi.org/10.28919/cmbn/7538
[2] Ghazali, Shamihah Muhammad, Norshahida Shaadan, and Zainura Idrus. "Missing data exploration in air quality data set using R-package data visualisation tools." *Bulletin of Electrical Engineering and Informatics* 9, no. 2 (2020): 755-763. https://doi.org/10.11591/eei.v9i2.2088
[3] Middya, Asif Iqbal, and Sarbani Roy. "Multiview data fusion technique for missing value imputation in multisensory air pollution dataset." *Journal of Ambient Intelligence and Humanized Computing* 15, no. 8 (2024): 3173-3191. https://doi.org/10.1007/s12652-024-04816-9
[4] Shakya, Kaushlesh Singh, and Prashant Kumar. "Selection of statistical technique for imputation of single site-univariate and multisite–multivariate methods for particulate pollutants time series data with long gaps and high missing percentage." *Environmental Science and Pollution Research* 30, no. 30 (2023): 75469-75488. https://doi.org/10.1007/s11356-023-27659-x
[5] Wardana, I. Nyoman Kusuma, Julian W. Gardner, and Suhaib A. Fahmy. "Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder." *Neural Computing and Applications* 34, no. 18 (2022): 16129-16154. https://doi.org/10.1007/s00521-022-07224-2
[6] Alsaber, Ahmad R., Jiazhu Pan, and Adeeba Al-Hurban. "Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of Kuwait environmental data (2012 to 2018)." *International Journal of Environmental Research and Public Health* 18, no. 3 (2021): 1333. https://doi.org/10.3390/ijerph18031333
[7] Hadeed, Steven J., Mary Kay O'rourke, Jefferey L. Burgess, Robin B. Harris, and Robert A. Canales. "Imputation methods for addressing missing data in short-term monitoring of air pollutants." *Science of the Total Environment* 730 (2020): 139140. https://doi.org/10.1016/j.scitotenv.2020.139140
[8] Kwak, Sang Kyu, and Jong Hae Kim. "Statistical data preparation: management of missing values and outliers." *Korean journal of anesthesiology* 70, no. 4 (2017): 407. https://doi.org/10.4097/kjae.2017.70.4.407
[9] Libasin, Zuraira, Wan Suhailah Wan Mohamed Fauzi, Nur Azimah Idris, and Noor Azizah Mazeni. "Evaluation of Single Missing Value Imputation Techniques for Incomplete Air Particulates Matter (PM 10) Data in Malaysia." *Pertanika Journal of Science & Technology* 29, no. 4 (2021). https://doi.org/10.47836/pjst.29.4.46
[10] D'Agostino McGowan, Lucy<? show [AQ ID= GQ1, Sarah C. Lotspeich, and Staci A. Hepler. "The "Why" behind including "Y" in your imputation model." *Statistical Methods in Medical Research* 33, no. 6 (2024): 996-1020. https://doi.org/10.1177/09622802241244608
[11] Shaadan, N., and N. A. M. Rahim. "Imputation analysis for time series air quality (PM10) data set: A comparison of several methods." In *Journal of Physics: Conference Series*, vol. 1366, no. 1, p. 012107. IOP Publishing, 2019. https://doi.org/10.1088/1742-6596/1366/1/012107
[12] Xu, Xueying, Leizhen Xia, Qimeng Zhang, Shaoning Wu, Mingcheng Wu, and Hongbo Liu. "The ability of different imputation methods for missing values in mental measurement questionnaires." *BMC medical research methodology* 20 (2020): 1-9. https://doi.org/10.1186/s12874-020-00932-0
[13] Gad, Ibrahim, Doreswamy Hosahalli, B. R. Manjunatha, and Osama A. Ghoneim. "A robust deep learning model for missing value imputation in big NCDC dataset." *Iran Journal of Computer Science* 4 (2021): 67-84. https://doi.org/10.1007/s42044-020-00065-z
[14] Sukatis, Fahren Fazzer, Norazian Mohamed Noor, Nur Afiqah Zakaria, Ahmad Zia Ul-Saufie, and Suwardi Annas. "Estimation of missing values in air pollution dataset by using various imputation methods." *International Journal of Conservation Science* 10, no. 4 (2019): 791-804.

[15] Woźnica, Katarzyna, and Przemysław Biecek. "Does imputation matter? Benchmark for predictive models." *arXiv preprint arXiv:2007.02837* (2020). https://doi.org/10.48550/arxiv.2007.02837
[16] Patel, Dharmendra, Octavio Loyola-González, Arpit Trivedi, Hardik Rajgor, Tushar Mehta, Sanskruti Patel, Pranav Vyas, Nilay Ganatra, and Hardik I. Patel. "Single and Multiple Imputation Techniques to Treat Missing Numerical Variables (MNV) in Perspectives of Data Science Project-A Case Study." https://doi.org/10.14445/22315381/ijett-v70i5p202
[17] Zhang, Zhongheng. "Missing data imputation: focusing on single imputation." *Annals of translational medicine* 4, no. 1 (2016): 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38
[18] Hughes, Rachael A., Jon Heron, Jonathan AC Sterne, and Kate Tilling. "Accounting for missing data in statistical analyses: multiple imputation is not always the answer." *International journal of epidemiology* 48, no. 4 (2019): 1294-1304. https://doi.org/10.1093/ije/dyz032
[19] Jakobsen, Janus Christian, Christian Gluud, Jørn Wetterslev, and Per Winkel. "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts." *BMC medical research methodology* 17 (2017): 1-10. https://doi.org/10.1186/s12874-017-0442-1
[20] Little, Roderick JA, and Donald B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.
[21] Hossie, Thomas J., Jenilee Gobin, and Dennis L. Murray. "Confronting missing ecological data in the age of pandemic lockdown." *Frontiers in Ecology and Evolution* 9 (2021): 669477. https://doi.org/10.3389/fevo.2021.669477
[22] Pazhoohesh, Mehdi, Zoya Pourmirza, and Sara Walker. "A comparison of methods for missing data treatment in building sensor data." In *2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE)*, pp. 255-259. IEEE, 2019. https://doi.org/10.1109/sege.2019.8859963
[23] Peña, Mario, Patricia Ortega, and Marcos Orellana. "A novel imputation method for missing values in air pollutant time series data." In *2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pp. 1-6. IEEE, 2019. https://doi.org/10.1109/la-cci47412.2019.9037053
[24] Rumaling, Muhammad Izzuddin, Fuei Pien Chee, Jedol Dayou, Jackson Hian Wui Chang, Steven Soon Kai Kong, and Justin Sentian. "Missing value imputation for PM10 concentration in sabah using nearest neighbour method (NNM) and expectation-maximization (EM) algorithm." *Asian Journal of Atmospheric Environment* 14, no. 1 (2020): 62-72. https://doi.org/10.5572/ajae.2020.14.1.062
[25] Noor, Norazian Mohamed, Mohd Mustafa Al Bakri Abdullah, Ahmad Shukri Yahaya, and Nor Azam Ramli. "Comparison of linear interpolation method and mean method to replace the missing values in environmental data set." In *Materials science forum*, vol. 803, pp. 278-281. Trans Tech Publications Ltd, 2015. https://doi.org/10.4028/www.scientific.net/msf.803.278
[26] Yoon, Jinsung, William R. Zame, and Mihaela Van Der Schaar. "Estimating missing data in temporal data streams using multi-directional recurrent neural networks." *IEEE Transactions on Biomedical Engineering* 66, no. 5 (2018): 1477-1490. https://doi.org/10.1109/tbme.2018.2874712
[27] Zakaria, Nur Afiqah, and Norazian Mohamed Noor. "Imputation methods for filling missing data in urban air pollution data formalaysia." *Urbanism. Arhitectura. Constructii* 9, no. 2 (2018): 159.
[28] Chen, Mei, Hongyu Zhu, Yongxu Chen, and Youshuai Wang. "A novel missing data imputation approach for time series air quality data based on logistic regression." *Atmosphere* 13, no. 7 (2022): 1044. https://doi.org/10.3390/atmos13071044
[29] De Souza, Amaury, Soetânia S. De Oliveira, Flavio Aristone, Zaccheus Olaofe, Shiva Prashanth Kumar Kodicherla, Milica Arsić, Nabila Ihaddadene, and Ihaddadene Razika. "Modeling of the function of the ozone concentration distribution of surface to urban areas." *European Chemical Bulletin* 7, no. 3 (2018): 98-105.
[30] Efe-eyefıa, Eferhonore, Joseph Thomas, and Samuel Chiabom Zelıbe. "Theoretical analysis of the Weibull alpha power inverted exponential distribution: properties and applications." *Gazi University Journal of Science* 33, no. 1 (2020): 265-277. https://doi.org/10.35378/gujs.537832
[31] Salim, Omar M., Hassen Taher Dorrah, and Mahmoud Adel Hassan. "A generalized cascaded approach to estimate missing wind data using multivariate weibull distribution network." In *2020 12th International Conference on Electrical Engineering (ICEENG)*, pp. 68-72. IEEE, 2020. https://doi.org/10.1109/iceeng45378.2020.9171741
[32] Austin, Peter C., and Stef van Buuren. "The effect of high prevalence of missing data on estimation of the coefficients of a logistic regression model when using multiple imputation." *BMC Medical Research Methodology* 22, no. 1 (2022): 196. https://doi.org/10.1186/s12874-022-01671-0
[33] Morelli, Davide, Alessio Rossi, Massimo Cairo, and David A. Clifton. "Analysis of the impact of interpolation methods of missing RR-intervals caused by motion artifacts on HRV features estimations." *Sensors* 19, no. 14 (2019): 3163. https://doi.org/10.3390/s19143163