



OPEN ACCESS

EDITED BY

Zhongfan Zhu,
Beijing Normal University, China

REVIEWED BY

Pradeep Kumar,
Manipal University Jaipur, India
Arti Choudhary,
Banaras Hindu University, India

*CORRESPONDENCE

Muhammad Asad Khan,
✉ muhammad.a.khan@strath.ac.uk
Ahmad Alsaber,
✉ aalsaber@auk.edu.kw

RECEIVED 05 January 2026

REVISED 29 January 2026

ACCEPTED 30 January 2026

PUBLISHED 19 February 2026

CITATION

Khan MA, Pan J, Alshatti A, Alsaber A and Gray A (2026) Development of a novel imputation framework for PM2.5 particle data in Pakistani cities using machine learning and statistical techniques. *Front. Environ. Sci.* 14:1775982. doi: 10.3389/fenvs.2026.1775982

COPYRIGHT

© 2026 Khan, Pan, Alshatti, Alsaber and Gray. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Development of a novel imputation framework for PM2.5 particle data in Pakistani cities using machine learning and statistical techniques

Muhammad Asad Khan^{1*}, Jiazhu Pan¹, Amani Alshatti², Ahmad Alsaber^{3*} and Alison Gray¹

¹Department of Mathematics and Statistics, University of Strathclyde, Glasgow, United Kingdom,

²Department of Health Sciences, Public Authority of Applied Education and Training (PAAET) College of Health Sciences, Safat, Kuwait, ³Management Department, College of Business and Economics, American University of Kuwait (AUK), Salmiya, Kuwait

Introduction: Missing PM2.5 observations in environmental monitoring systems, caused by sensor malfunctions, communication failures, maintenance issues, and coverage gaps, compromise public health assessments and evidence-based air quality policymaking. Reliable imputation strategies are therefore essential to preserve data integrity and analytical validity.

Methods: This study evaluated five imputation techniques: Bayesian Regression (BR), K-Nearest Neighbors (KNN), missForest, Predictive Mean Matching (PMM), and Random Forest (RF), using daily PM2.5 measurements collected between May 2019 and December 2024 from monitoring stations in Islamabad, Karachi, Lahore, and Peshawar, Pakistan. Three missing data mechanisms, MCAR, MAR, and MNAR, were simulated at missing rates ranging from 5% to 25%. Model performance was assessed using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Results: Imputation under the MAR mechanism consistently yielded lower error values as missingness increased. Across all mechanisms and missing rates, missForest and KNN demonstrated superior performance. Notably, missForest achieved the lowest RMSE and MAE values overall and effectively preserved the temporal structure, range, and variability of the PM2.5 series.

Discussion: The findings suggest that machine-learning-based approaches, particularly missForest, provide robust and reliable imputation for PM2.5 datasets with varying missingness patterns. These results support the use of missForest as a preferred method for handling incomplete air quality data in similar monitoring contexts, thereby strengthening the reliability of environmental health analyses and air quality policy development.

KEYWORDS

air quality monitoring, machine learning, missForest, Pakistan, PM2.5 missing data imputation

1 Introduction

In data-driven environmental health, incomplete monitoring records bias inference, reduce model accuracy, and impede policy decisions; addressing missing data through imputation and machine learning (ML) is essential for reliable PM2.5 particulate matter assessment and intervention planning (Jäger et al., 2021). Combining machine-learning

forecasts and statistical imputation utilizes multi-source inputs (meteorology, traffic, remote sensing) to reconstruct high-resolution PM_{2.5} time series, improving predictive performance while remaining computationally efficient for local and regional applications (Fan et al., 2023; Saheer et al., 2022). Benchmarking imputation approaches and adapting methods to the temporal and spatial structure of urban PM_{2.5} maximizes reconstruction fidelity and supports robust exposure–health analyses in Pakistani cities, thereby strengthening evidence for air quality and public health policy (Darji et al., 2024; Mendes et al., 2022).

Ambient air pollution represents one of the most significant environmental and public health challenges of the 21st century. Among the various pollutants, particulate matter with an aerodynamic diameter of size than 2.5 μm (PM_{2.5}) (Liang et al., 2016) which poses significant health risks and environmental challenges globally. Accurate PM_{2.5} data is essential for effective monitoring, policy-making, and public health interventions. In developing countries like Pakistan, rapid industrialization, urbanization, and increased vehicular emissions have led to a significant deterioration of air quality, especially in major urban centers like Lahore, Karachi, Peshawar and Islamabad. A major and predominant source of fine particles in the studied region is the vehicular activity, biomass fires and industry (Ngangmo et al., 2023; Mezoue et al., 2023). The country frequently ranks among those with the highest levels of PM₂ pollution globally, posing a severe threat to the economy and to the health of millions of its citizens leading to estimate an annual cost of 6.5% of GDP per year due to health cost, reduce productivity and agricultural degradation (Suleman, 2022). Recognizing this crisis, government has installed air quality monitoring stations across the country to generate time-series data that is fundamental for various research and regulatory purposes.

However, the data collected from these monitoring stations are notoriously prone to incompleteness. Missing values in time-series data presents an inherent challenge, often arising from a confluence of factors: instrument malfunctions (e.g., sensor drift, calibration periods), harsh environmental conditions, sensor failures, power outages, communication failures, and routine maintenance schedules (Hua et al., 2024; Sun et al., 2023; Arnaut et al., 2024). The presence of missing data creates a significant impediment for data scientists and environmental researchers, as most statistical models and machine learning algorithms require complete datasets for training and inference. The critical question, therefore, shifts from whether data is missing to how to handle its absence effectively.

(Wijesekara and Liyanage, 2023) describe three different approaches to dealing with missing data: data deletion, imputation, and predictive estimation. In univariate time-series data, where a single variable is studied across time, the handling of missing data is especially problematic for the purpose of maintaining temporal dependency structures within imputation. Missing values in time series, if not handled properly, may often lead to biased results, subdue statistical power, and distort findings (Wijesekara and Liyanage, 2021). Avoiding these gaps or using assumptions such as mean imputation generates datasets that do not adequately model reality, resulting in erroneous insights. Consequently, many researchers have developed sophisticated imputation techniques, each with their unique advantages and disadvantages.

The missing data in PM₁₀ time series has been shown to lead to underestimating pollution level exceedances, making the assessment and management of air quality more problematic (Albano et al., 2018). This poses major difficulties in fulfilling public health protection obligations outlined in European law. Failing to acknowledge missing data can result in inadequate sampling, faulty measurements and data collection problems (Junninen et al., 2004). Some of the literature reviews the principal techniques developed to accommodate missing values within the context of univariate time series data, assessing their effectiveness, theory, and application. Two common approaches to cope with missing data include a single imputation (SI) approach and a multiple imputation (MI) approach (Noor et al., 2015). It has been shown that the choice of imputation technique is determined by the attributes of the time series, the proportion of missing data, and the acceptable margin of error (Ribeiro and Castro, 2022). Single imputation methods are more straightforward and quicker to execute, yet multiple imputation and model-driven methods are advanced, trustworthy and precise. As noted by Chhabra (2023), the approach to deal with the missing values in a univariate time series is either simple imputation-based or model-based strategies. Imputation-based techniques involve direct estimation of missing values with mean, median or mode while methods that solve equations, based on likelihoods are model-based methods (Armina et al., 2017; Aljuaid and Sasi, 2016; Pereira et al., 2024).

To understand missing data in detail, it is necessary to acknowledge the various missing data mechanisms. These include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), all of which need custom approaches to formulate effective imputation strategies (Kleinke et al., 2020).

The primary objective of this study is to develop and evaluate a robust imputation framework for addressing missing values in PM_{2.5} air quality datasets using advanced statistical and machine learning methods. Thus, this paper provides the first systematic evaluation of missing data methods such as (BR), (KNN), missForest, (PMM), and (RF), with a focus on the multi-year PM_{2.5} data from four major Pakistani cities. By simulating realistic missing data mechanisms and measuring the accuracy of imputations through metrics like RMSE and MAE, the researcher provides a solid framework to combat the gaps in the gaps in sparse air quality monitoring networks. These findings will be useful to policymakers and researchers that rely on proper estimates of PM_{2.5} in health risk assessment and environmental planning.

2 Related works

Here we discuss how scholars try to mitigate the still overwhelming issue of high rates of missing data. Several methodological frameworks are discussed, highlighting both the advances and ongoing challenges of the problem. Some basic imputation methods, such as simple mean imputation and spline interpolation, often perform poorly in univariate settings due to their reliance on inter-attribute dependencies (Moritz et al., 2015). Likewise, Lien et al. (2023), mean imputation, in which missing data points were replaced with the mean of the available data, has been

TABLE 1 Descriptive statistics of pollutant PM_{2.5} concentrations (May 2019–December 2024) for the four studied Pakistani cities.

Pollutant	City			
	Islamabad	Karachi	Lahore	Peshawar
PM _{2.5}				
Valid observations	2,046	2,046	2,046	2,046
Mean	114.436	108.576	185.981	149.887
Median	107.000	95.000	162.000	142.000
Std. Deviation	43.486	42.103	87.787	58.558
Skewness	0.679	0.741	1.147	1.147
Range	288.000	268.000	575.000	551.000
Minimum	10.000	9.000	5.000	6.000
Maximum	298.000	277.000	580.000	557.000
No. of missing observations	152	245	272	137

shown to yield inadequate results in datasets with more than 10% missing data because of temporal variability in PM_{2.5} data. Niako et al. (2024) applied and compared multiple imputation methods, including Kalman filtering, linear interpolation, and moving averages, quantifying their effects on forecasting accuracy of the ARIMA and LSTM models. Advanced methods such as ARIMA state-space models with Kalman smoothing have been proven to be robust against shifts of means and heavy-tailed distributions (Zainuddin et al., 2022; Haile et al., 2024; Sharma et al., 2025) used ARIMA model to predict PM_{2.5} concentration in Indian satellite cities and the model exhibited a high level of accuracy (Kumar et al., 2024). Implemented the machine learning models by building intricate relationships with various metrological variables and considered the linear regression model as the most favorable methods to PM_{2.5} concentration. Some studies employed various predictive modelling techniques to predict PM_{2.5}, PM₁₀ and other leading pollutants. For instance, Researchers used machine learning and data mining algorithms including Independent component regression (ICR), ElasticNet (ENET), boosted tree (BT), Random Forest (RF), Support Vector Machine (SVM), Bagged Multivariate Adaptive Regression Splines (MARS), and Bayesian Regularized Neural Networks (BRNN) to predict the distribution of pollutants like PM_{2.5}, PM₁₀, NO₂, and SO₂ (Kumar P. et al., 2025; Choudhary et al., 2023).

Further to this Wijesekara and Liyanage (2020) found that Kalman smoothing on Structural Time Series showed strong performance compared to six other methods of imputation on air quality data. A study by Arnaut et al. (2024) investigated the application of the Random Forest (RF) algorithm for bi-directional imputation of missing values for air quality data, with emphasis on PM_{2.5} concentrations, and analyzed its performance compared to the rather simple approaches of imputation such as mean and median imputation.

A study by Tyagi et al. (2021) analyzed air quality data and investigated five imputation techniques: K-Nearest Neighbors (KNN), Linear Interpolation (LI), Expectation-Maximization (EM), Multiple Imputation by Chained Equations (MICE), and

Random Forest (RF). The outcomes suggest that Random Forest imputation is the optimal method for filling in absent data (Alsaber, Pan, and Al-Hurban, 2021). Used imputation methods for missing values in environmental data sets collected in Kuwait and evaluates the performance of missForest, K-Nearest Neighbors (KNN), and Bayesian principal component analysis (Bayesian PCA) in a systematic way. Results showed that missForest has the lowest imputation error for varying degrees of missing data. The R statistical software has several packages designed for missing data, including imputation. A prominent R package, mice (Multivariate Imputation by Chained Equations), while primarily designed for multivariate data, can also be applied to univariate time series by treating lagged values of the series as predictors (Buuren and Groothuis-Oudshoorn, 2011). Bayesian regression also offers a robust statistical approach for dealing with the missing data. According to Aßmann et al. (2023), a Bayesian regression statistical approach, which incorporates prior knowledge and manages uncertainty during the imputation process is very useful for dealing with missing data in air quality datasets.

Several studies highlight the role of evaluation metrics in comparing imputation methods. Rantou et al. (2017) advocate for using Mean Root Squared Error (MRSE) and Mean Absolute Percentage Error (MAPE) to assess imputation accuracy, particularly in predictive modelling contexts. The approach with the lowest expected mean square error (EMSE) for each type of missing data for six imputation techniques across various scenarios, intended to improve logistic regression models. Junninen et al. (2004) assesses different techniques for imputing missing values in air quality data sets, classifying them into univariate, multivariate, and hybrid methods, as well as multiple imputation approaches. The main objective was to evaluate their efficacy in addressing data shortages. The effectiveness of each imputation method was assessed using the Root Mean Square Error (RMSE) and the Normalized Root Mean Square Error (NRMSE), which compare predicted values to actual observed values (Tyagi et al., 2021). Lien et al. (2023) applied classical imputation methods and advanced approaches on a

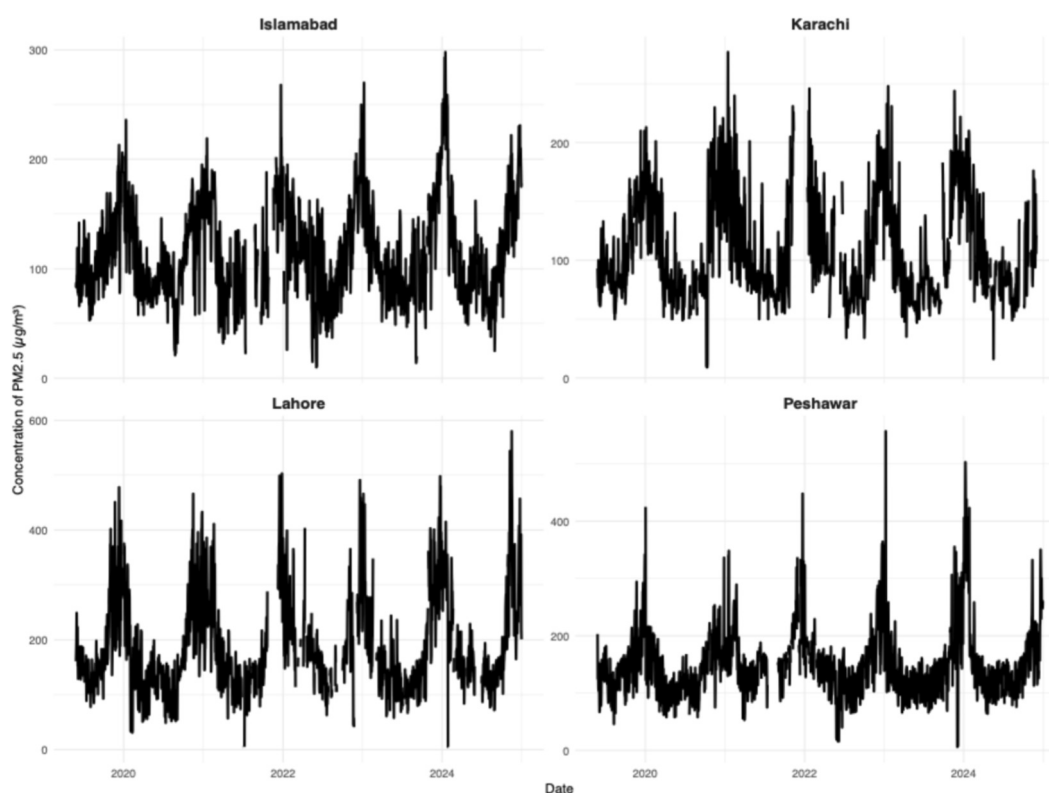


FIGURE 1
Time series plot of PM_{2.5} concentrations from May 2019 to December 2024, having missing values from four different monitoring stations in Pakistan.

crowd-sourced Fitbit data set, revealing that KNN performs best under low to moderate missingness (<30%).

This paper evaluates multiple imputation methodologies for PM_{2.5} air quality data, emphasizing their effectiveness in addressing different missing data scenarios. The analysis systematically compares established and emerging techniques. These include multiple imputation methods, as well as spectral and machine learning-based approaches. The study considers each method under a range of missing data mechanisms to inform the selection of optimal imputation strategies (Libasin et al., 2020; Zhang and Thorburn, 2022). The data set, from four monitoring stations or US consulates in Pakistan, has missing data, for various possible reasons. One reason is that there were many changes to the routine maintenance at the monitoring sites. Human error is a second reason. Thirdly, the PM_{2.5} data-sharing network needed to publish the information has been put on hold because of a lack of funds, resulting in missing data in the studied data set.

3 Objectives of the study

- To investigate the trend and percentage of missing in PM_{2.5} data in urban cities of Pakistan.
- To analysis the pattern of missingness through statistical and machine learning approaches.
- To compare and measure the precision of each imputation approach.

- To suggest the most appropriate imputation method for PM_{2.5} in the context of urban Pakistani data.

4 Methods

This section consists of data collection, analysis of missing data and implementation of various imputation methods.

4.1 Data description and missingness

In Pakistan air pollution data is collected through government monitoring stations, low-cost sensor networks, satellite remote sensing and research based intermittent sampling. Despite the fact that the Pakistan environmental protection agency (Pak-EPA) and provincial environmental protection Departments (EPDs) have relatively low volumes of such monitoring stations, the reliance on such monitoring stations and the density of such stations is relatively low, hence requiring a significant reliance on other modalities to provide the complete, real-time information, particularly with reference to PM_{2.5}. In this study the daily average PM_{2.5} concentration data sourced from four urban air quality monitoring stations situated in four major Pakistani cities: Lahore, Karachi, Islamabad, and Peshawar, collectively operated by federal government department (EPA) with provincial (EPA) and US consulate monitoring networks. The data were acquired from the publicly accessible Air Quality Index (AQI) platform (<https://www.>

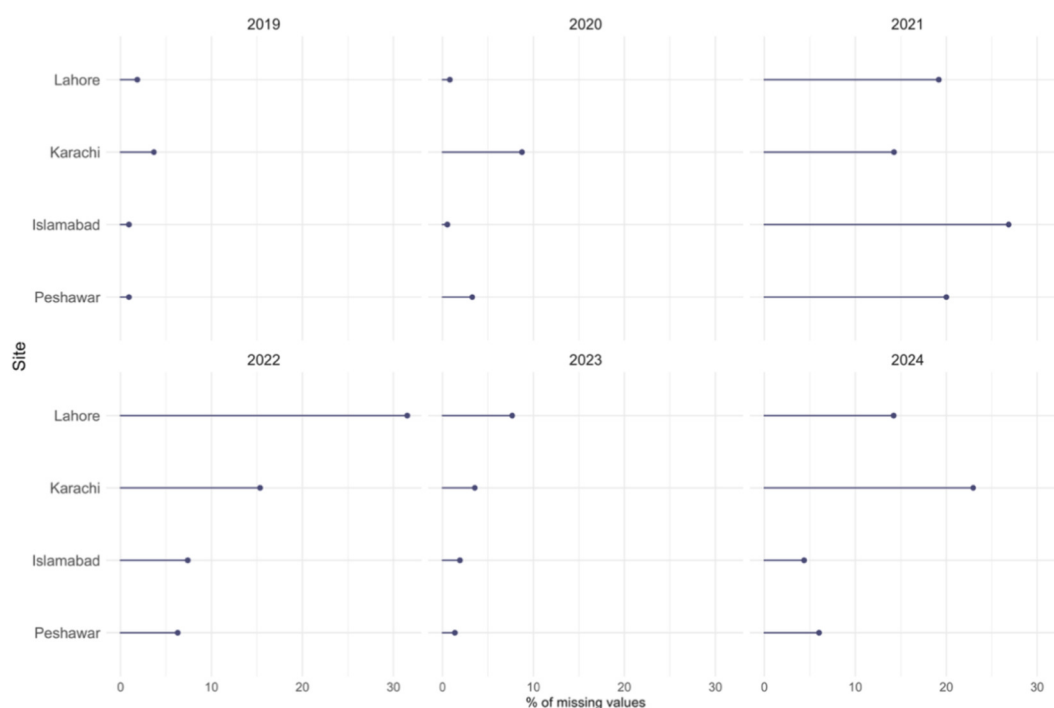


FIGURE 2 Missing values of PM2.5 per year (May 2019 to December 2024).

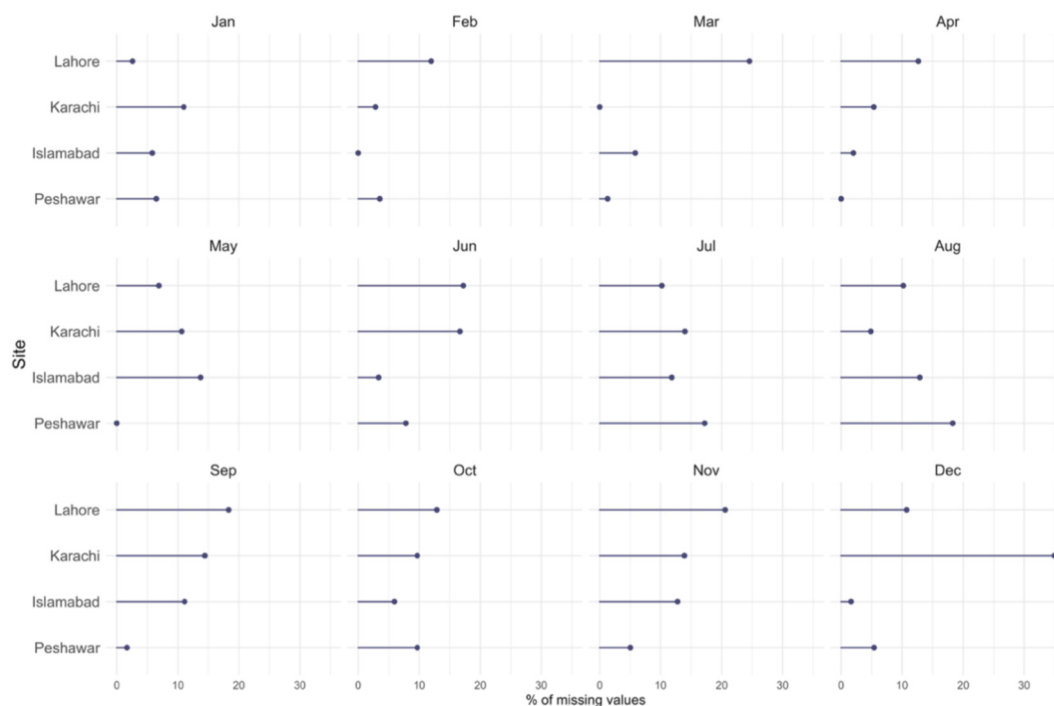
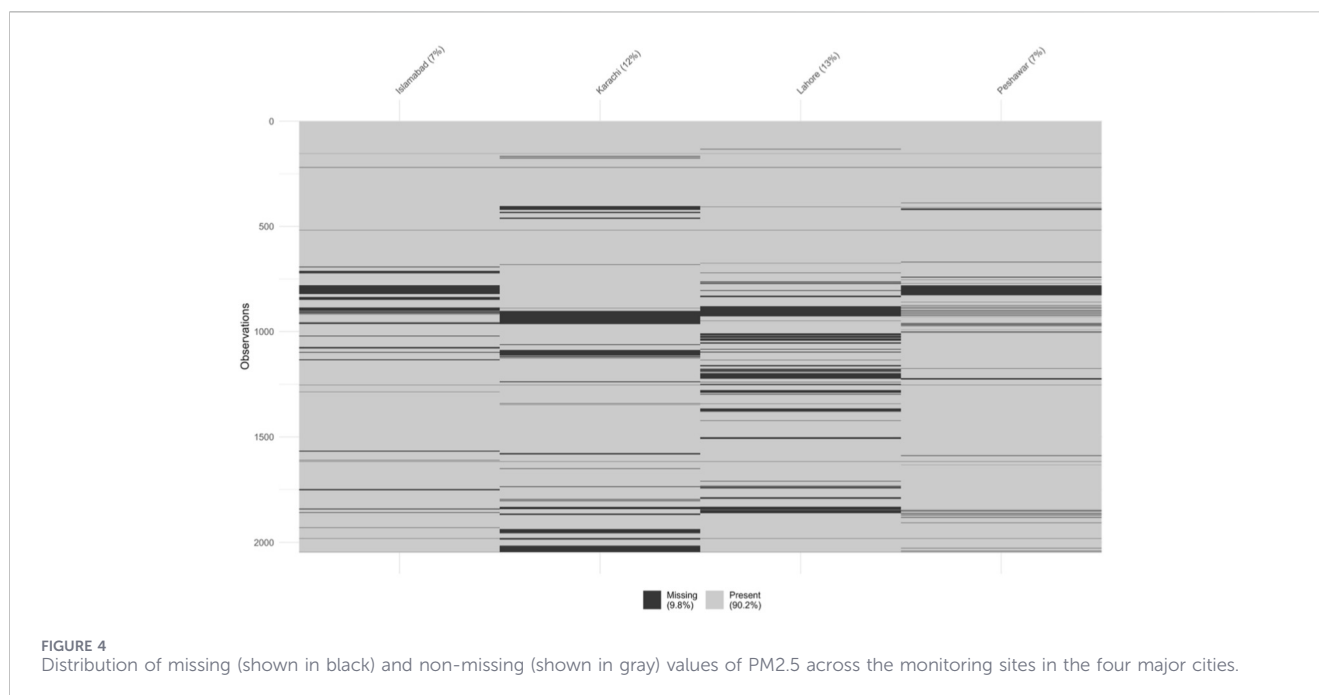


FIGURE 3 Missing values of PM2.5 per month (May 2019 to December 2024).



<https://aqicn.org/map/pakistan/>) which compiles real-time sensor. The data set encompasses almost a 6-year period from 27 May 2019, to 31 December 2024, comprising over 2,046 daily average observations per station, assuming complete temporal coverage. Each record includes a timestamp (local time), PM_{2.5} concentration values (in $\mu\text{g}/\text{m}^3$), and metadata such as station ID and location coordinates. Supplementary meteorological parameters temperature, humidity, and wind speed were also recorded from <https://www.visualcrossing.com/weather-data/> for multivariate analysis but does not have any missing values. Before analysis, data set for each city was cleaned and quality checked. Duplicate in records were removed and incorrect time stamps were corrected to maintain the chronological consistency of the data set. Preliminary summary analysis and time series plots were used to ensure the seasonal variation and data integrity. These processing steps were necessary for accurate imputation analysis.

Table 1 shows the descriptive statistics of the PM_{2.5} concentrations across the four major Pakistani cities, which exhibited substantial spatial and statistical variability, reflecting differing pollution dynamics and urban environmental pressures. Lahore recorded the highest mean concentration ($186 \mu\text{g}/\text{m}^3$) and the widest range ($575 \mu\text{g}/\text{m}^3$), with a pronounced right-skewed distribution (skewness = 1.147), indicating frequent extreme pollution episodes. Peshawar followed with a mean of $149.9 \mu\text{g}/\text{m}^3$ and the highest skewness (1.704), suggesting even more asymmetrical pollution patterns. In contrast, Islamabad and Karachi showed relatively lower mean values ($114.4 \mu\text{g}/\text{m}^3$ and $108.6 \mu\text{g}/\text{m}^3$, respectively) and moderate skewness, pointing to more stable but still elevated pollution levels. Overall, the mean concentration was higher than the median. Standard deviations ranged from $42.1 \mu\text{g}/\text{m}^3$ in Karachi to $87.79 \mu\text{g}/\text{m}^3$ in Lahore, underscoring the heterogeneity in pollutant dispersion.

Out of a total of 2,046 daily observations, a substantial percentage, approximately 39.4%, of the data were missing, with the highest incidence of missing entries reported in Lahore (272 records) and Karachi (245 records). Such data gaps have implications for the accuracy and reliability of longitudinal trend analyses and necessitate robust imputation strategies, as explored in subsequent sections.

Figure 1 illustrates the time series of the PM_{2.5} concentrations from 2019 to 2024, including seasonal variations and the completeness of the data from the four urban monitoring stations. Some of the more notable patterns of the data spikes above the trend line are in the winter months, which may be the result of emissions resulting from temperature inversion, and other associated emissions. The seasonal variability is high; during the post-monsoon season and winter, high concentrations and the incidence of the poor air quality can be observed, which can be explained by the burning of stubble, the increase of anthropogenic emissions, and specific meteorological circumstances (Kumar et al., 2025a). It is important to note that seasonal and spatial heterogeneity is observed in the distribution of air pollution with PM_{2.5}. Five recorded to occur at the highest levels in winter in some Indian urban cities (Kumar et al., 2025b).

The discontinuities in the series signal periods of the captured data that are often to be found in clumps. Figure 2 shows the volume of missing PM_{2.5} data recorded by year and it is seen that a greater percentage of data is missing during the years of 2021 and onward, and the greatest percentage of missing data is for the cities of Lahore and Karachi. Figure 3 breaks the information down further by month and indicates that the greatest loss of data happens during the months of the summer season, from June to August, and the first months of the winter season, in November and December. Figure 4 presents a simplified view of the available data in which colours gray and black are used, where gray stands for the available data and black

for missing data. Overall, the missing data only constitute 9.8% of the entire data set. The missing data is not uniformly distributed, and it is more concentrated in the data from the cities of Lahore and Karachi. In contrast, the data from the cities of Islamabad and Peshawar have a more even distribution.

4.2 Imputation methods

4.2.1 Bayesian regression

Bayesian regression imputation puts the missing-data imputation problem within a Bayesian framework. Prior distributions are applied to the model parameters after assuming a regression model for a variable with missing values conditional on other variables. Next, using the observed data, the posterior distribution of the parameters and missing values is obtained. Assume the linear regression model having missing data.

$$Y = X\beta + \varepsilon, \varepsilon \sim N(0, \delta^2 I)$$

where Y is $n \times 1$ vector of the observed and missing values of the target variable, X is the matrix of predictors (observed or imputed before), β represents the regression coefficients, and δ^2 denotes the residual variance. The response Y and the predictor matrix X are partitioned into observed and missing components as

$$Y = \begin{pmatrix} Y_{obs} \\ Y_{mis} \end{pmatrix}, X = \begin{pmatrix} X_{obs} \\ X_{mis} \end{pmatrix}$$

As in the case of observed data, the likelihood function can be expressed as

$$p(Y_{obs}|X_{obs}, \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(Y_{obs} - X_{obs}\beta)^T(Y_{obs} - X_{obs}\beta)\right)$$

The model parameters are then determined as prior distributions, usually with conjugate priors:

$$\beta \sim N(\beta_0, \Sigma_0), \sigma^2 \sim \text{Inverse-Gamma}(a_0, b_0)$$

The joint distribution of likelihood and priors yields the joint posterior distribution:

$$p(\beta, \delta^2|Y_{obs}, X_{obs}) \propto p(Y_{obs}|X_{obs}, \beta, \delta^2) \cdot p(\beta|\delta^2) \cdot p(\delta^2)$$

Based on this, the posterior parameters distributions can be obtained as follows:

$$\beta|\delta^2, Y_{obs}, X_{obs} \sim N(\beta_n, \Sigma_n)$$

where

$$\Sigma_n = \left(\Sigma_0^{-1} + \frac{1}{\delta^2} X_{obs}^T X_{obs}\right)^{-1}, \beta_n = \Sigma_n \left(\Sigma_0^{-1} \beta_0 + \frac{1}{\delta^2} X_{obs}^T Y_{obs}\right)$$

The posterior distribution for the residual variance δ^2 is given by

$$\delta^2|Y_{obs}, X_{obs} \sim \text{Inverse-Gamma}(a_n, b_n)$$

with

$$\begin{aligned} a_n &= a_0 + \frac{n_{obs}}{2}, b_n \\ &= b_0 + \frac{1}{2}(Y_{obs} - X_{obs}\beta)^T(Y_{obs} - X_{obs}\beta) + \frac{1}{2}(\beta - \beta_0)^T \Sigma_0^{-1}(\beta - \beta_0) \end{aligned}$$

TABLE 2 Imputation performance of BR, KNN, missForest, PMM, and RF under MCAR, MAR, and MNAR mechanisms for PM2.5 data from Islamabad (May 2019 – December 2024).

Islamabad						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
5% missingness rate						
BR	6.9459	4.4185	2.9759	0.7660	0.3688	0.2541
KNN	4.7392	4.2329	2.8011	0.7349	0.3810	0.2505
missForest	4.2914	3.8774	2.6892	1.0311	0.3355	0.2461
PMM	6.0434	5.6554	4.2209	0.9765	0.5217	0.3941
RF	6.2198	3.8022	3.5838	1.1698	0.3547	0.2940
10% missingness rate						
BR	10.5227	5.4818	5.9691	2.6693	0.6393	0.8236
KNN	7.8141	4.7298	4.7332	1.9206	0.5487	0.5507
missForest	7.3325	4.2569	3.5428	1.8384	0.4721	0.4793
PMM	10.9324	5.7756	5.7082	2.6225	0.6756	0.7391
RF	10.1362	7.1249	7.6545	2.4417	0.8247	0.9434
15% missingness rate						
BR	13.1217	8.8642	7.3151	4.0718	1.2294	1.0944
KNN	10.8521	7.5404	6.2323	3.0283	1.0380	0.9641
missForest	10.1198	7.0496	5.3759	2.7971	0.9837	0.8639
PMM	14.1481	9.6377	8.3045	4.0973	1.4086	1.2298
RF	13.3478	8.0975	8.6110	3.7191	1.1836	1.2077
20% missingness rate						
BR	16.2800	7.8382	10.2836	5.4275	1.4104	1.8043
KNN	13.0344	6.5182	7.7775	4.3299	1.1028	1.3368
missForest	12.1552	6.0432	7.6423	3.9848	1.0232	1.2814
PMM	17.0911	7.5663	10.7591	5.7978	1.3554	1.8668
RF	16.2007	7.7744	9.0320	5.2650	1.3368	1.5431
25% missingness rate						
BR	18.3932	9.7033	10.4893	7.0511	1.9052	2.0177
KNN	15.5873	8.5390	7.6341	5.7405	1.6798	1.4141
missForest	14.5335	7.1138	8.7437	5.2957	1.4006	1.6884
PMM	18.5346	9.4831	12.5926	6.9013	1.9476	2.3768
RF	18.5208	10.2792	11.4883	6.8847	2.0048	2.1470

The posterior distribution once obtained, the missing values for the conditional posterior distribution are imputed. As

$$Y_{mis}|X_{mis}, \beta, \delta^2 \sim N(X_{mis}\beta, \delta^2 I)$$

This process is repeated M times, and each time new values of parameter $\beta^{(m)}$ and $\delta^{2(m)}$ are sampled according to their respective posterior distributions, both generating multiple imputations.

TABLE 3 Imputation performance of BR, KNN, missForest, PMM, and RF under MCAR, MAR, and MNAR mechanisms for PM2.5 data from Karachi (May 2019 – December 2024).

Karachi						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
5% missingness rate						
BR	11.3127	7.0097	5.3521	1.9791	0.6712	0.4623
KNN	7.7584	4.4153	5.8571	1.2802	0.3699	0.5059
missForest	6.8086	4.8447	4.8891	1.1308	0.4232	0.4030
PMM	10.1285	7.7706	5.7794	1.7081	0.7067	0.4790
RF	7.7644	6.0794	6.9224	1.3154	0.5128	0.5638
10% missingness rate						
BR	16.0114	7.2181	8.1367	4.1237	0.8513	1.0355
KNN	12.3929	6.9778	5.8515	2.9586	0.8647	0.7329
missForest	10.3800	6.4588	6.1914	2.4900	0.7603	0.7656
PMM	13.9272	9.9383	7.4271	3.1946	1.2422	0.9096
RF	14.6627	7.4785	8.9750	3.6184	0.8834	1.0614
15% missingness rate						
BR	17.9959	9.5874	11.3960	5.3865	1.4883	1.8134
KNN	14.1040	9.6130	9.7382	4.0559	1.4983	1.5480
missForest	13.6361	8.7902	8.8851	4.0320	1.3439	1.3593
PMM	16.6428	11.1686	11.4660	4.8820	1.7239	1.6501
RF	18.2083	10.1163	12.5136	5.3975	1.5383	1.8910
20% missingness rate						
BR	20.9656	12.5430	14.9225	7.5508	2.2805	2.7069
KNN	16.2445	10.5427	11.9095	5.1774	1.9703	2.1132
missForest	15.0468	9.3889	11.1508	5.0122	1.7235	1.9517
PMM	20.2340	12.2224	13.3238	6.8903	2.1415	2.4389
RF	19.6015	10.5827	13.0786	6.5335	1.8516	2.3658
25% missingness rate						
BR	23.5024	14.2354	13.9488	8.9869	3.0380	2.8647
KNN	18.9021	13.2710	13.0453	7.1463	2.7564	2.6522
missForest	17.6292	12.2496	11.9908	6.6553	2.4354	2.4850
PMM	24.5560	14.6149	14.9728	9.3561	2.9731	3.1201
RF	22.1069	15.1012	14.1526	8.5100	2.9331	2.8309

TABLE 4 Imputation performance of BR, KNN, missForest, PMM, and RF under MCAR, MAR, and MNAR mechanisms for PM2.5 data from Lahore (May 2019 – December 2024).

Lahore						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
5% missingness rate						
BR	18.5552	9.0194	10.0925	3.1451	0.7154	0.9324
KNN	11.2296	4.6633	8.6900	1.7757	0.4369	0.8199
missForest	11.3943	4.1519	8.6085	1.7225	0.3594	0.7884
PMM	18.9745	10.6164	11.9413	2.8399	0.7909	1.0738
RF	14.0462	7.4573	11.6312	2.3533	0.6963	1.0538
10% missingness rate						
BR	28.1091	17.7309	12.6096	6.8083	1.9197	1.5397
KNN	20.3444	12.4769	10.3735	4.5162	1.2698	1.2105
missForest	18.9415	12.8041	10.2677	4.0208	1.2937	1.2859
PMM	27.1513	15.0401	15.3854	5.8999	1.6763	1.7088
RF	27.1839	18.7879	16.2533	5.9986	1.9655	2.0407
15% missingness rate						
BR	32.1213	18.3188	24.4118	9.6584	2.5032	3.5942
KNN	27.0447	17.9144	18.0595	7.2643	2.0849	2.6253
missForest	25.6894	15.8219	18.0701	7.0947	1.8859	2.5703
PMM	31.1574	16.1361	23.9724	8.4679	2.0428	3.4741
RF	32.3940	22.8152	22.9764	8.9641	3.0545	3.4824
20% missingness rate						
BR	36.0795	18.4557	24.2248	12.5334	2.9019	4.2682
KNN	26.6868	13.8071	18.1267	8.8261	2.0683	3.2905
missForest	24.8515	16.3134	18.7313	8.2427	2.3614	3.3073
PMM	32.3884	18.0584	26.7171	10.9255	2.7888	4.8247
RF	34.7324	22.2056	26.2488	11.4934	3.1077	4.5114
25% missingness rate						
BR	42.1727	18.1403	27.4497	16.3832	2.5914	5.5036
KNN	32.9418	15.2492	20.8084	11.5562	2.4167	3.9593
missForest	33.6900	13.7558	22.0600	11.7553	3.6135	4.2723
PMM	43.1453	21.8732	30.3646	15.3009	3.5583	5.9593
RF	44.0573	23.1512	30.6094	15.6708	3.0048	5.6211

To generate multiple imputations, this process is repeated M times, each time drawing new parameter from their respective posterior distributions and the resulting multiple imputations are as follows.

$$Y_{mis}^m \sim N(X_{mis}\beta^m, \delta^{2(m)}I)$$

4.2.2 KNN

The KNN imputation approach relies on the k -nearest neighbours algorithm which is based on calculation of the pairwise distances between observations to select the k nearest most similar records. The main idea of this approach is that the

TABLE 5 Imputation performance of BR, KNN, missForest, PMM, and RF under MCAR, MAR, and MNAR mechanisms for PM2.5 data from Peshawar (May 2019 – December 2024).

Peshawar						
Method	RMSE			MAE		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR
5% missingness rate						
BR	11.2081	7.6705	8.0021	1.9966	0.5647	0.7226
KNN	7.9243	5.9647	5.9912	1.1739	0.4334	0.5528
missForest	6.9446	5.4055	5.3786	1.1072	0.4185	0.5155
PMM	12.1558	9.0391	6.7441	2.0656	0.7315	0.5321
RF	11.4600	7.3479	8.2715	1.7088	0.5576	0.6915
10% missingness rate						
BR	16.4352	11.3951	9.4832	3.9666	1.1088	1.0723
KNN	13.2767	9.4284	9.5037	2.8502	0.8854	1.0373
missForest	13.1888	8.4275	9.9840	2.7981	0.7615	1.2020
PMM	16.4851	13.6884	14.0755	3.6508	1.2733	1.5652
RF	19.9262	10.7153	10.2943	3.7840	1.0455	1.2402
15% missingness rate						
BR	22.9630	10.6287	11.9480	6.8402	1.3216	1.8298
KNN	15.6158	8.6463	10.2389	4.4507	1.1215	1.5141
missForest	15.1271	7.6521	9.4985	4.1924	0.9912	1.4417
PMM	19.6347	13.6666	15.2880	5.6253	1.7129	2.0952
RF	21.6763	13.5531	14.0440	5.3575	1.5342	1.8958
20% missingness rate						
BR	22.9622	13.5042	17.1455	7.8244	2.0311	2.8959
KNN	17.0087	10.3588	16.6286	5.5155	1.4831	2.6625
missForest	18.2668	9.8270	15.3086	5.5439	1.3722	2.4490
PMM	20.7053	12.4517	18.1341	6.8792	1.7219	3.0952
RF	21.2274	15.5859	17.4297	6.6266	2.2533	2.8820
25% missingness rate						
BR	26.9482	15.0989	22.1357	10.2556	2.3193	4.0627
KNN	21.7375	15.9218	19.2263	7.5155	2.2070	3.2616
missForest	20.3368	13.2474	19.7537	7.0825	1.8272	3.3543
PMM	29.8099	15.6488	22.3431	10.5790	2.3623	3.9924
RF	29.0618	15.5781	24.7987	9.8489	2.4658	4.1201

missing datum can be estimated by examining the closest neighbours which are closest to the target datum. The similarity, or proximity, between observations is measured using a number of distance measures; the most common distance measure when using continuous variables is the Euclidean distance, which is defined as follows.

$$d(i, j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

where,

i = the observation on which the distance is to be calculated, which is usually the observation with missing data.

j = Another observation that is going to be compared, typically one with fully observed data.

$k \in \{1, 2, \dots, p\}$, where, p is the total number of features/attributes.

4.2.3 Random forest

Random Forest was first introduced by Breiman (2001), is ensemble learning algorithms which builds a series decision trees, each of which is trained on a bootstrap sample of the observed data. The approach consider each variable with missing values as a dependent variable and uses the rest of the variables as predictors to estimate the missing values. The resulting final imputed value is determined by adding up predictions across all trees-using the mean with continuous variables or majority vote with categorical variables. In regression tasks, the final prediction is derived by

$$\hat{Y}_i = \frac{1}{B} \sum_{b=1}^B T_b(x_i),$$

where,

\hat{Y}_i = imputed value for observation i

B = number of trees in the forest

$T_b(x_i)$ = prediction from tree b based on the observed predictor values x_i for observation i .

4.2.4 MissForest

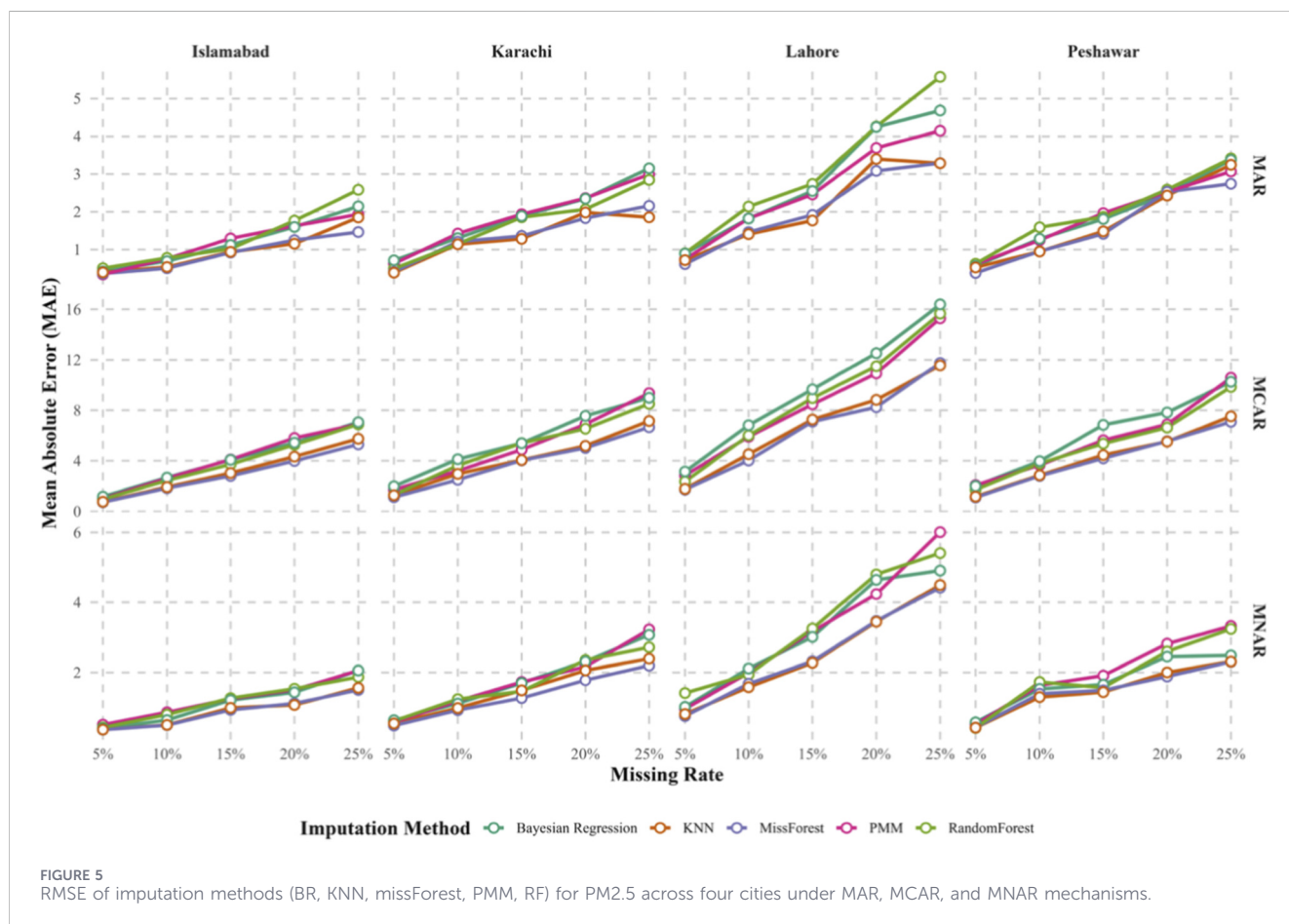
Stekhoven and Bühlmann (2012) introduced an iterative imputation method known as MissForest based on the Random Forest algorithm that is specifically designed to handle incomplete data both in continuous and categorical variables. In each iteration and for every variable X_s with missing values ($s = 1, \dots, p$), the data is partitioned into four subsets: 1. the observed values of X_s , denoted $y_{obs}^{(s)}$; 2. the missing values of X_s , denoted $y_{mis}^{(s)}$; 3. the corresponding observed values of the remaining variables, $x_{obs}^{(s)}$; and 4. the corresponding missing values of the remaining variables, $x_{mis}^{(s)}$. A Random Forest model $f_s(\cdot)$ is then trained on the observed data pairs $(x_{obs}^{(s)}, y_{obs}^{(s)})$, obtaining an estimator \hat{f}_s . This trained model is used subsequently to make predictions of the missing values as $\hat{y}_{mis}^{(s)} = \hat{f}_s(x_{mis}^{(s)})$, and the imputed entries are updated accordingly in matrix X . The process is repeated until the convergence criterion for all variables are met.

4.2.5 Predictive mean matching (PMM)

Predictive Mean Matching (PMM) is being commonly used statistical method for imputing missing data, particularly in the context of multiple imputations procedures. The approach was first proposed by Donald B. Rubin and R. J. A. Little in the late 1980s

TABLE 6 Consise summary of multiple imputation approaches.

Methods	RMSE/MAE accuracy	Stability as missingness rates	Sensitivity to missing mechanism	Performance across locations
BR	Moderate to higher errors; competitive with RF and PMM in number of cases	Stability declines at higher missingness	Highly sensitive to MCAR	Performance varies across all locations
KNN	Comparable to missForest in some cases	Stability is reasonable	Stronger to some cases of MCAR, MNAR and MAR	Good performance across all locations in some cases even results better than missForest
missForest	Lowest/near lowest errors in most cases	Higher stability	Mostly robust across MCAR, MNAR and MAR	Mostly strong and consistent across all locations
PMM	Moderate to high errors; overlaps with BR and RF	Lower stability	Sensitive to MNAR and MCAR	Performance varies across all locations
RF	Moderate accuracy and is competitive with BR and PMM depending on situation	Lower stability	Highly sensitive to MCAR, MNAR and MAR	Performance varies across all locations

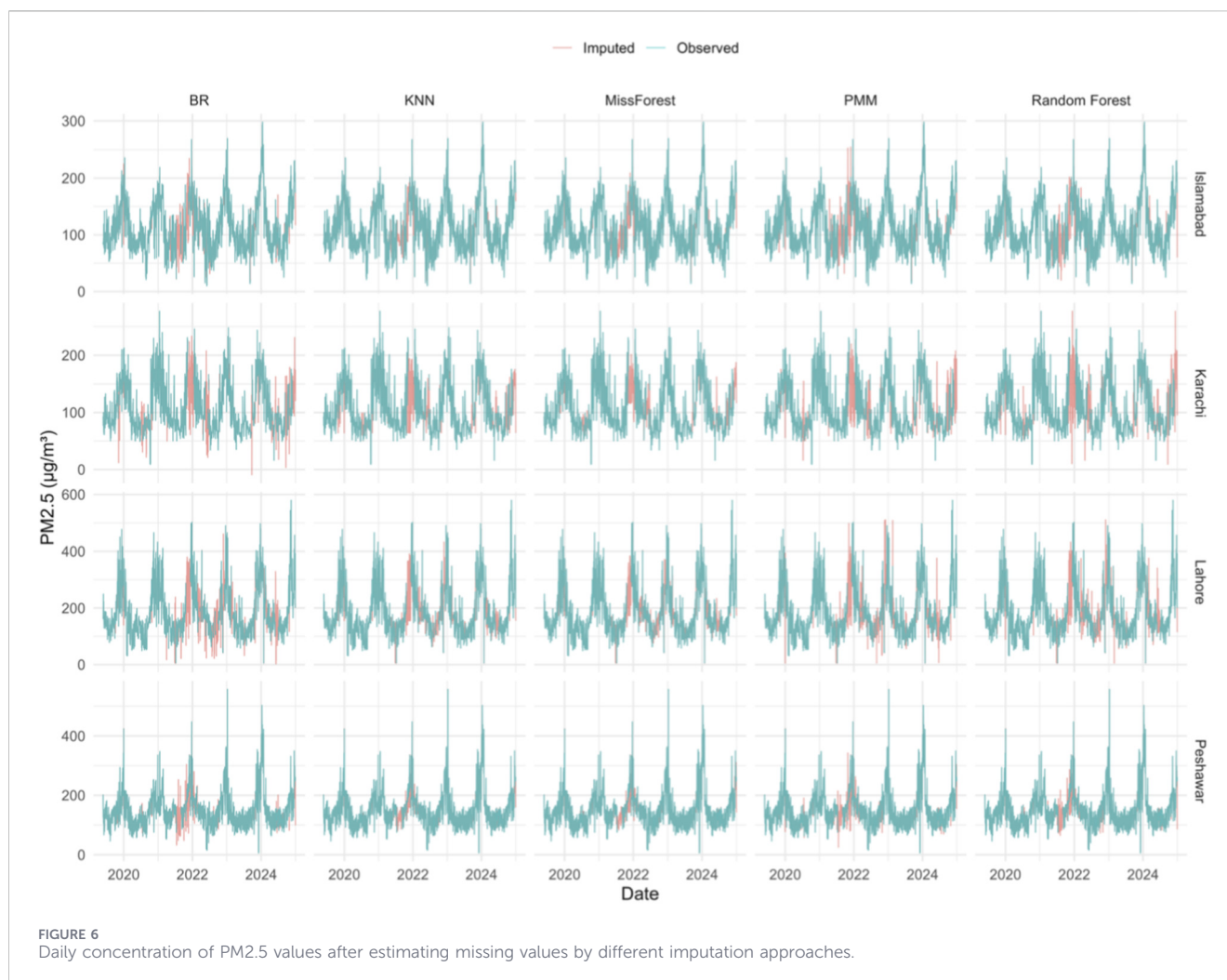


(Sugden and Rubin, 1988; Little, 1988). PMM of its ability to produce plausible imputed values through a use of available data, therefore maintaining the integrity of the original data. PMM works by Fitting regression models $\hat{Y}_{mis,i} = X\hat{\beta}$, the predicted value for those with Y_i missing and $\hat{Y}_{obs,j} = X\hat{\beta}$, the predicted value for those with Y_j observed. For each missing case i , an absolute distance metric $d(i, j) = |\hat{Y}_{mis,i} - \hat{Y}_{obs,j}|$ is computed. One donor is Y_j^* randomly selected from a set of $d(i, j) = \operatorname{argmin} |\hat{Y}_{mis,i} - \hat{Y}_{obs,j}|$ and is used to fill in the missing value. The process is repeated

until the predetermined criteria for the number of iterations is defined. A mice package in R is implemented and serves as default imputation for continuous variables (Buuren and Groothuis-Oudshoorn, 2011).

4.2.6 Results and discussion

This section discusses the empirical results across cities, interprets the implications of RMSE and MAE variations, and



highlights the conditions under which specific algorithms, particularly missForest and KNN, exhibit superior robustness and predictive fidelity.

Tables 2–5 report the performance of the five imputation methods, namely, Bayesian Regression (BR), K-Nearest Neighbors (KNN), missForest, Predictive Mean Matching (PMM), and Random Forest (RF) applied to the PM2.5 data of the four studied major cities under varying missingness mechanisms (MCAR, MAR, MNAR) and missingness rates 5%–25%. Performances were evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as accuracy metrics, for which lower values indicate superior imputation accuracy.

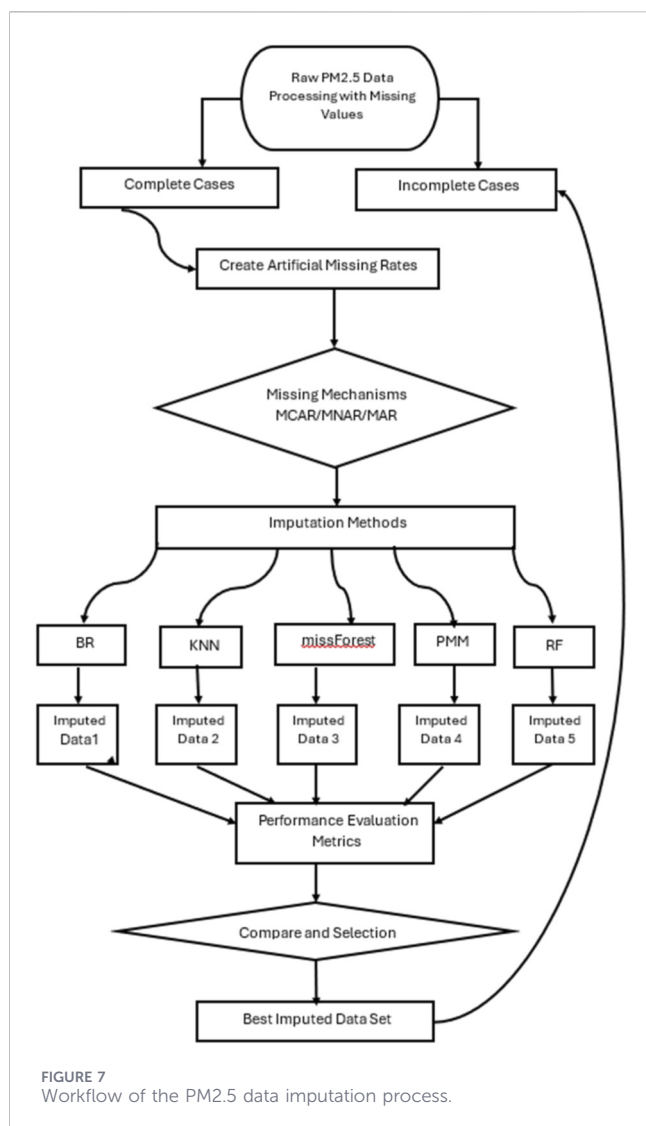
Consistently, both RMSE and MAE increased monotonically with the missingness rate across all cities and methods, reflecting the deteriorating reliability of imputations. Errors were generally lowest under MAR, higher under MNAR, and highest under MCAR for each mechanism. Of all methods employed, missForest was the most consistent and reliable across the cities, followed by KNN. In contrast, BR and PMM had low robustness and RF had intermediate performance and instabilities at higher missingness levels.

The results for Islamabad demonstrated the lowest overall error magnitudes across all methods and mechanisms. At low missingness

(5%–10%), missForest achieved the best performance (RMSE \approx 2.7–7.3, MAE \approx 0.24–1.83). KNN is also able to perform well, particularly under MAR and MNAR, though not quite as accurately as missForest.

As missingness increased to 20%–25%, errors rose sharply for all methods, but missForest and KNN remained comparatively stable while BR, PMM, and RF deteriorated substantially, suggesting that the Islamabad dataset is less prone to extreme variation, making it relatively easier to impute.

Karachi displayed higher error levels than Islamabad across all rates, particularly under BR and PMM. At 5% missingness, missForest again outperformed others. At low missingness (5%), missForest produces the most accurate imputations (RMSE = 4.89–6.81; MAE = 0.40–1.13), closely followed by KNN (RMSE = 4.42–7.76; MAE = 0.37–1.28) while RF performs moderately well under MCAR, and BR and PMM lag with larger errors. At 10%–15% missing rates, overall errors increase; although, missForest outperforms its counterparts and KNN did reasonably well. In contrast, RF, BR, and PMM deteriorate much more notably, particularly under MCAR, whereas RF, BR, and PMM deteriorate, especially under MCAR. With higher missingness (20%–25%) rates, missForest and KNN continue to provide relatively robust imputations, while



BR, PMM, and RF declines sharply, with RMSE values above 20 and MAE values exceeding 2.5.

Lahore consistently demonstrates the greatest imputation challenge, with markedly higher error magnitudes across all methods even at low missingness rates (RMSE > 10 for several approaches) under MCAR and severe deterioration at 10%–25%, where even missForest and KNN record RMSE values above 30 and MAE exceeding 3.0; these results underscore the city's volatile and irregular PM2.5 dynamics. Peshawar exhibits moderate levels of difficulty. Under all three mechanisms and at the 5% missingness rate, missForest and KNN again outperform other approaches, yielding lowest errors. Conversely, BR, RF, and PMM demonstrate markedly poorer performance which is contrast to the study conducted by Choudhary et al. (2023), stated that RF is more effective model for PM2.5 prediction and is superior to other data mining algorithms.

Mechanism-specific comparisons confirm that MAR consistently yields the lowest errors, whereas MCAR produces the highest errors across cities, reflecting the difficulty of imputing values tied to unobserved processes. Overall, the cross-city analysis highlights that Islamabad is the least challenging environment for imputation of air quality data, Peshawar and Karachi present moderate difficulty, and

Lahore poses the most severe challenge, while across all locations, non-parametric ensemble and neighbors-based methods (missForest and KNN) are far more effective approaches for recovering incomplete PM2.5 time series. These findings corroborate with the previous research that demonstrates the effectiveness of different imputation approaches for environmental data with high rates of missing values. For instance, the Kuwait air-quality data analysis by Alsaber et al. (2021) and Ritthewa and Samart (2024) stated that missForest demonstrated the lowest RMSE and MAE in various conditions of missing data. Similarly Umar and Gray (2023), in a different environmental context, who also found that missForest and KNN performed strongly in comparison of imputation approaches using RMSE and MAE for evaluation of imputed time series water level data. A comparative summary Table 6 explicitly compares the performance of various imputation methods concisely.

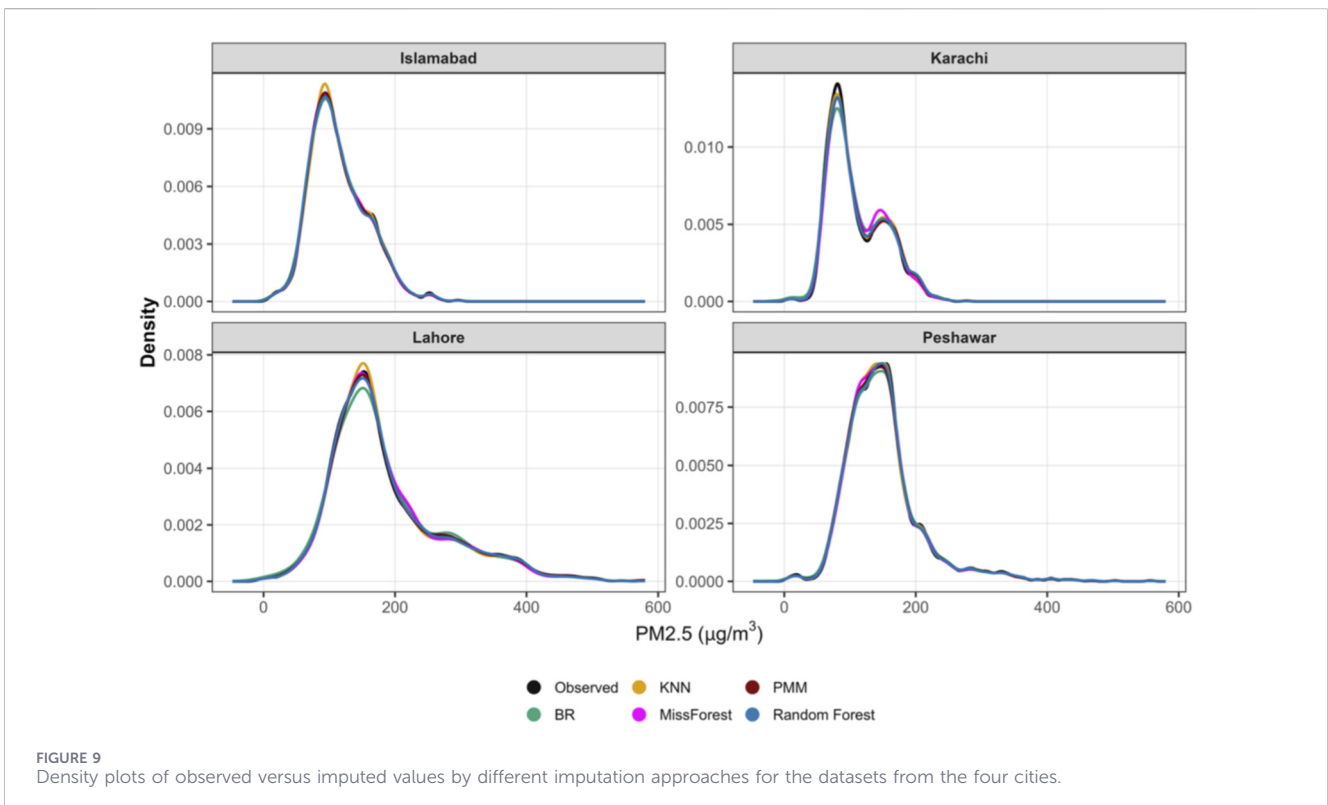
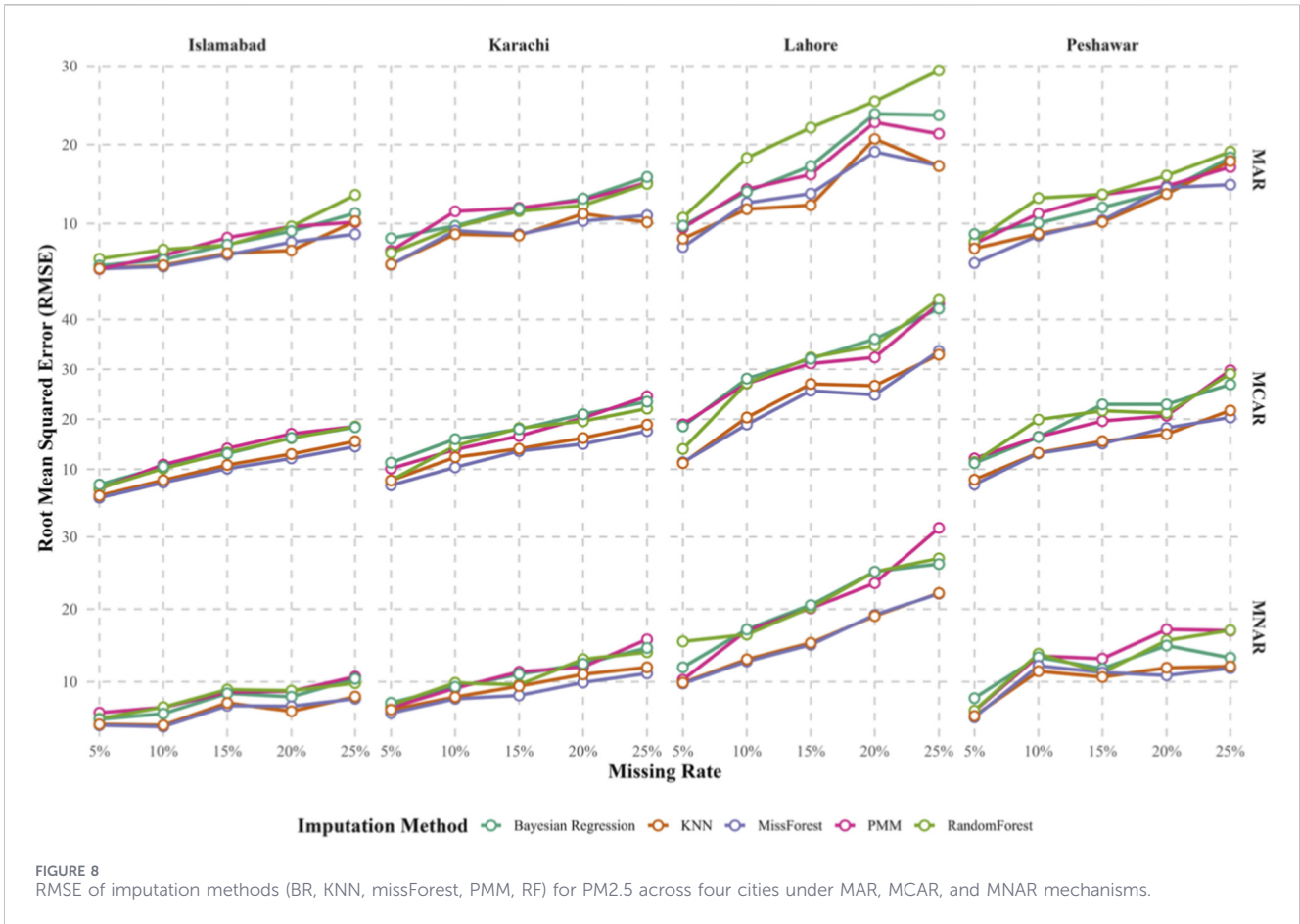
To evaluate the reliability of different imputation methods, missing values were imputed into the original datasets and then the alignment with observed PM2.5 concentrations was assessed. Figure 5 shows that gaps in missing data were effectively filled in a manner consistent with historical patterns, indicating that advanced approaches can provide robust estimates. Across Islamabad, Karachi, Lahore, and Peshawar (2019–2024), all methods preserved seasonal cycles and winter peaks, but their ability to capture variability differed and KNN and missForest performed well by closely tracking observed fluctuations. Similarly, the density distributions of observed versus imputed values of PM2.5 for the different methods of imputation shown in Figures 6, 7, illustrate that the values from all approaches are in broad correspondence to the observed data, validating their capabilities for estimating the missing values of PM2.5 concentration. Based on both visual inspection and error metrics (RMSE and MAE; Tables 2–5 and Figures 8, 9), missForest emerged as the most accurate method, slightly more accurate than KNN, making missForest the most suitable technique for imputing daily PM2.5 concentrations in these environmental time series.

5 Limitations and future work

Although this study provides valuable insights into the performance of various imputation techniques for PM2.5 data, certain limitations should be acknowledged. The analysis focused primarily on daily averages from four urban monitoring stations, which may not capture short-term variability or spatial heterogeneity in air quality. Additionally, only five imputation methods were evaluated, excluding emerging deep learning-based approaches that may further enhance accuracy. Future research should extend this framework to incorporate high-frequency and multi-pollutant datasets, explore hybrid or ensemble deep learning models, and examine the influence of meteorological and socioeconomic factors to develop more adaptive and scalable imputation strategies for environmental monitoring systems.

6 Conclusion

The results of this study provide critical insights into the comparative performance of statistical and machine learning-based imputation techniques applied to PM2.5 air quality datasets across major Pakistani cities. By systematically evaluating five



methods—Bayesian Regression (BR), K-Nearest Neighbors (KNN), missForest, Predictive Mean Matching (PMM), and Random Forest (RF) under varying missingness mechanisms (MCAR, MAR, MNAR) and rates (5%–25%), the analysis reveals distinct patterns of imputation accuracy and reliability. The findings underscore the sensitivity of model performance to both the volume and mechanism of missing data, demonstrating that improper handling of missingness can lead to substantial estimation errors and distort temporal pollution trends. The study used data from four major cities of Pakistan from May 2019 to December 2024, which allowed for an in-depth analysis of the various missing data mechanisms (MCAR, MAR, MNAR) and missing data rates (5%–25%) considered. The results indicate that as the amount of missing data increases, the quality of the data also diminishes. Moreover, data that is classified as Missing at Random (MAR) was found to have relatively less imputation error than MCAR (Missing Completely at Random) and MNAR (Missing Not at Random) because of the inherent temporal or spatial correlations within the datasets. Among all the methods, missForest had the lowest error overall, in terms of RMSE and MAE measures, and KNN also had relatively good results. These two techniques maintained the trends, cycles, and extremes of the time series data that are necessary for the robust analysis of environmental data. On the other hand, the Bayesian Regression, PMM technique and Random Forest technique, although more stable, lacked adequate reliability especially at higher levels of missingness. These examples confirmed the importance of customizing imputation strategies for the missingness mechanism and the attributes of the data set. In particular, imputation methods have not, to our knowledge, been studied for PM_{2.5} data from Pakistan and the use of missForest is recommended for imputation of missing PM_{2.5} data in that context.

In short, this study demonstrates that ensemble missForest and neighbor-based methods, the missForest, is the most effective machine learning approach for partial PM_{2.5} data set imputation relatively to KNN. Accurate imputations are necessary for improving the statistical credibility of environmental assessments and to ensure effective policy formulation for air quality management.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: All datasets used in this study are publicly available through established online air-quality and meteorological data platforms.

References

Albano, G., La Rocca, M., and Perna, C. (2018). “On the imputation of missing values in univariate \$PM_{10}\$ P M 10 time series,” in *Computer aided systems theory – EUROCAST 2017*, 12–19. *Lecture notes in computer science* (Cham: Springer International Publishing).

Aljuaid, T., and Sasi, S. (2016). “Proper imputation techniques for missing values in data sets,” in *2016 international conference on data science and engineering (ICDSE)* (IEEE), 1–5.

Alsaber, A. R., Pan, J., and 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). *Int. J. Environ. Res. Public Health* 18 (3), 1333. doi:10.3390/ijerph18031333

Author contributions

MK: Methodology, Resources, Writing – original draft, Software, Investigation, Visualization, Formal Analysis, Data curation, Validation, Conceptualization, Writing – review and editing. JP: Project administration, Validation, Supervision, Writing – review and editing, Visualization, Investigation. AmA: Investigation, Writing – review and editing, Validation. AhA: Writing – review and editing, Methodology, Software, Formal Analysis, Validation. AG: Supervision, Writing – review and editing, Project administration, Validation.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Armina, R., Zain, A. M., Ali, N. A., and Sallehuddin, R. (2017). A review on missing value estimation using imputation algorithm. *J. Phys. Conf. Ser.* 892 (September), 012004. doi:10.1088/1742-6596/892/1/012004

Arnaut, F., Đurđević, V., Kolarski, A., Srećković, V. A., and Jevremović, S. (2024). Improving air quality data reliability through Bi-Directional univariate imputation with the random forest algorithm. *Sustainability* 16 (17), 7629. doi:10.3390/su16177629

Aßmann, C., Gaasch, J.-C., and Stingl, D. (2023). A bayesian approach towards missing covariate data in multilevel latent regression models. *Psychometrika* 88 (4), 1495–1528. doi:10.1007/s11336-022-09888-0

- Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi:10.1023/a:1010933404324
- Buuren, S., and Groothuis-Oudshoorn, K. (2011). MICE: multivariate imputation by chained equations in R. *J. Stat. Softw.* 45 (December), 1–67. doi:10.18637/jss.v045.i03
- Chhabra, G. (2023). Comparison of imputation methods for univariate time series. *Int. J. Recent Innovation Trends Comput. Commun.* 11 (2s), 286–292. doi:10.17762/ijritcc.v11i2s.6148
- Choudhary, A., Kumar, P., Pradhan, C., Sahu, S. K., Chaudhary, S. K., Joshi, P. K., et al. (2023). Evaluating air quality and criteria pollutants prediction disparities by data mining along a stretch of urban-rural agglomeration includes coal-mine belts and thermal power plants. *Front. Environ. Sci.* 11 (November), 1132159. doi:10.3389/fenvs.2023.1132159
- Darji, J., Biswas, N., Padul, V., Gill, J. M., Kesari, S., and Ashili, S. (2024). Efficient use of binned data for imputing univariate time series data. *Front. Big Data* 7, 1422650. doi:10.3389/fdata.2024.1422650
- Fan, K., Dhammapala, R., Harrington, K., Lamb, B., and Lee, Y. (2023). Machine learning-based ozone and PM_{2.5} forecasting: application to multiple AQS sites in the Pacific northwest. *Front. Big Data* 6, 1124148. doi:10.3389/fdata.2023.1124148
- Haile, T. T., Tian, F., AlNemer, G., and Tian, B. (2024). Multiscale change point detection for univariate time series data with missing value. *Mathematics* 12 (20), 3189. doi:10.3390/math12203189
- Hua, V., Nguyen, T., Dao, M.-S., Nguyen, H. D., and Nguyen, B. T. (2024). The impact of data imputation on air quality prediction problem. *PLoS One* 19 (9), e0306303. doi:10.1371/journal.pone.0306303
- Jäger, S., Allhorn, A., and Bießmann, F. (2021). A benchmark for data imputation methods. *Front. Big Data* 4, 693674. doi:10.3389/fdata.2021.693674
- Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmos. Environ. Oxf. Engl.* 1994 38 (18), 2895–2907. doi:10.1016/j.atmosenv.2004.02.026
- Kleinke, K., Reinecke, J., Salfrán, D., and Spiess, M. (2020). *Applied multiple imputation: advantages, pitfalls, new developments and applications in R. 2020th ed. Statistics for social and behavioral sciences.* Cham, Switzerland: Springer Nature.
- Kumar, R. P., Prakash, A., Singh, R., and Kumar, P. (2024). Machine learning-based prediction of hazards fine PM_{2.5} concentrations: a case study of Delhi, India. *Discov. Geosci.* 2 (1), 34. doi:10.1007/s44288-024-00043-z
- Kumar, R. P., Jahan, A., Singh, R., Kumar, P., Bag, R., Bhatla, R., et al. (2025a). Spatio-temporal analysis of air pollution and meteorological influences in Western Uttar Pradesh using geospatial techniques: insights for policy and management. *Int. J. Remote Sens.* 00 (00), 1–28. doi:10.1080/01431161.2025.2529601
- Kumar, R. P., Rana, R., Choudhary, A., and Singh, R. (2025b). Spatiotemporal variability and source attribution of PM_{2.5}/PM₁₀ ratios: aerosol type classification and AQI evaluation across seventy monitoring stations in Delhi and Haryana, India. *Phys. Chem. Earth* 140 (104005), 104005. doi:10.1016/j.pce.2025.104005
- Kumar, P., Choudhary, A., Joshi, P. K., Kumar, R. P., and Bhatla, R. (2025). Machine learning models for estimating criteria pollutants and health risk-based air quality indices over eastern Coast coal mine complex belts. *Front. Environ. Sci.* 13 (May), 1589991. doi:10.3389/fenvs.2025.1589991
- Liang, X., Li, S., Zhang, S., Huang, H., and Chen, S.X. (2016). PM_{2.5}Data reliability, consistency, and air quality assessment in five Chinese cities: CONSISTENCY IN CHINA'S PM_{2.5}DATA. *J. Geophys. Res. Atmos.* 121 (17), 10220–10236. doi:10.1002/2016jd024877
- Libasin, Z., Ul-Saufie, A. Z., Ahmat, H., and Shaziyani, W. N. (2020). Single and multiple imputation method to replace missing values in air pollution datasets: a review. *IOP Conf. Ser. Earth Environ. Sci.* 616 (1), 012002. doi:10.1088/1755-1315/616/1/012002
- Lien, P.L., Do, T. T., and Nguyen, T. (2023). "Data imputation for multivariate time-series data," in *2023 15th international conference on knowledge and systems engineering (KSE)* (IEEE), 1–6.
- Little, R. J. A. (1988). Missing-data adjustments in large surveys. *J. Bus. and Econ. Statistics A Publ. Am. Stat. Assoc.* 6 (3), 287–296. doi:10.1080/07350015.1988.10509663
- Mendes, L., Monjardino, J., and Ferreira, F. (2022). Air quality forecast by statistical methods: application to Portugal and Macao. *Front. Big Data* 5, 826517. doi:10.3389/fdata.2022.826517
- Mezoue, C. A., Cedric Ngangmo, Y., Choudhary, A., and Monkam, D. (2023). Measurement of fine particle concentrations and estimation of air quality index (AQI) over northeast douala, Cameroon. *Environ. Monit. Assess.* 195 (8), 965. doi:10.1007/s10661-023-11582-2
- Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., and Stork, J. (2015). Comparison of different methods for univariate time series imputation in R. *arXiv preprint arXiv:1510.03924*. Available online at: <http://arxiv.org/abs/1510.03924>.
- Ngangmo, Y. C., Mezoue Adiang, C., Choudhary, A., and Monkam, D. (2023). Road traffic-induced particle matter dispersion in a calm wind environment at the main roundabout in douala, Central Africa. *J. Air Pollut. Health.* doi:10.18502/japh.v8i1.12030
- Niako, N., Melgarejo, J. D., Maestre, G. E., and Vatcheva, K. P. (2024). Effects of missing data imputation methods on univariate blood pressure time series data analysis and forecasting with ARIMA and LSTM. *BMC Med. Res. Methodol.* 24 (1), 320. doi:10.1186/s12874-024-02448-3
- Noor, M. N., Yahaya, A. S., Ramli, N. A., and Al Bakri Abdullah, M. M. (2015). Filling the missing data of air pollutant concentration using single imputation methods. *Appl. Mech. Mater.* 754–755 (April), 923–932. doi:10.4028/www.scientific.net/amm.754-755.923
- Pereira, R. C., Abreu, P. H., Pereira Rodrigues, P., and Figueiredo, M. A. T. (2024). Imputation of data missing not at random: artificial generation and benchmark analysis. *Expert Syst. Appl.* 249 (123654), 123654. doi:10.1016/j.eswa.2024.123654
- Rantou, K. E. (2017). Missing data in time series and imputation methods (Master's thesis) (Samos, Greece: University of the Aegean).
- Ribeiro, S., and Castro, C. (2022). "Missing data in time series: a review of imputation methods and case study," in *Learning and nonlinear models.* doi:10.21528/lnlm-vol20-no1-art3
- Ritthewa, T., and Samart, K. (2024). Performance of different imputation methods in logistic regression with multicollinearity. *Philipp. J. Sci.* 153 (3). doi:10.56899/153.03.05
- Saheer, L. B., Bhasya, A., Maktabdar, M., and Zarrin, J. (2022). Data-driven framework for understanding and predicting air quality in urban areas. *Front. Big Data* 5, doi:10.3389/fdata.2022.822573
- Sharma, V., Ghosh, S., Mishra, V. N., and Kumar, P. (2025). Spatio-temporal variations and forecast of PM_{2.5} concentration around selected satellite cities of Delhi, India using ARIMA model. *Phys. Chem. Earth* 138 (103849), 103849. doi:10.1016/j.pce.2024.103849
- Stekhoven, D. J., and Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28 (1), 112–118.
- Sugden, R. A., and Rubin, D. B. (1988). Multiple imputation for nonresponse in surveys. *J. R. Stat. Soc. Ser. A, Statistics Soc.* 151 (3), 567. doi:10.2307/2983027
- Suleman, K. (2022). World bank report.
- Sun, Y., Li, J., Xu, Y., Zhang, T., and Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: a review and comparative study. *Expert Syst. Appl.* 227 (120201), 120201. doi:10.1016/j.eswa.2023.120201
- Tyagi, A., Koul, A., and Mahajan, M. (2021). "Performance analysis of imputation methods on air quality dataset," in *Smart computing* (London: CRC Press), 694–701.
- Umar, N., and Gray, A. (2023). Comparing single and multiple imputation approaches for missing values in univariate and multivariate water level data. *Water* 15 (8), 1519. doi:10.3390/w15081519
- Wijesekara, W. M. L. K. N., and Liyanage, L. (2020). "Comparison of imputation methods for missing values in air pollution data: case study on Sydney air quality index," in *Advances in intelligent systems and computing* (Cham: Springer International Publishing), 257–269.
- Wijesekara, L., and Liyanage, L. (2021). "Air quality data pre-processing: a novel algorithm to impute missing values in univariate time series," in *2021 IEEE 33rd international conference on tools with artificial intelligence (ICTAI)* (IEEE), 996–1001.
- Wijesekara, L., and Liyanage, L. (2023). Mind the large gap: novel algorithm using seasonal decomposition and elastic net regression to impute large intervals of missing data in air quality data. *Atmosphere* 14 (2), 355. doi:10.3390/atmos14020355
- Zainuddin, A., Hairuddin, M. A., Yassin, A. I. M., Latiff, Z. I. A., and Azhar, A. (2022). "Time series data and recent imputation techniques for missing data: a review," in *2022 international conference on green energy, computing and sustainable technology (GECOST)* (IEEE), 346–350.
- Zhang, Y., and Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: a system and a review of selected methods. *Future Gener. Comput. Syst. FGCS* 128 (March), 63–72. doi:10.1016/j.future.2021.09.033