Imputation Analysis for Time Series Air Quality (PM10) Data Set: A Comparison of Several Methods

N Shaadan\textsuperscript{1,2} and N A M Rahim\textsuperscript{1}

\textsuperscript{1}Centre of Statistical and Decision Science Studies, Faculty of Computer and Mathematical Sciences, UiTM, 40450, Shah Alam, Selangor, Malaysia
\textsuperscript{2}Advanced Analytics Engineering Centre, Faculty of Computer and Mathematical Sciences, UiTM, 40450, Shah Alam, Selangor, Malaysia

1E-mail: shahida@tmsk.uitm.edu.my

Abstract. Good quality data is important to guarantee for the best quality results of research analysis. However, the quality of the data often being impacted by the existence of missing values that bring bad implication on the accuracy of analysis and subsequently lead to biased results. In air quality data set, missing values problem often caused by various reasons, for example machine malfunction and errors, computer system crashes, human error and insufficient sampling used. In the case for time series modelling, complete series of data is very important to enable for the model construction. This paper aims to highlight a systematic statistical procedure and analysis on how to investigate the performance of several missing values imputation methods to solve for the problem of missing value existence when data are time series. The knowledge could help researchers to implement a comprehensive procedure in deciding a type of imputation method that suits with their data. A case study was conducted using real data set from Shah Alam air quality monitoring station. The results have shown that the missing data at the monitoring station is completely at random (MCAR). Among six imputation methods compared and based on the performance of indicators such as RMSE, MAE, AI and $R^2$ it is shown that imputation using Kalman Filter using ARIMA model is the best appropriate method for the data set.

1. Introduction

Time series is a sequence of measurements of some quantity of interest taken at different time points consecutively. Most of the time, the measurements are observed at constant spaced time intervals which producing a univariate discrete-time time series. The ultimate objective of univariate time series analysis is to find the dynamic dependence of $y_t$ on its past values \{$y_{t-1}, y_{t-2}, \ldots$\} for the purpose of forecasting and trend investigation. However, there is a problem that might arise in gathering time series data which is the existence of missing values. Missing values occur when the values are unavailable due to several reasons. Consequently, missing values will cause the problem in the modelling of time series predictive model due to unavailable values. According to [1, 9], missing values has become a common problem in almost all research studies including environmental and air pollution studies.

In particular, air quality monitoring is carried out to provide data for the use of the environmentalists and researchers to study air pollution problem as well as to predict its future occurrence. Thus, analysis of the recorded data plays an important role to provide information for mitigation and adaptation purposes. However, such analysis is complicated by the frequently large proportions of observations missing from the data. Missing values in air quality data set often reported to be due to malfunctioned equipment, routine maintenance, and human errors or even the system might stay off-line for several
days. There are three major problems that may arise when dealing with incomplete data [2,14,15]. The first problem is there is a loss of information due to missing values and, as a consequence, a loss of efficiency. Second, there are several complications related to data handling, computation and analysis, due to irregularities in data structure and the impossibility of using standard software [13]. Lastly, the results may be biased due to systematic differences between observed and unobserved data. For the case of Time series data, missing value problem will prevent time series analysis to be conducted if it is not being treated. This is because time series analysis requires complete series of time-point data [3]. Therefore, missing values need to be replaced with reasonable values. In statistics this process and technique of replacing with reasonable values is called imputation. Up to date, only a limited number of studies have taken a closer look at the special case of univariate time series imputation [4]. Due to the missing values problems, the estimation of missing values becomes the first priority in the data preparation process. Various replacement methods have been used to tackle the problem and are profoundly discussed in the literature.

Generally, methods available for imputing missing values can be divided into two main categories of experimentation which are single imputation and multiple imputations. Single imputation works by filling in one value for each missing one; they have many appealing features, because standard complete-data methods can be applied directly and because imputation needs to be carried out only once. While multiple imputation generates multiple simulated values for each missing value, in order to reflect the uncertainty attached to missing data. Besides that, these methods require a full specification of the distributional form of data in order to derive the conditional distribution of the missing data given the observed data. The impact of missing data on the result of statistical analysis depends on the mechanism that made the data to be missing and on the way the data analyst deals with them as well as the choice of an appropriate method for handling missing data including the pattern of missingness and the missing mechanism. [5] had classified missing data mechanism can be in three categories including missing at random (MAR), missing completely at random (MCAR) and lastly missing not at random (MNAR). For most air quality data set, usually the missing data mechanism of air quality data is assumed to be MAR [6,7] discusses on the methods applicable to air quality data sets in the context of whether the air quality data is univariate and multivariate forms of data representation. The imputation methods applied includes the univariate, multivariate as well as hybrid. In a context of univariate data, linear, spline and nearest neighbourhood interpolation has been applied. While, for multivariate, Regression-based imputation (REGEM), nearest neighbour (NN), Self-organizing map (SOM) and multi-layer perception (MLP) are used. For multivariate air quality data in Malaysia, [8] had applied mean and medium substitution method; expectation maximization (EM) method, singular value decomposition (SVD), and K-nearest neighbour (KNN) method for multivariate air quality data sets. Their results showed that EM, KNN and SKNN methods were consistently superior irrespective of the different air quality monitoring stations and percentage of missingness. Another study by [1] has shown that EM imputation was the best method when compared to mean substitution and hot deck methods. [2] has found mean –before – after method is the best method compared to several Interpolation methods (i.e. Linear, Quadratic and Cubic) when applied for univariate time series air quality (PM10) data set in Seberang Perai, Penang, Malaysia. Other than that, a machine learning approach was also found actively considered for imputing missing values in multivariate environmental data set [17,18,19]. A recent study by [20] has incorporated a multiple imputation approach in solving missing values for multiple time series medical data set using Amelia package in R. The developed technique incorporates prior or bound argument in the imputation approach.

Noticeably, from the literatures, most of the studies conducting missing values imputation methods were in the context of multivariate air quality data set in comparison to univariate perspectives. On top of that, often in practical application, with the same kind of data set, researchers used a chosen technique that they feel appropriate based on the literature review obtained without knowing whether the method is really appropriate for their data set. Preliminary investigation to confirm about the type of missing mechanism is also often being neglected. Thus, for the better results of estimation accuracy of the imputed values when applying a chosen imputation method, this paper discusses on a comprehensive
procedure and experimentation to prior investigate the type of missing mechanism and to analyse the performance of several considered imputation techniques. Since in Malaysia, the most indicative and significant element of air pollutant is PM$_{10}$, hence, this study will be specified into missing values imputation methods in the context of univariate PM$_{10}$ time series data for experiment illustration. The knowledge could help researchers to decide on the best imputation method for their data set.

2. Methodology

2.1. Missing Value Investigation Process and Framework

After the data set to be studied was in hand at phase 1, investigation on the appropriate imputation technique for a data set should begin with the knowledge on the kind of missing mechanism whether the missingness is MAR, MCAR or MNAR. Based on this knowledge, next step investigation is the searching for the available techniques in the literature that have been used within the same kind and type of data set. These investigation activities are considered as the preliminary analysis (phase 2). When several techniques under MAR per say, has been identified, now the third step is to determine which one is the best technique. A comparison analysis must be conducted comprehensively and is conducted using missing data imputation analysis (phase 3). The determination of which method is the best can be based on the performance of the methods using several indicators such as Root Mean Square Error (RMSE), coefficient of determination ($R^2$), Mean Absolute Error (MAE), Agreement Index (AI) (phase 4) and etc. The framework of the investigation procedure is shown in figure 1.

![Figure 1. Framework of missing values investigation process.](image)

Figure 2 describes the process which is generally comprises of four important stages of analysis within the missing data imputation analysis phase (phase 3). The first stage is obtaining benchmark data. What is benchmark data and how to obtain the data? Benchmark data is a real observed or recorded data set. It is a set of complete data taken as a sample from the population of studied data and is made as a reference data set [9]. The data set is used to assess the performance of imputation method based on the amount of difference between the observed and the imputed values (i.e. MAE) as well as their degree of relationship (i.e. $R^2$ and AI).
Figure 2. Stages involved in missing values imputation analysis.

After obtaining the benchmark data, simulation process will take place in stage 2. This simulation process will produce several sets of simulated data of several patterns of missingness. Missing data pattern can be generated according to various criteria including the amount or percentage of data being missing, for example 5%, 10% and 15%, the size of the sequence of the missingness and whether the missingness is random or not random depending on the prior analysis conducted in figure 1. Next, the chosen imputation methods are implemented to impute the missing observations in the simulated data sets at stage 3 and the performance of the methods is assessed in stage 4 using several indicators [9,10]. For the case study conducted in this paper, the performance indicators involved are RMSE, MAE, $R^2$ and AI. Let $Y$ denotes the observed value that is assumed missing and $\hat{Y}$ is the estimated or imputed value then the formula for RMSE, MAE, $R^2$ and AI [15,16] are as follows.

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

\[
MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{2}
\]

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} y_i^2} \tag{3}
\]

\[
AI = 1 - \left[ \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} |\hat{y}_i - \bar{y}| + |y_i - \bar{y}|^2} \right] \tag{4}
\]

2.2. Imputation Methods

Missing values imputation analysis are conducted using several imputation techniques for univariate data set including the methods that have been used by previous researchers for Malaysia air quality data set such as Linear Interpolation (M1), Spline Interpolation (M2), Exponential Moving Average (M3) and mean before after (M6). Two other methods including Kalman Filter) using ARIMA Model (M5) and Random replacement (M4) are introduced to be applied for the data set. The analysis is conducted using R software. The mathematical models and algorithm of the techniques can be obtained from the following review sources; [2, 7, 8, 11].
2.2.1. Linear Interpolation Method (M1)
When two data points are connected with a straight line, hence the interpolation function is as follow:

\[ y_i(t) = b_0 + b_1(t - t_0) \]  

(5)

\( t \) is the independent variable, \( t_i \) \((i=0,1,2,...)\) is a known value of the independent variable, and \( b_i \) are unknown coefficients. Then from (4),

\[ b_0 = y(t_0) \]  

(6)

and

\[ b_1 = y(t_1) - y(t_0)/t_1 - t_0 \]  

(7)

in which in this case \( y_{\text{missing}} = y_i \), in this case \( y_i \) will be the imputed value.

2.2.2. Spline Interpolation (M2)
Given a tabulated function \( y_k = y(t_k), k = 0, 1, ..., N \), a spline is a polynomial between each pair of tabulated points, but one whose coefficients are determined “slightly” non-locally. The non-locality is designed to guarantee global smoothness in the interpolated function up to some order of derivative. We focus on one particular interval \((t_k, t_{k+1})\). The linear spline interpolation in that interval gives the interpolation formula

\[ y = Ay(t_k) + By(t_{k+1}) \]  

(8)

where

\[ A \equiv (t_{k+1} - t)/(t_{k+1} - t_k), B \equiv 1 - A = (t - t_k)/(t_{k+1} - t_k) \]  

(9)

Thus, the imputed values for \( y_{\text{missing}} = y \).

2.2.3. Exponential Moving Average (M3)
A moving average not just a regular mean, as new values become available, the oldest data points must be dropped from the set and new data points must come in to replace them. Thus, the data set is constantly "moving" to account for new data as it becomes available. This method of calculation ensures that only the current information is being accounted for. The exponential moving average is a type of moving average that gives more weight to recent observed value in an attempt to make it more responsive to new information. Below is the formula for imputation by exponential moving average:

\[ Z_i = (E(y) \cdot \alpha) + (Z_{i-1} \cdot (1 - \alpha)) \]  

(10)

Given that, \( y \): observed value, \( \alpha \): smoothing factor = \( 2/(1+N) \), \( N \): Number of time periods. Thus, the imputed values for \( y_{\text{missing}} = EMA \).

2.2.4. Random Sample (M4)
The simplest approach is to impute missing values of earnings based on the observed data for this variable. This method works by replacing missing values with a random value that is drawn between the minimum and maximum of the variable. For variable \( y \), assume that the data set is \( \{y_i\} \) where \( i=1,2,...,n \). Let \( \min(y) = \min_{i=1,n} \{y_i\} \), and let \( \max(y) = \max_{i=1,n} \{y_i\} \). The random value

\[ R = \min(y) + (\max(y) - \min(y)) \cdot \text{randuni}(SEED) \]  

(11)
is a function that takes a SEED (number) as input and returns a random value from a uniform distribution between 0 and 1.

2.2.5. Kalman Filter Using ARIMA Model (M5)

In a state space model, we have a (potentially unobserved) state variable, \( \alpha_t \), and measurements, \( y_t \). Let \( Y_{t-1} \) be all measurable \( \{y_1, y_2, \ldots, y_{t-1}\} \) variables up to time \( t - 1 \). The state space is characterized by State Equation, describing the evolution state:

\[
(a_t | a_{t-1}, Y_{t-1})
\]

(12)

And measurement Equation, describing how the measurable variables relate to state variables.

\[
f(y_t | a_t, Y_{t-1})
\]

(13)

The Kalman Filter is a recursive algorithm for obtaining predictions of future observations and quickly provides the one step ahead prediction errors and the variances.

For an ARIMA model,

\[
y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t + \theta \epsilon_{t-1}
\]

(14)

\[
\alpha_t = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ \theta y_{t-2} + \theta \epsilon_{t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \epsilon_t,
\]

(15)

\[
y_t = [10] \alpha_t
\]

(16)

Kalman filter can be used for the following task:

1. Construct the conditional distribution of \( y_t \) and \( \alpha_t \) given all past observations \( |y_{t-1} \). We know that it is normal, Kalman filter calculates the conditional mean \( y_t|_{t-1}, \alpha_t|_{t-1} \) and conditional variances.
2. Calculate a likelihood of \( \{ y_1, y_2, \ldots, y_T \} \)
3. Construct the distribution of \( \alpha_t \) given all observations of \( y_t, \alpha_t|y_t \)
4. Calculate \( y_t|_{t-1} \), which is the best linear forecast of \( y_t \) given \( Y_{t-1} \) even if error are non-normal.

2.2.6. Mean Before After Method (M6)

Replacing the missing value by averaging both before and after values of that specific missing value.

\[
y_{\text{imputed}} = 0.5*(y_{t-1} + y_{t+1}),
\]

(17)

Given that \( y_t \) is the missing value.

3. Application and Case Study

For an illustration purposes, a case study analysis is conducted using series of four months period of hourly recorded PM10 data (June 2015 –September 2015) from Shah Alam air quality monitoring station. The data were obtained from the Department of Environment (DOE), Putrajaya Malaysia.

The original data of the Air Quality Data (PM10) from June 2015 to September 2015 from Shah Alam Station is a daily hourly recorded PM10 data which is also categorized as Univariate Time Series Data which consists of missing values. From this dataset, a complete case data of 1009 observations (consist of 42 days of hourly data) which cover from July 2015 up to August 2015 were selected for simulation. This data set is treated as benchmark data and will be used as reference data set. For the missing imputation analysis, several pattern of missing data set were simulated. In simulation, we make the comparison of the errors and performances for each of the methods listed above based on the missing percentage of 5%, 10% and 15% by generating 200 simulated patterns for each missing percentage.
3.1. Results of Case Study Analysis

Overall, based on the results in figure 3, it shows that the error (MAE and RMSE) would be increased and the measure of performances (R² and d) decreased as the percentages of simulated missing data increases. This was justified by the statement stated [12, 16] that the validity of the estimates would be decreased when the missing values increased.

M4 shown the highest RMSE among the six methods of imputation. M4 gave the highest error in the range of 34.90 and 35.38. Error in M4 shown a big difference compared to other methods which probably will make M4 is the least preferable for imputing missing values. M5 has the least MAE of the range of 7.07 and 7.16. Then it followed by M1 and M3 which has MAE of the range of 8.38 and 9.17. M2 and M6 have a slightly higher MAE values compared to M1, M3 and M5 which are at range of 8.38 and 9.17. Again, M4 shown a great difference in terms of error (MAE) compared to other methods of imputation used with the values of in the range of 30.67 and 31.16. M1 has the highest Agreement Index (AI) of the range of 0.925 and 0.927 which is approaching 1. Then it followed by M5 and M3 which also shown a good Agreement Index (AI) of the range of 0.916 to 0.924. M2 and M6 have a moderate performances of Agreement Index (d) in the range of 0.883 and 0.909. M4 shown a weak performance of Agreement Index (AI), which is in the range of 0.5021 and 0.5029. M4 gave a big gap in term of performance with the other methods of imputation. M5 and M1 represented the good performance in terms of R² in the range of 0.755 and 0.760. While M3 gave a moderately performance of R² with the values in the range of 0.916 and 0.917. The least performed method in terms of R² is M4 which having a value in the range of 0.5021 and 0.5029 which are too low compared to the other methods used.

Based on the summary result in table 1, the performances are parallel according to its percentage of missingness. The best imputation method for estimating the simulated missing data is M5 (Kalman Filter using ARIMA Model) method. This is because M5 give the smallest values of MAE and RMSE and the highest values for R² and Agreement Index (AI) in almost all parameters and percentages of the simulated missing data. Hence, M5 is the best method used for imputing missing values in this study.

**Figure 3.** Results of the performance of imputation methods based on performance indicators.

**Table 1.** Summary of the Best Imputation Methods.
| Imputation Method | Best Method |
|-------------------|-------------|
| Percentage of Missing | 5% | 10% | 15% |
| RMSE | M5 | M5 | M5 |
| MAE | M5 | M5 | M5 |
| Agreement Index (d) | M1 | M1 | M1 |
| R² | M5 | M5 | M5 |

According to figure 4, the boxplots have clearly shown that there is a big difference between RMSE for M4 compared to the other methods based on respective percentages of missing values. Moreover, the greatest value of RMSE for respective missing percentage increases as the percentage of missing value increasing (from 5% to 10% and 15%). RMSE for M4 averagely is between 25 to 45, while the other methods only have RMSE values between 5 to 17. Hence, M4 should be prevented since this method might give a greatest error and produced biasness. M1, M3 and M5 are the methods that seems to be the good methods among the other methods. As shown in the figure above, these methods produced a fairly good values of RMSE from the range of 9 to 11 on average. However, among these three methods, it can be seen that M5 gave the least value of RMSE and should be considered as the best method with least error.

Figure 4. The consistency of performance indicators based on box-plot analysis.

Mean Absolute Error (MAE) for M4 are very high compare to the remaining five methods regardless of its percentage of missing values. The boxplots for M1, M3, and M5 looked alike and the values behave the same within the range of 6.5 and 8. These three methods can be considered the good methods so far. Therefore, we should look a bit details on how these three differ between each other. For M5, it is clearly shown that it’s median for MAE (average of 7) are the lowest among those three methods. M1, M3 and M5 are the methods that seem to be the good methods among other methods. As shown in the figure above, these methods produced a quite good values of Agreement Index (AI) from the range of
0.90 to 0.95 on average. However, among these three methods, it can be seen that M5 gave the highest value of Agreement Index (AI) and should be considered as the best method with a good performance in terms of Agreement Index (AI) which is approaching to 1.

By referring to figure 5, Linear Interpolation, Exponential Moving Average and Kalman Filter using ARIMA Model are the methods that can be considered as good methods. These three methods are good because the scatterplots explained that almost all points aligned in a line which are having a strong positive relationship between imputed values and observed values and consistent compared to the other three methods used. Therefore, these three methods of imputation (Linear Interpolation, Exponential Moving Average and Kalman Filter using ARIMA Model) are recommended since they imputed those missing values with the values that approaching the same to its actual values.

![Figure 5](image-url)  
*Figure 5. The relationship between the observed and estimated value based on scatter plot.*

4. Conclusion
The main aim of this paper is to discuss and highlight the framework and the step by step approach needed in the investigation of missing values problem focuses on time series data set. To illustrate the methodology, a case study to assess and compare several imputation methods for PM$_{10}$ time series from Shah Alam Air Quality Monitoring Station were conducted. The experimentation in determining the best appropriate methods were conducted involving Linear Interpolation, Spline Interpolation, Exponential Moving Average, Random Sample, Kalman Filter using ARIMA Model and Mean Before and After. The results of the analysis show that Kalman Filter using ARIMA Model is the best imputation method for PM$_{10}$ time series data at Shah Alam air quality monitoring station. However, this identified method cannot be generalized to other locations. The same experiment and analysis will need to be conducted for the data set at different location. This is because, the data might have different behaviour due to different background and would be influenced by different pattern of meteorological factors and geography.

Acknowledgement
Authors wishing to acknowledge the Department of Environment (DOE) Malaysia for providing the data for the study analysis.

References
[1] Razak N A, Zubairi Y Z and Yunus R M 2014 Sains Malaysiana. 43 (10) 1599
[2] Norazian M N, Shukri Y A, Azam R N and Bakri A M 2008 ScienceAsia 34 341
[3] Noor MN, Abdullah AB, Yahaya AS and Ramli NA 2015 Material Science Forum 803 278
[4] Beck M W, Bokde N, Asencio-Cortés G and Kulat K 2018 The R Journal 10 (1) 218
[5] Rubin D B 1976 Biometrika 63 581
[6] Plaia A and Bondi A L 2006. Atmospheric Environment 40 667
[7] Junninen H, Niska H, Tuppurainen K, Ruuskanen J and Kolehmainen M 2004 Atmospheric Environment 38 2895
[8] Zainuri N A, Jemain A A and Muda N 2015 Sains Malaysiana 3 (44) 449
[9] Shaadan N, Deni S M and Jemain A A 2015 Sains Malaysiana 44 (10) 1531
[10] Mazlan N, Rahman NA and Deni S M 2015 Int. Journal of Applied Mathematics and Statistics 53 6
[11] Dermirhan H and Renwick Z 2018 Applied Energy 225 998
[12] Junger W L and Leon A P 2015 Atmospheric Environment 102 96
[13] Ramli N A, Khairi S S M and Razlan N A 2018 Int. Journal of Academic Research in Business & Social Science 8 (4) 1185
[14] Burhnuddin S N Z A , Deni S M and Ramli N M 2016 Scientific Research Journal 13 (1) 84
[15] Jamaludin S, Deni S M and Jemain A A 2008 Asia-Pacific Journal of Atmospheric Sciences 44 (2) 93
[16] Burhnuddin S N Z A , Deni S M and Ramli N M 2017 International Journal of Geomate 13 (36) 131
[17] Shukur O B and Lee M H 2015 Modern Applied Science 9 (11) 1
[18] Che Z, Purushotham S, Cho K, Sontag D and Liu Y 2018 Scientific Report 8 6085
[19] Bertsimas D, Powlowski C and Zhuo Y D 2018 Journal of Machine Learning Research 18 1-35
[20] Zhang Z 2016 Annals of Translational Medicine 4 (3) 56