



# Selection of statistical technique for imputation of single site-univariate and multisite–multivariate methods for particulate pollutants time series data with long gaps and high missing percentage

Priti K<sup>1,2</sup> · Kaushlesh Singh Shakya<sup>1,2</sup> · Prashant Kumar<sup>1,2</sup>

Received: 2 January 2023 / Accepted: 11 May 2023 / Published online: 23 May 2023  
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

## Abstract

Monitoring air contaminants has become essential to exposure science, toxicology, and public health research. However, missing values are common while monitoring air contaminants, especially in resource-constrained settings such as power cuts, calibration, and sensor failure. In contaminants monitoring, evaluating existing imputation techniques for dealing with recurrent periods of missing and unobserved data are limited. The proposed study aims to perform a statistical evaluation of six univariate and four multivariate time series imputation methods. The univariate methods are based on inter-time correlation characteristics, and the multivariate approach considers multi-site to impute missing data. The present study retrieved data from 38 ground-based monitoring stations for particulate pollutants in Delhi for 4 years. For univariate methods, missing values were simulated under 0–20% (5%, 10%, 15%, and 20%), and high 40%, 60%, and 80% missing levels having long gaps. Before evaluating multivariate methods, input data underwent pre-processing steps: selecting the target station to be imputed, choosing covariates based on the spatial correlation between multiple sites, and framing a combination of target and neighbouring stations (covariates) under 20%, 40%, 60%, and 80%. Next, the particulate pollutants data of 1480 days is provided as input to four multivariate techniques. Finally, the performance of each algorithm was evaluated using error metrics. The results show that the long interval time series data and spatial correlation of multiple stations significantly improved outcomes for univariate and multivariate time series methods. The univariate Kalman\_arima performs well for long-missing gaps and all missing levels (except for 60–80%), yielding low error and high  $R^2$  and  $d$  values. In contrast, multivariate MIPCA performed better than Kalman-arima for all target stations with the highest missing percentage.

**Keywords** Imputation techniques · Missing values · Particulate pollution · Multiple imputation · Univariate imputation · Missing at random · Spatial correlation · Time series data

## Introduction

In common parlance, the issue of missing data exists across many areas of research, including exposure science, statistical surveys, epidemiological studies, and occupational

health research (Junninen et al. 2004, Huisman 2009, Aslan 2010, Eekhout et al. 2012, Ramli et al. 2013, Sukatis et al. 2019, Kumar 2022). Air pollution monitoring has become an essential aspect of atmospheric exposure, public health, policy framework research, and risk communication to the public. The regulators have focused on strengthening the ground monitoring station networks to check air quality compliance with standards and for research applications. As a result, data on air quality is becoming more widely available, and the research underlying the associated health effects is also significantly growing.

Air pollution has been a significant concern in recent years due to unprecedented growth in urban centres. The rise in acute air contaminants incidences in many urban centres

Responsible Editor: Marcus Schulz

✉ Prashant Kumar  
prashantkumar@csio.res.in

<sup>1</sup> Academy of Scientific & Innovative Research (AcSIR), Ghaziabad 201002, India

<sup>2</sup> CSIR-Central Scientific Instruments Organisation, Sector 30-C, Chandigarh 160030, India

globally has increased the burden of disease, carrying culpability for around one out of every nine deaths annually. The WHO report of 2016 suggests that exposure to ambient and household air pollution contributes to mortality of over 7 million annually (WHO 2016). Also, air pollution is a direct indication of sustainable growth since sources of air pollution also produce climatic pollutants such as CO<sub>2</sub> or black carbon (Chan 2015). Monitoring of air pollutants has various spatial and temporal scales, and sampling methods can range from ground-based continuous monitoring stations as part of an ambient monitoring network (AMN) to low-cost or personal wearable sensors employed in the household or occupational settings that cover data from hours to days. Irrespective of sampling techniques, missingness can arise due to a variety of uncontrollable factors, including instrument malfunction, sensor sensitivity, maintenance or repair, calibration, and many other reasons (Wardana et al. 2022). Automated air quality monitoring stations preclude both regular and unscheduled shutdowns for maintenance. For illustration, many air monitors require 1 or 2 h every 2 weeks to examine (adjust) the zero value and air input flow, as well as a full calibration/maintenance every 6 months (i.e. a full working day shutdown) (Gómez-Carracedo et al. 2014, Little and Rubin 2019). Furthermore, remote stations experience lengthy outages due to power supply failures, problems with air aspiration pumps, or electronic processing malfunction. Such uncontrolled factors can impact time series data like air pollution as empty values can distort temporal information such as autocorrelation, trends, and seasonality. Therefore, mechanism and pattern of missingness are the most significant aspects to consider when choosing the optimal solution for imputation through preventing poor handling of missing data and misleading data interpretation.

Rubin and Little gave a mathematical explanation of the missing data mechanism and divided it into three types based on the relationship of values of attributes with unobserved/missing values (Rubin 1976). Missing at random (MAR) pattern is frequently observed in air pollution research data (Junninen et al. 2004, Plaia and Bondi 2006, Ghazali et al. 2020). Using the MAR mechanism, the cause of data missingness is explainable as missingness for an attribute is defined by the observed data, not by missing values. Missing information is recovered by using other variables for which the sample lacks missing data. In air pollution studies, missing data has a MAR pattern if data loss is due to power failure or system shutdown. When data is lost due to the inability of sensors to detect lower concentration limits of pollutants, it is missing not at random (MNAR) (Gómez-Carracedo et al. 2014). In a complex missingness structure, both MAR and MNAR missingness mechanisms exist. Furthermore, when data is missing due to unknown reasons, then it is missing completely at random (MCAR) (Hadeed et al. 2020). However, most empty values exist due

to explainable conditions (Ghazali et al. 2020). Thus, understanding the selectivity of missing data and the corresponding mechanism is a pivotal step for dealing with missing data adequately.

Imputation techniques are broadly classified into two categories: Single imputation (SI) and multiple imputations (MI) to completely filled the missing data. SI provides single values for missing data items, whereas MI imputes multiple values for a given missing datum. Under the SI category, mean replacement (unconditional mean imputation) is the most widely used method in research studies. However, this method provides varied results depending on the pattern and type of missingness, pattern of missingness percentage, and gape size (short or long period of time the data are missing from the datasets) of missing values. Under the MAR mechanism, the estimated results through mean imputation provide inconsistent results through the variance of the regression coefficients, whereas the MCAR mechanism holds consistent variance but underestimates (Junger and De Leon 2015). Also, mean imputation gives more relevant results for normal data distribution than for skewed data. Median is another simple imputation technique that provides better results for skewed distribution, ameliorating the mean imputation gap (Junger and De Leon 2015, Hadeed et al. 2020). The univariate single imputation mean method tends to alter the tails of distribution as the number of observations is higher at the centre of normal distribution (Little and Rubin 2019). Another single imputation method that fills the missing values of univariate time series data of air pollutants using the last observed value of the same variable is last observation carried forward (LOCF). This method is convenient as it fixes the entire univariate time series dataset. However, even in MCAR conditions, this method can generate biased estimates (Molenberghs and Kenward 2007). Spline interpolation is another univariate inter-time imputation technique that imputes NAs values using the `na.interpolation` function in `TS` imputes package. The fitted function is a piecewise nonlinear polynomial that uses current available pollutants data to impute missing values (Wijesekara and Liyanage 2020). Other univariate methods considered for long-missing gaps, highly complex time series with trend and seasonality, and low autocorrelation are Kalman smoothing and seasonal decomposition using `na.Kalman` and `na.seadec` function, respectively. The `seadec` algorithm performs seasonal decomposition as a pre-processing step (Liu et al. 2020).

Multivariate time series imputation is another class of imputation techniques that handles multiple variables simultaneously, which provide single imputation or multiple imputations for missing data points. For the proposed work, all the techniques are multiple imputations based that rely on correlations between different co-variables to estimate values for missing data. Random forest imputation

is a typical example of a multivariate imputation method implemented under the mice package (Little and Rubin 2019). This multivariate imputation approach fills NAs data points of the target variable using the time series of neighbouring variables. Other widely used multivariate imputation methods are predictive mean matching, multiple imputations through chain equation (MICE), expectation-maximisation, weighted predictive mean matching, random imputation, and multiple imputation principal component analysis. Predictive mean matching (PMM) is considered for multivariate imputation under the mice package of R. PMM is a hot deck-based algorithm where missing values are filled using covariates with the same distribution characteristics (Kleinke 2018). In their study, Marshall et al. (2010a, b) reported that PMM provides less biased estimates and significantly improved performance metrics in the context of missing covariate data (Marshall et al. 2010a, b). Another skewed data imputation study recommended that PMM performs better when missing data is MAR and < 50% (Marshall et al. 2010a, b). Another algorithm that stimulates multiple variate data points is midastouch (multiple imputation by automatic, distance-aided donor selection) under the MICE package. The method is based on a hot deck iterative algorithm with distance-based donor selection, and this measure of distance controls the trade-off between bias and variance of estimates (Siddique and Belin 2008, Siddique and Harel 2009). This method replaces PMM within the mice package by ‘midastouch’ (Gaffert et al. 2018). Multiple imputation PCA is another stimulation technique that estimates NAs data points and considers the similarity between individuals (rows) and the link between variables across all individuals. This method performs better when the correlation structure between variables is stronger (John et al. 2019). Regularised terms fix the problem of overfitting by penalising relatively unreasonably large parameters. In addition, these multiple imputation methods can take account of extra variability, yielding more precise results; however, such methods are only addressed a little (Schafer 1997).

Despite the fact that a large number of imputation tools are present in different statistical packages but used scarcely in air pollution studies as the most common approach is to ignore them. However, data continuity is important in time series modelling as excluding incomplete values influences temporal characteristics such as autocorrelation, seasonality, and trends. Moreover, deep learning algorithm requires a large amount of data, and some algorithms are sensitive to missing values; therefore, there are better options than complete case analysis to handle especially long missing data. Though some studies have used different imputation techniques, may it be single or multiple imputation-based methods for air quality datasets. However, its overall application still needs to be improved, with few tests of performance in

real-world scenarios and little guidance regarding the imputation of air quality data.

This study presents the comparative analysis of six univariate time series methods used to impute low (0–20%) and high missing (40%, 60%, and 80%) percentage along with long missing gaps. Different approaches were used to filter the temporal components. Kalamam\_arima and seadec method considered the temporal component of time series. In addition, we use spatial correlation component, i.e. multi-site, to select covariates for multiple imputation methods like PMM, MIDAS, RF, and MIPCA to impute high missing percentages under four missing categories (20%, 40%, 60%, and 80%). We examine the performance of single univariate imputation and multivariate multi-imputation methods for the short and long consecutive missing periods and provide future guidance to implement in other study settings.

## Methodology

### Study area

The study area considered for the proposed work is Delhi, the national capital of India, located in the northern part of Indo-Gangetic plain, which lies between the latitude of 28°24'17" and 28°53'00" north and longitude of 76°50'24" and 70°20'37" of the east with the total geographic area of 1483 Km<sup>2</sup> (700 km<sup>2</sup> urban and 783 km<sup>2</sup> rural). The climate of Delhi makes it favourable for pollution stagnation, especially during winter when the temperature reaches 22 °C to 5 °C and with no dispersion (Budhiraja et al. 2019). According to the Koppen classification system of climate Delhi has five seasons with the extreme type of climate and witnesses 714 mm of annual rainfall. From a demographic perspective, Delhi has the highest population density of 11,297 persons per sq. km as per the 2011 census compared to other Indian states/U.T. (Census, 2011; <http://census2011.co.in>). The key factors that cause alarming situation for Delhi's air quality include the city's landlocked geographical location, residual crop burning in neighbouring states of Punjab, Haryana, Uttar Pradesh, and Rajasthan, vehicular emissions, industrial pollution, and large-scale construction activities (Chatterji 2021). The city is contending with the capital's escalating air pollution problem and associated health risks.

### Data information

The empirical data of 38 continuous air quality monitoring stations (CAQMS) of Delhi is retrieved from the central pollution control board website (<http://www.cpcb.nic.in/>). The stations IDs have a series such as D\_01, D\_02, D\_03, D\_05..... D\_40. It is important to mention over here that the data of two stations, namely, D\_04, and D\_12

are not available. The geospatial distribution of these sites is shown in Fig. 1. Air pollutants data are in matrix form, ( $m \times n$ ), with each column  $n$  (representing variables) containing a time series of two air pollutants ( $PM_{2.5}$  particulate matter with an aerodynamic diameter of less than  $2.5 \mu m$ ,  $PM_{10}$  particulate matter with an aerodynamic diameter of less than  $10 \mu m$ ), considered and each attribute per station are studied. The particle pollutants are initially in hourly measurements but later aggregated on a 24-h basis. Thus, 1480 individuals (rows) are considered for two pollutants from February 2018 to February 2022 (4 years and 9 days). These stations are deployed in traffic, industrial, and residential zones to monitor important pollutants. While selecting the stations for study, some conditions are considered: the accessibility of particle pollution with at least 3 years from the same timespan is included.

### Descriptive statistics of all monitoring stations

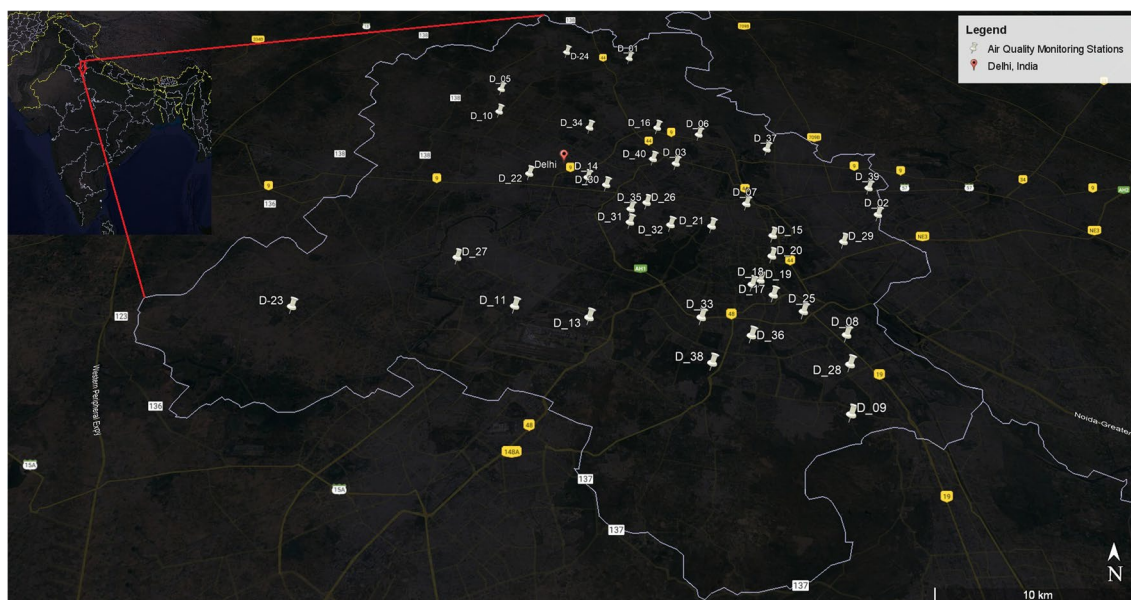
Appendix 1 shows descriptive statistics of all stations of the Delhi region for the  $PM_{2.5}$  and  $PM_{10}$  datasets. The detailed information can be found in Appendix 1 of five statistical characteristics: mean, 25% quantile, median, and 75% quantile, and standard deviation for each station ( $PM_{2.5}$ ,  $PM_{10}$ ), excluding missing values. The standard deviation is a measure of statistical dispersion that indicates how far observed values are dispersed from the mean of each variable for different locations. For example, small  $\sigma$  value for a monitoring station indicates that the monitored values of contaminants are close to the average; in contrast, a high standard deviation value depicts that actual values are far

dispersed from the mean. Another measure of variability in descriptive statistics is quartiles, dividing the structured observed data (from lowest to highest) into four parts. The first quartile (25%) is defined as the value in the middle of the range between the minimum and the median, while the third quartile (75%) is defined as the value in the middle of the range between the median and the maximum (Wardana et al. 2022).

### Missing data mechanism

Knowledge of the selectivity of missing data and the corresponding mechanism is a crucial starting point for dealing with missing data effectively. A correct approach to missing data is determined by whether the data is MAR or missing, not at random (MNAR). Observed data can be used to estimate missing values where the missing data pattern is MAR (or can be assumed to be MAR) (Rubin 1976). Such estimation is impossible when the data is MNAR, as estimation depends on unobserved/missing data. MNAR condition of missingness is nonignorable and generally viewed as a condition that results in biased predicted values. A third mechanism missing completely at random (MCAR) assumes that subjects with missing data are a random subset of the entire study sample, making them less susceptible to bias.

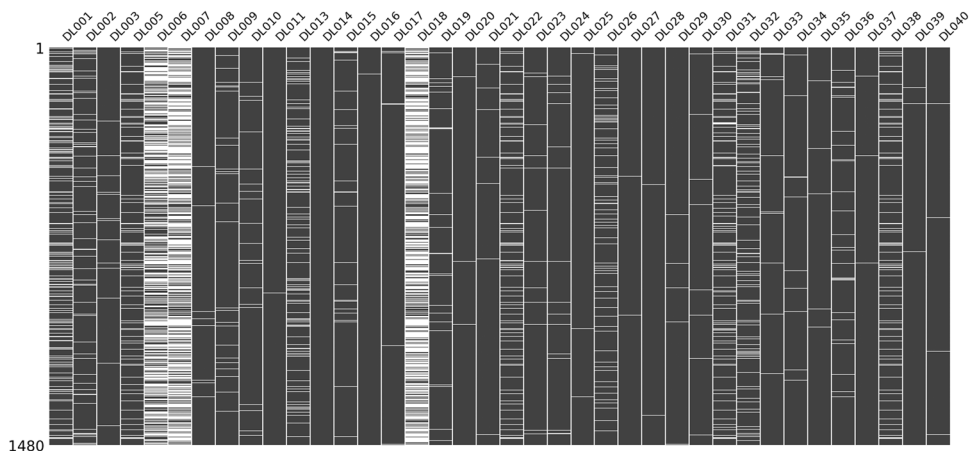
In most cases, missing data is neither MCAR nor MNAR. In the case of air quality data, the missingness condition most often observed is MAR, as reasons for missingness are known, such as routine maintenance, sensor malfunction, and power outages (Gómez-Carracedo et al. 2014). In some cases, the causes of missingness are unexplainable



**Fig. 1** Study area map showing spatial distribution of 38 ground monitoring stations in Delhi. The map scale is 10 Km/cm



**Fig. 2** Missing data pattern for 1480\*38 data matrix of PM<sub>2.5</sub>



(e.g. MCAR). Furthermore, complete-case analysis for both MAR and MNAR data can produce biased results as such conditions have higher missing values and reduce the analysis’s precision. For the MCAR mechanism, complete case analysis (which allows for excluding incomplete observations) produces unbiased results. However, as air pollution is time series data, the listwise deletion method may break the data structure, resulting in the loss of valuable information. MAR’s missing data mechanism is more general and realistic than MCAR’s. Modern missing data methods typically begin with the MAR assumption (Van Buuren 2018). Thus, under the proposed study, we assumed MAR condition for air quality data imputation from available information.

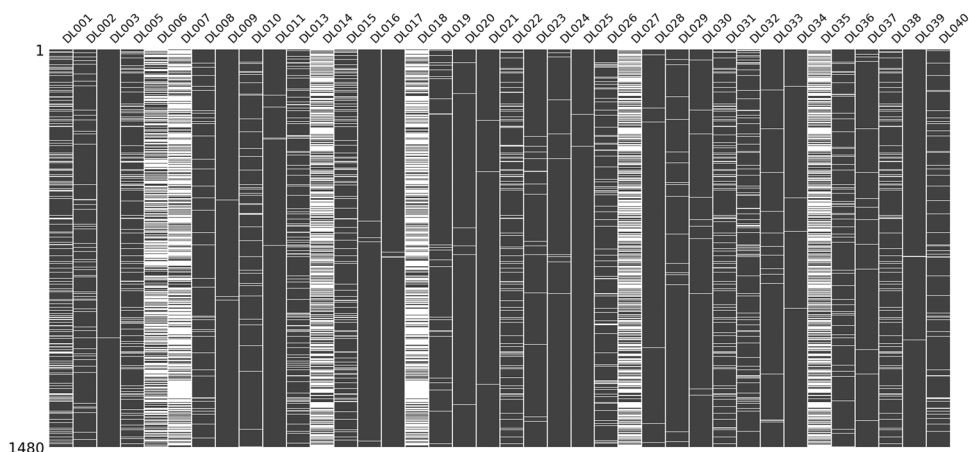
**Missing data statistics**

It is important to understand the missing statistics of a dataset before selecting imputation techniques (Appendix 2). The higher percentage of missing value affects the time series of air pollutants. It would be much more challenging to perform time-related analyses, such as identifying weekly patterns or autoregressive models that predict from previous days. The percentage of missing across the two variables PM<sub>2.5</sub> and PM<sub>10</sub> in this study varied between 0 and 80%.

Figures 2 and 3 depict that the dark areas represent the span of the monitored values of pollutant with no missing values, while the white stripes represent the missing values for PM<sub>2.5</sub> and PM<sub>10</sub>, respectively.

Depending on the causes of the air pollution data missing, the gap size of the missing pattern can be divided into the short interval and long period missing. The missingness due to temporary power outages and routine maintenance causes short gaps in data collection, while sensor malfunctions and other critical reasons cause long-interval data missing. Missing values appears both in the discrete and consecutive missing pattern. For short-interval broken data, different level of missing percentage is considered as input for univariate time series imputation methods (20% missing rate category divided into four subcategories of 5%, 10%, 15%, and 20%) and for long-interval of missing levels 40%, 60%, and 80% is considered. Furthermore, we applied different multiple imputation methods based on the MICE algorithm in the case of long-interval missing levels. Before multivariate imputation, the pre-processing steps are applied to consider the spatial characteristics of monitoring stations while imputing missing values. The target stations to be filled under four categories, 20%, 40%, 60%, and 80%

**Fig. 3** Missing data pattern for 1480\*38 data matrix of PM<sub>10</sub>



is, chosen based on the highest missing percentage (missing rates varied for each target station).

The univariate time series methods consider both low and high missing percentage monitoring stations for PM<sub>2.5</sub> and PM<sub>10</sub>, and for multivariate imputation techniques, high missing percentage target monitoring sites are selected. The neighbouring sites were selected based on spatial characteristics of monitoring sites with a missing percentage of < 20%.

## Imputation methods

The methods for imputation of time series can be broadly classified into three types: inter-time (time series), inter-variable (cross-sectional), and inter-variable + inter-time (TS cross-sectional). This study considered the first two approaches.

### Univariate time series methods (inter-time series)

The univariate time series method uses temporal characteristics such as trend, seasonality, autocorrelation, and periodicity for air pollutants data. Temporal substitution techniques replace missing air pollutants data either with the mean of the whole column (time series) or stimulate the mean of neighbouring values using lagged data to fill missing gaps. One of the simplest and most popular methods to fix missing data is to replace them with the overall mean. There are better ideas than putting the overall mean in time series data like air pollution, which shows seasonality characteristics. Thus, mean imputation is generally not recommended by statisticians; it is only used in case of rapid fix only when less missing data is present. In case of large number of missing values, a fix value (i.e. mean) can change the distribution shape, standard deviation becomes smaller for imputed data compared to actual values. For high missing percentage with long consecutive gaps, mean imputation can cause shrinkage in standard deviation. Some studies have illustrated that mean imputation can work better with conditions like < 10% missing values in dataset and low correlation between variables (Raymond 1986, Tsikriktsis 2005). Though data stratification into smaller groups can slightly enhance this method (Norazian et al. 2008, Junior et al. 2016).

Another method, median, is a standard univariate imputation algorithm that does not consider the time series characteristics of air pollutants. However, it can give better results than the mean in a skewed distribution. Despite their popularity, these statistical techniques have the evident constraint of ignoring important temporal information and further affect correlation structure reasonably (Weerakody et al. 2021, Iodice D'Enza et al. 2022). However, these univariate time series substitution methods are relatively simple and do not account for any temporal variation in air pollution data.

LOCF is the univariate time-series longitudinal data method. The concept of the LOCF imputation technique is to substitute the most recently observed value for the entire missing data of the same variable. LOCF must be followed by a proper statistical analysis method that distinguishes between actual and imputed data. In time series data, the first-order Markov model assumes a similar condition as the LOCF imputation method. The probability of an observation at time 't' depends on the last observation at a time (t-1) (Canales 2004). This method is simplest that takes advantage of time series characteristics. However, this algorithm has a drawback with a time series dataset with seasonality, when there are considerable differences between observations at t and the last Observation at the time (t-1) (Moritz et al. 2015). Another univariate time series method is the spline that replaces missing values based on the interpolation algorithm. This temporal interpolation method fits the observed data with a curve to stimulate NAs in air pollution data (Cho et al. 2020). The imputed air pollutants dataset can accurately simulate the temporal trend for short gap size. However, in case of more extended gap size, these simple imputation techniques like LOCF and interpolation might provide significant bias between imputed values and corresponding true values as data structure during an unobserved period is unpredictable from a univariate standpoint (Liu et al. 2020). This single imputation method ignores uncertainty in imputed values.

'Seasonally Decomposed Imputation' and 'Kalman Smoothing on Structural Time Series Models' are complex methods for imputation of missing data in univariate series since their ability to capture time-dependent characteristics like trend, seasonality, and cycles. However, they necessitate longer computation times (Moritz and Bartz-Beielstein 2017). In this study, Kalman Smoothing on the state space representation of an ARIMA model is used for imputation of time series data. However, to obtain more accurate results of imputation, smoothing algorithm is used (Agbailu et al. 2020). This technique has been developed by (Kalman 1960, Welch 2006). The algorithm imputes missing values based on lagged data points of the same variable. Let  $z_{(t-1)}$  represents a set of lagged values and assume conditional distribution of  $\mu_t$  is  $N(\mu_t, p_t)$ ,  $\mu_t$ ,  $p_t$  are assumed to have been determined (Agbailu et al. 2020). The function `na_kalman` uses `KalmanSmooth` (as `KalmanSmoothen`) and operates on state space representation of an ARIMA model obtained by `auto.arima` (Harvey 1990). This imputation technique based on the recursive data assimilation system of the Kalman filter (Moritz and Bartz-Beielstein 2017). To fill missing values, structural model fitted with maximum likelihood is used (Han et al. 2023).

Another univariate time series method is Seasonal decomposition (Secdec) combined with moving average to impute missing values. The Secdec algorithm removes the

seasonal component of time series and afterwards algorithm like moving average perform imputation on deseasonalised dataset (Moritz and Bartz-Beielstein 2017). The air pollutant time series contains various levels of time dependency structures, such as 1 and 2 autocorrelation lags which would have been poorly captured by straightforward techniques like interpolation, mean values, random values. For this reason, among other imputation options, we choose moving average filters. In actuality, earlier Seadec tests that utilised these options produced subpar outcomes. Seadec. is therefore described as a useful technique for seasonal and autocorrelated series (Lloret et al. 2000, Benavides et al. 2022).

### Multivariate timeseries methods (cross-sectional inter-variables)

Multivariate is another class of imputation method that accounts for temporal characteristics of a set of observed predictor variables. Compared to a single imputation (such as the mean), having multiple imputations eliminates uncertainty in missing values. In this case, covariates are selected as the correlation structure of the data is also an important factor in determining the performance of some multiple imputation methods (Dray and Josse 2015). The mice function of the R package MICE is a multiple imputation method that estimates the missing data and the variability associated with the imputation. MICE is an iterative algorithm; at each iteration, all variables except one are fixed and will be used as predictors, while an imputation method is chosen for the unfixed variable.

Predictive mean matching (PMM) was applied using the MICE package. The PMM method is especially effective for numerical variables without a normal distribution (Allison 2001). PMM is a semi-parametric method random case is chosen from complete cases with a predictive value that is close to the missing case (Little and Rubin 2019). Since the method uses data set values, they are accurate and there are no arbitrary imputed values (Schenker and Taylor 1996, Abayomi et al. 2008). The number of multiple imputations is 6 and maxit iterations is 20. Each target variable in the dataset is imputed using the donor variables in each iteration. These iterations should be repeated until convergence appears to have been achieved (Li et al. 1999). Several PMM variants were developed using various predictor variables (neighbouring stations) combinations. The number of imputations generated in PMM is six, and donor selection,  $d = 5$ , i.e. in each case of target stations missing value to be predicted, is matched to five cases with closest values. One of the donor values is chosen randomly, and its value is assigned to the target station with missing data (Allison 2015) and 24-h mean across the 20 iterations were used to impute daily concentrations of target stations. PMM Imputes missing values of target stations from observed values of

covariates using linear regression coefficients. The advantage of PMM is that the imputed values are drawn from a range of observed values, and the possibility of imputing unrealistic values, such as negative concentrations, is eliminated (Van Buuren 2018). Similarly, two other multiple imputation methods considered for this study in the MICE package are MIDAS and Random Forest, which uses covariates to fill the target stations missing values giving multiple imputed datasets. RF method is based on Breiman's algorithm as presented in (Breiman 2001, Doove et al. 2014). This non-parametric method predicts missing instances of features using values of other attributes. In this manner, the missing values for that feature are discovered. The process is repeated until it converges. For this study, value of parameter ntree chosen was 10.

Lastly, the multivariate imputation method was implemented on the air pollution dataset is MIPCA, implemented in the missMDA package. Josse and Husson (2009) proposed a regularised version of the iPCA algorithm (RPCA) to overcome the problem the overfitting (Iodice D'Enza et al. 2022). MIPCA provides multiple imputed datasets using the PCA model. The observed values are consistent across datasets, whereas the imputed values vary. The variability in imputed values reflects the variability in missing value prediction. According to Little and Rubin (2002), multiple imputations are appropriate because it accounts for parameter variability (Josse and Husson 2011). The Bayesian method is used for multiple imputations to reflect variability caused by missing values. The argument used for its implementation is a method.mi='Bayes'. The threshold for the criteria convergence is  $1e^{-04}$  (Audigier et al. 2016). Three steps are followed to perform MIPCA on the incomplete dataset: estimate the number of dimensions using the function estim\_ncpPCA by k-fold cross-validation method to select tuning parameters. The second step is to use the function impute PCA to run the (regularised) iterative PCA algorithm with the number of dimensions selected in the first step. The third step is analysing the precision of each method by comparing the sample variance of the imputed values with the original variance (Li et al. 1999). These steps are repeated to achieve convergence.

### R packages Functions used for missing values imputation

The R software (Crawley 2012) is a free and open-source statistical programming language. All single and multiple imputation methods are implemented under different R packages and are freely available from the Comprehensive R Archive Network (CRAN) at <http://www.R-project.org/>. We use the TS stats time series objects from base R to represent univariate time series (Zeileis and Grothendieck 2005). We implemented multiple univariate time series-based

algorithms using TS packages such as mean, median (na.mean function), last observation carried forward (na.locf), spline (na.interpolation), Kalman Smoothing on the state space representation of an arima model (na.Kalman), seasonally decomposition model (na.seadec) (Moritz and Bartz-Beielstein 2017). Another R package used for multivariate imputation by chain equation is mice. The algorithms used are PMM, MIDAS, and random Forest to determine the missing values present in target variables (Van Buuren and Groothuis-Oudshoorn 2011). MissMDA package is used to for imputing missing values through MIPCA algorithm, (Stekhoven and Stekhoven 2013, Josse and Husson 2016). The evaluation metric to validate the imputation model through comparison of imputed and actual observed time series data using the goodness of fit hydroGOF package of R (Moriassi et al. 2007).

### Methods for evaluation metric

There are several methods used for the evaluation of considered imputation methods. The accuracy of imputation methods is measured based on the difference between actual and imputed data through different error metrics across four levels of missingness (i.e. 0–20%, 20–40%, 40–60%, and 60–80 missing). Generally, five indicators are used to assess the performance of imputation algorithms as discussed below (Quinteros et al. 2019). In common parlance, low missing percentage result in lower RMSE/MAE/PBIAS error and higher  $R^2$ , and d scores. The RMSE/MAE values may differ significantly due to the physical nature of each pollutant.  $R^2$  score is introduced to provide a more intuitive view of performance.

**Root mean square error: RMSE** is an error metric computed between the imputed value and the respective actual value time series, i.e. the standard deviation is calculated as referred to in Eq. 1.

$$RMSE(\bar{x}, x) = \sqrt{\frac{\sum_{i=1}^n (\bar{x}_i - x_i)^2}{n}} \quad (1)$$

This method has been extensively used in the literature to evaluate the performance of imputation methods (Junger and De Leon 2015, Moritz et al. 2015, Wardana et al. 2022). In air pollution data, seasonality and trends are observed; summer values for particulate pollutants are low compared to winter. The upward trend can be seen during winter, but error metrics like RMSE may not be appropriate for such data with significant differences (Moritz et al. 2015). A lower value indicates that the model performed better.

**Mean absolute error: MAE** is an important error metric for datasets showing strong trends. The evaluation metric is based on the difference between imputed values and actual

observed values in time series. This metric is least affected due to the dataset having a strong trend, i.e. the difference between datasets is high. Unlike RMSE, MAE is the more natural measure of average error that is unambiguous as RMSE varies with the variability of the error magnitude distribution, the square root of the number of errors ( $n^{1/2}$ ), and the average-error magnitude (Willmott and Matsuura 2005) Eq. 2.

$$MAE(\bar{x}, x) = \frac{\sum_{i=1}^n |\bar{x}_i - x_i|}{n} \quad (2)$$

**Coefficient of determination:  $R^2$**  is commonly used for evaluating models as a goodness-of-fit metric.  $R^2$  is calculated by squaring the correlation coefficient between two columns and evaluating the variance between observed and predicted concentrations (Quinteros et al. 2019, Hadeed et al. 2020) Eq. 3.

$$R^2 = \left( \frac{\sum_{i=1}^n (x_i - \bar{x}_i) * (\bar{x}_i - \bar{x}_i)}{\sqrt{\sum_{i=1}^n x_i - \bar{x}_i} * \sqrt{\sum_{i=1}^n \bar{x}_i - \bar{x}_i}} \right)^2 \quad (3)$$

where  $x_i$  &  $\bar{x}_i$  are the  $i_{th}$  value of actual and imputed observations, respectively, and  $\bar{x}_i$  &  $\bar{x}_i$  are the mean of actual and imputed datasets. It indicates how well the model explains the variance in the observations when compared to using the mean of the observations as the prediction.

**Percentage bias: PBIAS:** The goodness of fit measures the average tendency of an imputation algorithm, i.e. if the imputed 24-h aggregated pollutants concentration within each station is larger or smaller than their observed values. A low value of PBIAS indicates accurate model simulation (zero implies optimal value; positive values indicate overestimation bias, while negative values indicate underestimation bias in the model). The metric is calculated in percentage as given in Eq. 4.

$$PBIAS(\bar{x}, x) = \frac{\sum_{i=1}^n x_i - \bar{x}_i}{\sum_{i=1}^n x_i} \times 100 \quad (4)$$

**Index of agreement: Willmott (1981)** developed the Index of Agreement (d) as a standardised measure of model prediction error that ranges from 0 to 1 (Willmott and Matsuura 2006). A value of one indicates a strong positive match, while a value of zero indicates no agreement at all. The index of agreement can identify additive and proportional differences in observed and simulated means and variances; however, due to squared differences, it is extremely sensitive to extreme values (Legates and McCabe Jr 1999) given in Eq. 5.



$$d = 1 - \left( \frac{\sum_{i=1}^n (x_i - \bar{x}_i)}{\sum_{i=1}^n (x_i - \bar{x}_i) + (\bar{x}_i - \bar{x}_i)} \right)^2 \quad 0 \leq d \leq 1 \quad (5)$$

where  $x_i$  &  $\bar{x}_i$  are the  $i_{th}$  value of actual and imputed observations respectively and  $\bar{x}_i$  is the mean of actual datasets.

## Result and analysis

Initially, 38 monitoring stations and two pollutants per station were analysed. Descriptive statistics were applied on all the stations for  $PM_{2.5}$  and  $PM_{10}$ , including mean, 25% percentile, median, 75% percentile, and standard deviation, as mentioned in Appendix 1. The statistical dispersion around the mean for particulate pollutants varies across all the stations, with the highest of all stations being  $135.28 \pm 108.83$  for  $PM_{2.5}$  and  $308.78 \pm 143.58$  for  $PM_{10}$ . Next, to understand the complexity of time series data for NAs, it is necessary to know the statistics of missing components like the number of missing values, percentage of missing, number of gaps, average gap size, longest NA gap (series of consecutive NAs), and most frequent gap size (series of consecutive NA series). Appendix 2 provides detailed missing statistics information on the air pollutants dataset. A total of 1480 observations are considered for each station. The percentage of missing across 38 stations varied between 0 and 80%. The missing percent range, 0–80, is divided into four categories (0–20%, 20–40%, 40–60%, and 60–80%). Most of the stations lie between 0 and 20, missing percentage. Therefore, we further categorised 0–20% into four ranges 0–5%, 5–10%, 10–15%, and 15–20%. Furthermore, some stations were eliminated from the univariate time series experimental, with a missing percentage of < 1.0%. The largest missing gap size (NAs in rows) is 1016 days continuous, with an average gap size of 70.2 for  $PM_{2.5}$  and  $PM_{10}$ . Furthermore, air pollutants data are omitted in a random manner because data may be missed due to a variety of explainable circumstances. For example, many air pollutants analysers require 1 to 2 h every 2 weeks to verify and analyse the air input flow. The unanticipated events, such as power supply, pump, and electronic processor failure, occur randomly, resulting in missingness, as shown in Figs. 2 and 3. The missing pattern is missing at random (MAR) as NAs data can be imputed based on observed data.

## Univariate imputation techniques

Tables 1 and 2 show imputation results of different univariate (mean, median) and univariate time series methods (LOCF, spline, Kalman\_arima, seadec) under various measurement

indicators (MAE, RMSE, PBIAS, d,  $R^2$ ), low (0–20%) and higher missing (at 40%, 60%, and 80%) percentage, and air pollutants ( $PM_{2.5}$ ,  $PM_{10}$ ) through comparison of imputed value with actual observed value within each station. PBIAS measures the difference between observed and imputed mean concentrations of pollutants (Tables 1 and 2). The results estimated through univariate substitution methods that replace the missing values of air pollutant data with the mean and median of the single imputation value for each station show that the median imputation method consistently resulted in low PBIAS% across all levels of missingness compared to mean substitution for all sites. In the case of the mean imputation method, a decline of PBIAS% is observed across all stations with increasing missing values. Also, mean and median methods do not account for any temporal characteristic of  $PM_{2.5}$  and  $PM_{10}$  time series data. Thus, both method performance depends only on the percentage of missing, not merely on missing interval gaps (short gap/ longer consecutive missing gap).

For  $PM_{2.5}$  low missing percentage, i.e. for 0–20%, the Kalman imputation yields absolute bias, i.e. the difference between actual and simulated 24-h mean by  $1.87 \mu\text{g}/\text{m}^3$  (for station D\_36 at 5% missing),  $4.43 \mu\text{g}/\text{m}^3$  (D\_26 at 10% missing),  $1.83 \mu\text{g}/\text{m}^3$  (D\_13 at 15% missing),  $0.90 \mu\text{g}/\text{m}^3$  (D\_31 at 20% missing), and for higher missing percentage at 40% missing, this difference is around  $3.64 \mu\text{g}/\text{m}^3$  (D\_01), at 60% missing,  $4.46 \mu\text{g}/\text{m}^3$  (D\_06), and at 80% missing, imputed mean differed from observed mean by  $67.39 \mu\text{g}/\text{m}^3$  (D\_18). The biasness between the estimated and actual values increases with an increase in missing percentages. However, in the case of Kalman imputation, the biasness also depends on missing gap size. For instance, D\_31 has a higher missing percentage than D\_26. However, the missing interval in the case of D\_31 has longer consecutive values (consecutive gap of 153 values) compared to D\_26 (longest gap of 33 continuous missing values) refers to Appendix 2. The univariate time series Kalman\_arima method accounted for the temporal characteristics of continuous readings of 4 years of data like autocorrelation and seasonal trends that potentially improved the performance of the Kalman method for long consecutive missing intervals. The method performed well for stations with long consecutive missing gaps, irrespective of the percentage of missing values.

Furthermore, compared with the mean and median, LOCF performs well for < 20% level of missing; however, for higher missing percentage, the difference between imputed and observed mean increases, and the performance decrease significantly with high values of MAE, RMSE, and PBIAS%. The NAs values filled with the spline interpolation method accurately stimulate the temporal trend of particulate pollutants at the low level of missing < 5%; however, with increasing missing percentage, the evaluation metric PBIAS% provides negative values indicating high bias between imputed values and corresponding actual values.

**Table 1** Comparative evaluation of univariate time series imputations method using five different metrics for PM<sub>2.5</sub>. The stations with highest missing percentage under each category are considered and for other stations results are in supplementary file (Appendix 3)

% of missing		Stations ID's	Error metric	Mean	Median	LOCF	Spline	Kalman	Seadec
		<i>Sub-category of missing %</i>							
0–20%	0–5%	D_36 (4.05%)	MAE	3.28	2.44	2.00	1.75	1.87	2.13
			RMSE	18.03	13.38	13.51	11.63	12.02	13.58
			PBIAS%	3.40	2.50	2.10	1.60	1.95	2.20
			d	0.99	0.99	0.99	0.99	0.99	0.99
			R <sup>2</sup>	0.95	0.97	0.97	0.98	0.98	0.97
	5–10%	D_26(8.11%)	MAE	7.65	5.67	4.55	15.51	4.43	4.73
			RMSE	27.92	20.70	20.31	85.74	18.81	21.28
			PBIAS%	8.10	6.00	4.80	−6.00	4.70	5.00
			d	0.97	0.98	0.98	0.80	0.99	0.98
			R <sup>2</sup>	0.89	0.94	0.94	0.52	0.95	0.94
	10–15%	D_13 (12.57%)	MAE	9.85	7.39	5.55	53.98	1.84	5.08
			RMSE	29.41	22.07	17.06	204.41	8.26	16.51
			PBIAS%	12.60	9.50	7.10	−65.20	2.40	6.50
			d	0.95	0.97	0.98	0.47	1.00	0.99
			R <sup>2</sup>	0.85	0.92	0.95	0.31	0.99	0.95
	15–20%	D_31 (15.54%)	MAE	12.06	8.29	16.54	121.94	1.66	10.67
			RMSE	34.69	23.84	49.31	435.92	9.57	36.28
			PBIAS%	13.80	9.50	18.90	−137.50	1.90	12.20
			d	0.96	0.98	0.91	0.27	1.00	0.95
			R <sup>2</sup>	0.86	0.93	0.72	0.21	0.99	0.84
20–40%	NA	D_01 (25.00%)	MAE	22.37	16.17	38.68	39006.46	3.64	43.79
			RMSE	49.60	35.85	88.46	104243.60	23.05	98.88
			PBIAS%	25.50	18.50	44.10	44515.70	4.20	50.00
			d	0.91	0.95	0.77	0.00	0.98	0.73
			R <sup>2</sup>	0.76	0.88	0.36	0.15	0.94	0.28
40–60%	NA	D_06 (57.23%)	MAE	70.17	57.68	43.40	3205.77	4.46	61.67
			RMSE	92.75	76.24	61.31	5770.05	19.81	85.74
			PBIAS%	133.80	110.00	82.80	−5733.40	8.50	117.60
			d	0.70	0.78	0.85	0.02	0.99	0.74
			R <sup>2</sup>	0.45	0.64	0.72	0.17	0.94	0.47
60–80%	NA	D_18 (72.64%)	MAE	94.95	82.50	97.11	2454435.87	79.38	87.82
			RMSE	128.90	123.76	134.56	3466196.80	110.14	126.65
			PBIAS%	259.80	244.00	286.50	−8048964.10	241.00	251.60
			d	0.33	0.41	0.24	0.00	0.46	0.39
			R <sup>2</sup>	0.24	0.29	0.15	0.08	0.31	0.26

Correlation metric R<sup>2</sup> was used to estimate the relationship between daily observed and 24-h imputed concentrations of PM<sub>2.5</sub> and PM<sub>10</sub>. A high R<sup>2</sup> value indicates the significant performance of imputation methods, and a low value signifies the weak relationship between observed and imputed values. Mean, median, LOCF, Kalman\_arima, and sedec imputation method all produced higher R<sup>2</sup> values than the spline method, with Kalman producing the highest R<sup>2</sup> (Table 1). R<sup>2</sup> values for Kalman are around 0.99 at 0–5% missingness (D\_15), 0.95 at 5–10% missingness (D\_26),

0.99 at 10–15% missingness (D\_13), and 0.94 at 15–20% missingness (D\_31). At 40%, 60%, and 80% missingness, R<sup>2</sup> dropped significantly to around 0.85, 0.94, and 0.43, respectively for PM<sub>2.5</sub>. Similar results can be observed for PM<sub>10</sub> for the Kalman method in Table 2. The performance of spline is significant only for ≤ 5% missing percentage.

The index of agreement (d) measures the prediction error that varies from 0 to 1. A value of 0 indicates no agreement between the simulated and observed values, and 1 indicates a perfect match. Spline results for both pollutants

**Table 2** Comparative evaluation of univariate timeseries imputations method using five different metrics for PM10. The stations with highest missing percentage under each category are considered and for other stations results are in supplementary file (Appendix 4).

% of missing		Stations ID's	Error metric	Mean	Median	LOCF	Spline	Kalman	Seadec
<i>Sub-Category of missing %</i>									
0–20%	0–5%	D_08 (4.66%)	MAE	10.29	9.28	6.28	62.55	6.47	6.44
			RMSE	47.63	42.97	34.03	363.04	32.85	37.23
			PBIAS %	4.90	4.40	3.00	–27.20	3.10	3.10
			d	0.97	0.97	0.98	0.49	0.99	0.98
			R <sup>2</sup>	0.89	0.91	0.94	0.30	0.95	0.93
	5–10%	D_32 (9.08%)	MAE	15.73	13.61	8.48	36.00	3.66	9.39
			RMSE	50.25	43.49	30.16	158.06	20.80	33.00
			PBIAS%	10.90	9.40	5.90	–15.10	2.50	6.50
			d	0.94	0.96	0.98	0.72	0.99	0.98
			R <sup>2</sup>	0.82	0.86	0.93	0.45	0.97	0.92
	10–15%	D_31 (13.51%)	MAE	24.75	22.82	32.50	328.87	3.62	23.44
			RMSE	71.17	65.63	97.23	1224.22	20.24	74.77
			PBIAS%	13.80	12.70	18.10	182.60	2.00	13.00
			d	0.92	0.93	0.86	0.04	0.99	0.91
			R <sup>2</sup>	0.75	0.79	0.55	0.08	0.98	0.72
	15–20%	D_15 (17.23%)	MAE	23.34	21.40	19.11	32.27	15.76	19.49
			RMSE	64.63	59.25	69.13	99.14	52.62	62.61
			PBIAS%	15.00	13.70	12.30	20.50	10.10	12.50
			d	0.90	0.92	0.90	0.81	0.94	0.91
			R <sup>2</sup>	0.72	0.76	0.67	0.41	0.80	0.73
20–40%	NA	D_01 (25.81%)	MAE	40.86	36.80	50.32	112392.56	4.92	65.58
			RMSE	91.36	82.28	112.71	300312.52	31.22	150.85
			PBIAS%	25.00	22.50	30.80	68774.30	3.00	40.10
			d	0.87	0.90	0.82	0.00	0.99	0.72
			R <sup>2</sup>	0.66	0.73	0.50	0.22	0.95	0.25
40–60%	NA	D_27 (59.93%)	MAE	133.89	128.44	95.65	139.43	93.99	110.89
			RMSE	172.95	165.91	123.56	206.03	121.59	143.80
			PBIAS%	149.60	143.50	106.90	148.80	105.00	123.90
			d	0.65	0.67	0.78	0.57	0.79	0.73
			R <sup>2</sup>	0.35	0.41	0.71	0.12	0.74	0.57
60–80%	NA	D_18 (73.11%)	MAE	121.57	112.74	94.91	3885226.13	123.69	104.19
			RMSE	142.21	130.56	113.94	5486955.24	148.25	125.18
			PBIAS%	257.40	235.60	196.10	–8615733.80	274.30	218.90
			d	0.41	0.53	0.39	0.00	0.53	0.40
			R <sup>2</sup>	0.18	0.28	0.25	0.05	0.21	0.10

(Tables 1 and 2) show a significant drop in d values after 0–5 missing percentage categories. In some cases, a 0 value is observed for D\_22 and D\_05 (Table 1 and Appendix 3) and PM<sub>10</sub> stations D\_05 and D\_01 (Table 2 and Appendix 4). Mean, median, locf, and seadec all performed well except for 40–60% and 60–80%, missing the percentage category. Kalman gives the highest d values across all missing categories.

The error metric RMSE performed well for mean, median, locf, Kalman, and seadec for low missing percentage

categories (0–5%, 5–10%). RMSE, like other error metrics, increased with the level of missingness (Table 1). At 10–15%, 15–20%, 20–40%, 40–60%, and 60–80% missing, Kalman and the median performed well; however, spline imputation produced the highest RMSE values for the high missing category (Table 1 and Appendix 3). In the case of PM<sub>10</sub>, locf, median, and Kalman perform equally well for higher missing percentages (Table 2 and Appendix 4).

Another error metric is MAE which is less sensitive to outliers in predicted values. Similar to RMSE, the lower the

value, the better the performance of the simulated method. At 0–8% missing percentage, spline performed moderately well; however, performance declined after. Mean method performance yield high MAE values as the missing percentage increases. Seadec provides fluctuating results across all levels of missing. Median, locf, and Kalman performed well across all levels of missing, with Kalman at 20–80% missing, performing best, yielding the lowest MAE values of 1.66 (D\_31), at 15–20% missing, 3.64 (D\_01) at 20–40% missing, 4.46 (D\_06) at 40–60% missing, and 47.11 (D\_18) for 60–80% missing (Table 1). Similar results for PM<sub>10</sub> are observed in Table 2.

Based on univariate time series results, the worst performing method was spline, which consistently produced high errors at low and high missing percentages except for values  $\leq 5\%$  missing level. Even in some cases, PBIAS gives negative values and MAE, and RMSE is 3–4 times higher than considered univariate methods (Tables 1 and 2). Across all levels of missingness, kalamn\_arima provided the best estimates of 24-h mean PM<sub>2.5</sub> and PM<sub>10</sub> concentrations. At low levels of missingness (0–20%), the mean method performed well when evaluating the error metric for 1480 days. However, as missingness increases, performance becomes low. Median and locf imputation performed on average at higher levels of missingness (20–60%) (Tables 1 and 2). Henceforth, Kalman\_arima is the better-performing imputation method for use with PM<sub>2.5</sub> and PM<sub>10</sub> data from all univariate methods across all metrics.

We further worked on different multiple imputation methods using another characteristic of time series data, i.e. inter-variable/cross-sectional (described in the “Evaluation of spatial characteristics” section and the “Multivariate imputation” section). First, the target stations were selected for both variables with the highest missing percentage under four categories at 20%, 40%, 60%, and 80%. Then, to fill in missing values in target stations, neighbouring stations were chosen based on spatial correlation values using the Pearson coefficient. Finally, after framing the data matrix of 1480\*4, different MICE-based imputation methods were applied and evaluated based on five error metrics.

## Evaluation of spatial characteristics

### Selecting target station

Air pollutants possess strong spatial relationship, but such characteristic is scarcely used in imputing missing air pollutants data. A total of 38 monitoring stations deployed in Delhi can have highly stochastic spatial correlations among each other. To reconstruct incomplete data, the proposed study used spatial characteristics of all monitoring stations for two variables (PM<sub>2.5</sub> and PM<sub>10</sub>) to select

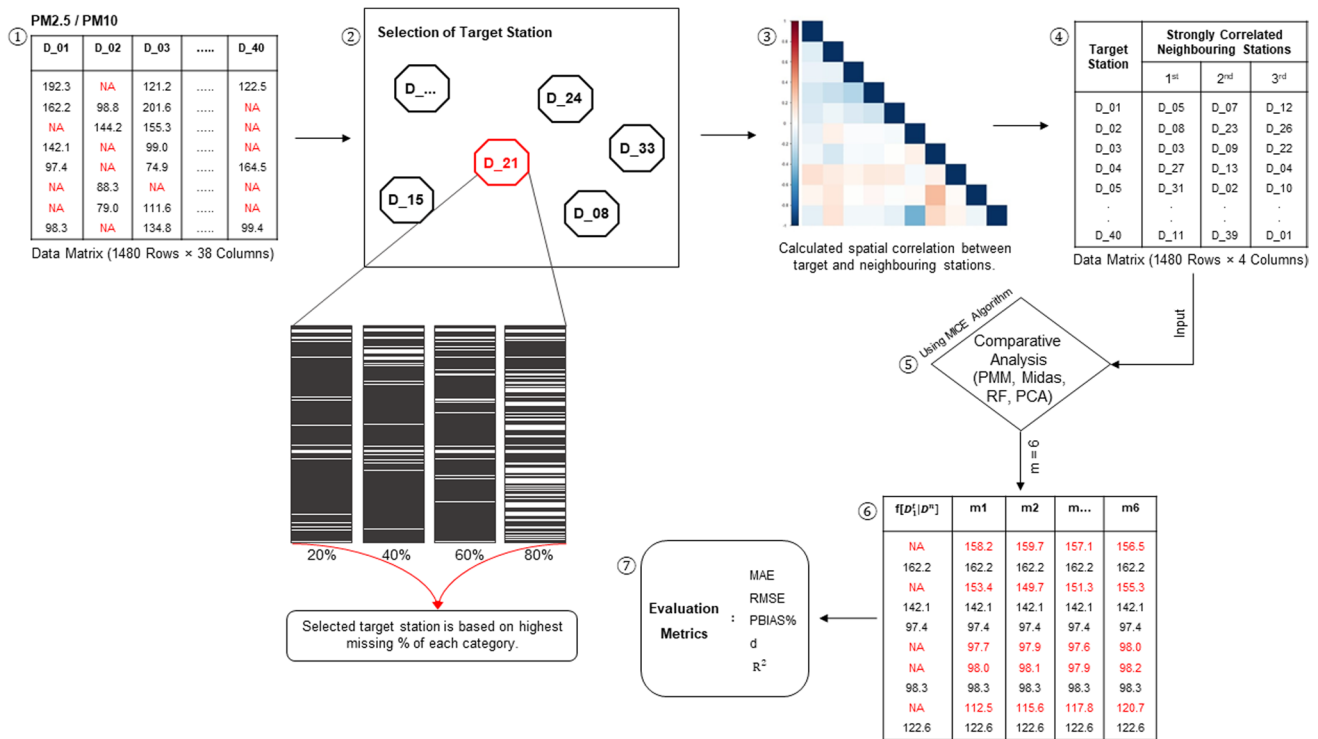
covariates and concurrently use these co-variables for multiple imputations to yield less biased estimates for high missing percentages (process mentioned in Fig. 4). The considered multivariate methods take advantage of the spatial relationship by using highly correlated sites as covariates to fill the missing values of target stations. There are two assumptions while selecting covariates, i.e. the combination of neighbouring stations has a significant correlation ( $> 0.50$ ) with the target site and has a low missing percentage compared to the target stations to be filled. This pre-processing step is advantageous when the target station’s sensors fail to get data on air pollutants. Then the lost values can be estimated from the neighbouring station’s data. Even though the raw dataset has several pollutants, we consider two variables, PM<sub>2.5</sub> and PM<sub>10</sub>, as target pollutants from a health perspective. The same pollutant PM<sub>2.5</sub>/PM<sub>10</sub> is combined from 38 monitoring stations, and the correlation coefficient is calculated to examine the spatial characteristics of air contaminants.

The selection of the target station is based on the highest missing percentage from each category, i.e. 20%, 40%, 60%, and 80%. For instance, target stations selected for PM<sub>2.5</sub> based on the above criteria are D\_31 (15.54%), D\_01 (25.00%), D\_06 (57.23%), and D\_18 (72.64%) and similarly for PM<sub>10</sub> D\_15 (17.23%), D\_01 (25.81%), D\_27 (59.93%), and D\_18 (73.11%) under 20%, 40%, 60%, and 80% missing percent category as mentioned in Table 3. Furthermore, to fill in missing data in target stations, a combination of covariates neighbouring stations is chosen based on their coefficient correlations with the target station.

### Spatial correlation coefficient of pollution data to select covariates for MICE algorithm

To analyse the overall spatial correlation of air pollutants data, a length of 1480 individuals and 38 columns are considered, with a significance level ( $\alpha$ ) of 0.01, to compute Pearson coefficient  $\rho$ . Examining PM<sub>2.5</sub> and PM<sub>10</sub> target pollutants coefficient of correlation among air monitoring stations is an important task in cross-sectional imputation. First, the data of the same target pollutant (PM<sub>2.5</sub>/PM<sub>10</sub>) is combined for all stations into a single data matrix of  $m \times n$ , where  $m = 1480$  (no. of individuals) and  $n = 38$  (representing columns as the number of monitoring stations) given in Fig. 4. Once the data is prepared in 1480\*38 for a single pollutant further, correlation coefficients are calculated through Pearson correlation formulae for each pollutant as given in Eq. 6. Pearson correlation determines the linear relationship between a pair of monitoring stations for two time series for the same pollutant of the same timestamp.





**Fig. 4** General description of spatial characteristics of monitoring stations to fill missing values using multivariate imputation techniques based on MICE algorithm

$$r(D^t, D^n) = \frac{\sum (D_i^t - \mu_t) \sum (D_i^n - \mu_n)}{\sqrt{\sum ((D_i^t - \mu_t)(D_i^n - \mu_n))^2}} \quad (6)$$

Suppose the  $D^t$  and  $D^n$  are two time series with sequences as  $(D_1^t, D_2^t, D_3^t, \dots, D_{i=1480}^t)$  and  $(D_1^n, D_2^n, D_3^n, \dots, D_{i=1480}^n)$  of  $PM_{2.5}/PM_{10}$  data for the target and neighbouring station of Delhi with the same time range (1 to 1480 days), respectively. Equation 6  $r(D^t, D^n)$  indicates the correlation coefficient of the target and neighbouring stations' time series for the same attribute.  $D_i^t$  and  $D_i^n$  represents  $i^{th}$  sampling points for  $D^t$  and  $D^n$ , respectively. Further  $\mu_t$  and  $\mu_n$  denotes the mean of both time series  $D^t$  and  $D^n$ , respectively, representing  $\mu_t = \frac{1}{1480} \sum_{i=1}^{1480} D_{1...1480}^t$  and

$\mu_n = \frac{1}{1480} \sum_{i=1}^{1480} D_{1...1480}^n$ . The numerator represents the covariation of both time series  $D^t$  and  $D^n$  from their mean values, and the denominator denotes the variance of  $D^t$  and  $D^n$ . Correlation coefficient value range between  $-1$  and  $1$ . If the value of  $r = 1$ , for  $D^t$  and  $D^n$  then stations are highly correlated, and if it is  $-1$  then they are negatively correlated, and the  $0$  value of  $r$  represents no obvious linear relationship between the two stations.

Appendices 6 and 7 depict the target station's correlation with different air monitoring stations for  $PM_{2.5}$  and  $PM_{10}$ , respectively. While analysing the correlation matrix, the results show that target station particulate pollution significantly correlates with neighbouring stations. The target stations having Pearson coefficient value of  $> 0.5$  suggests that particle pollution has significant spatial correlation and is little affected by meteorological and social activities. The correlation matrix for  $D^t$  and  $D^n$  shows that particle pollution in Delhi is not uniformly dispersed as the spatial correlation between  $D^t$  and  $D^n$  is independent of the geographical distance between the two as mentioned in Appendix 5. The obtained coefficient infers the strength of the relationship between  $D^t$  and  $D^n$  for  $PM_{2.5}$  and  $PM_{10}$  (Appendix 6 and 7). The negative coefficient values in  $PM_{2.5}$  and  $PM_{10}$  datasets correspond to a negative relationship. Appendix 6, the  $r$  value among  $PM_{2.5}$  monitoring stations range between  $-0.43$  and  $0.98$ . Pollutant  $PM_{2.5}$  correlations are

**Table 3** Criteria for selection of target station to impute missing values

Missing % category	Highest missing percentage under each category (%)		Target stations	
	PM <sub>2.5</sub>	PM <sub>10</sub>	PM <sub>2.5</sub>	PM <sub>10</sub>
0–20%	15.54	17.23	D_31	D_05
20–40%	25.00	25.81	D_01	D_01
40–60%	57.23	59.93	D_06	D_27
60–80%	72.64	73.11	D_18	D_18

strong for target station D\_06 with 31 neighbouring stations, except D\_01, D\_07, D\_18, and D\_31 have negative values. In contrast, the correlations with target stations D\_31, D\_01, and D\_18 are weaker with most of the paired neighbouring stations except for three (presented in bold). Appendix 7 shows that the correlation coefficients between  $D^f$  and  $D^n$  stations measuring  $PM_{10}$ , which range from  $-0.63$  to  $0.98$ , are more varied. Furthermore, based on the correlation coefficient analysis of each station, we draw different combinations of  $D^n$  for each  $D^f$ , as given in Table 4, which are eligible to deliver data to impute target sites. We carefully selected the three  $D^n$  with the highest correlation coefficients ( $> 0.50$ ) to  $D^f$ . Table 4 shows the selected neighbouring stations for  $PM_{2.5}$  and  $PM_{10}$ , sorted from highest to lowest  $\rho$  values.

### Selecting neighbouring station (covariates) to frame combination

After analysing the overall spatial correlation between the target and neighbouring stations (covariates) for the same pollutants, the next step is selecting neighbouring stations based on  $\rho$  values of paired stations ( $D^f$  and  $D^n$ ) and frame combinations. We selected the three highest  $\rho$  values between  $D^f$  and  $D^n$  and built a data matrix of  $1480 \times 4$  (one  $D^f$  and three  $D^n$ ) for multivariate imputation techniques. The selected neighbouring sites may differ for  $PM_{2.5}$  and  $PM_{10}$  as correlation coefficients vary, as given in Appendix 6 and 7. Based on Appendix 6 and 7 results, the number of neighbouring stations  $D^n$  (as covariates) are selected as input for multivariate methods based on the MICE algorithm. Three neighbouring stations are selected for each target column (in total, four columns are used as input)  $D^f$  based on the first three highly correlated neighbouring sites with  $D^f$  as given in Table 4. The columns of the input set contain data from the target station as well as data from nearby stations. Before finalising three neighbouring stations as input in multiple imputation methods, we evaluated the imputation results for the varying number of neighbouring stations (2 to 5) for  $PM_{2.5}$  and  $PM_{10}$ . It was analysed that performance is only significant if the number of neighbouring stations increased and decreased. However, it increases the computation load. Therefore, we use three neighbouring stations  $D^n$  in addition to the target station  $D^f$  for multivariate imputation techniques.

### Multivariate imputation

The data matrix obtained after spatial correlation analysis is  $1480 \times 4$  (1480 represents daily concentration values, and four refers to one  $D^f$  and three  $D^n$  columns) to impute target stations with the highest missing percentage from each category (20%, 40%, 60%, 80%) for  $PM_{2.5}$  and  $PM_{10}$  (target pollutants). The multivariate imputation methods like PMM, MIDAS, RF, and MIPCA are based on the MICE algorithm that uses target air pollutants data of  $D^n$ , neighbouring sites, which are chosen based on correlation coefficient ( $> 0.5$ ) to impute high missing percentage data of target stations. The highly correlated neighbouring sites (as covariates) are used to estimate high missing percentage data of ( $D^f$ ). The missing percentage for covariates is fixed at  $< 20\%$ . MICE-based methods provide multiple imputation values to missing data instead of just one, as in the case of the univariate method. Multiple imputation methods were used for each missing category, and evaluation metrics were used to compare the accuracy of the estimated results for  $PM_{2.5}$  and  $PM_{10}$  concentrations at each target station (Tables 5 and 6). One of the crucial features of MICE-based methods is it gives multiple imputed values for each target station missing values, as mentioned in Fig. 4 of step 6.

PMM, MIDAS, RF, and MIPCA consistently resulted in the lowest PBIAS values across 0–20% and 20–40% levels of missingness; however, for  $PM_{10}$  at the same missing levels, the values of PBIAS were slightly higher. At 40–60% and 60–80%, the PBIAS increases as missingness increases. The multivariate method outperformed in terms of low PBIAS error compared to univariate methods across all missing levels except for kalman\_arima in D\_06 (Table 1).  $R^2$  and d values are high at 20% and 40% missing levels for both variables, but the performance decreases as the missing percentage increases. The MIPCA method performed well for both variables. PMM, MIDAS, RF, and MIPCA showed good performance for both target pollutants  $PM_{2.5}$  and  $PM_{10}$  at 20% and 40%, with low error and high  $R^2$ . However, the performance of methods decreases as the level of missing increases.

From a spatial correlation perspective, PMM, MIDAS, RF, and MIPCA decrease substantially with a decrease in correlation (as given in Appendix 5). For all target stations

**Table 4** Covariates selection for each target station based on highest correlation coefficient

	PM <sub>2.5</sub>			PM <sub>10</sub>				
	D <sup>f</sup>	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	D <sup>f</sup>	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>
D_31	D_38	D_05	D_22	D_05	D_38	D_22	D_31	D_15
D_01	D_31	D_38	D_14	D_01	D_38	D_31	D_28	D_38
D_06	D_34	D_02	D_40	D_27	D_34	D_28	D_22	D_38
D_18	D_31	D_22	D_37	D_18	D_31	D_22	D_38	D_38

$D^n$  (1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup>) having significant spatial correlation with  $D^f$

**Table 5** Evaluation of multivariate imputation methods for PM<sub>2.5</sub> under four missing categories with highest missing percentage *D<sup>f</sup>*

% of missing	Target stations	Combination of neighbouring stations	Error metric	PMM	MIDAS	RF	PCA
0–20%	D_31	D_38, D_05, D_22	MAE	0.93	0.89	0.98	0.95
			RMSE	7.16	6.93	8.23	7.23
			PBIAS%	1.00	0.90	1.00	1.00
			d	1.00	1.00	1.00	1.00
			R <sup>2</sup>	0.99	0.99	0.99	0.99
20–40%	D_01	D_31, D_38, D_14	MAE	1.69	1.90	1.85	1.74
			RMSE	15.78	17.59	18.49	13.85
			PBIAS%	1.60	1.80	1.70	1.60
			d	0.99	0.99	0.99	0.99
			R <sup>2</sup>	0.97	0.96	0.96	0.98
40–60%	D_06	D_34, D_02, D_40	MAE	47.66	45.58	47.54	45.20
			RMSE	76.58	74.35	79.82	72.33
			PBIAS%	90.90	86.90	90.60	86.20
			d	0.78	0.79	0.77	0.80
			R <sup>2</sup>	0.49	0.51	0.45	0.54
60–80%	D_18	D_31, D_22, D_37	MAE	74.16	73.87	68.86	75.10
			RMSE	109.25	115.20	100.75	89.03
			PBIAS%	243.50	242.60	226.10	246.60
			d	0.54	0.51	0.58	0.63
			R <sup>2</sup>	0.14	0.11	0.18	0.44

**Table 6** Performance comparison of four MICE based imputation methods for PM<sub>10</sub> four missing categories with highest missing percentage *D<sup>f</sup>*

% of missing	Target Stations	Combination of neighbouring stations (covariates)	Error Metric	PMM	Midas	RF	PCA
0–20%	D_05	D_38, D_22, D_31	MAE	7.82	7.84	8.02	8.08
			RMSE	49.19	49.74	51.30	50.25
			PBIAS%	3.30	3.30	3.40	3.40
			d	0.97	0.97	0.97	0.97
			R <sup>2</sup>	0.90	0.89	0.89	0.89
20–40%	D_01	D_38, D_31, D_15	MAE	40.82	40.09	41.21	40.99
			RMSE	107.43	105.97	107.45	91.49
			PBIAS%	25.00	24.60	25.20	25.10
			d	0.84	0.85	0.84	0.87
			R <sup>2</sup>	0.53	0.54	0.53	0.66
40–60%	D_27	D_34, D_28, D_38	MAE	134.41	135.57	134.37	132.35
			RMSE	197.21	195.44	196.37	190.80
			PBIAS%	150.20	151.50	150.10	147.90
			d	0.59	0.59	0.59	0.65
			R <sup>2</sup>	0.15	0.16	0.15	0.36
60–80%	D_18	D_31, D_22, D_38	MAE	207.80	210.19	203.16	191.25
			RMSE	276.10	280.02	269.37	234.72
			PBIAS%	218.40	220.90	213.50	200.60
			d	0.52	0.51	0.53	0.60
			R <sup>2</sup>	0.09	0.08	0.10	0.30

(Tables 5 and 6), MIDAS and MIPCA imputation methods consistently produced the lowest MAE, RMSE, and PBIAS error across all levels of missingness; however, the error metrics values increased at 60–80% missingness. MIPCA performed best in all methods for all high missing percentage target stations. PMM performed well for target stations with < 50% missing percentage (Table 5).

### Statistical significance test (Wilcoxon Signed Rank Test): comparison of imputation methods

Since the Kolmogorov-Smirnov test revealed that many of our features originated from populations with non-normal distributions, the statistical Wilcoxon Signed Rank Test was used to demonstrate which method performed better (Cohen, Cohen et al. 2013). We have performed Wilcoxon signed rank test (one-tail) for comparison of univariate (mean, median, LOCF, spline, kalman\_arma, seadec.) and multivariate (PMM, MIDAS, RF, PCA) methods with varying missing percentages at the significance level = 0.05. Based on the five evaluation metrics results, we have formulated two null hypotheses to evaluate the consistency of performance of Kalman and the MIPCA imputation method for long gaps and high missing percentages. Null hypothesis ( $H_0^1$ ) is there is no significant difference in the performance of Kalman\_arma imputation methods compared to five univariate methods (mean, median, LOCF, spline, seadec.). Null hypothesis ( $H_0^2$ ) is there is no significant difference in the performance of MIPCA compared to all nine considered methods (six univariate and three multivariate), especially for high missing percentages and long gaps. To determine the likelihood of Kalman and MIPCA imputation methods performance on PM<sub>2.5</sub> and PM<sub>10</sub> datasets and we compared nonparametric test results of target methods (Kalman and MIPCA) with different imputation techniques at the missing percentage of 20% (5%, 10%, 15%, 20%), 40%, 60%, and 80%. This resulted in a total of 142 pairwise test (70 for Kalman and 72 for MIPCA) is included in Table 7 and 8. In all tests, the pairs the sets are ( $K_i, K_j$ ), where  $K_i$  is the target method (Kalman) and  $K_j$  ( $j = 1, 2, \dots, 5$ ) are the rest five univariate methods. For multivariate method testing, the pairs are ( $L_i, L_j$ ), where  $L_i$  is the target method (MIPCA) and  $L_j$  ( $j = 1, 2, 3, \dots, 9$ ) are nine univariate and multivariate methods. In Table 7, three conditions are considered (1)  $K_i > K_j$  (2)  $K_i \cong K_j$  (3)  $K_i < K_j$  and similarly in Table 8, (1)  $L_i > L_j$  (2)  $L_i \cong L_j$  (3)  $L_i < L_j$ , these conditions respectively denote the performance of target method ( $K_i$  and  $L_i$  is significantly better than  $K_j$  and  $L_j$ ; (2) there is no significant difference between the performances of  $K_i$  and  $K_j$ ,  $L_i$  and  $L_j$ ; and (3) the performance of  $K_i$  and  $L_i$  is significantly less than  $K_j$  and  $L_j$ . Table 7 illustrates the results of Kalman imputation; it was found that at the low

**Table 7** Comparison of kalman with five univariate methods based on Wilcoxon signed rank test

Missing percentage	$K_i > K_j$	$k_i \cong K_j$	$K_i < K_j$
5%	2	6	2
10%	3	5	2
15%	4	5	1
20%	4	4	2
40%	6	3	1
60%	8	2	0
80%	9	1	0
Total	<b>36</b>	<b>26</b>	<b>8</b>
Percentage of cases	<b>51.43%</b>	<b>37.14%</b>	<b>11.43%</b>

$p$ -value frequency ( $\alpha=0.05$ )

**Table 8** Comparison of MIPCA with nine imputation methods based on Wilcoxon signed rank test

Missing percentage	$L_i > L_j$	$L_i \cong L_j$	$L_i < L_j$
20%	10	6	2
40%	12	7	1
60%	14	4	0
80%	18	0	0
Total	<b>54</b>	<b>17</b>	<b>3</b>
Percentage of cases	<b>75%</b>	<b>23.61%</b>	<b>&lt;1%</b>

$p$ -value frequency ( $\alpha=0.05$ )

missing percentage (5%, 10%, 15%, and 20%) out of 40 tests, Kalman is statistically significant for 14 cases and in 20 cases, it performed equally well as Wilcoxon test statistic closely match null hypothesis, and in 6 instances Kalman showed  $p$ -value > 0.05. From Table 7, it is observed that at 5% missing class, from 10 tests, Kalman performed statistically significant with  $p$ -value < 0.05 for 2 cases ( $p$ -value = 0.034, 0.028) and performed equally well for 6 cases and in 2 cases, it performed less than  $K_j$  ( $p$ -value = 0.425, 0.163). In 10% missing class, out of 10 tests, the low performance of Kalman remains constant ( $p$ -value = 0.855, 0.879), the number of times Kalman performs equally with others decreases to 5, while the number of wins over others increases to 3 ( $p$ -value = 0.002, 0.012, 0.020). At 15% missing level, win cases for Kalman are 4 ( $p$ -value = 0.001, 0.018, 0.027, 0.010), loss cases decrease to 1 ( $p$ -value = 0.171), and performed equally for 5 cases. At 20%, the number of cases the Kalman lost increased to 2 ( $p$ -value = 0.171, 0.228), equally performed with others decreased to 4, but the winning frequency remains the same as previous ( $p$ -value = 0.025, 0.013, 0.001, 0.001).

For 40%, 60%, and 80% (high missing percentage with long gaps), the kalman method performed better with winning frequencies of 6, 8, and 9 at 40%, 60%, and 80%,



respectively. On the contrary, Kalman shows insignificant results for 1, 0, 0 times ( $p$ -value = 0.866) at 40%, 60%, and 80%, respectively. Table 8, the multivariate method (MIPCA) considers four classes 20%, 40%, 60%, and 80%. A total of 72 tests were implemented; at 20% onwards, MIPCA performed significantly better than others for 10 cases, equally well for 6 cases, and in 2 cases it lost ( $p$ -value = 0.879, 0.855). At 40%, 60%, and 80% missing classes, the winning cases are 12, 14, and 18, respectively, performed equally well for 7, 4, and 0 cases, respectively, and found statistically insignificant results for 1 ( $p$ -value = 0.526), 0, and 0 at 40%, 60%, and 80%, respectively. To summarise, in the case of Kalman (Table 7), beyond 20%, the performance increases with a rising missing percentage. The overall performance of Kalman is 51.43%, significantly better than other univariate methods; in 37.14% of cases, it is comparable to the other five univariate methods and of 11.43% cases, it performed insignificant to others. Similarly, for multivariate (MIPCA), the overall performance is quite significant compared to all nine methods from 20% onwards with a 75% success rate, 23.61% cases performed equally with others, and < 1% cases performed worse than the nine considered methods.

## Discussion

The missing data are commonly encountered in air pollution data due to calibration, repair of instruments, inconsistent power supply, and voltage fluctuation, which affects routine monitoring. Regular air monitoring is crucial to understand the impacts of policy changes and health impacts due to exposure and designing the daily alert system. In this context, we investigated approaches for dealing with missing data for low and high missing percentages. Among the six univariate methods evaluated across several metrics, the Kalman\_arima method performed well across all missing levels except for 60–80%. Low missing percentage < 5% spline produced the best 24-h mean estimates with the lowest error and highest  $R^2$  and  $d$  values. Mean, median, and lofc imputation methods performed inconsistently for input time series in terms of percent missing. Seadec. It also provides uncertain results for stations with a high missing percentage. Kalman\_arima imputation method performed exceptionally well for time series data for both variables with strong seasonality and trend. It may be a realistic option for imputing time-series data of this type. Effective univariate time series imputation algorithms utilised inter-time correlation characteristics. The performance of imputation is always highly dependent on the characteristics of the input time series, and therefore, for time series with strong seasonality, Kalman\_arima performed well with statistically significant cases for about 51.43% cases, where  $p$ -value for

higher missing percentage and long gaps were lower than 0.05 (alpha value).

The multivariate imputation methods PMM, MIDAS, RF, and MIPCA performed much better for target stations, with the highest missing percentage under 20% and 40% missing categories. At 60% and 80% performed of PMM, MIDAS, and RF significantly dropped. However, MIPCA performed highly significant than Kalman\_arima at 60% and 80%. This is because multivariate imputation methods impute missing values from observed concentrations of neighbouring stations (covariates) with similar spatial characteristics. Comparison of MIPCA with nine imputation methods based on Wilcox signed rank test  $p$ -value frequency ( $\alpha = 0.05$ ), also suggest that above results based on evaluation metrics are significant. MIPCA performed highly statistically significant compared to other univariate and multivariate methods, as  $p$ -value were lower than < 0.05.

Interestingly, MIPCA performed well across all levels of missingness. One possible explanation for the superior performance of MIPCA imputation from univariate imputation is the spatial correlation characteristics of the monitoring station. The MICE algorithm captures the spatial characteristics of the neighbouring station's time series data to fill target stations missing values. This would explain why univariate methods using partial data observed within each station (Inter-time correlations) yielded low estimates of PM concentrations, especially for high missing percentages, compared to multivariate methods using spatial characteristics (cross-sectional relationship) between target and neighbouring stations.

Relatively median showed good performance was an unexpected finding given that this method can yield biased estimates under MAR and is generally discouraged. One reason for mean imputation's success could be that the partially observed data within each station is distributed positively skewed (Miettinen 2012, Junger and De Leon 2015, Hadeed et al. 2020). The mean imputation method performed well for low missing levels, but the results dropped as the missing percentage increased. Moreover, the method provides a constant value for a missing gap, which brings biasness in a time series with seasonal characteristics like an air pollution dataset. Therefore, this method is highly discouraged for the MAR dataset (Quinteros et al. 2019). In time series data, the first-order Markov model assumes a similar condition as the LOCF imputation method (Canales 2004). LOCF fills the missing values with  $t-1$  observed values. LOCF performed well for short interval gaps; however, for long gap size, especially in the case of time series data having temporal characteristics like seasonality and trend, this method's performance drops (Moritz and Bartz-Beielstein 2017). The performance of seadec. fluctuates and needs to consider the temporal component of the time series accurately as the level of missing increases. The method utilized partial

data observed within each station to predict concentrations missing within the same station, which may be effective for monitoring stations located in areas completely different from one another.

Kalman\_arima univariate time series approach can work efficiently for monitoring stations located in a resource-constrained location with one/ two stations. The temporal characteristics to impute values based on inter-time correlation, like seasonality, may explain this method's success. In addition, the method considers some aspects of the structural nature of the time series and computes the probability of concentrations based on lagged observed values. In addition, to the inter-time correlation component of time series, the multivariate time series algorithms consider correlation among covariates. PMM performed better for target stations with < 50% missing percentage (Marshall et al. 2010a, b). MIDAS performed better than PMM as it selects donor values based on distance and is a higher variant of PMM. RF performed slightly better than PMM and MIDAS for 40–60% and 60–80%. The following high iteration number ensures the algorithm imputes 24-h  $PM_{2.5}$  and  $PM_{10}$  missing data with low error and high  $R^2$  and  $d$  values.

One of the limitations of the proposed study is that it used the time series characteristics like inter-time correlation and inter-variable (cross-sectional) correlation separately for univariate time series and multivariate time series. The performance for a high missing percentage can be improved if both the characteristics, i.e. TS cross-sectional (inter-variable + inter-time) taken into account simultaneously. Also, univariate methods like mean, median, and locf fail to capture temporal events like trend and seasonality. This could have consequences in designing an alert system where specific tasks may be associated with high pollutant concentrations. Such a univariate imputation method may only account for these high pollutant tasks if the concentration of pollutants is observed. Imputed values are based on partial observations, which can result in under- or overestimation. Thus, considering this limitation of simple univariate methods, kalamn\_arima appears to be a viable option for imputing missing data in samples and populations with high heterogeneity, and multivariate imputation method MIPCA used spatial characteristics of monitoring stations to give better results for high missing percentage target stations.

There are only a few comparable studies in the context of ambient air monitoring data over the long-time interval. The existing studies are tailored to impute missing data, which are frequently preceded and followed by observed data for short periods. However, in our proposed, there are long consecutive periods of missing data at a fixed time interval, and also, in some cases,

the missing data have random start points in a column. Thus, this unique nature of missing data can explain the difference between our proposed study's and previous studies' imputation performance. The advantage of our study is its large sample size; however, it is limited to ambient particulate pollution in densely populated urbanised regions. Therefore, the results for other pollutants can vary depending on the contaminant, sampling duration, and co-variates selected.

For instance, (Hadeed et al. 2020) compared imputation methods for short-term monitoring of  $PM_{2.5}$ . Their study used datasets with varying degrees of missingness to assess the performance of mean, median, random, Markov, LOCF, Kalman, PMM, and RMM. The three univariate methods, random, Markov, and mean outperformed all missing levels. The multivariate method used predictor variables such as type of fuel, geographic location, relative humidity, and temperature to impute  $PM_{2.5}$  missing concentration. However, PMM, and RMM performed worse as the covariates selected are heterogeneous. However, our study contains the following: Continuous data of 4 years which allows the seasonality and trend to be accounted for, potentially improving the univariate time series method Kalman\_arima as it requires inter-time correlation to work efficiently. In addition, the study employed spatial correlation characteristics of multiple stations to select covariates to impute target stations missing concentration, and MIPCA performed well across all missing levels.

(Junger and De Leon 2015) explored missing levels ranging from 5 to 40% in their study. Twelve imputation methods, including univariate and multivariate methods, were evaluated (complete case analysis, mean, median, nearest neighbour, EM models, ARIMA, general additive models, and spline models). For 366 days, each day concentrations were collected from 10 sampling sites, and found concentrations were strongly correlated between stations. Even with low levels of missingness, conditional mean performed well, whereas unconditional imputation (median, mean) performed poorly. Multivariate methods such as conditional mean and EM-models performed remarkably well as the frequency and interval of missing data increased. Our study results were similar to those published by (Junger and De Leon 2015); however, in our study the distribution is skewed, the median performed better at low missing levels than mean. The improved performance is attributed to the inclusion of a longitudinal component and the high inter-site correlation among monitoring stations. The study of long-term monitoring data from multiple sites in our analysis could explain the commonalities in imputation performance. More complex multivariate imputation methods like PMM, MIDAS, RF, and MIPCA impute concentrations performed well based on information observed at other monitoring sites.

## Conclusion

Imputation methods presented in this study yield reliable results for high missing concentrations and long consecutive gaps. The overall performance of kalamn\_arima methods is influenced by missing concentration statistics like missing gap size and missing percentage. For a long consecutive missing gaps, Kalman\_arima performs well irrespective of missing percentage, and long gaps specially for 1016 consecutive values and high missing percentages, MIPCA gives significant results. The relevant imputation technique for missing concentration should be chosen based on missing data statistics and variables temporal and spatial characteristics. The imputation methods can be useful in resource-constrained developing nations where still manual monitors are used, or a smaller number of continuous monitoring stations are deployed, power cuts, sensor failures, etc. However, there has been little work done in evaluating current methods for filling missing data for long consecutive gaps and high missing percentages. Our finding can help reconstruct long missing gaps and high missing percent data for deep learning prediction algorithms that requires continuous time series data without gaps for designing the alert system. Regardless of the finding, more studies are required to investigate and identify appropriate imputation technique that are pervasive to various settings.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11356-023-27659-x>.

**Acknowledgements** The authors are thankful to CSIR-CSIO Chandigarh for providing the necessary infrastructure and support to carry out our work and especially thanks to the Central Pollution Control Board for making air pollutants data openly available to their official website.

**Author contribution** All authors have contributed substantially to the following: Priti K performed writing—original draft, conceptualization, data curation, formal analysis, and methodology. Kaushlesh Singh Shakya performed data curation, formal analysis, and visualization. Prashant kumar: supervision, validation, and review and editing.

**Data availability** Air pollutants data is freely available on the website of Central Pollution Control Board, India. Any question/ inquires pertaining to the type of data or the attributes of datasets, can be directed to the corresponding author, upon reasonable request.

**Code availability** NA

## Declarations

**Ethics approval** NA

**Consent to participate** NA

**Consent for publication** NA

**Conflict of interest** The authors declare no competing interests.

## References

- Abayomi K, Gelman A, Levy M (2008) Diagnostics for multivariate imputations. *J R Stat Soc Ser C Appl Stat* 57(3):273–291
- Agbailu AO, Seno A, Clement OO (2020) Kalman filter algorithm versus other methods of estimating missing values: time series evidence. *Studies* 4(2):1–9
- Allison P (2015) Imputation by predictive mean matching: promise & peril. *Statistical Horizons*
- Allison PD (2001) *Missing data*. Sage publications
- Aslan S (2010) Comparison of missing value imputation methods for meteorological time series data. MS thesis, Middle East Technical University
- Audigier V, Husson F, Josse J (2016) Multiple imputation for continuous variables using a Bayesian principal component analysis. *J Stat Comput Simul* 86(11):2140–2156
- Benavides IF, Santacruz M, Romero-Leiton JP, Barreto C, Selvaraj JJ (2022) Assessing methods for multiple imputation of systematic missing data in marine fisheries time series with a new validation algorithm. *Aquac Fish J*
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Budhiraja B, Gawuc L, Agrawal G (2019) Seasonality of surface urban heat island in Delhi city region measured by local climate zones and conventional indicators. *IEEE J Sel Top Appl Earth Obs Remote Sens* 12(12):5223–5232
- Canales RA (2004) The cumulative and aggregate simulation of exposure framework. Stanford University
- Chan M (2015) Achieving a cleaner, more sustainable, and healthier future. *The Lancet* 386(10006):e27–e28
- Chatterji A (2021) Air pollution in delhi: filling the policy gaps. *Massach Undergr J Econ* 17
- Cho B, Dayrit T, Gao Y, Wang Z, Hong T, Sim A, Wu K (2020) Effective missing value imputation methods for building monitoring data. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE
- Cohen J, Cohen P, West SG, Aiken LS (2013) *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge
- Crawley MJ (2012) *The R book*. John Wiley & Sons
- Doove LL, Van Buuren S, Dusseldorp E (2014) Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput Stat Data Anal* 72:92–104
- Dray S, Josse J (2015) Principal component analysis with missing values: a comparative survey of methods. *Plant Ecol* 216(5):657–667
- Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW (2012) Missing data: a systematic review of how they are reported and handled. *Epidemiology* 23(5):729–732
- Gaffert P, Meinfelder F, Bosch V (2018) Towards multiple-imputation-proper predictive mean matching. *JSM*:1026–1039
- Ghazali SM, Shaadan N, Idrus Z (2020) Missing data exploration in air quality data set using R-package data visualisation tools. *Bull Electr Eng Inform* 9(2):755–763
- Gómez-Carracedo MP, Andrade J, López-Mahía P, Muniategui S, Prada D (2014) A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometr Intell Lab Syst* 134:23–33
- Hadeed SJ, O'Rourke MK, Burgess JL, Harris RB, Canales RA (2020) Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Sci Total Environ* 730:139140
- Han H, Sun M, Han H, Wu X, Qiao J (2023) Univariate imputation method for recovering missing data in wastewater treatment process. *Chin J Chem Eng* 53:201–210
- Harvey AC (1990) Forecasting, structural time series models and the Kalman filter
- Huisman M (2009) Imputation of missing network data: some simple procedures. *J Soc Struct* 10(1):1–29

- Iodice D'Enza A, Markos A, Palumbo F (2022) Chunk-wise regularised PCA-based imputation of missing data. *Stat Methods Appt* 31(2):365–386
- John C, Ekpenyong EJ, Nworu CC (2019) Imputation of missing values in economic and financial time series data using five principal component analysis approaches. *CBN J Appl Stat (JAS)* 10(1):3
- Josse J, Husson F (2009) Gestion des données manquantes en analyse en composantes principales. *Journal de la société française de statistique* 150(2):28–51
- Josse J, Husson F (2011) Multiple imputation in principal component analysis. *Adv Data Anal Classif* 5(3):231–246
- Josse J, Husson F (2016) missMDA: a package for handling missing values in multivariate data analysis. *J Stat Softw* 70:1–31
- Junger W, De Leon AP (2015) Imputation of missing data in time series for air pollutants. *Atmos Environ* 102:96–104
- Junior JRB, do Carmo Nicoletti M, Zhao L (2016) An embedded imputation method via attribute-based decision graphs. *Expert Syst Appl* 57:159–177
- Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. *Atmos Environ* 38(18):2895–2907
- Kalman RE (1960) A new approach to linear filtering and prediction problems. *Trans ASME J Basic Eng* 82:35–45
- Kleinke K (2018) Multiple imputation by predictive mean matching when sample size is small. *Methodology: Euro J Res Methods Behav Res Methods* 14(1):3
- Kumar P (2022) A critical evaluation of air quality index models (1960–2021). *Environ Monit Assess* 194(4):1–45
- Legates DR, McCabe GJ Jr (1999) Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour Res* 35(1):233–241
- Li KH, Le ND, Sun L, Zidek JV (1999) Spatial–temporal models for ambient hourly PM10 in Vancouver. *Environmetrics: the official journal of the Int Environ Sci* 10(3):321–338
- Little RJA, Rubin DB (2019) *Statistical analysis with missing data*, vol 793. John Wiley & Sons
- Little RJA, Rubin DB (2002) Single imputation methods. *Statistical analysis with missing data*. p 59–74. <https://doi.org/10.1002/9781119013563.ch4>
- Liu X, Wang X, Zou L, Xia J, Pang W (2020) Spatial imputation for air pollutants data sets via low rank matrix completion algorithm. *Environ Int* 139:105713
- Lloret J, Lleonart J, Solé I (2000) Time series modelling of landings in Northwest Mediterranean Sea. *ICES Mar Sci Symp* 57(1):171–184
- Marshall A, Altman DG, Holder RL (2010a) Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Med Res Methodol* 10(1):1–10
- Marshall A, Altman DG, Royston P, Holder RL (2010b) Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol* 10(1):1–16
- Miettinen OS (2012) Theoretical epidemiology: principles of occurrence research in medicine. *Theoretical epidemiology: principles of occurrence research in medicine*:359–359
- Molenberghs G, Kenward M (2007) *Missing data in clinical studies*. John Wiley & Sons
- Moriassi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans ASABE, Appl* 50(3):885–900
- Moritz S, Bartz-Beielstein T (2017) ImputeTS: time series missing value imputation in R. *R J* 9(1):207
- Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, Stork J (2015) Comparison of different methods for univariate time series imputation in R. *arXiv preprint arXiv:1510.03924*
- Norazian MN, Shukri YA, Azam RN, Al Bakri AMM (2008) Estimation of missing values in air pollution data using single imputation techniques. *SciAsia* 34(3):341–345
- Plaia A, Bondi A (2006) Single imputation method of missing values in environmental pollution data sets. *Atmos Environ* 40(38):7316–7330
- Quinteros ME, Lu S, Blazquez C, Cárdenas-R JP, Ossa X, Delgado-Saborit J-M, Harrison RM, Ruiz-Rudolph P (2019) Use of data imputation tools to reconstruct incomplete air quality datasets: a case-study in Temuco, Chile. *Atmos Environ* 200:40–49
- Ramli MN, Yahaya A, Ramli N, Yusof N, Abdullah M (2013) Roles of imputation methods for filling the missing values: a review. *Adv Environ Biol* 7(12 S2):3861–3870
- Raymond MR (1986) Missing data in evaluation research. *Eval Health Prof* 9(4):395–420
- Rubin DB (1976) Inference and missing data. *Biometrika* 63(3):581–592
- Schafer JL (1997) *Analysis of incomplete multivariate data*. CRC press
- Schenker N, Taylor JM (1996) Partially parametric techniques for multiple imputation. *Comput Stat Data Anal* 22(4):425–446
- Siddique J, Belin TR (2008) Multiple imputation using an iterative hot-deck with distance-based donor selection. *Stat Med* 27(1):83–102
- Siddique J, Harel O (2009) MIDAS: a SAS macro for multiple imputation using distance-aided selection of donors. *J Stat Softw* 29:1–18
- Stekhoven DJ, Stekhoven MDJ (2013) Package ‘missForest’. R package version 1
- Sukatis FF, Noor NM, Zakaria NA, Ul-Saufie AZ, Annas S (2019) Estimation of missing values in air pollution dataset by using various imputation methods. *Int J Conserv Sci* 10(4):791–804
- Tsikriktsis N (2005) A review of techniques for treating missing data in OM survey research. *J Oper Manag* 24(1):53–62
- Van Buuren S (2018) *Flexible imputation of missing data*. CRC press
- Van Buuren S, Groothuis-Oudshoorn K (2011) Mice: multivariate imputation by chained equations in R. *J Stat Softw* 45:1–67
- Wardana I, Gardner JW, Fahmy SA (2022) Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder. *Neural Comput Appl*:1–26
- Weerakody PB, Wong KW, Wang G, Ela W (2021) A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing* 441:161–178
- Welch G (2006) An Introduction to the Kalman Filter. Univ. of North Carolina [http://www.cs.unc.edu/~welch/media/pdf/kalman\\_intro.pdf](http://www.cs.unc.edu/~welch/media/pdf/kalman_intro.pdf). Accessed 10 Oct 2022
- Wijesekara W, Liyanage L (2020) Comparison of imputation methods for missing values in air pollution data: case study on Sydney air quality index. In: *Future of Information and Communication Conference*. Springer
- Willmott CJ (1981) On the validation of models. *Phys Geogr* 2:184–194
- Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Res* 30(1):79–82
- Willmott CJ, Matsuura K (2006) On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *Int J Geogr Inf Sci* 20(1):89–102
- World Health Organization (2016) *Ambient air pollution: A global assessment of exposure and burden of disease*
- Zeileis A, Grothendieck G (2005) zoo: S3 Infrastructure for Regular and Irregular Time Series. *J Stat Softw* 14(6):1–27

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.