Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3037

THE APPLICATION OF MULTIPLE IMPUTATION METHOD BASED ON HYBRID MULTI-STRATEGY IN HANDLING MISSING AIR QUALITY MONITORING DATA

ZhiQuan Zheng[1*], WenYong Zhang[1,2], ZhongChen Luo[3]
[1]*School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, Guizhou, China.*
[2]*School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, Guangdong, China.*
[3]*School of Nursing, Guizhou Medical University, Guiyang 550001, Guizhou, China.*
Corresponding Author: ZhiQuan Zheng, Email: zhengzhiquan@gzmu.edu.cn

Abstract: Air quality monitoring data is a crucial basis for assessing air pollution levels and formulating control measures. However, missing data is a prevalent issue due to instrument malfunctions, human factors, and other reasons, significantly compromising data integrity and usability. To address this problem, this study collected nearly 1 million air quality monitoring records from 12 monitoring stations between 2015 and 2023, summarizing and analyzing the mechanisms and characteristics of missing data in such datasets. Data imputation experiments were conducted using R. Through missing mechanism control and imputation experimental design strategies, the imputation performance of algorithms was evaluated under the criteria of MAE, RMSE, and WMAPE based on completely random missingness. Specifically, data imputation experiments under different missing scenarios were repeated N times, and the mean values were used to evaluate four multiple imputation algorithms, with 95% confidence intervals provided. The experimental results show that: (1) the hybrid multi-strategy imputation method MNPRF demonstrates significant advantages across all datasets, with the smallest confidence limits and interval widths; (2) this method not only inherits the strengths of parent algorithms, substantially improving data quality, but also mitigates the weaknesses of the original algorithms to some extent.
Keywords: Air quality monitoring; Missing mechanism; Data imputation; Confidence interval; Multiple imputation; Hybrid multi-strategy imputation

1 INTRODUCTION

At present, with the increasing application of information analysis, data mining, and neural network model training in various industries, data loss has become an important problem in most application fields, such as statistical investigation [1], environmental protection [2], medical research [3], etc. Some studies have shown that more than 40% of the data sets in the international open database UCI have missing observations [4-5]. There are many reasons for missing data. Taking air quality monitoring data as an example, data may be lost due to equipment failure, environmental conditions, human management and other factors, such as sensor failure or damage caused by extreme weather conditions, data transmission loss caused by communication equipment failure, and failure to properly process and collect caused by software errors. In addition, the particularity of geographical location will sometimes affect the stability of the monitoring equipment, resulting in the inability to continuously record data. Missing data will not only lead to deviations in statistical results, but also lead to the unavailability of the original model [6].

The processing methods of data missing mainly include deletion method and filling method. Considering different missing scenarios, there are many specific ways to delete data, such as deleting columns with missing values, or deleting observed samples with missing values. However, while the deletion method is simple, it can miss useful information in the original data set and lead to incorrect statistical results. For this reason, statisticians and scholars in related fields do not recommend the use of erasure to deal with missing data. Aiming at the problem of missing data, many scholars devote themselves to the research of missing value filling, and have laid the theoretical foundation of this problem in statistics [7-9].

For the same missing value, the Data Imputation Algorithms is classified according to the number of its filling values, and the Data Imputation Algorithms is divided into two categories: Single Imputation(SI) and Multiple Imputation(MI)[10]. Single value filling includes mean value filling, mode filling, regression filling, etc. These methods can effectively solve the problem of missing data due to their advantages of high computational efficiency and strong interpretability. However, single value filling fills only one possible estimate for each missing value, ignoring the uncertainty of the missing data, and such methods will change the original distribution of the data, resulting in the distortion of the statistical characteristics of the data (such as variance and covariance). Multiple Imputation can effectively reflect the uncertainty of the data while dealing with the missing data. As a method used to process missing data, the core principle of Multiple Imputation is to fill in missing data by generating multiple possible interpolation values, and improve the accuracy and reliability of statistical analysis. Based on Bayesian statistics and sampling theory, the method generates Multiple Imputation values from the posterior distribution of missing data and simulates the possible distribution of missing data, thus improving the accuracy of estimates. The implementation methods of

Multiple Imputation fall into two main categories: Joint Modeling (JM) and Fully Conditional Specification (FCS) [11]. Joint modeling assumes that all variables (including missing and observed variables) obey some joint distribution (such as a multivariate normal distribution) and generates interpolation values from that distribution [8,12-13]. The Complete conditional gauge (FCS) is an iterative interpolation method, also known as Multiple Imputation by Chained Equations (MICE) [14]. Its core idea is to construct conditional models for each missing variable, such as regression model, random forest, etc., and update the interpolation value through iteration [15]. As a filling idea, Multiple Imputation not only reflects the uncertainty of missing data, but also can handle multiple types of data and is applicable to different Missing Data Mechanisms. In subsequent studies, scholars have proposed different versions of JM and FCS methods [16-22].

2 ALGORITHM INTRODUCTION AND IMPROVEMENT

2.1 Introduction of R Mice Package

The R mice package is an implementation tool for Multivariate Imputation by Chained Equations that is very useful in working with missing data. The package allows users to interpolate different types of variables through multiple models, thus improving the quality and reliability of data analysis. On the R platform, the mice package is installed through the install.packages() function and loaded using the library() function. The mice package provides a series of methods for managing and analyzing datasets that contain missing values. The main process consists of several stages such as creating multiple interpolating objects, performing the actual interpolating process, and summarizing the results. The main parameters of the mice() function include dataframe (the data set to be filled), m (the number of interpolations), maxit (the upper limit of iterations), method (the name of the specific algorithm used), and seed (random seed setting to ensure reproducibility).

Once the mice() function has been filled, you can use the complete() function to obtain one of the complete datasets, or you can use the loop structure to traverse all possible combinations of results. This paper uses norm.predict (Regression Prediction method) and RF (Random Forest) filling algorithms in the mice package of R language to perform data filling experiments on air quality monitoring data, and tries to improve the algorithm.

2.2 Norm.Predict Filling Algorithm Based on Multiple Imputation (MNP)

The MNP method is a realization of the regression prediction method. When dealing with missing data, this method utilizes complete covariates to construct a linear model and predict the missing values. Specifically: for the target variable with missing values, a linear regression model under a multivariate normal distribution is first established using the data without missing values. Then, for each observation record with missing values, the expected mean μ and standard deviation σ are obtained by substituting the known covariate values into the above trained model. Finally, a random number is drawn from the normal distribution $N(\mu,\sigma^2)$ as the filling result. This method can well maintain the original data structure characteristics while introducing reasonable uncertainty estimation.

2.3 Random Forest Filling Algorithm Based on Multiple Imputation (MRF)

Random forest is widely used in the field of machine learning. As one of the classical classification algorithms, it has good robustness and accuracy. The algorithm evolved from the decision tree, reduces the risk of overfitting in the decision tree, and is not sensitive to noise or outliers in the data set, so it has good prediction and generalization ability. The MRF method combines the idea of Multiple Imputation with a random forest model in the field of machine learning to estimate missing values by building multiple decision trees. For each tree, the importance of multiple features is taken into account during the node splitting process, and the model is trained using the unmissing data. When a missing worth sample is encountered, the most likely value is deduced according to the existing feature information as the filling result. This process is repeated many times to produce a stable and reliable prediction.

2.4 Characteristics Analysis of Air Quality Monitoring Data

The incomplete air quality monitoring data is mainly caused by the failure of acquisition equipment and other reasons, which brings difficulties to further experimental analysis. It is very important to analyze the Missing Data Mechanisms and reason of the data and choose the appropriate filling algorithm for its processing. In 1987, Little and Rubin [8] proposed the concepts of Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR), classifying the complex and diverse reasons for data missing. However, in view of the specific data missing problem, it is often necessary to explore the correlation among variables and the distribution of missing values in the data set to be filled. For this reason, before filling in the data, this paper pre-analyzed the missing rate of about 1 million air quality monitoring data from 12 stations under different observation indicators, each of which had 7 monitoring indicators, and plotted a stack bar chart (left of Figure 1) and a percentage stack bar chart (right of Figure 1). The experimental results are shown in Figure 1:

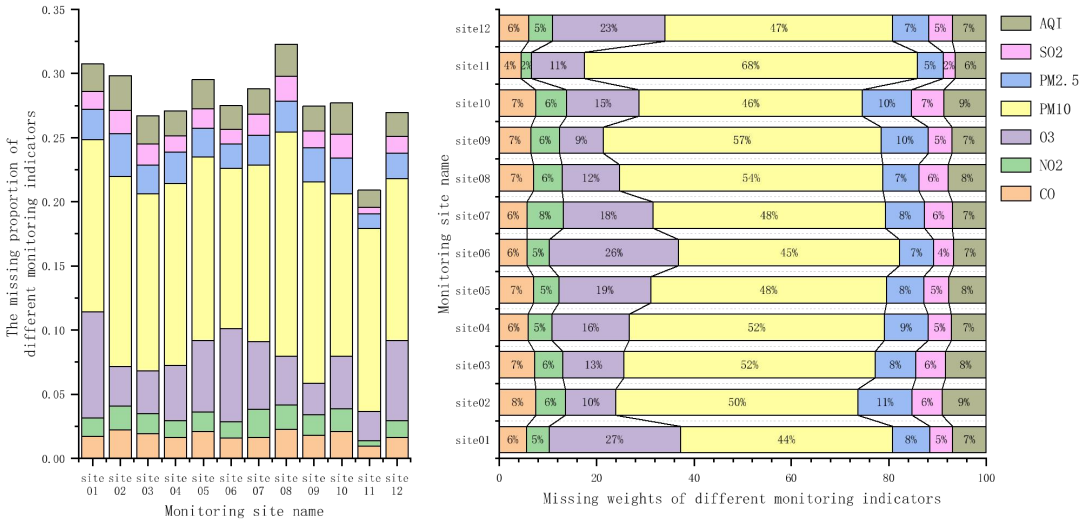


Figure 1 The Proportion of Missing Values of Different Observation Indicators at Different Monitoring Stations

The results of Figure 1 show that: First, in the air quality monitoring data, not only do all monitoring dimensions have missing data, but the missing rates of different monitoring dimensions are different in the air quality monitoring data of each station; Second, there are certain similarities in the missing rate of different stations under the same observation index. As can be seen from the experimental results in right of Figure 1, the missing rate of observation indicators CO, NO2, PM2.5, SO2 and AQI accounted for 5%-10% of the cumulative missing rate in all stations, while the missing rate of PM10 accounted for half of the total missing rate. Third, the experimental results of left of Figure 1 show that in all sites, the missing rate of monitoring indicator PM10 is above 15%, while the missing rate of the remaining vast majority of monitoring indicators is below 5%. In summary, for the air quality monitoring data, the missing trend of the internal observation indicators follows a similar rule, that is, in the air quality monitoring data of each station, only a few missing values of the monitoring dimension account for a relatively high proportion, more than 15%, while the missing values of the remaining most dimensions account for a very low proportion, less than 5%.

2.5 Hybrid Multi-strategy Interpolation Method Based on MICE

According to the missing proportion of different observed variables in the data set to be filled, a more targeted Data Imputation Algorithms is adopted to fill in different dimensions of the original data, which is the core idea of the algorithm improvement in this paper. Since the missing proportion of different monitoring dimensions of air quality data is highly differentiated, and there is a certain correlation between some observed variables, combined with the proportion distribution of missing values in each monitoring dimension of the data set to be processed, the observed variable with the largest missing proportion is given priority to be filled, and based on the initially filled data set, Another Data Imputation Algorithms was used to fill in the remaining observed variables. Considering the general filling effect of the algorithm, this paper improves the algorithm based on MRF and MNP. For the column with the largest proportion of missing in the same data set, the above two algorithms are used to fill in the first stage, and the result set after the initial filling is filled with another algorithm to fill in the missing values in the remaining observed variables. According to the processing order of the same data set by different filling algorithms, MRFNP(Random forest imputations and Linear regression mixed filling based on Multiple Imputation) and MNPRF(Linear regression and Random forest based on Multiple Imputation) are proposed imputations Mixed filling) algorithms are proposed. For ease of description, M represents the data set to be processed, $M_{i,j}, i \in [1,n], j \in [1,m]$ represents the element in the i th column of the j th row in M, n represents the number of sample points, m represents the total number of variables, $P_j, j \in [1,m]$ represents the proportion of missing values of the observation variable in column j of M. The specific filling process of the algorithm is as follows:

Step 1: Find the column with the largest missing rate by formula (1) :

$$j_{\max} = \left\{ j \mid P_j = \max\left(\frac{is.na(M_{i,j})}{nrow(M)} \right), i \in [1,n], j \in [1,m] \right\} \tag{1}$$

In formula (1), $\max(\cdot)$ is the maximum function, $is.na(\cdot)$ is a function for counting the number of observations in the jth column variable of M that are marked as NA, and $nrow(\cdot)$ is the sample quantity function in statistics;

Step 2: Generate the initial dataset M' to be filled, where $M' = M_{i,(-j_{\max})}$, and $M_{i,(-j_{\max})}$ represents the removal of the $j_{\max}$ column from M ;

Step 3: The data of M' is filled in by using MRF or MNP algorithms, and a complete data set M'_c is obtained;
Step 4: Concatenate M'_c and $M_{i,j_{\max}}$ column-wise to obtain a new dataset M'' that needs to be filled in;
Step 5: Based on the filling algorithm adopted in the third step, corresponding to another algorithm (namely MNP or MRF), data filling processing is carried out on M'' respectively, and then the final complete data set M_c is obtained.

3 RESEARCH METHODS

3.1 Data Sources and Experimental Environment

The operating system of all experiments in this paper is Windows 11, the program writing software is RStudio 2023.03.0 Build 386, the program execution kernel is R version 4.2.3, and the drawing tool is Origin. Considering that there may be some differences in air quality Monitoring data in different regions, in order to verify the effect of the algorithm under different data sets, this paper obtained data from CNEMC (China National Environmental Monitoring Centre, https://www.cnemc.cn/en/) collected air quality data of 12 monitoring stations. The monitoring indexes were CO, NO2, O3, PM10, PM2.5, SO2 and AQI, and the data collection frequency of each station was once an hour. The period is from 0:00 on Jan 1, 2015 to 23:00 on Dec 31, 2023. In the real data collection process, due to equipment failure, extreme weather impact, communication transmission problems and other factors, the collected data contains missing values, and even blank data for a period of time. In this paper, the collected data will be preprocessed, and the processing method is divided into the following two parts:
(1) For the data without recording time, this paper does delete processing; As for the data with time records, even if all monitoring indicators are missing, they are still retained, so as to obtain 12 data sets with missing values to be filled, and a total of 914,100 air quality monitoring records are obtained. In order to facilitate the description of subsequent experiments, this group of data is marked as $A_i, i \in [1,12]$;
(2) The records containing missing observed values were deleted for processing, so as to obtain 12 complete data sets, with a total of 741,679 air quality monitoring records. In order to facilitate the subsequent experimental description, this set of data was marked as $B_i, i \in [1,12]$. As the B_i records are all real data values, therefore, in the third part of the paper "comparison of experimental results", this set of data will provide a comparison basis for the advantages and disadvantages of MRF, MNP, MRFNP and MNPRF algorithms under different evaluation criteria. Taking into account the reasons for the absence of air quality monitoring data, this paper will conduct missing data processing for B_i based on complete random missingness by using computer simulation. The simulation method steps of the Missing Data Mechanisms are presented in Section 2.2.

3.2 Simulation of Missing Data Mechanisms

The absence of observed values in air quality monitoring data meets the definition of completely random absence. Therefore, all computer simulation experiments in this paper are conducted on the premise of completely random absence. The detailed simulation steps of the Missing Data Mechanisms are as follows:
Step 1: According to the experimental results shown in Figure 1, the number of records with missing values in different data sets varies. To better simulate the proportion of missing values in actual problems, this paper sets a general range for the proportion of missing values in the complete data set M, while the overall missing rate P_n^t of data set M is randomly generated within this range. That is to say, the proportion of records with missing values is determined from the row perspective. The formula is as follows:

$$P_n^t = runif\left(p_n \mid p_n \in [p_{\min}, p_{\max}]\right) \tag{2}$$

In Formula (2), P_n^t represents the proportion of observation records with missing values among the total number of records; $runif(\cdot)$ is a random number generation function; $p_{\min}$ and $p_{\max}$ respectively denote the upper and lower limits of the values that P_n^t can take, indicating the random selection of n values of p from $[p_{\min}, p_{\max}]$. Here, $n=1$, that is $P_n^t = \left(p_1^t\right)$.

Step 2: Based on the value of P_1^t, randomly select data rows from dataset M. The set of row numbers corresponding to these observation records is denoted as R_m, and R_m is determined by formula (3):

$$R_m = sort(sample(row(M), floor(p_t \times nrow(M)))) \tag{3}$$

In formula (3), $row(\cdot)$ is the extraction row number function, which is used to obtain the row number set corresponding to each observation data in M. $row(M)$ is a vector and is denoted as $V_\alpha, \alpha \in [1,n]$. $nrow(\cdot)$ is used to obtain the total number of records in M, $floor(\cdot)$ represents the floor function, $floor(p_t \times nrow(M))$ indicates the total number of

records in M that contain missing observations, and this is counted as N_m. $sample(\cdot)$ is a random sampling function, indicating the random extraction of N_m elements without replacement from V_α; $sort(\cdot)$ represents a sorting function, here in ascending order, used to arrange the randomly selected elements in the order from small to large.

Step 3: In real-world application scenarios, the randomness of whether each observation record is missing a certain observation indicator is also completely random. Based on this, in this paper, from the perspective of columns, the random missing processing is carried out for each observation indicator of each record corresponding to R_m, in order to simulate the missing situation of the air quality monitoring dataset in real scenarios to the greatest extent. There are a total of 11 observed variables in M. Among them, 4 are time-related records, namely year, month, day and hour; there are 7 air quality indicators, and the corresponding column numbers are denoted as $V_\phi^m, \phi \in [5,11]$. During the experiment execution, only the missing values of the air quality indicators were handled. Firstly, through formula (4), a set of random missing weight combinations P_β^w is generated for the 6 observed variables in M. The larger the value is, the greater the possibility of missingness for the current record in the corresponding observed variable is; conversely, the smaller the value is, the lower the possibility is.

$$P_\beta^w = runif\left(p_\beta^w \mid p_\beta^w \in \left[p_{\min}, p_{\max}\right]\right), \beta \in [1,6] \tag{4}$$

In formula (4), $runif(\cdot)$ is a random number generation function, while $p_{\min}$ and $p_{\max}$ respectively represent the upper and lower limits of the values that P_β^w can take, that is, randomly select 6 values of p^w from $\left[p_{\min}, p_{\max}\right]$, and at this time $P_\beta^w = \left(p_1^w, p_2^w, \cdots, p_6^w\right)$.

Step 4: Taking into account the missing characteristics of the air quality monitoring data shown in Figure 1, it is necessary to assign a higher probability of missingness to one of the remaining observation variables in M, so as to make the computer simulation experiment more closely resemble the real application scenarios. The initial probability value P_β^w is optimized through formula (5) to obtain the probability value $P_{\beta+1}^w$.

$$P_{\beta+1}^w = sample\left(\frac{c\left(P_\beta^w, a \times sum\left(P_\beta^w\right)\right)}{(a+1) \times sum\left(P_\beta^w\right)}\right) \tag{5}$$

In formula (5), $sample(\cdot)$ is a random sampling function, $sum(\cdot)$ is a summation function, $c(\cdot)$ is a vector concatenation function, a is a constant term, and $a > 0$. In the subsequent experiments, by controlling the value of a, the superiority of the proposed hybrid imputation algorithm in this type of missing scenario can be verified under different combinations of missing columns.

Step 5: For each record corresponding to R_m, the actual missing column $C_\varphi^{R_m}, \varphi \le 7$ is generated through formula (6).

$$\begin{cases} C_\varphi^{R_m} = sort\left(sample\left(U, \varphi, W_1\right), l_1\right) \\ \varphi = sample\left(\left[1, length\left(U\right)\right], 1, sort\left(W_2, l_2\right)\right) \\ W_2 = sample\left(c\left(P_\beta^w, sum\left(P_\beta^w\right)\right) / 2sum\left(P_\beta^w\right)\right) \end{cases}, U = V_\phi^m, W_1 = P_{\beta+1}^w, l_1 = 0, l_2 = 1 \tag{6}$$

In formula (6), $C_\varphi^{R_m}$ represents the specific missing columns in each row record corresponding to R_m; U represents the data source to be extracted; φ represents the number of samples extracted from U; and the value of φ determines the number of elements contained in C_φ^R. W_1 represents the missing weights or probability distribution of each element in U. In this paper, non-uniform random selection operations are realized through W_1. $sum(\cdot)$ represents the summation function; $sort(\cdot)$ represents the sorting function. When the value of parameter l is 0, it indicates ascending order; when it is 1, it indicates descending order. It is used to arrange the randomly selected elements in the prescribed order.

Step 6: To facilitate the comparison of subsequent experimental results, first, the corresponding observed values in M are saved in sequence. Then, the missing values in the data are handled by setting the corresponding values in $M_{i,j}, i = R_m, j = C_\varphi^{R_m}$ as "NA".

3.3 Experimental Methods

In real life, the overall proportion of missing data is not fixed, and there is uncertainty about whether the observed index is missing in each record. In order to verify the accuracy and stability of the Data Imputation Algorithms, this paper conducts experiments in five steps:

Step 1: Select an arbitrary data set from for the experiment, and denote this data set as M.
Step 2: On the R platform, the Missing Data Mechanisms method described in this paper is adopted to conduct a completely random missing treatment on the overall missing proportion P_n^t of M within the range of proportions $[3\%,7\%]$, $[13\%,17\%]$ and $[23\%,27\%]$, thereby obtaining a non-complete data set M^*.
Step 3: Step 3: For each M^* under current P_n^t, apply 4 different imputation algorithms to fill in the missing values once. After the imputation is completed, evaluate the deviation degree of the imputation results from the original values in M under different evaluation criteria for different imputation algorithms when the current missing rate is considered.
Step 4: Taking into account the randomness of computer simulation, for each different range of P_n^t, this paper repeats the third step experiment operation 100 times for each case, thereby obtaining the mean, standard deviation and confidence interval of the evaluation results after Multiple Imputations [23].
Step 5: To verify the effectiveness of the data imputation algorithm, the experimental method from step 2 to step 4 was repeated for each data set in $B_i, i \in [1,12]$. The mean and median of the evaluation results after multiple imputations were given to verify the superiority of the algorithm under different data sets.

3.4 Evaluation Criteria

In order to compare the effects of different filling algorithms, this paper assumes $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n\}$, $Y = \{y_1, y_2, \cdots, y_n\}$, where $\hat{Y}$ represents the filling value, Y represents the original value, and n represents the number of missing values in the dataset M^* that needs to be filled. Based on this, the experiment makes a comparison of the results from two aspects of absolute error and relative error. Three evaluation criteria, namely MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and WMAPE (Weighted Mean Absolute Percentage Error), are selected to evaluate the algorithm presented in the paper. Among them, MAE and RMSE respectively represent the absolute error between $\hat{Y}$ and Y. The formulas are defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{8}$$

In formula (7)-(8), it can be seen that the ranges of MAE and RMSE are both within $[0,+\infty]$. The closer the values are to 0, the smaller the deviation between the imputed values and the true values, which indicates that the performance of the data imputation algorithm is better. Compared with MAE, RMSE also reflects the stability of the deviation degree. It is more sensitive to outliers. A smaller value not only indicates a better imputation effect but also reflects the stability of the imputation algorithm.
In real life, since the value range of the observed variables to be filled may be very different, if only MAE and RMSE are considered, the degree of error between the filled value and the true value is often unable to reflect the degree of error relative to the true value itself. Based on this, this paper adopts the statistical quantity WMAPE, which can represent the relative error between $\hat{Y}$ and Y, to evaluate the superiority of different algorithms. The definition of WMAPE is as follows:

$$WMAPE = \sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \Big/ \sum_{i=1}^{n} y_i \tag{9}$$

Compared with $MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$, WMAPE is less sensitive to outliers in Y. That is, when y_i approaches 0 or equals 0, it can still effectively reflect the relative error results among different algorithms. The range of WMAPE remains $[0,+\infty]$, and the closer the value is to 0, the better the filling effect is.

Taking into account the experimental methods of this paper, after Multiple Imputations under the same P_n^t range, different evaluation result vectors will be obtained, as follows:

$$MAE_\lambda = \{MAE_1, MAE_2, \cdots, MAE_N\} \tag{10}$$

$$RMSE_\lambda = \{RMSE_1, RMSE_2, \cdots, RMSE_N\} \tag{11}$$

$$WMAPE_\lambda = \{WMAPE_1, WMAPE_2, \cdots, WMAPE_N\} \tag{12}$$

The evaluation criteria used in this paper are derived from traditional MAE, RMSE and WMAPE calculations, and are defined as follows:

$$\overline{MAE} = \frac{1}{N}\sum_{\lambda=1}^{N} MAE_{\lambda} \tag{13}$$

$$\overline{RMSE} = \frac{1}{N}\sum_{\lambda=1}^{N} RMSE_{\lambda} \tag{14}$$

$$\overline{WMAPE} = \frac{1}{N}\sum_{\lambda=1}^{N} WMAPE_{\lambda} \tag{15}$$

In Formulas (13)-(15), N represents the number of fillings within the current P_n^t range, λ represents the λ th filling experiment, MAE_{λ}, $RMSE_{\lambda}$, and $WMAPE_{\lambda}$ respectively represent the values of MAE, RMSE and WMAPE obtained in the λ th experiment, and $\overline{MAE}$, $\overline{RMSE}$, and $\overline{WMAPE}$ respectively represent the mean values of the results under different evaluation criteria in multiple rounds of experiments.

4 RESULTS AND DISCUSSION

4.1 Experimental Results Based on Air Quality Monitoring Data in a Certain Region

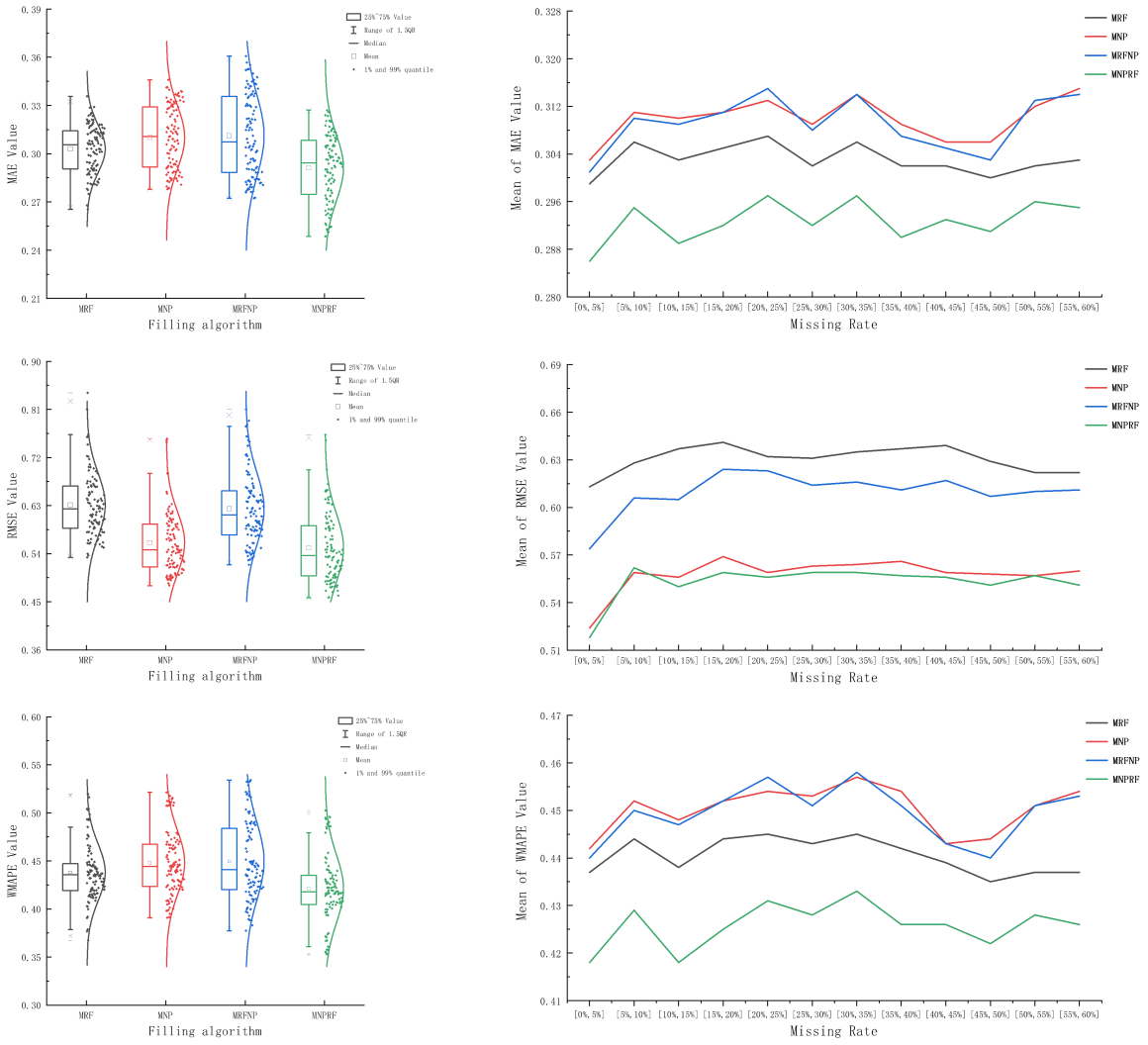
To facilitate the subsequent experimental description, let the mean value of the overall missing rate range be $\tau = \frac{p_{\min} + p_{\max}}{2}$, and the maximum weight of the missing column be $\omega = \frac{a}{a+1}$. Considering that in real scenarios, the overall missing rate of the dataset to be filled is not fixed, in order to simulate the effectiveness of the imputation algorithm under different missing rate ranges, this paper stipulates $p_{\max} - p_{\min} = 4\%$, and the values of $p_{\min}, p_{\max}$ are all accurate to 0.01, and a is a positive integer. During the experiment, $\tau = 0.05,0.15,0.25$, $\omega = 0.5,0.8,0.9$, the number of experimental repetitions $N = 100$, and the experimental results are as shown in Table 1:

Table 1 Under Different Loss Rates and Loss Weights, Four Filling Algorithms Fill 100 Times under Different Evaluation Criteria

Weight	Algorithm	$\tau = 0.05$			$\tau = 0.15$			$\tau = 0.25$		
		$\overline{RMSE}$	$\overline{MAE}$	$\overline{WMAPE}$	$\overline{RMSE}$	$\overline{MAE}$	$\overline{WMAPE}$	$\overline{RMSE}$	$\overline{MAE}$	$\overline{WMAPE}$
$\omega = 0.5$	MRF	0.634	0.306	0.442	0.625	0.300	0.437	0.634	0.301	0.436
	MNP	0.550	0.309	0.447	0.556	0.306	0.445	0.560	0.309	0.448
	MRFNP	0.616	0.310	0.449	0.613	0.304	0.442	0.616	0.306	0.444
	MNPRF	0.548	0.297	0.430	0.556	0.294	0.428	0.556	0.297	0.430
$\omega = 0.8$	MRF	0.634	0.307	0.446	0.631	0.305	0.441	0.636	0.303	0.442
	MNP	0.548	0.309	0.449	0.553	0.306	0.443	0.552	0.307	0.448
	MRFNP	0.606	0.309	0.449	0.614	0.307	0.444	0.605	0.307	0.449
	MNPRF	0.541	0.292	0.423	0.547	0.290	0.420	0.549	0.290	0.423
$\omega = 0.9$	MRF	0.633	0.306	0.445	0.627	0.301	0.438	0.629	0.303	0.439
	MNP	0.555	0.308	0.447	0.553	0.307	0.446	0.554	0.310	0.449
	MRFNP	0.610	0.308	0.448	0.608	0.309	0.448	0.610	0.311	0.451
	MNPRF	0.545	0.287	0.418	0.542	0.287	0.417	0.548	0.292	0.423

The experimental results in Table 1 show that under 3 different miss rate ranges and weight distributions, 4 algorithms fill 100 randomly generated incomplete data sets respectively, and then obtain the corresponding mean value of evaluation results. The above experimental results show that: (1) under MAE, RMSE and WMAPE evaluation criteria, the filling effect of MRF, MNP, MRFNP and MNPRF algorithms will not vary widely with the gradual increase of the missing rate, which preliminarily proves that the filling effect of the multi-filling algorithm based on the multi-interpolation idea is stable under different missing rates; (2) Under different ranges of missing rates and weight distributions, the MNPRF algorithm achieved the lowest A, B, and C results, and its overall filling effect was the best. The MNP algorithm's filling effect was second-best; (3) When $\omega = 0.5$ is true, the superiority of the MNPRF algorithm is not significant under the three evaluation criteria. However, as ω increases, the filling advantage of this algorithm becomes greater. In summary, the proposed algorithm improvement idea can inherit and extend the advantages of the existing algorithm, and has certain portability, and improve the filling optimization method from different perspectives.

In order to further verify the filling effects of the four algorithms under different missing rates, in this paper, the range of missing rates $[p_{\min}, p_{\max}]$ of the data set to be filled is gradually increased from $[0\%,5\%]$ to $[55\%,60\%]$. The step size of the upper limit $p_{\max}$ and the lower limit $p_{\min}$ of the missing rate range is set to 5% respectively, and the number of experiments for each missing rate range is $N=100$. The results of $\overline{MAE}$, $\overline{RMSE}$ and $\overline{WMAPE}$ are plotted under three criteria, and the experimental results are shown in Figure 2 (right). More importantly, in order to explore the stability of multiple experiments of different algorithms under MAE, RMSE and WMAPE criteria, this paper presents the box plot, result drop point and distribution curve of the miss rate similar to the real data. The experimental results are shown in left of Figure 2.



(a) **(b)**

Figure 2 (**a**) When $\tau=0.15, \omega=0.8, N=100$ is Set, Box Plots of Four Filling Algorithms under Different Evaluation Criteria; (**b**) When $\omega=0.8, N=100$ is Set, the Mean Line Graphs of MAE, RMSE and WMAPE Results Of The Four Filling Algorithms Under Different Missing Rates Are Presented

Figure 2 of (**a**) presents the box plots and distribution curves of the evaluation results obtained from 100 data imputation experiments conducted by MRF, MNP, MRFNP and MNPRF respectively when the range of missing rate is $[13\%,17\%]$ and the maximum weight of the missing column is 80%. It can be seen that: (1) MNPRF algorithm always maintains the minimum mean, median, upper quartile, lower quartile and other statistics under MAE, RMSE and WMAPE evaluation criteria, and has the best filling effect; (2) Under the MAE evaluation criterion, the box length of the MNPRF algorithm is only greater than that of the MRF algorithm. Moreover, the MAE_{λ} distribution of the MNPRF algorithm is more similar to the normal distribution compared to that of the MRF algorithm. This further validates that although the MNPRF algorithm is based on the MNP algorithm, it still retains a significant amount of the filling advantages of the MRF algorithm; (3) Under the RMSE evaluation criteria, the distance between the top and bottom quarterback values of the MNP algorithm is the smallest, followed by the MNPRF algorithm, and the distribution curves of the evaluation results of the two algorithms are similar, which indicates the system stability of the

MNPRF algorithm; (4) WMAPE value reflects the relative error of the algorithm. Under this evaluation criterion, MNPRF not only obtains the smallest spacing between top and bottom quarterbacks, but also the mean and median WMAPE values of the algorithm are even smaller than the lower quarterback values of the other three algorithms, showing significant filling advantage.

The experimental results in Figure 2 of (**b**) more directly demonstrate: (1) the filling advantage of MNPRF algorithm under different miss rates is significantly better than the other three algorithms under MAE and WMAPE criteria, while the filling effect under RMSE criteria is slightly better than MNP algorithm, but still significantly better than MRF and MRFNP algorithms; (2) The filling effect of MRF, MNP, MRFNP and MNPRF algorithms based on the idea of Multiple Imputation does not increase significantly with the increase of missing columns, and the filling effect is stable.

In the process of the experiment, considering the factors such as the sample size and the time complexity of the algorithm, combined with the uncertainty brought by the computer random simulation, this paper only repeated the experiment 100 times under each missing rate range, which is slightly insufficient to evaluate the overall filling effect of the algorithm. To further verify the overall differences among the aforementioned imputation algorithms, this paper uses MAE_λ, $RMSE_\lambda$, and $WMAPE_\lambda$ as new samples to construct a 95% confidence interval, thereby evaluating the accuracy of the imputation effects of different algorithms. Since the overall variance of the new sample is unknown, and according to the experimental results in Figure 2, the data MAE_λ, $RMSE_\lambda$, and $WMAPE_\lambda$ generated by different algorithms in this sample do not fully satisfy the normal distribution. Therefore, in this paper, formulas (16)-(18) are used to construct the Confidence Interval for the mean absolute error (MAE_{CI}), the root mean square error ($RMSE_{CI}$), and the mean absolute percentage error ($WMAPE_{CI}$), and the confidence interval lengths CIL (Confidence Interval Length) are respectively given.

$$MAE_{CI}=\left(\overline{MAE}-Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{j=1}^{N}\left(MAE_\lambda-\overline{MAE}\right)^2\Big/N^2},\overline{MAE}+Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{j=1}^{N}\left(MAE_\lambda-\overline{MAE}\right)^2\Big/N^2}\right) \tag{16}$$

$$RMSE_{CI}=\left(\overline{RMSE}-Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{j=1}^{N}\left(RMSE_\lambda-\overline{RMSE}\right)^2\Big/N^2},\overline{RMSE}+Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{j=1}^{N}\left(RMSE_\lambda-\overline{RMSE}\right)^2\Big/N^2}\right) \tag{17}$$

$$WMAPE_{CI}=\left(\overline{WMAPE}-Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{j=1}^{N}\left(WMAPE_\lambda-\overline{WMAPE}\right)^2\Big/N^2},\overline{WMAPE}+Z_{1-\frac{\alpha}{2}}\sqrt{\sum_{j=1}^{N}\left(WMAPE_\lambda-\overline{WMAPE}\right)^2\Big/N^2}\right) \tag{18}$$

In Equations (16)-(18), the value of α is set at 0.05. By calculation, $Z_{1-\frac{\alpha}{2}}=1.96$ and N represent the sample quantities, which are the number of experiments conducted under different missing rate ranges in this paper. The experimental results are shown in Table 2 as follows:

Table 2 Under Different Missing Rates, Four Filling Algorithms Filled the Confidence Interval Generated under Different Evaluation Criteria 100 Times

Missing Rate	Algorithm	$RMSE_{CI}$	CIL	MAE_{CI}	CIL	$WMAPE_{CI}$	CIL
$\tau=0.05$	MRF	[0.618, 0.650]	0.032	[0.304, 0.310]	0.006	[0.440, 0.452]	0.012
	MNP	[0.533, 0.563]	0.030	[0.306, 0.313]	0.007	[0.442, 0.455]	0.013
	MRFNP	[0.590, 0.623]	0.034	[0.304, 0.314]	0.010	[0.441, 0.457]	0.016
	MNPRF	[0.524, 0.559]	0.035	[0.288, 0.296]	0.009	[0.416, 0.431]	0.015
$\tau=0.15$	MRF	[0.620, 0.643]	0.023	[0.300, 0.306]	0.006	[0.432, 0.444]	0.012
	MNP	[0.549, 0.572]	0.023	[0.306, 0.314]	0.008	[0.442, 0.455]	0.013
	MRFNP	[0.611, 0.637]	0.026	[0.306, 0.316]	0.010	[0.442, 0.458]	0.016
	MNPRF	[0.538, 0.564]	0.025	[0.287, 0.295]	0.008	[0.414, 0.428]	0.014
$\tau=0.25$	MRF	[0.627, 0.645]	0.018	[0.300, 0.306]	0.005	[0.436, 0.448]	0.012
	MNP	[0.544, 0.560]	0.017	[0.303, 0.310]	0.007	[0.441, 0.454]	0.014
	MRFNP	[0.597, 0.614]	0.017	[0.303, 0.312]	0.009	[0.441, 0.456]	0.015
	MNPRF	[0.537, 0.560]	0.023	[0.286, 0.294]	0.008	[0.416, 0.431]	0.015

Statistical analysis of the experimental results in Table 2 with 95% confidence shows that:The MNPRF algorithm obtained the minimum lower and upper bounds of $RMSE_{CI}$, MAE_{CI} and $WMAPE_{CI}$ respectively under different missing rate ranges, which further verified the experimental results in Table 1; (2) Under different ranges of missing rates, the upper limit values of MAE_{CI} and $WMAPE_{CI}$ in the MNPRF algorithm are significantly lower than those of the other three algorithms; (3) Under different ranges of missing rates, the upper limit value of $RMSE_{CI}$ in the MNPRF

algorithm is always smaller than the lower limit value of $RMSE_{CI}$ in both the MRF and MRFNP algorithms, and the starting point of the interval of the MNPRF algorithm is always smaller than that of the MNP algorithm; (4) From the perspective of CIL, under the 95% confidence level, the confidence interval length values of MNPRF and MRFNP algorithms are consistently greater than those of MRF and MNP algorithms in all three evaluation criteria. In conclusion, the improved MNPRF algorithm has some values of $RMSE_{CI}$, MAE_{CI}, and $WMAPE_{CI}$ increasing due to the influence of certain special values in the data set to be filled. This affects the overall filling accuracy of the algorithm. However, its overall filling effect is significantly better than that of MRF and MNP algorithms.

4.2 Experimental Results Based on Air Quality Monitoring Data of Different Stations

In order to verify the general applicability of the algorithm in the field of air quality monitoring data, this paper conducts experiments on 12 datasets collected by adopting the same experimental method. Considering the missing rate situation of the original real datasets, the P_n^t value of this part of the experiment refers to the missing rate of the real datasets, that is $p_{\min}=0.1, p_{\max}=0.16$. Each dataset is executed 100 times with the same experimental steps, and all experimental results are plotted under the MAE, RMSE, and WMAPE criteria. The details are shown in Figure 3, Figure 4, and Figure 5:

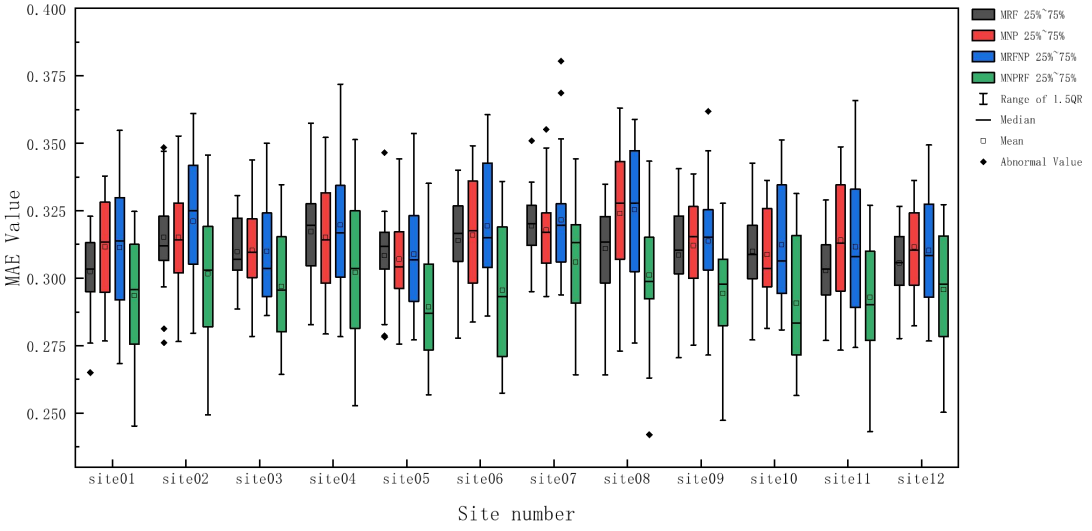


Figure 3 When $P_n^t \in [10\%,16\%], N=100$ is Set, Box Plot of MAE Results from 4 Algorithms Based on Different Datasets for Imputation Experiments

Figure 3 Experimental results show that: (1) Under MAE evaluation criteria, the MAE mean, median, upper quartile, lower quartile and 1.5x interval of the MNPRF algorithm in all regional air quality monitoring data sets are significantly lower than the corresponding statistics of the other three algorithms; (2) In the experimental results of all regions, the MNPRF algorithm has a longer interquartile interval, while the MRF algorithm has the shortest interquartile interval. To sum up, MNPRF algorithm has significant advantages in filling different data sets.

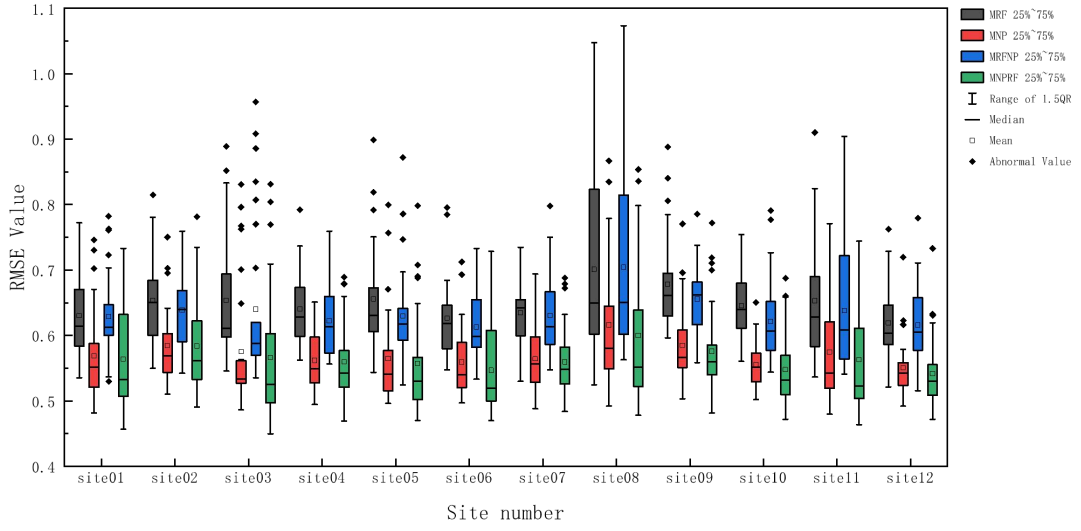


Figure 4 When $P_n^t \in [10\%,16\%], N = 100$ is set, Box Plot of RMSE Results from 4 Algorithms Based on Different Datasets for Imputation Experiments

Figure 4 Experimental results show that: (1) Under the RMSE evaluation criteria, the RMSE mean, median, upper quartile, lower quartile and 1.5x interquartile interval values of MNPRF algorithm in all regional air quality monitoring data sets are significantly lower than the corresponding statistics of MRF and MRFNP algorithms, and slightly lower than MNP algorithm; (2) In the experimental results of all monitoring stations, the interquartile spacing of the four algorithms had no obvious rule, and there were a few outliers in the RMSE experimental results of almost all stations. In summary, the MNPRF algorithm has the best system stability in different data sets.

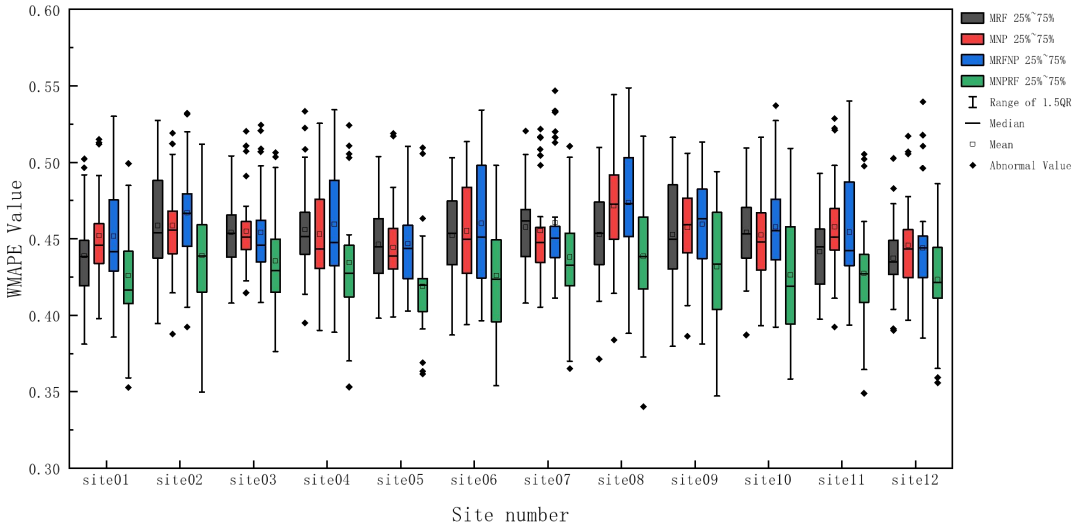


Figure 5 When $P_n^t \in [10\%,16\%], N = 100$ is set, Box Plot of WMAPE Results from 4 Algorithms Based on Different Datasets for Imputation Experiments

Figure 5 Experimental results show that: (1) Under the WMAPE evaluation criteria, the mean, median, upper quartile, lower quartile and 1.5x interquartile of WMAPE in all regional air quality monitoring data sets of MNPRF algorithm have the smallest values; (2) In almost all monitoring site experiments, the mean and median values of the evaluation results of the MNPRF algorithm were smaller than the lower quartile values of the other three algorithms; (3) In the experimental results of all monitoring sites, the quartile spacing of the four algorithms has no obvious rule, and the upper and lower quartile spacing of the MNPRF algorithm still has a good performance in the data filling experiments of some sites. In summary, the relative error of MNPRF algorithm in different data sets is the smallest, and the filling effect is the best.

5 CONCLUSIONS

With the application and popularity of machine learning and neural network algorithms in all walks of life, the scale and quality of data have become increasingly important, and missing value processing has become the most important part of data pre-processing. For the same data set to be filled, data analysts often need to choose a suitable Data Imputation Algorithms according to data characteristics, missing reasons, data scale and other factors. Considering the complexity of practical problems, high-dimensional data and multi-source heterogeneous data are becoming more and more common in the current real application scenarios, which leads to the cause of missing data of different dimensions becoming no longer single. Meanwhile, filling algorithms dealing with missing observed values of different dimensions in the same data set may also be different. Based on the original data imputation algorithms MRF and MNP, this paper makes improvements, tries to use different algorithms to deal with the missing value problem of different monitoring dimensions in air quality monitoring data, and proposes two algorithms MRFNP and MNPRF according to the kernel execution order of the improved algorithm. The experimental results of this paper show that the filling effect and accuracy of the algorithm can be greatly improved by selecting a more appropriate filling algorithm for different missing dimensions of the same data set. The algorithm improvement concept in this paper provides a new idea for the future development of the data filling field.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This research was funded by Guizhou Minzu University Fund Project (Youth Project) (No. GZMUZK[2023]QN12), the Guizhou Provincial Basic Research Program (Natural Science) Young Scholars Guidance Project (No. [2024]208), the Guizhou Provincial Basic Research Program(Natural Science) (No. zk2025-536).

REFERENCES

[1] Di Z, Guarnera U, Luzi O. Imputation through finite Gaussian mixture models. Computational Statistics and Data Analysis, 2007, 51: 5305-5316.
[2] Junninen H, Niska H, Tuppurainen K, et al. Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 2004, 38: 2895-2907.
[3] Sabine V, Karlien V B, Peter G. Sequential imputation for missing values. Computational biology and chemistry, 2007, 31, 320-327.
[4] Pedro J, Garcia L, Jose-Luis S G, et al. K nearest neighbours with mutual information for simultaneous classification and missing data imputation, Neurocomputing, 2009, 72: 1483-1493.
[5] Wu S, Feng X D, Shan Z G. Missing Data Imputation Approach Based on Incomplete Data Clustering.Chinese journal of computer, 2012, 35: 1726-1738.
[6] Sethia K, Gosain A, Singh J. Review of Single Imputation and Multiple Imputation Techniques for Handling Missing Values. Lecture Notes in Networks and Systems, 2023, 730: 33-50.
[7] Rubin D B. Inference and Missing Data. Biometrika, 1976, 63: 581-592.
[8] Little R, Rubin D B. Statistical Analysis With Missing Data; Wiley and Sons Inc: New York, USA, 1987.
[9] Enders C K. Applied Missing Data Analysis. Guilford Press: New York, USA, 2010.
[10] Sebastian J, Arndt A, Felix B. A Benchmark for Data Imputation Methods. Frontiers in big data, 2021, 4, 674-693.
[11] Hakan D. Flexible Imputation of Missing Data. Journal of Statistical Software, 2018, 85: 1-5.
[12] Schafer J L. Analysis of Incomplete Multivariate Data. Chapman & Hall: Oxfordshire, UK, 1997.
[13] Schafer J L, Yucel R M. Computational strategies for multivariate linear mixed-effects models with missing values. Journal of Computational and Graphical Statistics, 2002, 11: 437-457.
[14] Stef V B. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 2007, 16: 219-242.
[15] Van B S, Brand J P L, Groothuis-Oudshoorn C G M, et al. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 2006, 76: 1049-1064.
[16] Yusuke Y, Toshihiro M, Kazushi M. A comparison of multiple imputation methods for incomplete longitudinal binary data. Journal of Biopharmaceutical Statistics, 2018, 28: 645-667.
[17] Kim H J, Reiter J P, Wang Q, et al. Multiple Imputation of Missing or Faulty Values Under Linear Constraints. Journal of Business & Economic Statistics, 2014, 32: 375-386.
[18] Enders C K, Keller B T, Levy R. A fully conditional specification approach to multilevel imputation of categorical and continuous variables. Psychological methods, 2018, 23: 298-317.
[19] Vincent A, Ndeye N. Clustering with missing data: which equivalent for Rubin's rules? Advances in Data Analysis and Classification, 2023, 17: 623-657.
[20] Van B S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 2007, 16: 219-242.
[21] Goldstein H, Carpenter J, Kenward M G, et al. Multilevel models with multivariate mixed response types. Statistical Modelling, 2009, 9: 173-197.
[22] Yang Z. Diagnostic checking of multiple imputation models. AStA Advances in Statistical Analysis, 2022, 106: 271-286.
[23] Zhi Q Z, Yan C, Meng M W, et al. Research on Stability of Data Imputation Algorithms With Different Miss Rates. Statistics and The Decision, 2023, 33: 12-17.