Journal of Current Science and Technology, April-June 2026 Vol. 16 No. 2, Article 187
Copyright ©2018-2026, Rangsit University ISSN 2630-0656 (Online)

Cite this article: Saeyang, S., Notodiputro, K. A., & Wanishsakpong, W. (2026). Analysis of imputation methods for missing not at random (MNAR) data: A comparative study of air pollution data in Bangkok, Thailand. *Journal of Current Science and Technology*, *16*(2), Article 187. https://doi.org/10.59796/jcst.V16N2.2026.187

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Analysis of Imputation Methods for Missing Not at Random (MNAR) Data: A Comparative Study of Air Pollution Data in Bangkok, Thailand

Sattawat Saeyang[1], Khairil Anwar Notodiputro[2], and Wandee Wanishsakpong[1,*]

[1]Department of Statistic, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand
[2]Statistics and Data Science Study Program, School of Data Science, Mathematics and Informatics, IPB University, Bogor 16680, Indonesia

*Corresponding author; E-mail: wandee.w@ku.th

Received 22 November 2025; Revised 25 December 2025; Accepted 24 January 2026; Published online 25 March 2026

Abstract

Environmental datasets are often large in scale and frequently contain missing observations due to interruptions. Addressing these missing values is essential for ensuring the reliability of subsequent analyses. In many practical cases, missingness depends on the unobserved values or the technical issues themselves. This missing status is called Missing Not at Random (MNAR), which remains one of the most challenging missing data mechanisms. This study investigates the MNAR pattern in an air pollution dataset from Bangkok, Thailand. A simulation framework used the PM_{10} variable as the target variable for random missing with MNAR patterns at varying rates. The methods used for missing-value imputation were K-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), and the Expectation Maximization (EM) algorithm. Imputation accuracy was evaluated using Root Mean Square Error (RMSE). Furthermore, imputation efficiency was tested by conducting ARIMA, LSTM, and RF forecasting models, and model performance was evaluated using Mean Absolute Percentage Error (MAPE). The simulation results showed that KNN consistently achieved the lowest RMSE (1.66-2.68) for missing-value imputation at 10%-70% missing rates. In addition, KNN-based models showed better performance in ARIMA, and the LSTM models achieved the lowest MAPE (2.31–2.46) across all missing rates, while EM-based models excel at the RF model better than KNN. When applied to the actual air pollution dataset, KNN-based models also performed most effectively for variables containing MNAR. However, for other missingness types, MICE- and EM-based models outperformed KNN-based models. Overall, this study highlights practical and efficient approaches for handling MNAR missingness in environmental datasets and provides insights that can help future researchers better recognize, manage, and mitigate MNAR-related issues in real-world data collection.

Keywords: *air pollution; auto-regressive integrated moving average (ARIMA); imputation; long short-term memory (LSTM); missing values; missing not at random (MNAR); random forest (RF)*

1. Introduction

In recent years, Thailand has faced a significant increase in pollution levels, particularly in its urban centers. Rapid population growth, coupled with industrialization, has contributed to the rise in pollutant emissions from various sources such as vehicles, factories, and agricultural activities (Ahmad et al., 2022). As a result, air quality in major cities, including Bangkok, has deteriorated, with particulate matter (PM), such as $PM_{2.5}$and PM_{10}, becoming a major environmental concern (Cheewinsiriwat et al., 2022). While the detrimental health effects of $PM_{2.5}$ have received considerable attention, PM_{10} particles also present a substantial threat to public health (Hosamane et al., 2020). To fully understand the impact and natural patterns of PM concentration, robust forecasting models are essential. In practice, traditional time series models like Autoregressive Integrated Moving Average (ARIMA), deep learning methods such as Long Short-Term Memory (LSTM) networks and Random Forest (RF) are established models that offer distinct advantages and consistently demonstrate high efficacy in forecasting research (Ali et al., 2012; Ban & Shen, 2022; Nourmohammad & Rashidi, 2025). The most effective approach often involves multivariate modeling, leveraging other relevant predictive factors. It is widely established that various air quality elements and meteorological factors exhibit strong correlations with PM levels, making them critical components in comprehensive prediction models (Nguyen et al., 2024; Ponsawansong et al., 2023). However, a major challenge in environmental data acquisition is the limited availability and operational consistency of data collection infrastructure. This limitation frequently results in incomplete datasets and significant data gaps. The presence of these missing values severely affects the accuracy and reliability of air quality assessments. In the context of time series analysis, missing data introduce significant methodological complications, often leading to skewed analyses and unreliable model calibration. This data incompleteness is particularly problematic when working with long-term datasets, as it hinders the ability to track the evolution of pollutant levels over time, thereby affecting both forecasting and policymaking (Junger & De Leon, 2015). Therefore, innovative approaches are needed to handle these gaps effectively and ensure that the conclusions drawn from such datasets are robust and reliable. Ultimately, this type of dataset, with its inherent time-dependent nature and missing values, must be treated as time series data to better handle and analyze the complexities it presents.

Missing data can significantly impact time series analysis, as they leads to incorrect estimations, biased forecasts, and compromised results (Chhabra, 2024). The nature of missing data can vary, and researchers have categorized them into three main types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The type of missingness determines how much bias the missing data introduce into the analysis and how effectively it can be mitigated (Agiwal & Chaudhuri, 2024). For instance, with MCAR, simpler imputation methods like mean substitution or deletion of missing cases may suffice without compromising the dataset's integrity. In contrast, MAR demands more sophisticated techniques, such as multiple imputation or regression-based approaches, to account for the relationships between observed and missing variables. The most complex missing data pattern is MNAR. In general, this issue is especially relevant in large-scale datasets such as air pollutants and environmental datasets. The missing occurs when the probability of missingness is related to the unobserved data itself, making it more complex to address (Alruhaymi et al., 2021; Ranganathan & Hunsberger, 2024)

Unlike MCAR or MAR, MNAR is more complex to address (Pereira et al., 2024). Numerous studies (Kang, 2013; Li et al., 2024) have tested various imputation techniques. Specifically, methods such as k-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), and Expectation-Maximization (EM) are frequently found to perform effectively compared to simpler alternatives, particularly under the assumption of MCAR. Moreover, many researchers facing missing-data issues just handle them by imputation without knowing the missing pattern. In most situations, attention is paid to forecast modeling after imputing missing values. KNN, MICE, and other statistical methods are used to address missing data and then forecast with deep learning models like LSTM (Karnati et al., 2025; Utama et al., 2024). This study was designed to conduct a comprehensive air pollution dataset simulation to evaluate the efficacy of KNN, MICE, and the Expectation-Maximization EM imputation techniques in handling data with MNAR in time-series contexts. The research focuses on three core areas: first, assessing the impact of KNN, MICE, and EM on forecasting accuracy using a controlled simulation, where the Root Mean Square Error (RMSE) is utilized for optimal model hyperparameter selection; second, modeling the imputed data using the traditional ARIMA approach, a deep-learning-based LSTM network, and decision-tree-based RF with final forecasting performance evaluated by the Mean Absolute Percentage Error (MAPE); and third, applying the entire established methodology to an actual air pollution dataset to provide practical insights into handling MNAR missingness under field conditions. Ultimately, this study aims to deliver a robust methodological framework for environmental scientists and air quality researchers by rigorously testing these diverse imputation techniques combined with both classic and advanced forecasting models. This comprehensive approach is designed to maintain data integrity, maximize the accuracy and long-term predictive reliability of the data, and offer clear, actionable guidance for addressing severe MNAR data gaps in real-world environmental monitoring applications.

2. Objectives

The objective of this research is to study the MNAR pattern in an air pollution dataset and impute missing of MNAR using KNN, MICE, and EM methods. Additionally, the research seeks to analyze the forecasting efficiency of the imputed datasets by conducting time-series forecasting using ARIMA, LSTM, and RF models.

3. Materials and Methods

3.1 Data and Study Area

Bangkok is considered an urban heat island (UHI) due to its rapid urbanization, which has led to increased built-up areas and a significant reduction in natural habitats (Iamtrakul et al., 2024). This urban expansion, characterized by high population density and inadequate green spaces, contributes to elevated air pollution in the city. The data for this study were obtained from the Din Daeng station under Department of Pollution Control (DPC) in Bangkok, Ministry of Natural Resources and Environment, Bangkok, Thailand (Figure 1). The air pollution dataset includes daily measurements of air pollutants and various meteorological variables as numerical data types, covering the period from January 1, 2017, through April 30, 2023, with a total of 2,310 days of observations. The air pollutants consist of $PM_{2.5}$ concentration ($PM_{2.5}$, µg/m^3), PM_{10} concentration (PM_{10}, µg/m^3), carbon monoxide (CO, mg/m^3), sulfur dioxide (SO_2, µg/m^3), nitrogen dioxide (NO_2, µg/m^3) and ozone (O_3, µg/m^3). The meteorological variables consist of wind speed (WS, m/s), wind direction (WD, °), temperature (Temp, °C), relative humidity (RH, %), rainfall (Rain, mm), and atmospheric pressure (Pres, mb).

3.2 Missingness Type

Missing data is a common and persistent challenge in time series analysis, particularly in real-world datasets where continuous data collection is required. In fields such as climate studies, finance, and healthcare, missing values often arise due to equipment malfunctions, transmission errors, or external disruptions. The presence of missing data can severely impact the reliability of statistical analysis and forecasting models, leading to biased estimates and inaccurate predictions (Mukherjee et al., 2023). If not handled appropriately, missing values can distort patterns, weaken model performance, and reduce the overall effectiveness of decision-making processes based on time series data. Therefore, addressing missing data is a crucial step in ensuring the integrity and accuracy of time series analysis, requiring careful selection of imputation and modeling techniques to minimize the impact on results. Generally, (Rubin, 1976) introduced a foundational framework categorizing missing data into three types based on its underlying mechanism. Missing Completely at random (MCAR) occurs when the probability of missing data is entirely random and unrelated to any observed or unobserved variables. This means the missing values do not introduce systematic bias and can often be handled using simple methods like deletion or mean imputation. Missing at random (MAR) happens when the likelihood of missing data depends on other observed variables but not on the missing values themselves. In this case, more advanced techniques such as regression-based imputation or multiple imputation are necessary to account for dependency. Missing not at random (MNAR) is the most complex scenario, where missingness is directly related to the unobserved values, making it difficult to address without additional assumptions or external information. This type of missing data requires specialized models and sensitivity analyses to minimize bias in analysis (Joel et al., 2022). In this study, missingness occurred simultaneously across all air pollution variables at certain time points. These missing patterns are likely due to real-world limitations in data collection, such as equipment malfunction or transmission failure. Therefore, the observed pattern strongly suggests that the dataset exhibits characteristics consistent with an MNAR mechanism.

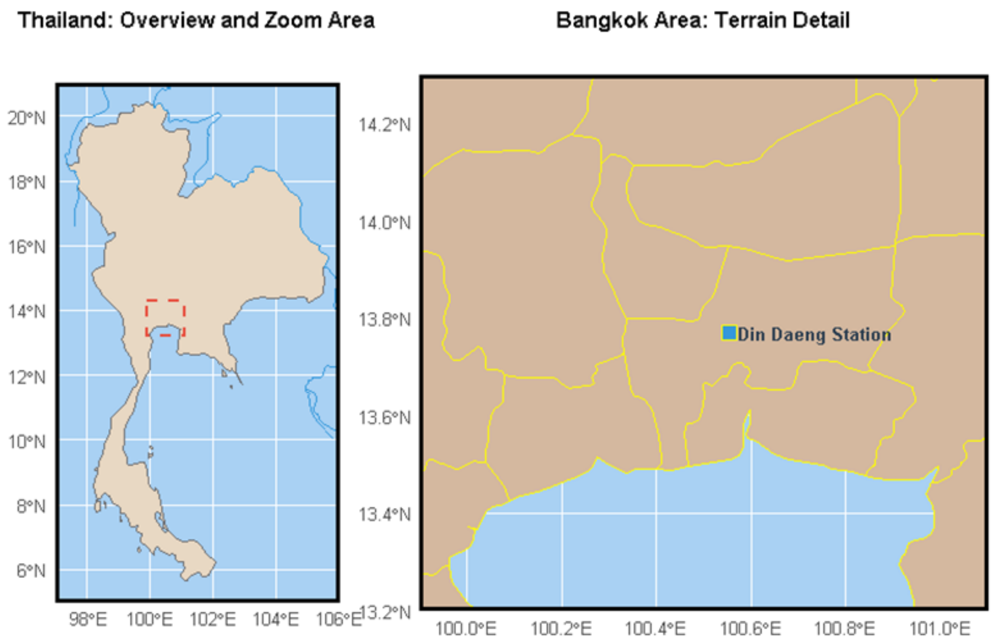

Figure 1 Location of the studied air pollution monitoring station in Bangkok, Thailand.

3.3 Imputation Techniques

Generally, There are various methods to handle missing data, each with different assumptions and effectiveness depending on the type of missingness (Heymans & Twisk, 2022). Complete Case Analysis (CCA), also known as listwise deletion, is a simple approach that removes any observation with missing values, but it can lead to biased results if data is not MCAR. Meanwhile, the air pollution dataset used in this study is MNAR. The imputation techniques are more reasonable than CCA and deletions. Based on findings from previous research (Kang, 2013; Li et al., 2024), these methods were selected for use in this study.

1) K-Nearest Neighbors (KNN). One effective way to estimate missing values is by leveraging similarity between observations. A commonly used method in this category finds missing values by identifying the most similar data points in the dataset based on predefined distance metrics. By averaging the values of the nearest neighbors, this method preserves local patterns but may be computationally expensive in large datasets (Fadlil et al., 2022). KNN imputation replaces missing values by finding the k most similar observations (neighbors) based on other available features. The missing value is then estimated using the average (for numerical data) or the most frequent category (for categorical data) of these neighbors. The k of this study was set as 5 which led the impute action take the 5 closest complete observations to calculated. The similarity is typically measured using a distance metric, such as Euclidean distance:

$$d(i,j)=\sqrt{\Sigma_{m=1}^{M}(X_{im}\text{-}X_{jm})^{2}} \tag{1}$$

where: $d(i,j)$ is the distance between instances i and j, X_{im} and X_{jm} are values of feature m for instances i and j, M is the number of features used for distance calculation. The missing value is then imputed as:

$$X_{imputed}=\frac{1}{k}\sum_{i=1}^{k}X_{i} \tag{2}$$

where: X_i represents the known values of the k nearest neighbors. KNN imputation is effective when data exhibits local patterns but can be computationally expensive for large datasets (Troyanskaya et al., 2001).

2) Multiple Imputation by Chained Equations (MICE). Another approach involves iteratively predicting missing values using regression models. Instead of relying on a single replacement, this method creates multiple possible values by considering relationships among variables. By repeating this process across different features and iterations, the approach provides robust imputations while accounting for variability (Alruhaymi et al.,

2021; Alruhaymi & Kim, 2021). One of the imputation methods available within the MICE is Predictive Mean Matching (PMM). PMM is a semi-parametric imputation technique that combines regression with nearest-neighbor matching. It first fits a regression model of the incomplete variable on selected predictors, then draws regression coefficients from the posterior distribution to generate predicted values (Laqueur et al., 2022).

$$\hat{y}_{i,j} = \beta_0 + \Sigma_{k \neq j} \beta_k x_{i,k} \quad (3)$$

where: $\hat{y}_{i,j}$ is the impute value, β_0 is the intercept, β_k is the coefficient assign to the predictor variable k , $x_{i,k}$ is the observed value of variable. For each missing entry, PMM identifies observed cases with the most similar predicted values and randomly selects one of their actual values for imputation (Morris et al., 2014). The imputation of missing in this method using PMM with 5 iterations (m = 5) and a convergence limit of 50 iterations.

3) Expectation-Maximization (EM). Some methods estimate missing values by maximizing the likelihood of observed data under an assumed probability distribution. This approach is particularly useful when data follows known distributions and when more accuracy is needed compared to simpler deterministic methods (Lin, 2010). EM is a statistical imputation method using Maximum Likelihood Estimator (MLE) to estimate missing values by iteratively optimizing the likelihood function. The EM algorithm consists of two steps: Expectation (E-Step): Estimates missing values using the expected value based on observed data and current parameter estimates. Maximization (M-Step): Updates the parameters by maximizing the likelihood function using both observed and estimated data. The procedure of EM imputation here has applied the "norm" function to identify the observed value and missing value as E-Step. After calculating the θ through MLE as an M-step, the next parameter was calculated too. If the parameters are still changing significantly, we will be back to step E. If they were similar, they would be placed in the missing gap. The iterative process follows:

$$Q(\theta|\theta^{(t)})=E[\log L(X_{observed}, X_{missing}|\theta)|X_{observed},\theta^{(t)}] \quad (4)$$

where $Q(\theta|\theta^{(t)})$ is the expected log-likelihood, and θ represents the parameters of the distribution. The process repeats until convergence, ensuring that missing values are estimated with minimal bias (Allison, 2009). In this study, Root Mean Square Error (RMSE) was used as an index to evaluate the precision of the imputation techniques for simulated PM_{10}. RMSE quantifies the average deviation between the imputed values and the true simulated PM_{10} values in the dataset.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2} \quad (5)$$

where is A_i is the simulated PM_{10} concentration at time i, F_i is the imputed simulated PM_{10} concentration at time i, n is the total number of data points. The RMSE result are further analyzed for Confidence Interval (CI) with the alpha 0.05 led Z-score equal 1.96 then multiply with margin of error ($s/\sqrt{n}$). This step ensures that the imputation performance is reliable (Cumming & Finch, 2005).

3.4 Models

1) ARIMA Model. The ARIMA process is a mathematical model used for forecasting time series data. ARIMA stands for Auto Regressive (AR), Integrated (I), and Moving Average (MA), with each component representing different aspects of the model. It has been extensively studied and remains a fundamental tool in time series analysis. The ARIMA process was popularized by George Box and Gwilym Jenkins in the early 1970s, leading to its alternative name, the Box-Jenkins Model. In their influential work, Bartholomew (1971) compiled in-depth information necessary for understanding and applying Univariate ARIMA processes. One of the simplest ARIMA models is the First-Order autoregressive model. Let Y_t be the observation at t time. Thus, AR (1) model is represented by the equation:

$$Y_t=c+\phi_1 Y_{t-1}+e_t \quad (6)$$

where c and ϕ_1 are constants, and e_t represents a random error at t time. This equation indicates that the next value in the series is determined by a constant $\boldsymbol{c}$, the weighted value of the previous observation, and a random error term. Higher-order autoregressive (AR) processes can be constructed by including additional past observations on the right-hand side of the equation.

$$Y_t=c+\phi_1 Y_{t-1}+\phi_2 Y_{t-2}+\ldots+\phi_p Y_{t-p}+e_t \quad (7)$$

Similarly, moving average (MA) processes are formed when past errors, rather than past observations, appear on the right-hand side of the equation.

$$Y_t=c+e_t-\theta_1 e_{t-1}-\theta_2 e_{t-2}-\ldots-\theta_p e_{t-p} \qquad (8)$$

ARIMA models also have extensions designed for more complex time series structures. Seasonal ARIMA models are tailored for seasonal time series data, capturing periodic patterns. Additionally, Vector ARIMA (VARIMA) is used for multivariate time series modeling, allowing multiple interrelated time series to be analyzed together. Other variations can incorporate explanatory variables to improve forecasting accuracy (Hyndman, 2001).

2) LSTM Model. Machine learning has revolutionized the field of data analysis, enabling computers to recognize patterns and make predictions without explicit programming. Among various machine learning techniques, deep learning has emerged as a powerful approach for handling complex datasets, particularly those with sequential dependencies (Takale et al., 2024). One of the most widely used deep learning models for sequential data is the Recurrent Neural Network (RNN). However, traditional RNNs suffer from the vanishing gradient problem, which limits their ability to capture long-term dependencies in sequences. To address this issue, Hochreiter & Schmidhuber (1997) introduced Long Short-Term Memory (LSTM) networks to enhance the standard RNN architecture by incorporating specialized gating mechanisms, namely the input gate, forget gate, and output gate, that regulate the flow of information. These gates enable LSTMs to retain relevant information over extended time periods while filtering out unnecessary details, making them particularly effective for time series forecasting and other applications requiring long-range memory retention.

3) The Random Forest (RF) algorithm, pioneered by Leo Breiman, is a highly effective ensemble learning method. It operates by constructing a multitude of unpruned decision trees, relying on two distinct randomization steps to achieve robust predictive performance. First, a unique, randomly drawn subset of the training data (a process known as bootstrapping) is used to train each individual tree. Second, when determining the optimal split at any given node, the model considers only a random subset of the available features (Breiman, 2001). By aggregating the forecasts from these decorrelated decision trees, the RF significantly reduces model variance while maintaining high predictive power (Lim et al., 2025).

The validation of the ARIMA and LSTM models was conducted using the Mean Absolute Percentage Error (MAPE). MAPE is a widely used metric for evaluating the accuracy of forecasting models, particularly in time series analysis, as it provides a clear understanding of the model's prediction performance in percentage terms. The primary advantage of MAPE is its simplicity and interpretability, making it an ideal choice for validating the models in this research. MAPE is calculated by comparing the difference between the predicted values and the actual values in the test dataset, divided by the actual values, and then averaged over all the data points (Kim & Kim, 2016).

$$MAPE= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i-F_i}{A_i}\right|\times 100 \qquad (9)$$

where is A_i is the observed PM$_{10}$ concentration at time i, F_i is the forecasted PM$_{10}$ concentration at time i, n is the total number of data points.

3.5 The Simulation and Analysis Procedures

The study began with a PM$_{10}$concentration dataset containing missing values classified as MNAR. Missing values were first imputed using KNN, the most effective method for time series data (Ahn et al., 2022) to obtain a complete series, which was then modeled using an ARIMA process to extract its parameters. These parameters were used to simulate a new complete dataset with the same temporal characteristics. Initially, KNN imputation was a necessary step to complete the dataset. Later in the simulation, the missing pattern was randomly generated. This approach can reduce the risk of overfitting for the KNN method because all missing values were re-randomized from a complete dataset. In the simulated air pollution dataset, PM$_{10}$ was selected as the primary variable for introducing artificial missingness. This decision was based on its relatively low missing rate in the original dataset and its clearly distinguishable temporal pattern compared with other variables. MNAR missingness was then simulated in PM$_{10}$ at varying proportions (10%, 30%, 50%, and 70%) to represent different levels of data degradation. The missingness mechanism was designed such that observations with higher PM$_{10}$

concentrations had a greater probability of being removed, reflecting realistic conditions in which extreme pollution values are more likely to be unrecorded due to sensor malfunction or automated data filtering. This probability-weighted deletion approach ensures that the simulated missingness aligns with the characteristics of Missing Not at Random (MNAR) behavior. The incomplete datasets were then imputed separately using KNN, MICE, and EM techniques. The imputation performance was evaluated using the Root Mean Square Error (RMSE).

As shown in **Figure 2**, The simulated PM_{10} series, containing missing data generated under the MNAR mechanism, was subsequently imputed using three distinct methods: KNN, MICE, and EM. To evaluate the predictive efficiency of the imputed data, each completed time series was then split into training and testing sets with an 80:20 ratio and subjected to three forecasting models which are the classical time series model ARIMA, LSTM and RF models. Model performance was evaluated using the Mean Absolute Percentage Error (MAPE) and repeated across 1,000 simulations to ensure result. consistency and robustness. Finally, the optimal imputation method and modeling process were applied to the original air pollution dataset to evaluate model performance under real-world conditions.

4. Results

The study utilized an air pollution dataset from Bangkok, Thailand, comprising 2,310 daily observations across 12 variables. All variables contained missing values at varying rates, as shown in **Figure 3**. The missingness pattern suggests that observations tend to be missing when the measured values are extremely low or high. This means the probability of missingness depends on the variable itself, which is characteristic of MNAR. However, identifying exact missing patterns requires more evidence to distinguish among the causes of MAR and MCAR.

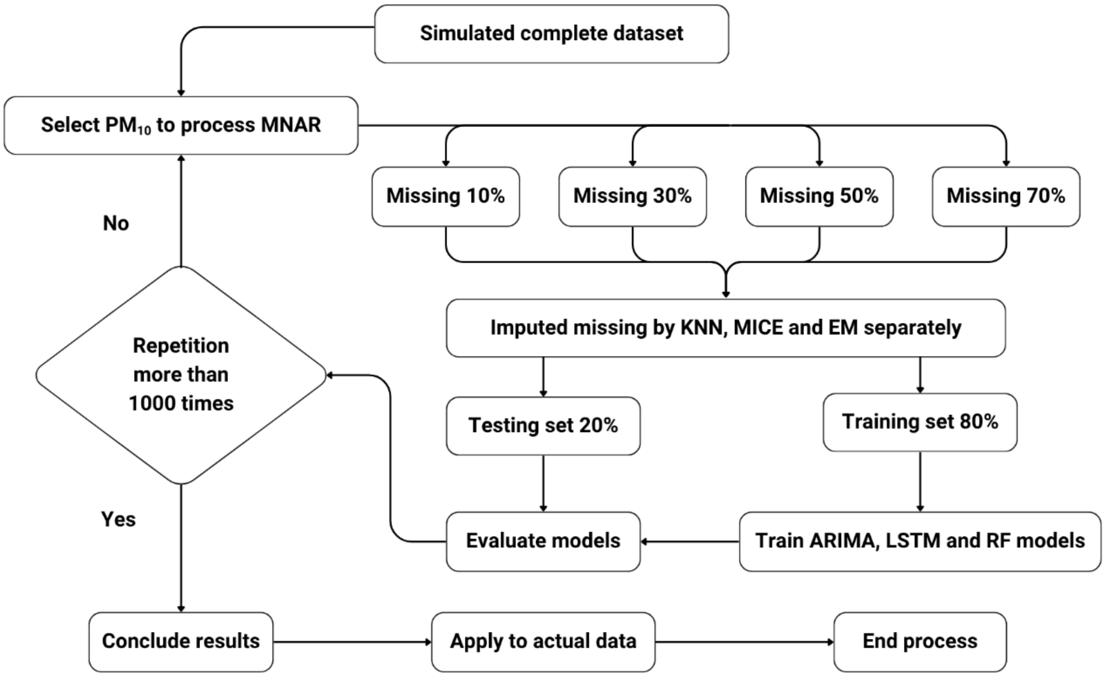
Simulated complete dataset
Select PM_{10} to process MNAR
Missing 10%
Missing 30%
Missing 50%
Missing 70%
Imputed missing by KNN, MICE and EM separately
Testing set 20%
Training set 80%
Train ARIMA, LSTM and RF models
Evaluate models
Repetition more than 1000 times
No
Yes
Conclude results
Apply to actual data
End process

Figure 2 Experiment process flowchart

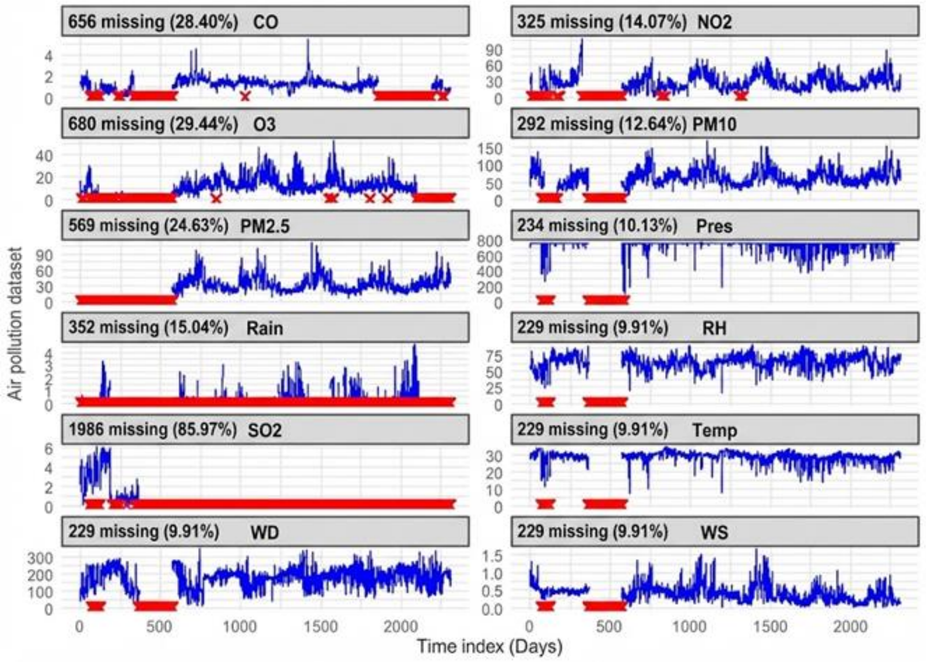

Figure 3 Time series plot of air pollution dataset (X' mark indicates missing positions)

Table 1 Summary of missing patterns in air pollution variables.

Variables	Min	Max	Average	Missing proportion	Missing types
CO (mg/m³)	0.02	5.51	1.24	0.28	MAR
NO_2 (µg/m³)	0.50	108.50	28.24	0.14	MNAR
O_3 (µg/m³)	0.67	52.63	13.07	0.29	MAR
SO_2 (µg/m³)	0.04	6.04	2.05	0.86	MAR
$PM_{2.5}$ (µg/m³)	7.13	111.96	32.97	0.25	MNAR
PM_{10} (µg/m³)	9.50	170.92	63.89	0.13	MNAR
Pres (mb)	94.38	765.42	739.81	0.10	MNAR
Rain (mm)	0.01	4.63	0.49	0.15	MCAR
RH (%)	17.17	91.46	64.85	0.10	MAR
Temp (°C)	7.13	34.87	29.09	0.10	MAR
WD (°)	11.25	349.42	179.32	0.10	MAR
WS (m/s)	0.02	1.67	0.40	0.10	MAR

The investigation of missingness patterns across all 12 air-pollution variables. **Table 1** shows the minimum and maximum values of each variable. The missing positions of NO_2 , $PM_{2.5}$, PM_{10}, and Pres were not related to those of the other variables, which is consistent with MNAR. Several variables also demonstrated characteristics consistent with MAR or even MCAR mechanisms. The actual air pollution dataset showed clear conditional relationships: when WS dropped below 0.25 m/s, SO_2 consistently became missing. 0_3 values began to disappear when temperature exceeded 31.16 °C. Missingness in CO occurred when WD was above 210.83° and Pres was more than 754.88 mb. When CO was above 1.38 ppm, simultaneous missingness occurred in WS, WD, RH, and temperature. Rain was the only variable whose missingness showed no statistical association with any other variable, suggesting a pattern closer to MCAR.

For the simulation pre-setup, the air pollution dataset was first imputed using the KNN method to create a complete dataset. An ARIMA model was then fitted to all variables in the dataset. The "auto.arima" function in the R program was used to select the proper model for each variable based on the lowest

Akaike Information Criterion (AIC), ensuring the best fit while balancing model complexity.

4.1 Simulation Study

The simulation process began with the original air pollution dataset, where missing values were first imputed to produce a fully complete dataset. An ARIMA model was then fitted to each variable individually to extract the corresponding model coefficients. These extracted parameters were used to generate a new completely simulated dataset that preserved the temporal characteristics of the original data. For the missingness experiment, only the PM_{10}variable in the simulated dataset was selected due to its clear temporal structure and relatively low missing rate (12.64%) in the real dataset, making it the most appropriate reference variable for time-series evaluation. To introduce MNAR missingness, a probability-weighted deletion strategy was applied. The probability of each PM_{10}value being removed was calculated by normalizing the absolute value of that observation over the sum of all PM_{10}values. Under this condition, higher concentration values were more likely to be removed, reflecting realistic MNAR behavior. Four missingness proportions were defined in the simulation: 10%, 30%, 50%, and 70%. The resulting missing distribution patterns are illustrated in **Figure 4**.

It illustrates the missing patterns applied to the simulated PM_{10}data across four missingness levels. The “X” markers indicate positions where values were removed. At 10% missingness, the overall time-series structure remains mostly visible. However, as the missing rate increases to 30%, 50%, and 70%, the PM_{10}pattern becomes progressively fragmented and difficult to interpret, demonstrating the increasing information loss under severe MNAR conditions. To address this issue, imputation techniques are required and will be applied to the incomplete simulated PM_{10} data for all missingness levels in the following sections.

The KNN imputation itself was carried out using the VIM: KNN function in R Studio, with the neighborhood size parameter explicitly set to k = 5. The KNN algorithm utilized all other available variables in the simulated dataset as predictors to find the 5 closest complete observations in the feature space. Once the missing PM_{10} values are estimated by averaging the neighbors' PM_{10} values, the points estimated through imputation were identified as shown in **Figure 5**.

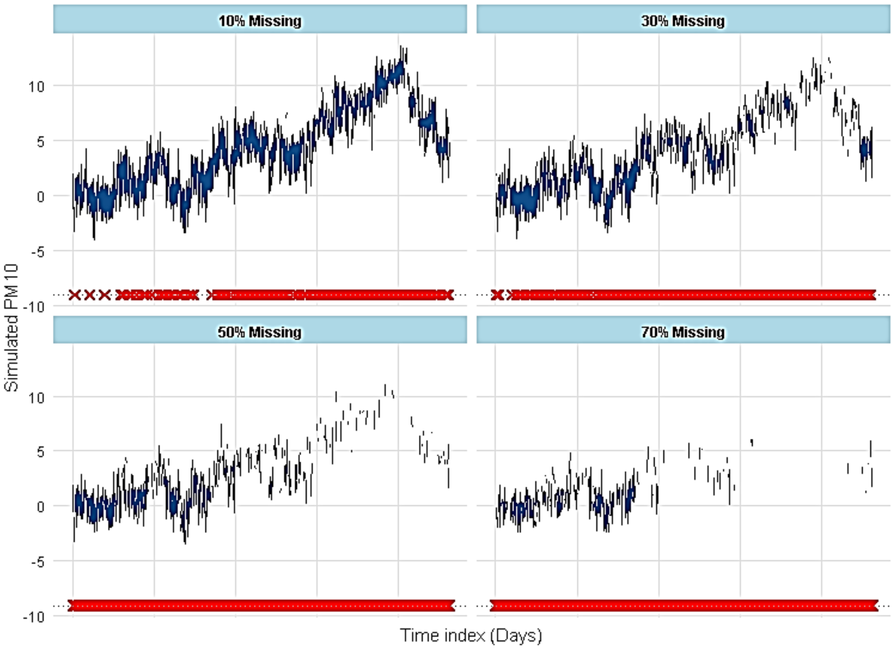

Figure 4 MNAR random missing on simulated PM_{10} data (X’ mark indicate missing positions)

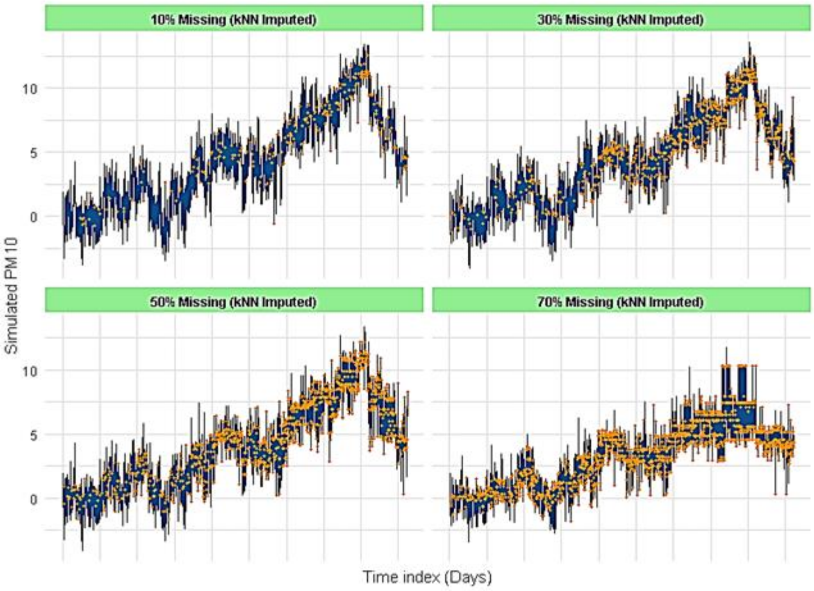

Figure 5 KNN imputation on simulated PM_{10} data (dot marks indicate impute points)

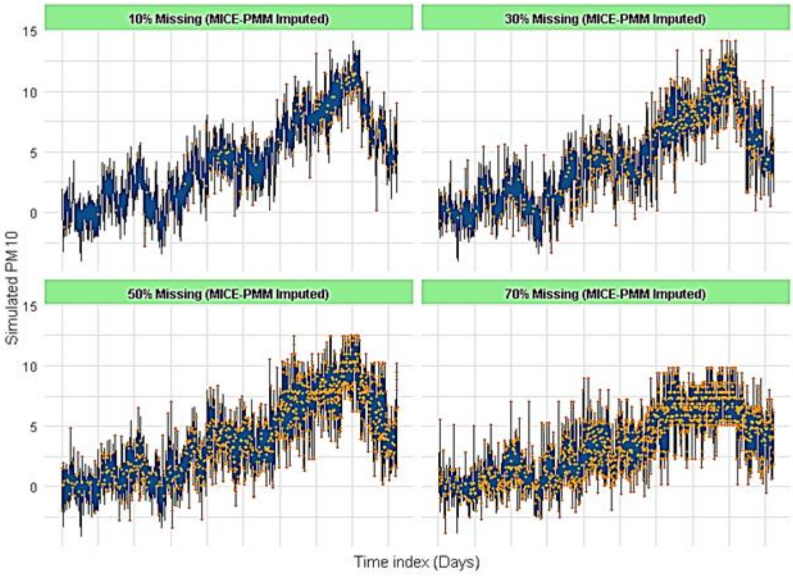

Figure 6 MICE imputation on simulated PM_{10} data (dot marks are impute points)

For MICE imputation utilizing the mice package, the imputation method parameter was explicitly set to PMM with 5 iterations and a maximum convergence limit of 50, allowing the model to impute missing PM_{10} values by regressing PM_{10} on all other complete variables and then matching the predicted value to the nearest observed PM10 value to generate an imputation, as shown in Figure 6.

Figure 7 reveals Expectation-Maximization (EM)-based technique, which is implemented in the R code using the mice package with the imputation method set to 'norm' (normal linear regression). This method serves as a statistically sound, package-consistent proxy for the EM algorithm's underlying principle, which assumes the data follow a multivariate normal distribution.

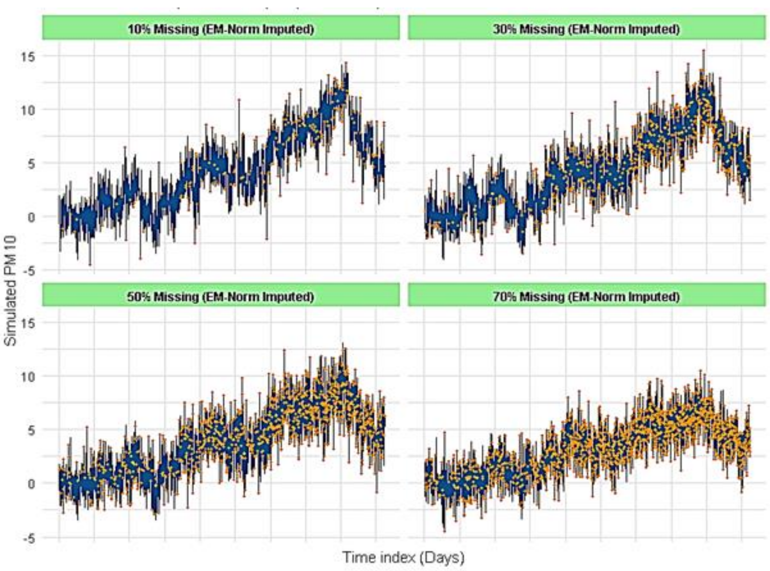

Figure 7 EM imputation on simulated PM_{10} data (dot marks indicate imputed points)

Table 2 Confidence intervals of RMSE (μg/m³) obtained from 1,000 simulations evaluating imputation techniques on simulated PM_{10} time-series data.

Imputation techniques	Missing rates			
	10%	**30%**	**50%**	**70%**
KNN	1.66 ± 0.02	1.75 ± 0.01	1.94 ± 0.01	2.68 ± 0.04
MICE	2.53 ± 0.02	2.59 ± 0.01	2.68 ± 0.01	3.02 ± 0.03
EM	2.56 ± 0.02	2.61 ± 0.01	2.72 ± 0.01	3.20 ± 0.03

Table 3 Cross-validation (CV) summary of LSTM hyperparameter tuning on simulated PM_{10} time-series data.

Lookback window	Units	Learning rate	Average RMSE
5	32	0.001 – 0.1	8.03 ± 0.001
5	64	0.001 – 0.1	8.03 ± 0.001
10	32	0.001 – 0.1	9.93 ± 0.001
10	64	0.001 – 0.1	9.93 ± 0.001
15	32	0.001 – 0.1	8.32 ± 0.001
15	64	0.001 – 0.1	8.33 ± 0.001

To ensure the robustness and reliability of the analysis, the missing data imputation process was repeated 1,000 times. Repetition reduces the impact of random variation that may arise from stochastic processes such as random missing value generation or model initialization. In the repeated simulation process, the complete dataset was preserved in each iteration, while only the positions of the missing values were randomly changed, maintaining the MNAR missingness condition. The imputation indices were calculated CI by analyzing the average of RMSE then ± the margin of error, as summarized in Table 2. The results indicate that the KNN method consistently outperformed the other two imputation techniques, achieving the lowest overall RMSE. Moreover, the CI of KNN aren't overlap to MICE and EM which explains that there were significant performance differences between KNN and the other techniques (Cumming & Finch, 2005).

Once the data had been imputed and completed, modeling served as a useful criterion for testing the efficiency and realism of the data. Later, two well-known time series models were introduced. The ARIMA setup involved generating the PM_{10} gold standard using a fixed ARIMA (1,1,2) in every one of the 1000 repetitions to ensure a statistically controlled ground truth. On the other hand, the LSTM model required appropriate parameter selection before modeling. The hyperparameter training setup for the LSTM model utilized Cross-Validation (CV) at k equal to 5 folds. The three critical hyperparameters being optimized were the Lookback Window (tested at 5, 10, and 15 timesteps), the number of LSTM Units (tested at 32 and 64), and the Learning Rate (tested on a logarithmic scale at 0.1, 0.01, and 0.001). The performance of each unique combination was quantified by the RMSE on the hold-out validation set of each fold.

The extensive hyperparameter search was quantified by the RMSE. The combination of a 5-timestep lookback window and 32 LSTM units consistently achieved the lowest overall mean RMSE. **Table 3** demonstrates the best generalization performance on unseen data. Furthermore, the evaluation across the logarithmic learning rate scale (0.1 to 0.001) revealed no statistically significant differences in performance. Therefore, based on these findings, the LSTM model setup for the KNN-based imputation simulation was fixed at 5 lookback steps, 32 units, and a 0.1 learning rate for all training runs. The RF models were configured with specific hyperparameters to ensure robust performance and stability. The number of trees was set at 500 to guarantee model convergence, and the number of features sampled at each split was set to 3. Model performance was measured by MAPE, and the process was repeated for 1000 repetitions.

The results in **Table 4** show that the KNN-based imputations yielded the strongest overall performance. Time series imputed using KNN consistently produced lower MAPE values in both the ARIMA and LSTM models. However, for the Random Forest (RF) model, EM and MICE imputations performed more efficiently than KNN. Regarding the forecasting models themselves, the ARIMA configuration demonstrated a slight overall advantage, outperforming the LSTM network, while the RF model showed the lowest efficiency among the three. Ultimately, the final simulation indicates that KNN-based imputations provide the most effective results when combined with the forecasting models.

Table 4 Average mean absolute percentage error (MAPE, %) obtained from 1,000 simulations of time-series models on simulated PM_{10} time-series data.

Models	**Missing rates**			
	10%	**30%**	**50%**	**70%**
KNN – ARIMA	2.30	2.32	2.37	2.45
KNN – LSTM	2.32	2.37	2.42	2.48
KNN – RF	72.97	54.64	38.26	26.87
MICE– ARIMA	4.17	4.18	4.17	4.17
MICE – LSTM	6.59	6.61	6.62	6.64
MICE – RF	31.53	30.36	34.36	65.16
EM – ARIMA	9.07	9.12	9.10	9.10
EM – LSTM	6.60	6.67	6.70	6.69
EM – RF	31.14	25.91	18.61	14.27

Table 5 Model performance indices by MAPE on the air pollution dataset

Variables	KNN			MICE			EM		
	ARIMA	**LSTM**	**RF**	**ARIMA**	**LSTM**	**RF**	**ARIMA**	**LSTM**	**RF**
CO	292.44	99.39	24.50	296.19	99.55	66.39	288.66	93.34	66.66
NO_2	52.46	46.60	27.72	52.83	46.66	28.91	53.46	46.58	28.86
O_3	36.41	25.26	23.76	34.83	24.43	41.04	35.67	24.01	40.94
SO_2	168.04	52.22	172.08	164.64	54.60	278.84	165.18	52.58	275.42
$PM_{2.5}$	31.28	26.67	10.99	32.18	26.69	11.62	32.04	27.13	11.52
PM_{10}	23.90	22.01	8.71	23.78	22.06	7.96	23.56	21.93	7.98
Pres	4.29	3.18	1.91	4.30	3.18	1.92	4.31	3.18	1.93
Rain	2527.51	909.53	1079.75	2609.85	863.61	672.97	2575.06	946.33	683.37
RH	13.84	12.02	5.37	13.82	11.95	5.62	14.07	12.16	5.57
Temp	13.70	8.08	2.80	13.08	8.06	2.95	13.51	8.00	2.94
WD	25.93	38.22	18.11	25.89	38.15	18.21	25.91	38.15	18.19
WS	2139.85	728.60	111.25	2228.88	693.73	122.07	2114.62	728.41	122.09

4.2 Application to Actual Data

The actual air pollution dataset comprised 12 features and 2,310 observations, with every variable exhibiting heterogeneous missingness rates. The core methodology involved first applying one of the three chosen imputation techniques (KNN, MICE, or EM) to the complete dataset, followed by separate time series modeling for each of the 12 variables. For the ARIMA framework, optimal model bounds were automatically selected based on the lowest AIC using the auto.arima function. Conversely, the LSTM network required an individualized CV process to train the best hyperparameters for each variable separately. The number of trees in the RF model was set at 500 for every variable.

Following the imputation of the actual air pollution data, a comparative forecast modeling exercise was conducted. As detailed in Table 5, the performance results for the four variables classified as MNAR, NO_2, $PM_{2.5}$, PM_{10}, Pres consistently validated the initial simulation findings. Specifically, the combined KNN based forecasting models (KNN-ARIMA, KNN-LSTM, KNN-RF) exhibited the strongest predictive performance for these four MNAR parameters. This high performance suggests that the KNN imputation methodology is highly effective at capturing the complex, non-random dependencies present in this missing data type. In contrast, the remaining variables, which were characterized by simpler MAR or MCAR patterns, showed notably poor forecasting accuracy when utilizing the same KNN-based methodology.

5. Discussion

Missing data constitutes a significant and pervasive obstacle in time series analysis, profoundly impacting the reliability of subsequent analytical findings. Among the mechanisms of data loss, MNAR is particularly crucial. It most accurately reflects the stochastic and systematic nature of real-world data collection in air pollution monitoring. Unlike MCAR or MAR, MNAR occurs when the probability of a value being missing is directly related to its own unobserved value. In environmental datasets, this mechanism commonly arises when sensors fail due to maintenance issues or when data collection is systematically interrupted during extreme measurement conditions, such as periods of unusually high pollution or adverse weather. These non-random data gaps introduce substantial bias into the dataset, which, if not appropriately addressed through imputation, compromises the accuracy and integrity of time series models.

To rigorously assess the impact of these biases, this study first constructed a simulated air pollution dataset in which missing data was deliberately introduced as MNAR at varying rates (10%, 30%, 50%, and 70%). The MNAR mechanism was specifically designed to detect missingness when simulated PM_{10} values were either extremely low or high. This ensured that the missing data pattern occurred precisely during conditions critical to the PM_{10} concentration itself. The imputation process across these four missing rates showed that the KNN method consistently outperformed both MICE and EM in recovering the true data, as measured by the lowest RMSE on the imputed points. This result is consistent with a previous study (Li et al., 2024).

Furthermore, to test how well the imputed data supported forecasting, the completed simulated dataset was split into training and test sets for subsequent modeling. The evaluation of the classic ARIMA model and the deep learning LSTM model confirmed the imputation results. The RF model showed different behavior from ARIMA and LSTM, potentially because RF is better suited to capturing short-term local patterns in the dataset rather than temporal sequences (Wang, 2024). Therefore, KNN-imputed series demonstrated the highest overall efficiency (lowest MAPE) compared with forecasts derived from MICE or EM imputation, reinforcing KNN's efficacy in handling high-volume MNAR gaps in a time series context. When the methodology was applied to the actual air-pollution dataset, the results revealed a clear distinction between variables exhibiting MNAR behavior and those characterized as MAR. For variables identified as MNAR (NO_2, $PM_{2.5}$, PM_{10}, and Pres), KNN-based models (KNN–ARIMA, KNN–LSTM, and KNN–RF) produced the best overall forecasting performance compared with MICE- and EM-based models. However, for variables whose missingness followed a MAR pattern (CO, O_3, SO_2, RH, Temp, WD, and WS), the performance trend shifted. In these cases, MICE and EM performed similarly and both outperformed KNN across the forecasting models. A few variables, such as Rain and WS, showed extremely high MAPE values, meaning poor model performance, which might be due to the variables having asymmetric distributions (Tayman & Swanson, 1999). Furthermore, the Random Forest (RF) model delivered the highest forecasting accuracy overall in the real dataset, followed by LSTM. An outcome that differs from the simulation results. This discrepancy likely reflects the greater complexity, noise, and inter-variable dependence present in real-world data compared with the controlled simulated environment. From a modeling perspective, the results are aligned with previous studies that highlight the strong predictive capability of RF models in environmental applications (Pan, 2024; Song et al., 2021). Regarding KNN, it is a non-parametric method that can identify the nearest data points and predict missing values, which makes it perform stably under any missingness type. In MNAR the missing data might be adjacent, which is difficult for other methods, but KNN can still detect the closest known value to make it work. The limitation of KNN is revealed when it deals with many features at once. Future work could further examine MAR mechanisms in environmental datasets and develop more targeted methods for diagnosing and handling mixed missingness structures.

6. Conclusion

This study developed a controlled simulation framework using an air pollution dataset to investigate the performance of three imputation techniques, KNN, MICE, and EM, under MNAR conditions, with $\mathbf{PM_{10}}$ selected as the primary variable due to its clear temporal patterns and relatively low missingness. Across the simulation experiments, KNN consistently produced the most accurate imputations, indicated by the lowest RMSE values, and yielded the strongest forecasting performance when paired with ARIMA, LSTM, and RF models, as measured by MAPE. When the methodology was applied to the real multi-variable air pollution dataset, KNN-based models continued to perform well for variables exhibiting MNAR behavior, aligning with the simulation findings, while variables driven by MAR or other mechanisms showed more variable performance across methods. Given that environmental and air-quality monitoring frequently encounters MNAR-related data gaps, the results underscore the importance of choosing imputation strategies that align with the underlying missingness mechanism. Overall, this study provides practical guidance for handling MNAR data in environmental applications and offers a foundation for future work exploring MAR and MCAR mechanisms, expanding imputation techniques, and refining forecasting strategies for complex real-world datasets.

7. Acknowledgements

This research was funded by Faculty of Science, Kasetsart University through the Graduate Fellowship Program and International SciKU Branding (ISB), Faculty of Science, Kasetsart University.

8. Abbreviations

Abbreviations	Full name
ARIMA	Autoregressive Integrated Moving Average
CO	Carbon Monoxide (mg/m³)
CV	Cross Validation
EM	Expectation–Maximization
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
MAPE	Mean Absolute Percentage Error
MAR	Missing at Random
MCAR	Missing Completely at Random
MICE	Multiple Imputation by Chained Equations
MNAR	Missing Not at Random
NO_2	Nitrogen Dioxide (µg/m³)
O_3	Ozone (µg/m³)
PM_{10}	Particulate Matter with size less than10 (µg/m³)
$PM_{2.5}$	Particulate Matter with size less than 2.5 (µg/m³)
Pres	Atmospheric pressure (mb)
Rain	Rainfall (mm)
RF	Random Forest
RH	Relative Humidity (%)
RMSE	Root Mean Square Error
SO_2	Sulfur Dioxide (µg/m³)
Temp	Temperature (°C)
WD	Wind Direction (°)
WS	Wind Speed (m/s)

9. CRediT Statement

Sattawat Saeyang: Writing – original draft, Supervision, Investigation, Resources, Software.
Khairil Anwar Notodiputro: Conceptualization, Methodology, Writing – review & editing.
Wandee Wanishsakpong: Conceptualization, Funding acquisition, Writing – review & editing.

10. References

Agiwal, V., & Chaudhuri, S. (2024). Methods and implications of addressing missing data in health-care research. *Current Medical Issues*, *22*(1), 60-62. https://doi.org/10.4103/cmi.cmi_121_23

Ahmad, M., Manjantrarat, T., Rattanawongsa, W., Muensri, P., Saenmuangchin, R., Klamchuen, A., ... & Panyametheekul, S. (2022). Chemical composition, sources, and health risk assessment of $PM_{2.5}$ and PM_{10} in urban sites of Bangkok, Thailand. *International Journal of Environmental Research and Public Health*, *19*(21), Article 14281. https://doi.org/10.3390/ijerph192114281

Ahn, H., Sun, K., & Kim, K. P. (2022). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials, & Continua*, *70*(1), 767-779. http://doi.org/10.32604/cmc.2022.019369

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 272-278.

Allison, P. D. (2009). Missing data. *The SAGE Handbook of Quantitative Methods in Psychology*, *23*, 72-89. https://doi.org/10.4135/9780857020994.n4

Alruhaymi, A. Z., & Kim, C. J. (2021). Why can multiple imputations and how (mice) algorithm work?. *Open Journal of Statistics*, *11*(5), 759-777. https://doi.org/10.4236/ojs.2021.115045

Alruhaymi, A. Z., Kim, C. J., Alruhaymi, A. Z., & Kim, C. J. (2021). Study on the missing data mechanisms and imputation methods. *Open Journal of Statistics*, *11*(4), 477-492. https://doi.org/10.4236/ojs.2021.114030

Ban, W., & Shen, L. (2022). PM2.5 prediction based on the CEEMDAN algorithm and a machine learning hybrid model. *Sustainability*, *14*(23), Article 16128. https://doi.org/10.3390/su142316128

Bartholomew, D. J. (1971). *Time series analysis forecasting and control*. *Journal of the Operational Research Society, 22*(2), 199–201. https://doi.org/10.1057/jors.1971.52

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cheewinsiriwat, P., Duangyiwa, C., Sukitpaneenit, M., & Stettler, M. E. (2022). Influence of land use and meteorological factors on $PM_{2.5}$ and PM_{10} concentrations in Bangkok, Thailand. *Sustainability*, *14*(9), Article 5367. https://doi.org/10.3390/su14095367

Chhabra, G. (2024). Handling Missing Data through Artificial Neural Network. *Communications on Applied Nonlinear Analysis*, *31*, 677-684. https://doi.org/10.52783/cana.v31.1429

Cumming, G., & Finch, S. (2005). Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170-180. https://psycnet.apa.org/doi/10.1037/0003-066X.60.2.170

Fadlil, A., & Herman, P. M. D. (2022). K nearest neighbor imputation performance on missing value data graduate user satisfaction. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *6*(4), 570-576. https://doi.org/10.29207/resti.v6i4.4173

Heymans, M. W., & Twisk, J. W. (2022). Handling missing data in clinical research. *Journal of Clinical Epidemiology*, *151*, 185-188. https://doi.org/10.1016/j.jclinepi.2022.08.016

Hosamane, S. N., Prashanth, K. S., & Virupakshi, A. S. (2020). Assessment and prediction of PM10 concentration using ARIMA. *Journal of Physics: Conference Series*, 1706(1), Article 012132. https://doi.org/10.1088/1742-6596/1706/1/012132

Hyndman, R. J. (2001). ARIMA processes. *Datajobs: Data Science Knowledge.* Retrieved from https://www.researchgate.net/publication/222105783_ARIMA_processes

Iamtrakul, P., Padon, A., & Chayphong, S. (2024). Quantifying the impact of urban growth on urban surface heat islands in the Bangkok Metropolitan Region, Thailand. *Atmosphere*, *15*(1), Article 100. https://doi.org/10.3390/atmos15010100

Joel, L. O., Doorsamy, W., & Paul, B. S. (2022). A review of missing data handling techniques for machine learning. *International Journal of Innovative Technology and Interdisciplinary Sciences*, *5*(3), 971-1005. https://doi.org/10.15157/IJITIS.2022.5.3.971-1005

Junger, W. L., & De Leon, A. P. (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

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402-406. https://doi.org/10.4097/kjae.2013.64.5.402

Karnati, H., Soma, A., Alam, A., & Kalaavathi, B. (2025). Comprehensive analysis of various imputation and forecasting models for predicting PM2.5 pollutant in Delhi. *Neural Computing and Applications*, *37*(17), 11441-11458. https://doi.org/10.1007/s00521-025-11047-2

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669-679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Laqueur, H. S., Shev, A. B., & Kagawa, R. M. (2022). SuperMICE: An ensemble machine learning approach to multiple imputation by chained equations. *American Journal of Epidemiology*, *191*(3), 516-525. https://doi.org/10.1093/aje/kwab271

Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., ... & Guo, H. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, *24*(1), Article 41. https://doi.org/10.1186/s12874-024-02173-x

Lim, A., Owusu, B. A., Thongrod, T., Khurram, H., Pongsiri, N., Ingviya, T., & Buya, S. (2025). Trend and association between particulate matters and meteorological factors: A prospect for prediction of $PM_{2.5}$ in Southern Thailand. *Polish Journal of Environmental Studies*, *34*(5), 5215-5223. https://doi.org/10.15244/pjoes/190787

Lin, T. H. (2010). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality & Quantity*, *44*(2), 277-287. https://doi.org/10.1007/s11135-008-9196-5

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(1), Article 75. https://doi.org/10.1186/1471-2288-14-75

Mukherjee, K., Gunsoy, N. B., Kristy, R. M., Cappelleri, J. C., Roydhouse, J., Stephenson, J. J., ... & Di Tanna, G. L. (2023). Handling missing data in health economics and outcomes research (HEOR): A systematic review and practical recommendations. *Pharmacoeconomics*, *41*(12), 1589-1601. https://doi.org/10.1007/s40273-023-01297-0

Nguyen, G. T. H., La, L. T., Hoang-Cong, H., & Le, A. H. (2024). An exploration of meteorological effects on $PM_{2.5}$ air quality in several provinces and cities in Vietnam. *Journal of Environmental Sciences*, *145*, 139-151. https://doi.org/10.1016/j.jes.2023.07.020

Nourmohammad, E., & Rashidi, Y. (2025). Ground data analysis for PM2.5 Prediction using predictive modeling techniques. *Journal of Air Pollution and Health*, *10*(1), 61-82. https://doi.org/10.18502/japh.v10i1.18095

Pan, X. (2024). The comparison between random forest and LSTM models based on the gold price prediction. *Advances in Economics Management and Political Sciences*, *94*(1), 102-108. https://doi.org/10.54254/2754-1169/94/2024ox0150

Pereira, R. C., Abreu, P. H., Rodrigues, P. P., & Figueiredo, M. A. (2024). Imputation of data missing not at random: Artificial generation and benchmark analysis. *Expert Systems with Applications*, *249*(B), Article 123654. https://doi.org/10.1016/j.eswa.2024.123654

Ponsawansong, P., Prapamontol, T., Rerkasem, K., Chantara, S., Tantrakarnapa, K., Kawichai, S., ... & Zhang, Y. (2023). Sources of $PM_{2.5}$ oxidative potential during haze and non-haze seasons in Chiang Mai, Thailand. *Aerosol and Air Quality Research*, *23*(10), Article 230030. https://doi.org/10.4209/aaqr.230030

Ranganathan, P., & Hunsberger, S. (2024). Handling missing data in research. *Perspectives in Clinical Research*, *15*(2), 99-101. https://doi.org/10.4103/picr.picr_38_24

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Song, J., Gao, Y., Yin, P., Li, Y., Li, Y., Zhang, J., ... & Pi, H. (2021). The random forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms. *Risk Management and Healthcare Policy*, *14*, 1175-1187. https://doi.org/10.2147/RMHP.S297838

Takale, G., Wattamwar, A., Saipatwar, S., Saindane, H., & Patil, T. B. (2024). Comparative analysis of LSTM, RNN, CNN and MLP machine learning algorithms for stock value prediction. *Journal of Firewall Software and Networking*, *2*(1), 1-10. https://doi.org/10.48001/jofsn.2024.211-10

Tayman, J., & Swanson, D. A. (1999). On the validity of MAPE as a measure of population forecast accuracy. *Population Research and Policy Review*, *18*(4), 299-322. https://doi.org/10.1023/A:1006166418051

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

Utama, A. B. P., Wibawa, A. P., Handayani, A. N., Irianto, W. S. G., & Nyoto, A. (2024). Improving time-series forecasting performance using imputation techniques in deep learning [Conference presentation]. 2*024 International Conference on Smart Computing, IoT and Machine Learning (SIML)*. IEEE, Surakarta, Indonesia. https://doi.org/10.1109/SIML61815.2024.10578273

Wang, W. (2024). Comparative analysis of ARIMA, random forest, and LSTM models for Mercedes-Benz stock price prediction. *Advances in Economics, Management and Political Sciences*, *134*(1), 1-8. https://doi.org/10.54254/2754-1169/2024.18718