Single and Multiple Imputation Method to Replace Missing Values in Air Pollution Datasets: A Review

Zuraira Libasin¹, Ahmad Zia Ul-Saufie¹ and Hasfazilah Ahmat², Wan Nur Shaziyani³

¹Department of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Pulau Pinang, Kampus Permatang Pauh, Malaysia
²Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kampus Shah Alam, Malaysia

E-mail: ahmadzia101@uitm.edu.my

Abstract. Imputation plays an essential role in handling the issue of missing data. The conventional techniques applied to overcome this problem are single imputation (SI) and multiple imputations (MI). These statistical strategies have their strengths and limitations in replacing missing data. This article reviews the state of the art of imputation methods employed in general publications in replacing missing values for air pollution data. A comprehensive review of the literature identifies the use of SI and MI slightly increases over the year. This paper concludes on the trend and the approaches used in the imputation methods. Subsequently, this paper put forward the gaps in imputation technique that less utilized a machine-learning approach in providing a substitute for missing values in air pollution data. The future direction of the research is to extend more machine-learning approach with higher accuracy with higher performance in imputing missing values.

1. Introduction

Air pollution is harmful gases and excessive trapped particles [1], which increase the risk factor associated with respiratory illness and health problems. According to [1], six common air pollutants affect human health and the environment. These six common air pollutants are particulate matter (PM), ozone (O₃), carbon monoxide (CO), sulfur oxides (SO₂), nitrogen oxides (NO₂) and lead. These pollutants are harmful and eventually become a real threat to the public's health. To overcome this issue, the researcher needs to evaluate, model and propose a prediction in the air pollution trends. Unfortunately, the air pollution data obtained from the continuous ambient air quality monitoring (CAAQM) station usually contained missing data. The justifications for such missing observations are due to machine failure, routine maintenance and human error.

Missing data is a prevalent problem in many scientific fields, including environmental research. It can cause serious problems that cause a significant difference in the research findings and hence lead to an incorrect conclusion [2]. Ignoring the missingness will lead to the situation of insufficient sampling, errors in measurements or faults in data acquisition [2].

Imputation plays an essential role in handling the issue of missing data. There are two methods in dealing with missing data: a single imputation method (SI) and multiple imputations method (MI). The researcher needs to carefully identify the pattern of missing data before engaging any suitable
imputation methods. This pattern is to ensure that the data imputed from any purposes will produce unbiased results [7]. The varieties of missing data and their descriptions [8] are as depicted in figure 1.

![Figure 1. Types of missing data and its description.](image)

This paper provides a comprehensive review of the recent studies in different applications related to imputation; single imputation (SI) and multiple imputations (MI).

2. Research Methodology
In this paper, only the Scopus database is used as the leading search engine to find the relevant articles since Scopus is the database that covers more journals than the other services [35, 36]. Figure 2 illustrates the number of documents available in the Scopus database for both SI and MI techniques used for replacing the missing values in air pollution data.

![Figure 2. The number of documents with SI and MI techniques.](image)
The number of documents for a single imputation method increased slightly in 2014 and 2019 with three papers each year. However, materials with MI technique recorded a higher amount than the SI technique. The database includes 12,599 documents for the 'imputation' search keyword itself.

3. Imputation Method
Imputation becomes a beautiful approach because it produces a complete data set by replacing missing data with substituted values. Two different types of imputation methods will review in detail in this paper; single imputation method (SI) and multiple imputations method (MI). Imputation can generate by using software packages such as SPSS, SAS, R, NORM, and many more.

3.1. Single Imputation (SI)
In a single imputation method, the imputed value is treating as the actual value [3]. It generates only one replacement value for each missing data point. According to [4], the resulting completed data set in the unique imputation method is used for inference proposes. Here comes the review in table 1.

| Reference | Year | Application | SI Method |
|-----------|------|-------------|-----------|
| [30]      | 2019 | Estimating an hourly NO₂ concentration | Spatial Interpolation, Linear Regression Models |
| [31]      | 2019 | Environmental contaminant on a health outcome | SI with a constant value, Likelihood-based estimation |
| [16]      | 2019 | Reconstruct incomplete air quality datasets | Mean imputation, Conditional mean imputation, KNN imputation |
| [32]      | 2015 | Functional data analysis | Means of curve estimation using a regression approach |
| [22]      | 2015 | The compositional approach of left-censored data | EM algorithm |
| [24]      | 2014 | Environmental epidemiological | Linear Regression Model (LM), Partial Least Squares Model (PLS) |
| [14]      | 2014 | Air quality datasets | Listwise deletion, Unconditional mean imputation, Principal component analysis (PCA) |
| [25]      | 2013 | Filling missing observations in short gap length. | EM algorithm, Mean imputation, Regression imputation, Stochastic regression imputation |
| [36]      | 2008 | Imputing missing value to annual hourly PM₁₀ concentrations | Linear Interpolation, Mean imputation |
| [33]      | 2006 | Imputing missing value on multilevel structure | SDEM method |
3.2. Multiple Imputation (MI)

Multiple imputations impute each replacement value with different plausible estimates of the missing data points. Since the SI ignores ambiguity and usually underestimates the variance, therefore the MI is more preferred than a SI which serves as a solution for both within-imputation ambiguity and between-imputation ambiguity. Here comes the review in table 2.

Table 2. List of MI reviewed papers.

| Reference | Year | Application | MI Method |
|-----------|------|-------------|-----------|
| [5]       | 2020 | Satellite AOD and PM2.5 | Deep Learning Approach |
|           |      |             | Censored Likelihood MI |
| [31]      | 2019 | Environmental contaminant on a health outcome | MI |
|           |      |             | Neural Network |
| [15]      | 2019 | Atmospheric pollutants | MI |
|           |      |             | Bayesian Principal Component Analysis Imputation |
| [16]      | 2019 | Reconstruct incomplete air quality datasets | Random forest |
| [17]      | 2018 | Predicting monthly PM$_{2.5}$ concentrations | MI by Markov Chain Monte Carlo (MCMC) |
| [6]       | 2018 | Urban air pollutant data | Multiple regression techniques |
|           |      |             | Artificial Neural Network Model |
| [18]      | 2018 | Ozone | MI – the combination of MAIAC and CTM |
| [19]      | 2017 | Estimating daily PM$_{2.5}$ concentrations | MI survey analysis method |
| [20]      | 2017 | Serum concentrations of dioxin-like compounds | Model-based MI |
| [21]      | 2015 | Chemical speciation | ML |
| [22]      | 2015 | Left-censored data under a compositional approach | |
| [23]      | 2014 | Spatio-temporally data | MI |
|           |      |             | Random forest |
| [24]      | 2014 | Environmental epidemiological | MI through LM and PLS |
|           |      |             | Bayesian approach |
| [14]      | 2014 | Complex missing data in air quality datasets | MI |
| [25]      | 2013 | Filling missing observations in short gap length. | KNN |
|           |      |             | SKNN |
|           |      |             | IA |
|           |      |             | BPCA |
| [26]      | 2013 | Environmental epidemiological | MI |
| [27]      | 2007 | Air toxics | MI |
|           |      |             | Model-based MI |
| [33]      | 2006 | Imputing missing value on multilevel structure | Optimal linear estimation |
| [28]      | 2004 | Environmental health | MI – regression |
| [29]      | 2001 | Incomplete multivariate time series | MI – Bayesian, MCMC |
4. Discussion

Either SI or MI, both have their strength and weaknesses. In choosing a suitable method, data must be in line with the varieties of missing data either MCAR, MAR or MNAR. Table 3 is a summary of the advantages and disadvantages of SI and MI methods in handling missing values [14].

| Method     | Advantages                                                                 | Disadvantages                                                                 |
|------------|----------------------------------------------------------------------------|------------------------------------------------------------------------------|
| SI (i.e. listwise) | • Simple to understand and apply  
• Fast  
• Any type of statistical analysis  
• No specific computational methods are required  
• Applied by default by much statistical software  
• Unbiased parameter estimates, under the MCAR assumption | • A substantial decrease in the sample size  
• Reduces statistical power  
• Doesn't use all facts  
• If data are MCAR, then there is a loss in command  
• Biased results if data are not MCAR  
• An association between variables can be affected |
| MI         | • General approach  
• Easy to understand, hard to program  
• Unbiased estimates provided that more validity than ad-hoc approaches  
• Conserves sample size and statistical power  
• Commercial software available  
• Standard errors of the estimates can be calculated | • Difficult to program (special software required)  
• Computing intensive  
• The analyst has little control over the imputation model.  
• The variance is more significant than for the single-imputation methods. |

Results in table 1 and table 2 provide an insight into the use of SI and MI. From the list, the applications of imputation techniques characterized into statistical analysis (SA) approach (i.e. linear interpolation, regression-based imputation) or/and machine learning (ML) approach (i.e. ANN).

Figure 3. The application of imputation techniques.

Figure 3 shows the application of imputation techniques using statistical analysis (SA) approach and machine learning (ML) approach in replacing the missing values in air pollution data. Majority of the articles applied SA approach for both imputation methods, whereas only one piece by [17] using the
ML approach in MI but none in the SI method. The findings will be of interest to enhance research work in the application of ML since ML is useful when dealing with big data. A total of three articles [5, 18, 30] utilized both approaches in imputing missing values. These findings provide the insights for future research to publish more articles in imputation using both methods where this combination has eventually come out with a hybrid model that is recommending highly in solving missing value problem, especially to the long gaps. Currently, there are three articles [37-39] in the environmental area use a combination of the hybrid method to solve for imputing missing values based on Scopus database.

5. Conclusions
The application of single imputation method (SI) and multiple imputation method (MI) has increased over the last decade. Along with the application of hybrids and machine learning, the process of imputing the missing value shows better result. Through these methods, the main objective of researchers is to get higher accuracy and efficiency.

Acknowledgement
The authors would like to extend their appreciation to the Department of Environmental Malaysia (DoE) for providing air quality data for this research. The research was funded by 600-IRMI/FRGS 5/3 (289/2019).

References
[1] Industrial Degreasers Ecolink Corporate 2018 Article 1. What are the six common air pollutants Retrieved June 7 2020 Available from https://ecolink.com/info/six-common-air-pollutants/
[2] Junninen H, Niska H, Tuppurainen K, Ruuskanen J and Kolehmainen M 2004 Methods for imputation of missing values in air quality data sets Atmospheric Environment 38 2895-2907
[3] Raghunath Arnab 2017 Survey Sampling Theory and Applications
[4] Mortaza Jamshidian and Matthew Mata 2007 Handbook of Latent Variable and Related Models
[5] Li L 2020 A Robust Deep Learning Approach for Spatiotemporal Estimation of Satellite AOD and PM$_{2.5}$ Remote Sens 12 264
[6] Zakaria N A and Noor N M 2018 Imputation methods for filling missing data in urban air pollution data for Malaysia Urbanism. Architecture. Constructions 9 159-166
[7] Sukanya G and Manoj K G 2018 A survey on different techniques for handling missing values in dataset International Journal of Scientific Research in Computer Science, Engineering and Information Technology p 295-301
[8] Hirabayashi S and Kroll C N 2017 Single Imputation Method of Missing Air Quality Data for i-Tree Eco Analyses in the Conterminous United States (Retrieved June 7, 2020. Available from https://ireetools.org/eco/resources/Simple_imputation_method_of_missing_air_quality_data_for_i-Tree_Eco_analyses_in_the_conterminous_United_States.pdf)
[9] Fairclough D L 2002 Multiple Imputation for Non-Random Missing Data in Longitudinal Studies of Health-Related Quality of Life In: Mesbah M., Cole B.F., Lee M.L.T. (eds) Statistical Methods for Quality of Life Studies (Boston: Springer)
[10] Fernández M P, Vallejo G, Livácic-Rojas P and Tuero Herrera E 2018 The (Ir)Responsibility of (Under)Estimating Missing Data Frontiers in Psychology 9 10.3389/fpsyg.2018.00556
[11] Little R J A and Donald B R 2019 Statistical Analysis With Missing Data 3rd ed. (John Wiley & Sons)
[12] Little R J A 1988 A test of missing completely at random for multivariate data with missing values Journal of American Statistical Association 83 1198-1202
[13] Dixon W J 1988 BMDP Statistical software (Los Angeles: University of California Press)
[14] Gómez-Carracedo, M P, Andrade, J M López-Mahía, P Muniategui S and Prada D 2014 A practical comparison of single and multiple imputation methods to handle complex missing
data in air quality datasets Chemometrics and Intelligent Laboratory Systems 134 23-33
[15] S Grisales et al. 2019 Imputación de datos perdidos en series de mediciones de contaminantes atmosféricos insuyo para la vigilancia en salud ambiental Congreso Colombiano y Conferencia Internacional de Calidad de Aire y Salud Pública (CASP) Barranquilla, Colombia 2019 p 1-7 doi: 10.1109/CASP.2019.8916686
[16] Quinteros M E, Lu S, Blazquez C, Harrison R M, Ruiz-Rudolph P. et al. 2019 Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile Atmospheric Environment 200 40-49
[17] Huang K, Xiao Q, Meng X, Gu D, Liu Y et al. 2018 Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain Environmental Pollution 242 675-683
[18] Arroyo Á, Herrero Á, Tricio V, Corchado E and Wóźniak M 2018 Neural models for imputation of missing ozone data in air-quality datasets Complexity 2018 7238015
[19] Xiao Q, Wang Y, Chang H H, Lyapustin A, Liu Y et al. 2017 Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China Remote Sensing of Environment 199 437-446
[20] Bichteler A, Wikoff D S, Loko F and Harris M A 2017 Estimating serum concentrations of dioxin-like compounds in the US population effective 2005–2006 and 2007–2008: A multiple imputation and trending approach incorporating NHANES pooled sample data Environment International 105 112-125
[21] Krall J R, Simpson C H and Peng R D 2015 A model-based approach for imputing censored data in source apportionment studies Environmental and Ecological Statistics 22(4) 779-800
[22] Palarea-Albaladejo J, Martin-Fernández J A, Palarea-Albaladejo J and Martin-Fernández J A 2015 Chemometrics and Intelligent Laboratory Systems 143 85-96
[23] Feng L, Nowak G, O’Neill T J and Welsh A H 2014 CUTOFF: A spatio-temporal imputation method Journal of Hydrology 519(PD) 3591-3605
[24] Roda C, Nicolis I, Monas I and Guihenneuc C 2014 New insights into handling missing values in environmental epidemiological studies PLoS ONE 9(9) e104254
[25] Norazian Ramli M N, Yahaya A S, Ramli N A, Yusof N F F M and Abdullah M M A 2013 Roles of imputation methods for filling the missing values: A review Advances in Environmental Biology 7 (SPEC. ISSUE 12) 3861-3869
[26] Beyea J, Stellman S D, Teitelbaum S, Mordukhovich I and Gammon M D 2013 Imputation method for lifetime exposure assessment in air pollution epidemiologic studies Environmental Health: A Global Access Science Source 12(1) 62
[27] Le H Q, Batterman S A and Wahl R L 2007 Reproducibility and imputation of air toxics data Journal of Environmental Monitoring 9(12) 1358-1372
[28] Lubin J H, Colt J S, Camann D, Bernstein L, Hartge P et al. 2004 Epidemiologic evaluation of measurement data in the presence of detection limits Environmental Health Perspectives 112(7) 1691-1696
[29] Hopke P K, Liu C and Rubin D B 2001 Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the arctic Biometrics 57(1) 22-33
[30] Van Roode S, Ruiz-Aguilar J J, González-Enrique J and Turias I J 2019 An artificial neural network ensemble approach to generate air pollution maps Environmental Monitoring and Assessment 191(12) 727
[31] Boss J, Mukherjee B, Ferguson K K, Meeker J D, Kim S et al. 2019 Estimating Outcome-Exposure Associations when Exposure Biomarker Detection Limits vary Across Batches Epidemiology 30(5) 746-755
[32] Shaadan N, Deni S M and Jemain A A 2015 Application of functional data analysis for the treatment of missing air quality data Sains Malaysiana 44(10) 1531-1540
[33] Paia A and Bondi A L 2006 Single imputation method of missing values in environmental
pollution data sets *Atmospheric Environment* 40(38) 7316-7330

[34] Burnham J F 2006 Scopus database: a review *Biomedical digital libraries* 3 1 https://doi.org/10.1186/1742-5581-3-1

[35] AlRyalat S A, Nassar A A, Tamimi F, Al-Fraihat E, Assaf L, Ghareeb R, Masoudi M and Al-Essa M 2019 The impact of the open-access status on journal indices: oncology journals *Journal of gastrointestinal oncology* 10(4) 777–782

[36] Norazian M N, Shukri Y A, Azam R N and Al Bakri A M M 2008 Estimation of missing values in air pollution data using single imputation techniques *ScienceAsia* 34(3) 341-345

[37] Sharma E, Deo R C, Prasad R and Parisi A V 2020 A hybrid air quality early-warning framework: An hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms *Science of the Total Environment* 709 135934

[38] Wu H, Liu H and Duan Z 2020 PM2.5 concentrations forecasting using a new multi-objective feature selection and ensemble framework *Atmospheric Pollution Research* Article in Press

[39] Qin S, Liu F, Wang J and Sun B 2014 Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models *Atmospheric Environment* 98 665-675