Vol. 16 No. 1 (2025) 23-31
Jurnal Riset
Teknologi Pencegahan Pencemaran Industri
Journal homepage: https://www.jrtppi.id

Addressing Missing Data in Environmental Technologies: Economic and Environmental Optimizing Air Quality Monitoring with Random Forest and MissForest

Titin Agustin Nengsih*[1], Indrawata Wardhana[2], M. Nazori Madjid[3]
[123]UIN Sulthan Thaha Saifuddin Jambi, Indonesia

ARTICLE INFO

Article history:
Received: February, 27 2025
Received in revised form March, 29 2025
Accepted: April, 25 2025
Available online: May, 28 2025

Keywords:
Air Quality
Imputation
Missing Values
Random Forest
missForest

ABSTRACT

Air quality monitoring often encounters missing data issues due to technical glitches, equipment malfunctions, or other causes. This study employs PM2.5 and PM10 datasets from station 6, calculating multiple weighted probabilities for imputation. The methodology employed in this study includes the simulation of missing data patterns using multivariate amputation techniques (MCAR, MAR, and MNAR), followed by the application of machine learning-based imputation methods—Random Forest and missForest. The performance of each method was assessed using statistical evaluation metrics: Root Mean Square Error (RMSE), Nash-Sutcliffe Efficiency (NSE), and Kling-Gupta Efficiency (KGE) with missing values introduced at rates of 10, 40, and 70 percents. The results show that missForest consistently outperforms Random Forest across all missingness levels and amputation types. For example, in the low missing data scenario (10%), MF achieves RMSE values as low as 0.83 (PM2.5) and 1.76 (PM10), with perfect NSE and KGE scores (1.00), while RF yields higher RMSEs and slightly lower efficiencies. Even under high missing data conditions (70%), MF maintains strong performance with RMSE values of 10.54 and NSE above 0.87. These findings highlight MF's superior accuracy and robustness for handling missing air quality data.

1. INTRODUCTION

Ambient air pollution poses a significant environmental concern, exerting adverse effects on human health. Exposure to particulate pollutants, including fine particulate matter and ozone, increases the risk of mortality. Notably, in 2015, the Global Burden of Disease report indicated a staggering toll, attributing 4.2 million deaths and 103.1 million disability-adjusted life years to PM2.5 worldwide (Burnett et al., 2014). To address this issue, ground-based air quality monitoring stations have been established, enabling real-time surveillance of air quality. However, these stations are often geographically dispersed, leading to gaps in the dataset. This gap is particularly prominent during the initial stages of station deployment (Chu & Bilal, 2019).

The evaluation of a missing data methodology in a practical context involves applying it to a realistic missing data issue. In this regard, (Brand, 1999) proposed a multivariate amputation concept that replicates the absence of actual data. While this idea was briefly mentioned twice in previous literature, (Schouten et al., 2018) expanded on schematic concepts and introduced an amputation procedure capable of generating complex missing data scenarios. To address missing data patterns across various data types, including continuous, discrete, binary, unordered categorical, and ordered categorical variables, the Multiple Imputation by Chained Equations (MICE) approach has been effective (White et al., 2011)(Deng et al., 2016)(Zhao

*Correspondence author.
E-mail: nengsih@uinjambi.ac.id (Titin Agustin Nengsih)
doi: https://10.21771/jrtppi.2025.v16.no1.p23-31
2503-5010/2087-0965© 2025 Jurnal Riset Teknologi Pencegahan Pencemaran Industri-BBSPJPPI (JRTPPI-BBSPJPPI).
This is an open acces article under the CC BY-NC-SA license (https://creativecommons.org/licenses/by-nc-sa/4.0/).
Accreditation number: (Ristekdikti) 158/E/KPT/2021

& Long, 2016). Furthermore, in the context of Missing at Random (MAR) and Missing Not at Random (MNAR) scenarios, the Bayesian Imputation approach has shown superior performance (Halme & Tannenbaum, 2018). For high-dimensional data with a predictive focus, Bayesian Linear Regression has been successful (Castillo et al., 2015). A notable solution, MissForest (Stekhoven & Bühlmann, 2012), offers a promising avenue for handling missing data.

To maximize the benefits derived from imputed data in MICE, where reimputing the data would not alter standard error estimates, it is recommended to perform multiple imputations. According to (Graham et al., 2007) the optimal number of imputations (m) is suggested as 3 into 5 imputations. Addressing the challenge of determining the number of imputations, (von Hippel, 2020) introduces a two-step approach. In the quest for identifying the best parameter for multiple imputation, (von Hippel & Bartlett, 2021) employ the Maximum Likelihood method, offering enhanced efficiency in point estimates. This approach is not only less computationally intensive but also quicker, resulting in slightly more efficient point estimates.

To assess the MCAR, MAR, and MNAR missingness mechanisms, we introduce missingness into the variable Y deliberately. Subsequently, we categorize the MAR and MNAR methods based on their consideration of different aspects: incomplete variable's left (LEFT/L), right (RIGHT/R), both tails (TAIL/T), or distributional center (MID). The occurrence of MAR-induced missingness in Y relies on X, as indicated in Figure 2, which portrays the four distribution functions (LEFT, RIGHT, MID, and TAIL) (Schouten & Vink, 2018)(Wardhana et al., 2021). In scenarios involving MNAR missingness, the presence or absence of the true value of Y influences the probability of Y itself being missing. Additionally, we generate three levels of missingness proportions: 0.1, 0.5, and 0.9. It is important to emphasize that these proportions represent the sampled ratio of incomplete cases in Y while keeping X as a constant covariate. To generate missing values across all conditions, we employ the multivariate amputation technique (Schouten et al., 2018).

In the context of data characterized by non-linearity and non-normality, a comparison was conducted between

2. METHODS

Random Forest Imputation and predictive maintenance mean (Hong & Lynn, 2020). The application of Machine Learning imputation techniques extends to diverse fields, including meteorology observation (Boomgard-Zagrodnik & Brown, 2022) as well as geostatistics (Li et al., 2020) and (Avalos & Ortiz, 2020).

Handling missing values in air quality data is a critical aspect of ensuring accurate and reliable analyses. In the realm of air quality assessment, missing data can arise due to various reasons such as sensor malfunctions (Zainuri et al., 2015), equipment downtime (Norazian et al., 2008), or data transmission issues (Junger & Ponce de Leon, 2015). These gaps in the data can potentially lead to biased or incomplete conclusions if not properly addressed. To tackle this challenge, several imputation techniques are commonly employed. These techniques involve replacing missing values with estimated values based on the available data. Methods like mean imputation (Junger & Ponce de Leon, 2015), interpolation (Norazian et al., 2008), nearest neighbor imputation (Zhou et al., 2021), regression-based imputation (Quinteros et al., 2019), and multiple imputation (Zainuri et al., 2015) provide ways to fill in the gaps and enable more comprehensive analyses. The choice of imputation method depends on factors such as the nature of the data, the extent of missingness, and the specific goals of the analysis.

This study introduces a novel approach by integrating multivariate amputation (MCAR, MAR, MNAR) with weighted probability distributions to generate more realistic missing data scenarios. The use of two imputation methods—Random Forest and missForest—was intentional, as they represent advanced tree-based algorithms suited for different strengths. Random Forest is widely recognized for its effectiveness in handling high-dimensional, nonlinear datasets, while missForest builds upon this by incorporating an iterative, non-parametric framework that enhances imputation accuracy. By comparing the two under various missingness conditions, this research provides a more comprehensive understanding of their performance in air quality data contexts.

Dataset

The dataset comprises air quality observations collected from 135 monitoring stations throughout Uganda. This continuous dataset encompasses calibrated hourly

PM2.5 and PM10 data derived from air quality monitoring devices and a reference-grade monitoring apparatus during the period between 2019 and 2020. It consists of two files, namely "hourly air quality data.csv" and "reference grade monitor hourly air quality data.csv." These files contain timestamps in UTC, PM2.5 and PM10 concentrations, unique site IDs for monitoring sites, and site coordinates (latitude and longitude). Analysis of the monitor dataset reveals mean PM2.5 and PM10 concentrations of 37.39g/m3 and 49.61g/m3, respectively. The reference-grade monitor employed for this data collection is the Met One Beta Attenuation Monitor Model 1022, specifically designed for hourly PM2.5 concentration measurement and recording. In contrast, the low-cost monitors utilize laser scattering technology and dual Plantower Sensors (PMS 5003) (Sserunjogi et al., 2022).

Method Imputation

Imputing missing air quality data is crucial for air pollution research and monitoring. Various methods exist: the simple single imputation replaces missing values with estimated ones using mean (Hirabayashi & Kroll, 2017), median, or regression; multiple imputation creates multiple simulated values to capture uncertainty(Schouten *et al.*, 2018); spectral methods employ discrete sampling for non-stationary time series(Alsaber *et al.*, 2021); logistic regression handles non-linear relationships in time series(Chen *et al.*, 2022), needing substantial data. Choosing a method depends on data specifics and resources. A comparative study can help identify the most effective approach.

Random Forest (RF) is an ensemble technique based on decision trees, designed to fill in missing data by consolidating outcomes from several decision trees (Deng *et al.*, 2016). These trees differ due to their creation from diverse datasets, leading to distinct outcome predictions for the same inputs. RF then combines these predictions through a voting process to yield a final result. This imputation method boasts strong classification capabilities and is well-suited for managing high-dimensional data. The process of RF as shown in algorithm 1.

Algorithm 1 Random Forest

Input : Data matrix $X = \{X_{obs}, X_{miss}\}$

Output : Imputed data matrix, $X = \{X_{obs}, X_{imputed}\}$

for i = 1 -> 4 (multiple imputations) **do**

$\{\bar{X}, \sigma^2\} \leftarrow X_{obs}$

Initial imputation, $X^0_{miss} \leftarrow N(\bar{X}, \sigma^2)$

for i = 1 -> N **do**

Estimate W^t , Equation 6

Estimate X^0_{miss} , Equation 7

$X^0_{miss} \sim P(X^t_{miss} \mid X_{obs}, X^{t-1}_{miss})$

end for

$X^i_{imputed} \leftarrow X^N_{misss}$

end for

return $X_{imputed} \leftarrow Aggregate(X^i_{imputed})$

MissForest (MF) employs an iterative method that utilizes the Random Forest algorithm to predict missing values. In each iteration, a Random Forest model is constructed for individual variables, leveraging observed data to estimate missing values. This process iterates until convergence, progressively refining imputed values with each iteration. (Zhang *et al.*, 2021)

Algorithm 2 missForest

Require: X an $n \; x \; p$ matrix, stopping criterion γ

Sort X by amount of missing values of stations descend;

Make an initial guess for missing values using another method;

While not γ **do**

$X^{imp}_{old} \leftarrow$ store previously imputed matrix;

For s in 1…p **do**

Fit a random forest : $y^{(s)}_{obs} \sim x^{(s)}_{obs}$

Predict $y^{(s)}_{obs}$ **using** $x^{(s)}_{obs}$;

$X^{imp}_{old} \leftarrow$ update impute matrix, using predicted $y^{(s)}_{obs}$;

Update γ;

Return the imputed matrix X^{imp};

Process of Missing Values

The process of generating missing values involves deliberately creating gaps in data to assess how a model performs on different complete datasets.To achieve this, missing values are introduced only to the testing instances,

while the training instances remain unaffected and complete. In cases where the original dataset has missing entries, the training instances with missing values are excluded, resulting in the construction of an RF model using fully observed training data. Three distinct missing mechanisms are introduced, each serving a unique purpose, and their specifics are outlined in (Karmitsa *et al.*, 2022)(Alsaber *et al.*, 2021):

- **MAR** : indicates that the likelihood of an attribute having missing values is influenced by the values of other attributes (Schouten & Vink, 2018).
- **MNAR** : In this scenario, the probability of an attribute having missing values is connected to the attribute's own value. Specifically, missing values are introduced in one attribute, with higher attribute values being removed at a certain proportion(Khan & Hoque, 2020).
- **MCAR**: Under this mechanism, a specific number of locations are chosen randomly, and the values at these chosen locations are removed. Importantly, the decision to introduce missing values is independent of the values of other attributes or the attribute itself (Idri *et al.*, **2018**).

To examine how the rate of missing data influences classification outcomes, portions of values within the datasets are randomly removed at fractions of 10%, 40%, and 70%, respectively. By systematically altering the missing rates, the study aims to gain insights into how the presence of missing data impacts classification results.

Evaluation Criteria

The performance of MissForest and Random Forest at air quality multiple imputation data was tested by comparing prediction data in different percentages with the observed data using the Normalized Root Mean Square Error (NRMSE), the Nash-Sutcliffe efficiency (NSE), and the Kling-Gupta efficiency (KGE) as can see in Eq (1-4).

$$NRMSE = \frac{RMSE}{\frac{1}{N}\sum_{i=1}^{N} X_i} \quad (1)$$

$$NSE(X, \hat{X}) = 1 - \frac{\sum_{i=0}^{N-1}(X_i - \hat{X}_i)^2}{\sum_{i=0}^{N-1}(X_i - mean(X))^2} \quad (2)$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \quad (3)$$

$$\beta = \frac{\mu_s}{\mu_o} \qquad \gamma = \frac{\sigma_s / \mu_s}{\sigma_o / \mu_o} \quad (4)$$

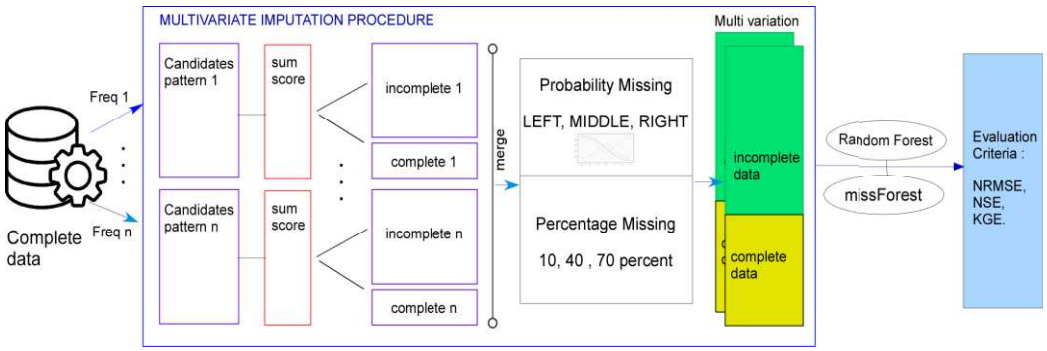

Figure 1. Framework of Multi Weight Probabilities

3. RESULT AND DISCUSSION

The multivariate continuous data, which was distributed across monitoring stations in Uganda with 0% missing data, was analyzed using Random Forest and missForest. The variable PM2 5, PM10, and MCAR missingness mechanisms were used to create the amputation. Considering element missingness is completely random and we can probably not predict that value from any other value in the data, it is assumed that some data is missing. The

Missing Completely at Random (MCAR) algorithm is utilized due to this. Multiple weight probabilities, along with the distributional center (DC), right tail (RT), and left tail (L), are employed to create incomplete variables (M). Several imputation percentages are also broken down into three groups: low (10%), middle (40%), and high (70%).

Figure 2 illustrates that among all stations, station 6 exhibited the lowest outlier values for both PM 2.5 and PM 10. Building upon this observation, station 6 was selected as a key input source, aggregating data from 135 monitoring stations. As detailed in Table 1, station 6 maintained a mean PM 2.5 concentration of 37.39 g/m³ without any data gaps, and a mean PM 10 concentration of 49.61 g/m³, also without any missing values.

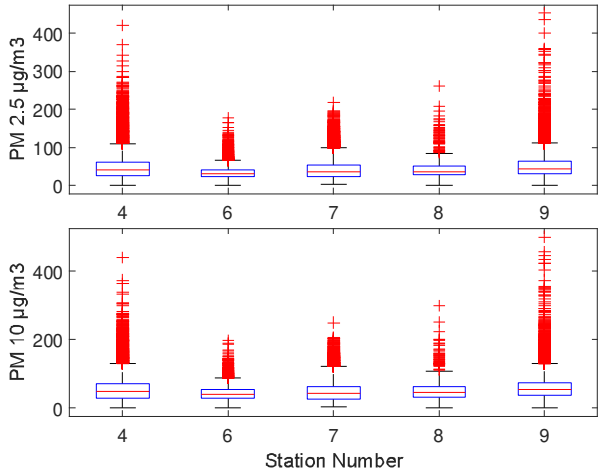

Figure 1. Station number with PM2.5 and PM10

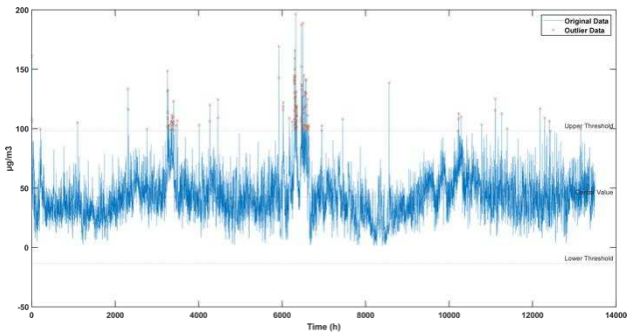

Figure 2. Outlier in Data Station 6

Fig 3 illustrates the distribution of data spanning between the lower and upper thresholds. The majority of the data points are within this range, with only a minimal portion classified as outliers. To detect outliers, a mean-based approach is employed. This technique involves calculating the mean of the data and identifying data points that deviate beyond a specific range from this mean.

From fig 4, we see that the kurtosis of both data PM 2.5 and PM 10 was leptokurtic with values: 3.3029 and 3.1156. The skewness was positive for both data with values : 0.4682 and 0.5168. The table 2 offers a comprehensive analysis of imputation techniques, specifically MF and RF, across various levels of missing data categorized as Low, Middle, and High. In the Low missing data scenario, both MF and RF exhibit favorable results with generally low RMSE values, indicating effective imputation. The NSE values are consistently high, implying a strong alignment between observed and imputed values. Additionally, the KGE values are notably high, reflecting robust model performance and accurate imputation.

Moving to the Middle missing data scenario, a nuanced trend emerges. Although MF tends to yield slightly higher RMSE values compared to RF, both methods maintain high NSE values, signifying proficient representation of observed values. The KGE values remain elevated, reinforcing the notion of reliable model efficiency even in moderately incomplete datasets.

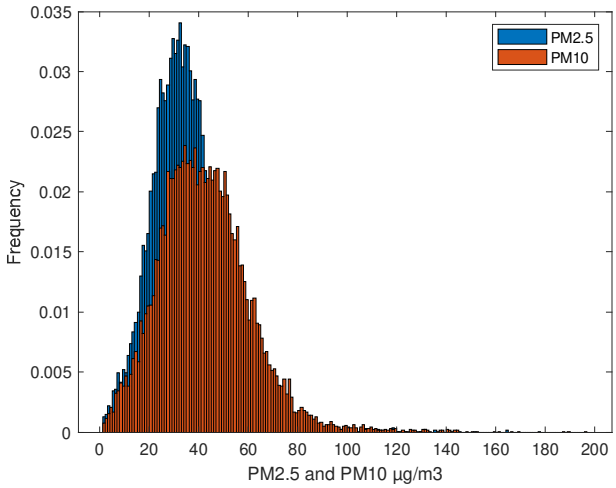

Figure 3. Data distribution of Station 6

Conversely, the high missing data scenario introduces more significant challenges. Imputation errors escalate for both MF and RF, as evidenced by elevated RMSE values. The NSE values experience a decline, particularly pronounced for RF, indicating a diminished concordance with observed data. A similar pattern emerges with the KGE values, illustrating reduced model efficiency in capturing variability under increased missing data conditions.

Table 1. Descriptive statistics of the Air Quality all station

Var	Unit	% miss	Range	Mean	Var	Std Dev
PM2.5	(μg/m³)	0	4.77 – 214.43	37.39	783.70	27.99
PM10	(μg/m³)	0	1 - 499.45	49.61	1482.66	38.50

In summation for PM 2.5 , the analysis underscores the effectiveness of both MF and RF imputation techniques, particularly in scenarios with lower levels of missing data. The outcomes in the High missing data scenario highlight the inherent difficulty of imputing highly incomplete datasets, with RF showing a marginally greater impact. Thus, tailored approaches may be necessary for addressing imputation challenges in varying missing data scenarios to ensure accurate and reliable results.

The RF method for the NSE and KGE values equal to one in type L, M, and R for 10 percent amputation may be observed in table 3 using the method imputation for PM10. For all performance evaluations, the MF still outperforms the RF even with a 40 percent amputation for type L, M, and R. The table's extensive analysis reveals a comprehensive comparison between the MF and RF models, examining their performance metrics across distinct categories and positions. Notably, Model MF consistently outperforms RF in terms of RMSE and KGE metrics, showcasing its ability to achieve a higher level of agreement between predicted and observed values. This superiority is evident across various categories, with Model MF demonstrating a marginal advantage in metrics like RMSE and KGE, particularly noteworthy in the high category. Additionally, both models exhibit robust predictive accuracy, as evidenced by consistently high NSE scores across most cases.

An intriguing observation emerges when considering data imputation under significant challenges. Despite a substantial 70% data amputation, Model MF showcases remarkable resilience in imputing data accurately, as reflected by its KGE scores nearing the ideal value of 1. Meanwhile, RF's imputation performance remains noteworthy, achieving up to 86% accuracy for types L, M, and R at the 70% amputation threshold.

In light of these findings, a compelling conclusion emerges: Model MF consistently demonstrates superior accuracy compared to RF across diverse types and percentages of data amputation. Its capacity to sustain high precision in data imputation even under severe conditions further reinforces its effectiveness. Ultimately, the comprehensive analysis underscores Model MF's efficacy and reliability in predictive modeling and data imputation scenarios.From figure 6, it shows that most of type L, M, and R in 70 percent missing values can be solved with MF. Most of error prediction imputation were held in range 40 – 60 as shown in orange circle.

Table 2. PM2.5 for Imputation missForest and Random Forest with evaluation criteria RMSE, NSE and KG

Percentage	type	RMSE		NSE		KGE	
		MF	RF	MF	RF	MF	RF
Low	L	1.74	3.48	1	0.98	1	0.98
	M	0.83	3.73	1	0.98	1	0.98
	R	1.82	3.18	1	0.99	1	0.98
Middle	L	2.7	5.18	0.99	0.97	1	0.96
	M	1.59	5.22	1	0.96	1	0.96
	R	2.26	5.76	0.99	0.96	1	0.95
High	L	7.43	7.81	0.93	0.91	0.97	0.89
	M	9.91	6.94	0.88	0.93	0.94	0.91

	R	10.54	8.55	0.87	0.89	0.93	0.88

Table 3. PM10 for Imputation missForest and Random Forest

Percentage	type	RMSE		NSE		KGE	
		MF	RF	MF	RF	MF	RF
low	L	2	5.14	1	0.99	1	0.98
	M	3.4	4.35	0.99	0.99	1	0.99
	R	1.76	5.15	1	0.99	1	0.98
middle	L	4.41	8.82	0.99	0.96	0.99	0.95
	M	4.96	8.27	0.99	0.96	0.99	0.96
	R	4.08	7.86	0.99	0.97	1	0.96
high	L	11.95	13.01	0.93	0.89	0.96	0.88
	M	14.49	12.11	0.89	0.91	0.94	0.89
	R	14.38	14.27	0.89	0.87	0.95	0.86

4. CONCLUSION

In conclusion, this research underscores the significance of accurate data imputation techniques, with MissForest proving to be a reliable and robust method for addressing missing data across varying levels of complexity in Air Quality Index. The findings emphasize the importance of tailored approaches and shed light on the limitations and strengths of different imputation strategies for enhancing data integrity and analysis.

In conclusion, this study highlights the critical role of accurate imputation in air quality monitoring. The missForest method consistently demonstrated superior performance across all missingness levels and types, outperforming Random Forest in terms of RMSE, NSE, and KGE. Notably, missForest achieved near-perfect results in low missingness scenarios, with RMSE as low as 0.83 (PM2.5) and 1.76 (PM10), and NSE and KGE values reaching 1.00. Even at a high missingness level of 70%, missForest maintained strong performance with RMSE up to 10.54 and NSE above 0.87. These findings underscore missForest's robustness and reliability in handling complex missing data, making it a highly recommended method for environmental data imputation

REFERENCE

Alsaber, A. R., Pan, J., & Al-Hurban, A. (2021). 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(3), 1–26.

Avalos, S., & Ortiz, J. M. (2020). Recursive convolutional neural networks in a multiple-point statistics framework. Computers and Geosciences, 141(May), 104522. https://doi.org/10.1016/j.cageo.2020.104522

Boomgard-Zagrodnik, J. P., & Brown, D. J. (2022). Machine learning imputation of missing Mesonet temperature observations. Computers and Electronics in Agriculture, 192(October 2021), 106580. https://doi.org/10.1016/j.compag.2021.106580

Brand, J. P. L. (1999). Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.

Burnett, R. T., Arden Pope, C., Ezzati, M., Olives, C., Lim, S. S., Mehta, S., Shin, H. H., Singh, G., Hubbell, B., Brauer, M., Ross Anderson, H., Smith, K. R., Balmes, J. R., Bruce, N. G., Kan, H., Laden, F., Prüss-Ustün, A., Turner, M. C., Gapstur, S. M., Diver, W. R., & Cohen, A. (2014). An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. Environmental Health Perspectives, 122(4), 397–403. https://doi.org/10.1289/ehp.1307049

Castillo, I., Schmidt-Hieber, J., & Van Der Vaart, A. (2015). Bayesian linear regression with sparse priors. Annals of Statistics, 43(5), 1986–2018. https://doi.org/10.1214/15-AOS1334

Chen, M., Zhu, H., Chen, Y., & Wang, Y. (2022). A Novel Missing Data Imputation Approach for Time Series Air Quality Data Based on Logistic Regression. Atmosphere, 13(7). https://doi.org/10.3390/atmos13071044

Chu, H. J., & Bilal, M. (2019). PM 2.5 mapping using integrated geographically temporally weighted regression (GTWR) and random sample consensus (RANSAC) models. Environmental Science and Pollution Research, 26(2), 1902–1910. https://doi.org/10.1007/S11356-018-3763-7/METRICS

Deng, Y., Chang, C., Seyoum Ido, M., & Long, Q. (2016). Multiple Imputation for General Missing Data Patterns in the Presence of High-dimensional Data OPEN. Nature Publishing Group. https://doi.org/10.1038/srep21689

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science, 8(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Halme, A. S., & Tannenbaum, C. (2018). Performance of a Bayesian Approach for Imputing Missing Data on the SF-12 Health-Related Quality-of-Life Measure. Value in Health, 21(12), 1406–1412. https://doi.org/10.1016/j.jval.2018.06.007

Hirabayashi, S., & Kroll, C. N. (2017). Single imputation method of missing air quality data for i-Tree Eco analyses in the conterminous United States. Environmetal Research Engineering, 1, 1–24.

Hong, S., & Lynn, H. S. (2020). Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Medical Research Methodology, 20(1), 1–12. https://doi.org/10.1186/s12874-020-01080-1

Idri, A., Abnane, I., & Abran, A. (2018). Support vector regression-based imputation in analogy-based software development effort estimation. Journal of Software: Evolution and Process, 30(12), 1–23. https://doi.org/10.1002/smr.2114

Junger, W. L., & Ponce de Leon, A. (2015). Imputation of missing data in time series for air pollutants. Atmospheric Environment, 102, 96–104. https://doi.org/10.1016/j.atmosenv.2014.11.049

Karmitsa, N., Taheri, S., Bagirov, A., & Makinen, P. (2022). MAR. IEEE Transactions on Knowledge and Data Engineering, 34(4), 1889–1901. https://doi.org/10.1109/TKDE.2020.3001694

Khan, S. I., & Hoque, A. S. M. L. (2020). SICE: an improved missing data imputation technique. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00313-w

Li, Y., Jiang, Y., Yang, C., Yu, M., Kamal, L., Armstrong, E. M., Huang, T., Moroni, D., & McGibbney, L. J. (2020). Improving search ranking of geospatial data based on deep learning using user behavior data. Computers and Geosciences, 142(October 2019), 104520. https://doi.org/10.1016/j.cageo.2020.104520

Norazian, M. N., Shukri, Y. A., Azam, R. N., & Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia, 34(3), 341–345. https://doi.org/10.2306/scienceasia1513-1874.2008.34.341

Quinteros, M. E., Lu, S., Blazquez, C., Cárdenas-R, J. P., Ossa, X., Delgado-Saborit, J. M., Harrison, R. M., &

Ruiz-Rudolph, P. (2019). Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. Atmospheric Environment, 200(November 2018), 40–49. https://doi.org/10.1016/j.atmosenv.2018.11.053

Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation, 88(15), 2909–2930. https://doi.org/10.1080/00949655.2018.1491577

Schouten, R. M., & Vink, G. (2018). The Dance of the Mechanisms: How Observed Information Influences the Validity of Missingness Assumptions: Https://Doi.Org/10.1177/0049124118799376, 50(3), 1243–1258. https://doi.org/10.1177/0049124118799376

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118. https://doi.org/10.1093/BIOINFORMATICS/BTR597

von Hippel, P. T. (2020). How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. Sociological Methods and Research, 49(3), 699–718. https://doi.org/10.1177/0049124117747303

von Hippel, P. T., & Bartlett, J. W. (2021). Maximum Likelihood Multiple Imputation: Faster Imputations and Consistent Standard Errors Without Posterior Draws. Statistical Science, 36(3), 400–420. https://doi.org/10.1214/20-STS793

Wardhana, I., Ariawijaya, M., Hasnur, R., Syafitri, R., & Nasuha, A. (2021). Design and analysis security architecture virtualization OpenVz. Journal of Physics: Conference Series, 1940(1). https://doi.org/10.1088/1742-6596/1940/1/012088

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30(4), 377–399. https://doi.org/10.1002/sim.4067

Zainuri, N. A., Jemain, A. A., & Muda, N. (2015). A comparison of various imputation methods for missing values in air quality data. Sains Malaysiana, 44(3), 449–456. https://doi.org/10.17576/jsm-2015-4403-17

Zhang, S., Gong, L., Zeng, Q., Li, W., Xiao, F., & Lei, J. (2021). Imputation of GPS coordinate time series using missforest. Remote Sensing, 13(12), 1–18. https://doi.org/10.3390/rs13122312

Zhao, Y., & Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. Statistical Methods in Medical Research, 25(5), 2021–2035. https://doi.org/10.1177/0962280213511027

Zhou, X., Liu, X., Lan, G., & Wu, J. (2021). Federated conditional generative adversarial nets imputation method for air quality missing data. Knowledge-Based Systems, 228, 107261. https://doi.org/10.1016/j.knosys.2021.107261