International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 6
ISSN: 2088-5334

Using Multiple Regression Model and RNN for Imputing the Missing Values of PM_{10} Datasets

Moamin Amer Hasan Alsaeegh[a], Osamah Basheer Shukur[a,1]

[a] *Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq*
E-mail: [1]drosamahannon@uomosul.edu.iq

***Abstract*—The missing value in time series data is a scientific problem that should be solved by imputing these values by following some statistical techniques. This problem is more complex due to the missing values that existed in the dependent (response) variable. Particular matter (PM_{10}) is a time series dataset used to scale air pollution as a dependent variable, while there are many types of pollutants used as independent variables. Malaysian datasets of PM_{10} and several climate pollutants are examined in this study. This study aims to impute the missing values for different missing rates in a dependent variable with minimum error. In this paper, the independent variables were supposed completed while the missing values have been replaced in different rates and different distributions within the dependent variable. Multiple linear regression (MLR) has been used as a traditional method to impute the different missing values of PM_{10}. Recurrent neural network (RNN) is combined with MLR and used to impute the missing values of PM_{10}. The results reflected that th hybrid method outperformed MLR for imputing the missing values of PM_{10}. In conclusion, the hybrid method MLR-RNN can be used to impute the missing values of PM_{10} accurately compared to other traditional methods.**

***Keywords*—multiple linear regression; MLR; missing values; recurrent neural network; RNN.**

I. INTRODUCTION

Air pollution studying and forecasting is necessary to control and reduce the damage to the environment and human health. Particulate matter (PM_{10}) is a dataset used to measure air pollution and can be regarded as meteorological time series data. Therefore, the missing values in PM_{10} data should be filled and imputed. PM_{10} missing values imputation can be accomplished using multiple linear regression (MLR) models as classical methods when several pollutants are independent variables.

MLR model studies the relationship between the dependent variable and several variables that affect the dependent variable. It is necessary to identify the variables that have a real relationship with the dependent variable. Before starting MLR analysis, the explanatory variables' multicollinearity problem has been detected and treated [1], [2]. MLR has been suggested to model the daily PM_{10} data, which is the basis for forecasting[3]. [4] used MLR to study meteorological data and to find the best models that express the relationship between the dependent variable and a number of explanatory variables.

In most meteorological time series data sets, nonlinearity is a problem that may hamper time series analysis using MLR. In particular, PM_{10} data suffer from nonlinearity in addition to the missing values problem. In recent papers, RNN is introduced to impute missing values and to handle the nonlinearity of meteorological time-series datasets. Several ANN types are introduced for infilling missing daily weather records such as daily precipitation and daily extreme temperature series [5]. In this study, a hybrid MLR-RNN method is proposed to solve the problem by imputing missing values and handling the nonlinearity problem.

RNN was used by several researchers to forecast air quality datasets [6] developed novel deep learning models, namely GRU-D, as early attempts. GRU-D is based on Gated Recurrent Unit (GRU), a state-of-the-art recurrent neural network. [7] proposed a dynamic L-RNN to predict any missing value in a simple, fast manner to save time and cost. In this paper, the hybrid MLR-RNN method was used by combining MLR and RNN in one method. Its results were compared to MLR model results. Root mean square error (RMSE) and mean absolute error (MAE) was computed to measure the error of missing values imputation for all imputation methods and all datasets as a statistical criterion to evaluate the adequacy and accuracy of these methods.

In this study, the missing values have been distributed into several different parts based on several different missing proportions (5%, 10%, and 25%) to obtain six datasets groups of PM_{10} and several climate independent variables. RNN toolboxes in MATLAB have been used to perform all of RNN procedures, While Minitab and Excel have been used to perform MLR model procedures.

II. Materials and Method

A. Data and Framework of the Study

In this study, two datasets of daily Malaysian PM_{10} with several climate independent variables for three hydrological years (1 January 2013 – 31 October 2015) were collected from Kuala Lumpur meteorological station. The missing values have been distributed into several different parts based on several different missing proportions (5%, 10%, and 25%). The framework of this study includes the following:

- Completed the missing values in the independent variable by the traditional imputation methods.
- Constructed the most appropriate MLR model and the hybrid MLR-RNN after deleting the missing values in the dependent variable and the corresponding observations in the independent variables.
- Imputted the missing values of dependent variable (PM_{10}) by applying MLR, and hybrid MLR-RNN on the complete independent variables.
- Comparing the error of missing imputation for MLR model, and the proposed hybrid MLR-RNN method to determine which method would provide the best adequacy.

B. Multiple Linear Regression (MLR) Model

MLR is used for modeling the relationship between multiple independent variables and one dependent variable. The general expression of the MLR model can be formulated as follows [8].

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e_i \qquad (1)$$

where y_i is the dependent variable, $x_1, x_2, \dots, x_p$ are independent variables, β_0 is the constant model, $\beta_1, \beta_2, \dots, \beta_p$ are the parameters of regression model, and e_i is the amount of random error. Equation (1) can be written in matrix form as follows:

$Y = X\beta + \varepsilon$

where $Y$is the size $(n \times 1)$and the matrix $X$of the degree $\big(n \times (p+1)\big)$ and the size of β is $((p+1) \times 1)$ and the degree ε is $(n \times 1)$defined such as follows:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad (2)$$

For all $i = 1,2,3, \dots, n$

There are a number of assumptions (analysis assumptions) about the distribution variables in the MLR such as follows.

- The average of random error term (ε_i) should be equaled zero.
- The variance of the random error term must be constant in all time periods. This hypothesis is called Homoscedasticity. But when the variance of the random errors is not constant, it is called Heteroscedasticity.

$$Var(e_i) = E(e_i{}^2) = \sigma^2 \qquad (3)$$

for all $i = 1,2,3, \dots, n$

- The random error follows a normal distribution by zero mean and constant variance.

$$\varepsilon_i \sim N(0, \sigma^2) \qquad (4)$$

for all $i = 1,2,3, \dots, n$

- The random variable (e_i) is independent of (x_i)variables. This means that the covariance of the random error and independent variable are equal to zero.

$$COV(e_i, x_i\, n) = 0 \qquad (5)$$

for all $i = 1,2,3, \dots, n$

- The covariance of any random errors in different lags such as e_i, e_j is equals to zero $\mathrm{COV}(e_i, e_j \quad) = 0$ $(i \neq j)i, j = 1,2, \dots, n$.
- The root mean square error (RMSE), and mean absolute error (MAE) was used to measure the accuracy of missing imputation results. RMSE and MAE can be written such as follows [9].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i|, \quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (e_i)^2} \qquad (6)$$

where e_i is the forecasting error, and n is the number of observations.

C. Hybrid MLR-RNN Method

In this study, a hybrid MLR-RNN method has been proposed to improve missing imputation results of PM_{10} and other meteorological datasets. The framework of the proposed approach are detailed as follows.

- Completing the missing values in the independent variable by the traditional imputation methods.
- Depending on the most appropriate MLR model, the hybrid MLR-RNN was constructed after deleting the dependent and independent variables' missing values.
- Performing training and testing processes to impute the missing values of the dependent variable (PM_{10}) using complete independent variables. (1)
- Compared the error of missing imputation for MLR-RNN to the traditional method.

A hybrid MLR-RNN is proposed for imputing the missing values. MLR model was used only for determining the input layer structure of RNN. Listwise deletion was used before MLR modeling. Hybrid MLR-RNN was also proposed to handle the nonlinearity of dataset. Determining the training functions and the transfer functions types of hidden and output layers and other requirements were necessary to create the most appropriate RNN structure.

The primary reason for using RNN is the nature of data set non-linearity. RNN contains one or more layers and this may handle the non-linearity of data and improve forecasting results. RNN also contains a delay layer that may solve data heterogeneity because it contains longer memory than other algorithms [10], [11].

In this study, RNN contains two layers in addition to the input layer. The first layer is hidden and other is the output layer. In the input layer, there are R of inputs weighted randomly. In each hidden layer, there are M of neurons. The best number of neurons in the hidden layer is usually R * 2+1[12]. Every input variable z is weighted randomly. The weights of R inputs and M neurons were combined with the biased value of b by the transfer function. The sum of input variables in the transfer function F can be formatted as follows [13][14].

$$\text{SUM} = \sum_{i=1}^{M} \sum_{j=1}^{R} w_{i,j} Z_j + b \qquad (7)$$

The common transfer functions in the hidden and the output layers are as follows.

- Tan-sigmoid: which generates outputs between -1, +1.
- Log-sigmoid: which generates outputs between 0, 1.
- Linear function: which generates outputs between -1, +1.

The nonlinearity of PM_{10} and other climate datasets requires determining nonlinear transfer function such as tan-sigmoid and log-sigmoid for the hidden layer to filter the nonlinearity. Figure 1 and Figure 2 demonstrate the structure of RNN and the transfer function types, respectively.

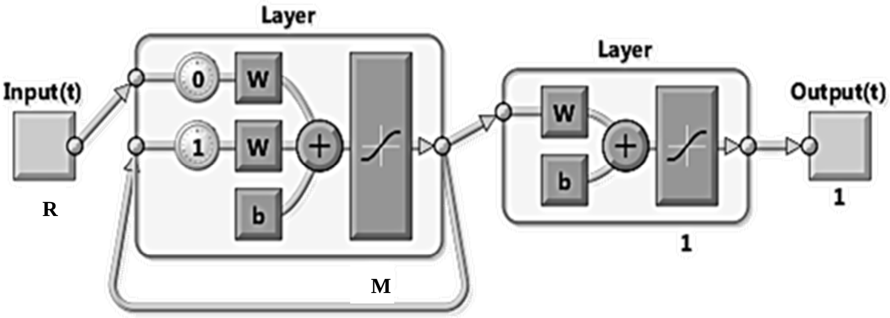

Fig. 1 The structure of RNN

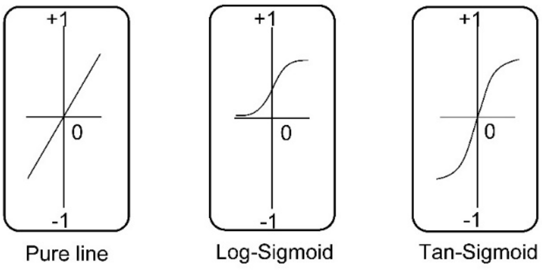

Fig. 2 The types of the transfer function

III. Results and Discussion

Malaysian PM_{10} (particulate matter less than or equal 10 micrometers) and several independent variables such as (CO: carbon monoxide, SO_2: sulfur dioxide, NO: nitric oxide, WS_{10}: wind speed 10 meters, and WS_{xx}: wind speed recorded at xx-meter height) for 34 months from 1 January 2013 till 31 October 2015 was investigated in this study. The total number of observations in the datasets is 1034. The missing values have been distributed into several different parts based on several different missing proportions (5%, 10%, and 25%).

A. MLR model

To obtain the best MLR model of PM_{10} and other independent variables after inserting the dependent and the independent variables for each of the six different data groups in Minitab program, the results is presented in this study. In the MLR model,, when the missing values are distributed discretely with a ratio 5% into a full period of PM_{10} variable, the formula is presented as follows.

$$y\ 107\ x_1\ +\ 1810\ x_2\ -\ 28x_3\ +\ 1.13\ x_4\ -\ .08\ x_5 \qquad (8)$$

where y is PM_{10}, x_1 is CO, x_2 is SO_2, x_3 is NO, x_4 is WS_{10}, and x_5 is WS_{xx}. The details of MLR coefficients in equation (8) is indicated in Table 1.

TABLE I
THE DETAILS OF MLR COEFFICIENT (MISSING VALUES DISTRIBUTED 5% DISCRETELY)

Term	β	Cal.t	P value
x_1	107.51	47.69	0.00
x_2	1810.00	3.64	0.00
x_3	-2288.00	-18.72	0.00
x_4	1.13	3.36	0.00
x_5	-0.08	-7.68	0.00

Table 1 shows that the coefficients of MLR model in equation (8) are significant because their corresponding p-values are less than the significant level 0.05. Therefore, the MLR model is suitable for the datasets of study. RMSE and MAE values are 18.2 and 13.5, respectively.

Figure (3) above explains the fitness between the missing values and the corresponding imputed values using MLR model where the missing ratio is (5%), and the missing distribution is discrete. In the MLR model, when the missing values are distributed successively with a ratio 5% into full period of PM_{10} variable, the formula is shown below.

$$y\ 105\ x_1\ +\ 192\ x_2\ -\ 24x_3\ +\ 0.33\ x_4\ -\ 0.08\ x_5 \qquad (9)$$

The details of MLR coefficients in equation (9) is presented in Table 2. Table 2 shows that the coefficients of MLR model in equation (9) are significant because their corresponding p-values are less than the significant level 0.05. Therefore, the MLR model is suitable for the datasets of study. RMSE and MAE values for the equation (5) are 18.3 and 13.7, respectively. Figure (4) shows the fitness between the missing values and the corresponding imputed values using MLR model where the missing ratio is (5%) and the missing distribution is successive.

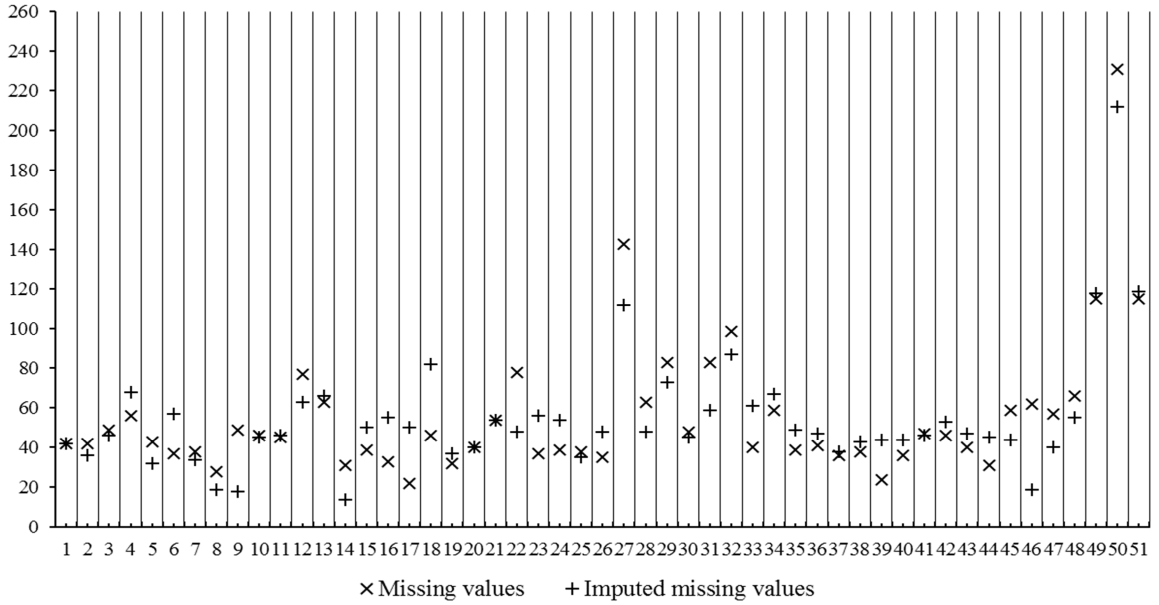

Fig. 3 The fitness between the missing values in PM_{10} and the corresponding imputed values by using MLR model where the missing ratio is (5%) and the missing distribution is discrete.

TABLE II
THE DETAILS OF MLR COEFFICIENT (MISSING VALUES DISTRIBUTED 5% SUCCESSIVELY)

Term	β	Cal.t	p-value
x_1	105.15	47.42	0.00
x_2	1952.00	4.02	0.00
x_3	-2247.00	-18.77	0.00
x_4	0.33	3.94	0.00
x_5	-0.08	-7.63	0.00

In the MLR model, when the missing values is distributed discretely with ratio 10% into a full period of PM_{10} variable, the formula is presented below.

$$y = 108x_1 + 184\ x_2 - 28x_3\ + 1.02x_4 - 0.09x_5 \qquad (10)$$

The details of MLR coefficients in equation (10) is displayed in Table 3. Table 3 shows that the coefficients of MLR model in equation (10) are significant because their corresponding p-values are less than the significant level 0.05. Therefore, the MLR model is suitable for the datasets of study.

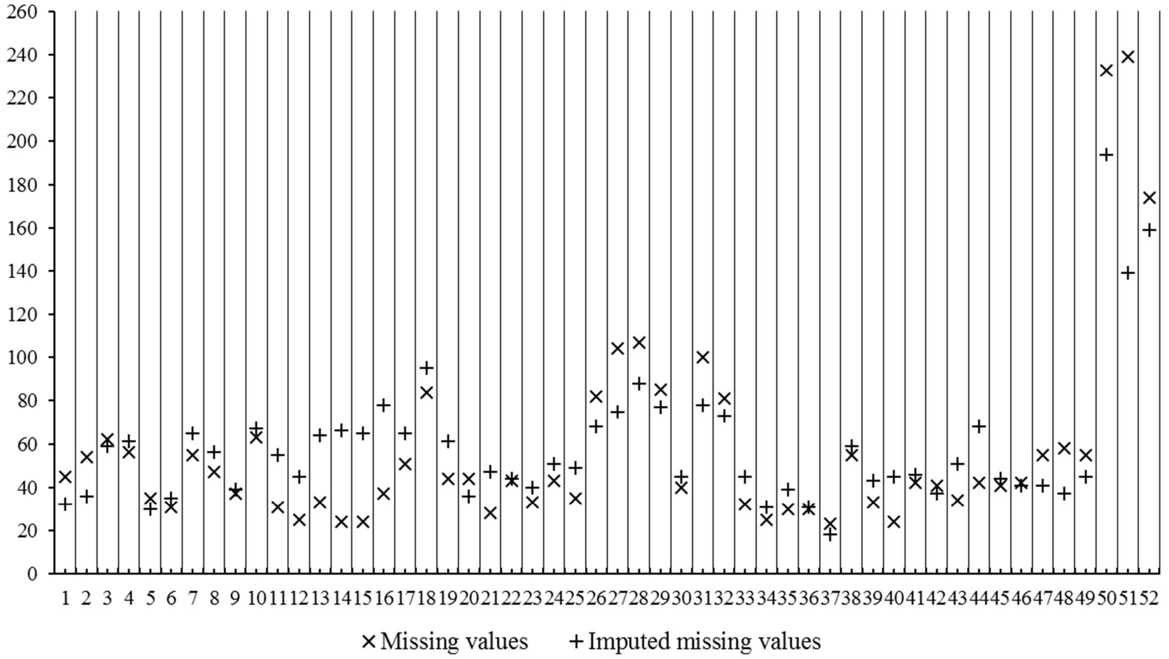

Fig. 4 The fitness between the missing values in PM_{10} and the corresponding imputed values by using MLR model where the missing ratio is (5%) and the missing distribution is successively.

RMSE and MAE values (6) are 18.2 and 13.5, respectively. Figure (5) describes the fitness between the missing values and the corresponding imputed values by using the MLR model where the missing ratio is (10%) and the missing distribution is discrete.

TABLE III
THE DETAILS OF MLR COEFFICIENT (MISSING VALUES DISTRIBUTED 10% DISCRETELY)

Term	β	Cal.t	P value
x_1	108.90	46.95	0.00
x_2	1840.00	3.57	0.00
x_3	-2286.00	-18.22	0.00
x_4	1.02	2.90	0.00
x_5	-0.09	-7.77	0.00

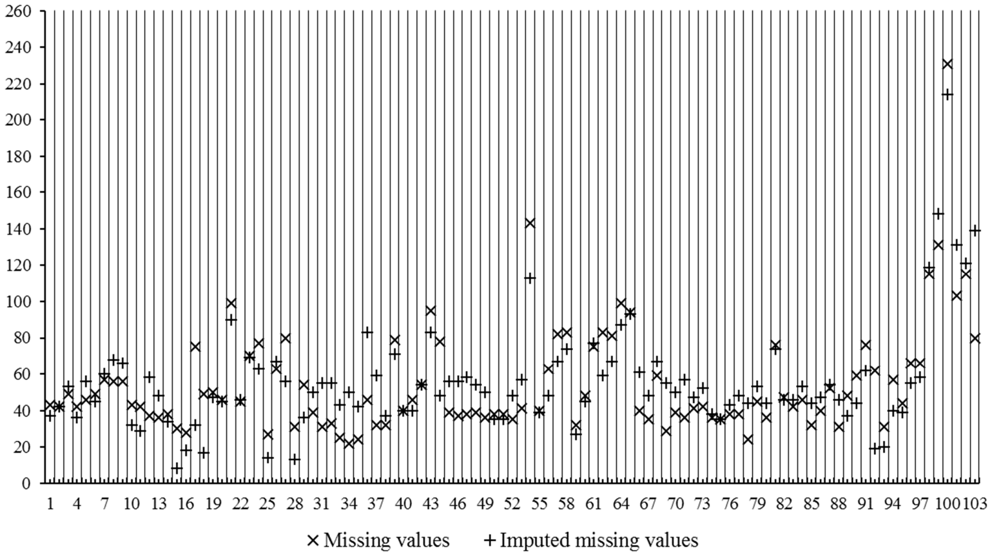

Fig. 5 The fitness between the missing values in PM_{10} and the corresponding imputed values by using MLR model where the missing ratio is (10%) and the missing distribution is discrete.

In the MLR model, when the missing values are distributed successively with a ratio 10% into a full period of PM_{10} variable, the formula is shown as follows.

$$y = 104.64\,x_1 - 22x_3 + 1.13x_4 - 0.08\,x_5 \qquad (11)$$

The details of MLR coefficients in equation (11) is displayed in Table 4. Table 4 shows that the coefficients of MLR model in equation (11) are significant because their corresponding p-values are less than the significant level 0.05. Therefore, the MLR model is suitable for the datasets of study. RMSE and MAE values are 18.3 and 13.4, respectively. Figure (6) shows the fitness between the missing values and the corresponding imputed values using MLR model where the missing ratio is (10%) and the missing distribution is successive.

TABLE IV
THE DETAILS OF MLR COEFFICIENT (MISSING VALUES DISTRIBUTED 10% SUCCESSIVELY)

Term	β	Cal.t	p-value
x_1	104.64	46.35	0.00
x_2	2123.00	4.25	0.00
x_3	-2252.00	-18.41	0.00
x_4	1.13	3.28	0.00
x_5	-0.08	-7.19	0.00

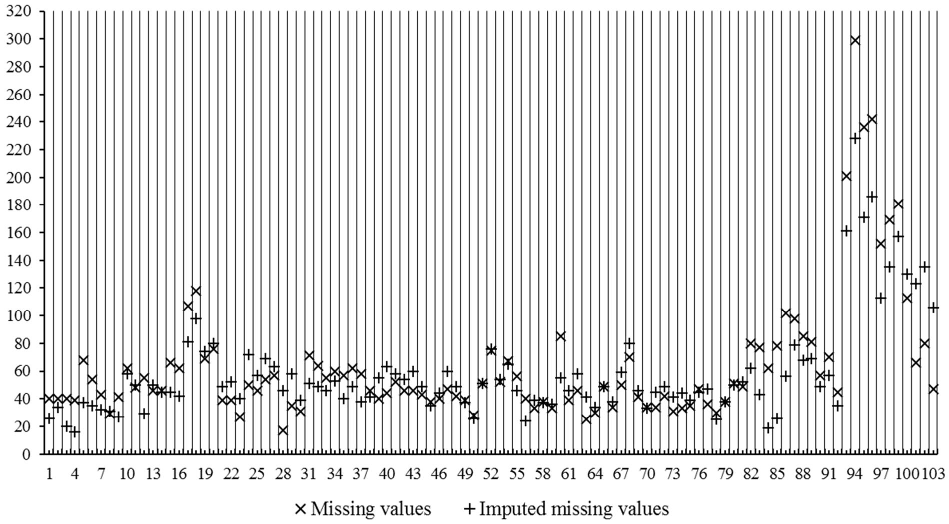

Fig. 6 The fitness between the missing values in PM_{10} and the corresponding imputed values by using MLR model where the missing ratio is (10%) and the missing distribution is successively.

In the MLR model, when the missing values are distributed discretely with ratio 25% into a full period of PM_{10} variable, the formula is presented as follows.

$$y = 107.12\, x_1 - 258 x_3 + 1.02 x_4 - 0.8 x_5 \qquad (12)$$

The details of MLR coefficients in equation (12) is presented in Table 5. Table 5 shows that the coefficients of MLR model in equation (12) are significant because their corresponding p-values are less than the significant level 0.05. Therefore, the MLR model is suitable for the datasets of study. RMSE and MAE values are 16.4 and 12.6, respectively. Figure (7) indicates the fitness between the missing values and the corresponding imputed values using MLR model where the missing ratio is (25%), and the missing distribution is discrete.

TABLE V
THE DETAILS OF MLR COEFFICIENT (MISSING VALUES DISTRIBUTED (25% DISCRETELY)

Term	$\boldsymbol{\beta}$	Cal.t	p-value
x_1	107.12	41.58	0.00
x_2	1956.00	3.47	0.00
x_3	-2258.00	-16.05	0.00
x_4	1.02	2.61	0.01
x_5	-0.09	-6.70	0.00

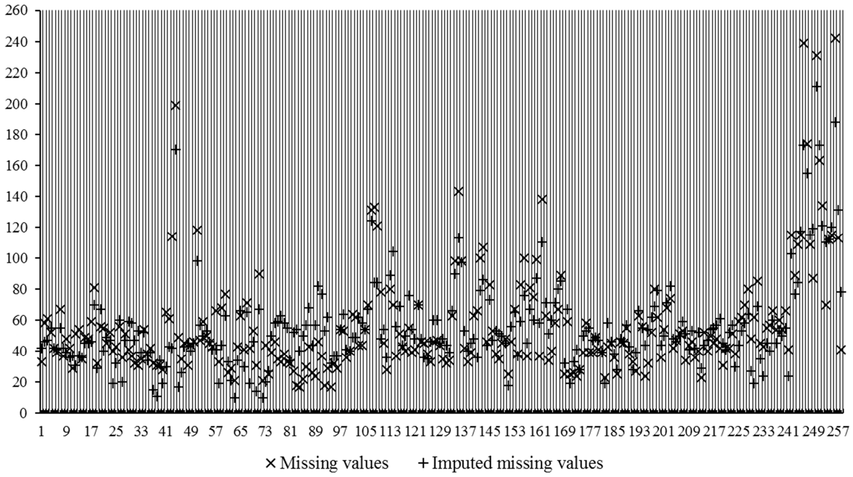

Fig. 7 The fitness between the missing values in PM_{10} and the corresponding imputed values using MLR model where the missing ratio is (25%) and the missing distribution is discrete.

In the MLR model, when the missing values are distributed successively with a ratio 25% into full period of PM_{10} variable, the formula is indicated below.

$$y = 1.29\, x_1 + 18 x_2 - 236 x_3 + 1.23\, x_4 - 0.10\, x_5 \qquad (13)$$

The details of MLR coefficients in equation (13) is presented in Table 6. Table 6 shows that the coefficients of MLR model in equation (13) are significant because their corresponding p-values are less than the significant level 0.05. Therefore, the MLR model is suitable for the datasets of study. RMSE and MAE values are 16.6 and 12.8, respectively. Figure (8) describes the fitness between the missing values and the corresponding imputed values using MLR model where the missing ratio is (25%) and the missing distribution is successive.

TABLE VI
THE DETAILS OF MLR COEFFICIENT (MISSING VALUES DISTRIBUTED (25% SUCCESSIVELY)

Term	$\boldsymbol{\beta}$	Cal.t	P value
x_1	111.29	44.06	0.00
x_2	1835.00	3.06	0.00
x_3	-2336.00	-16.52	0.00
x_4	1.23	2.95	0.00
x_5	-0.10	-7.53	0.00

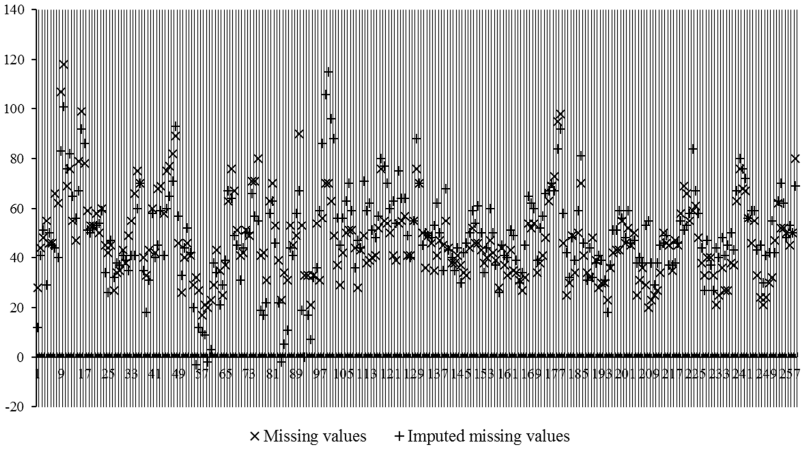

Fig. 8 The fitness between the missing values in PM_{10} and the corresponding imputed values using MLR model where the missing ratio is (25%) and the missing distribution is successive.

RMSE and MAE of the results of imputing the missing values by using MLR for different missing rates in two distribution methods (discretely and successively) are summarized as in Table 7.

TABLE VII
SUMMARY OF THE ERROR MEASUREMENTS OF MISSING VALUES IMPUTATION BY USING MLR MODEL.

Missing Values	**5%**		**10%**		**25%**	
Distribution	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Discretely	15.9	12.4	16.4	12.7	16.9	12.8
Successively	22.2	15.4	22.1	15.7	13.4	10.5

B. RNN method

RNN is a type of artificial neural network that is performed in several sequenced procedures. In RNN, all the inputs are independent of each other, while the inputs have direct effects on the outputs. In this paper, MATLAB has been used two perform the tasks of creating the neural network and forecasting. MATLAB toolboxes of artificial neural networks include several steps such as importing inputs and target variables, creating RNN, training, and exporting results. Layer recurrent is the name of network algorithm type in MATLAB, which refers to RNN. There are three types of transfer functions: log-sigmoid, linear, and tan-sigmoid, which can be chosen to produce better results.

The data was entered into MATLAB in rows form for the target variable and the input variables. Each row represents one variable and each column represent one observation. Missing values should be deleted with the corresponding observations in the independent variables. After deletion, multiplying each independent variable in MLR model by the value of the corresponding coefficient is necessary to specify the input variables in RNN structure. The framework of constructing RNN and training is as follows.

- After deletion and multiplying, the inputs variable and target variable should be entered into the workspace in MATLAB separately.
- The inputs variable and target variable should be imported from the workspace to neural network toolboxes.
- There are several essential requirements to perform the RNN construction, such as determining the training function, the number of hidden neurons, and the transfer functions of hidden and output layers.
- After constructing RNN, the complete variables before deletion was entered in the same way into the workspace. In this case, the target variable's size is 1×1034, and the size of the inputs variable is 5×1034.
- RNN that constructed in the previous stage was used to perform the training process again by using the complete variables after importing them from the workspace.
- The last training process's output variable is the estimated variable that contains the imputed missing values resulting from using MLR-RNN hybrid method.

Figure (9) describes the fitness between the missing values and the corresponding imputed values using MLR-RNN method where the missing ratio is (5%), and the missing distribution is discrete.

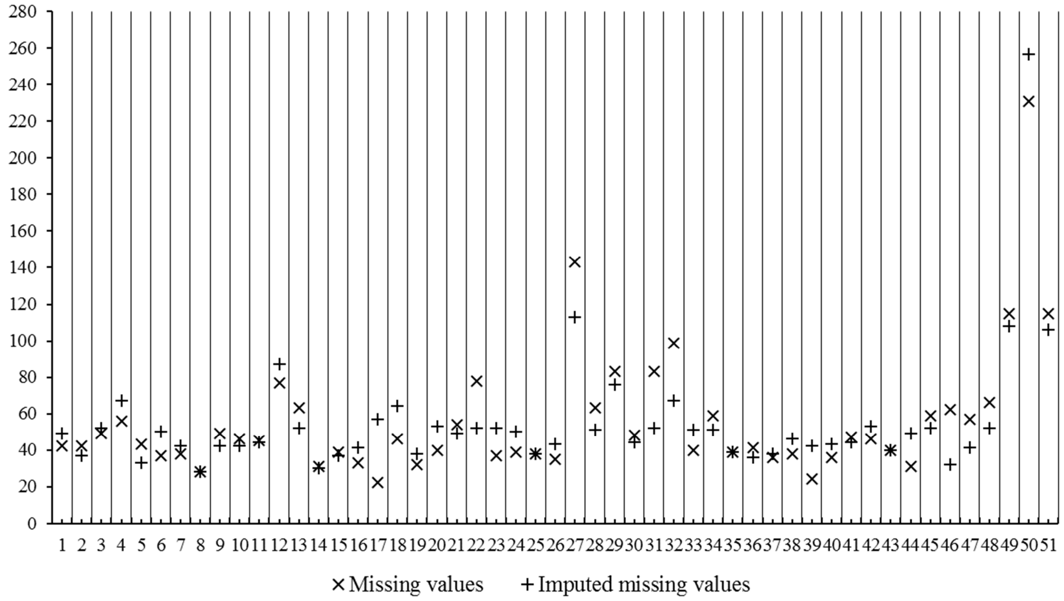

Fig. 9 The fitness between the missing values in PM_{10} and the corresponding imputed values using MLR-RNN method where the missing ratio is (5%) and the missing distribution is discrete.

Figure (9) proved that the fitness between the missing values in PM$_{10}$ and the imputed values where the missing ratio is (5%) and the missing distribution is discretely using the hybrid MLR-RNN method. It is better than the corresponding fitness using MLR model because the missing values are more fitted with the corresponding imputed values in the hybrid MLR-RNN method's fitness. Figure (10) shows the fitness between the missing values and the corresponding imputed values using MLR-RNN model where the missing ratio is (5%) and the missing distribution is successive.

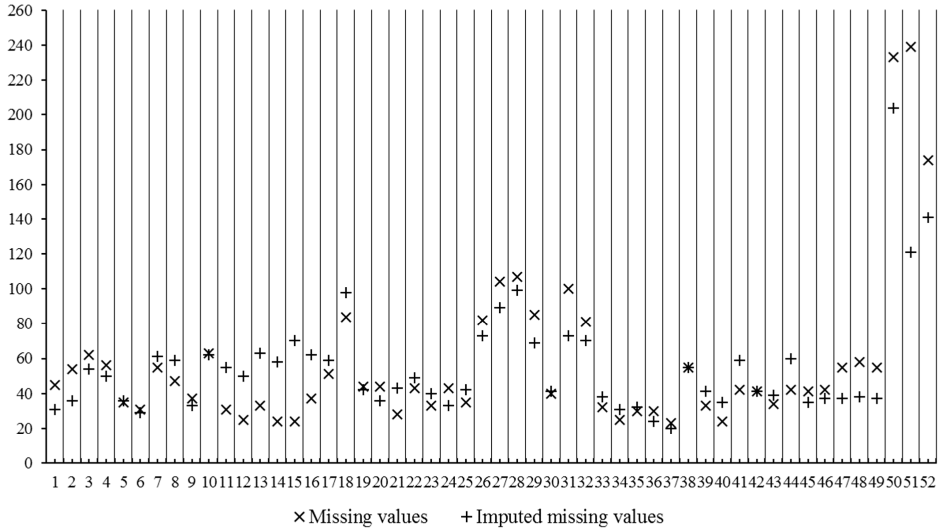

Fig. 10 The missing values of the original dependent variable (PM$_{10}$) and the imputed values using MLR-RNN method where the missing ratio is (5%) and the missing distribution are successive.

Figure (10) proved that the fitness between the missing values in PM$_{10}$ and the imputed values where the missing ratio is (5%) and the missing distribution is successively using the hybrid MLR-RNN method. It is better than the corresponding fitness by using MLR model because the missing values are more fitted with the corresponding imputed values in the hybrid MLR-RNNmethod's fitnessd. Figure (11) describes the fitness between the missing values and the corresponding imputed values using MLR-RNN method where the missing ratio is (10%), and the missing distribution is discrete.

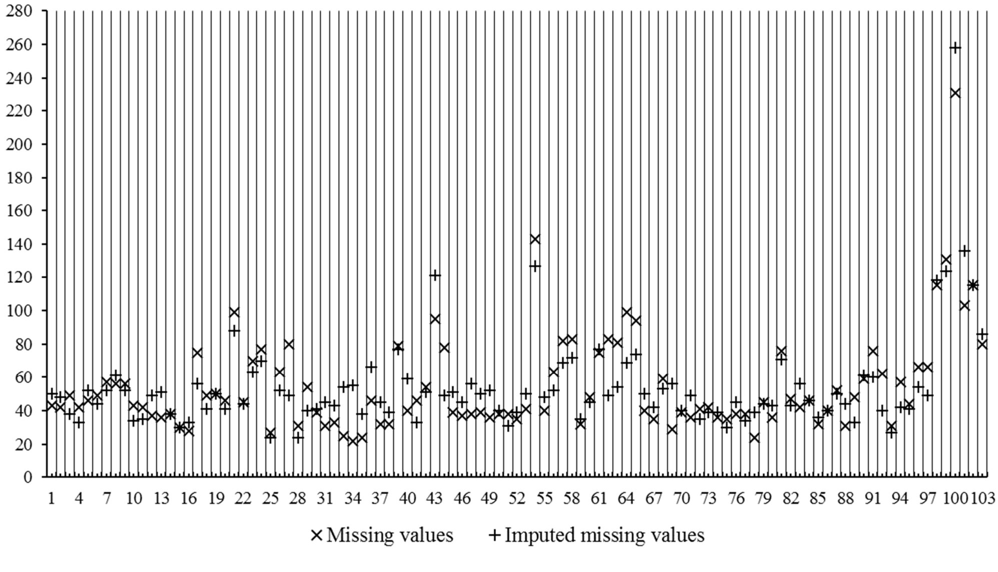

Fig. 11 The fitness between the missing values in PM$_{10}$ and the corresponding imputed values by using MLR-RNN method where the missing ratio is (10%) and the missing distribution is discrete.

Figure (11) proved that the fitness between the missing values in PM$_{10}$ and the imputed values where the missing ratio is (10%) and the missing distribution is discretely using the hybrid MLR-RNN method. It is better than the corresponding fitness by using MLR model because the missing values are more fitted with the corresponding imputed values in the hybrid MLR-RNN method's fitness. Figure (12) describes the fitness between the missing values and the corresponding imputed values using MLR-RNN method where the missing ratio is (10%) and the missing distribution is successive.

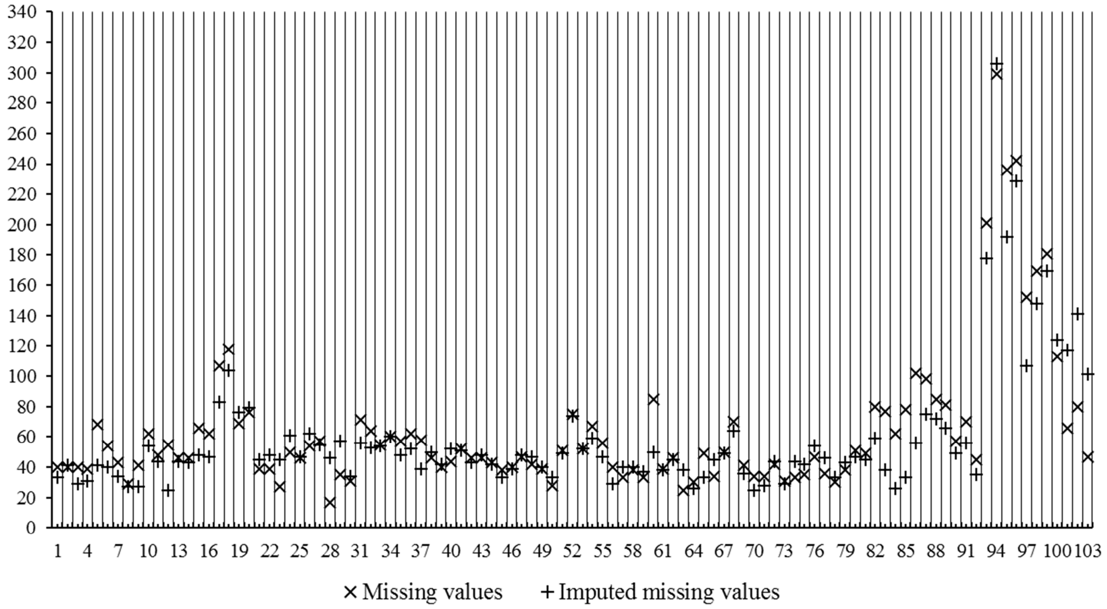

Fig. 12 The missing values of the original dependent variable (PM_{10}) and the imputed values by using MLR-RNN method where the missing ratio is (10%) and the missing distribution is successively.

Figure (12) proved that the fitness between the missing values in PM_{10} and the imputed values where the missing ratio is (10%) and the missing distribution is successively by using the hybrid MLR-RNN method. It is better than the corresponding fitness by using MLR model because the missing values are more fitted with the corresponding imputed values in the fybrid MLR-RNN method's fitness. Figure (13) indicates the fitness between the missing values and the corresponding imputed values using MLR-RNN method where the missing ratio is (25%), and the missing distribution is discrete.

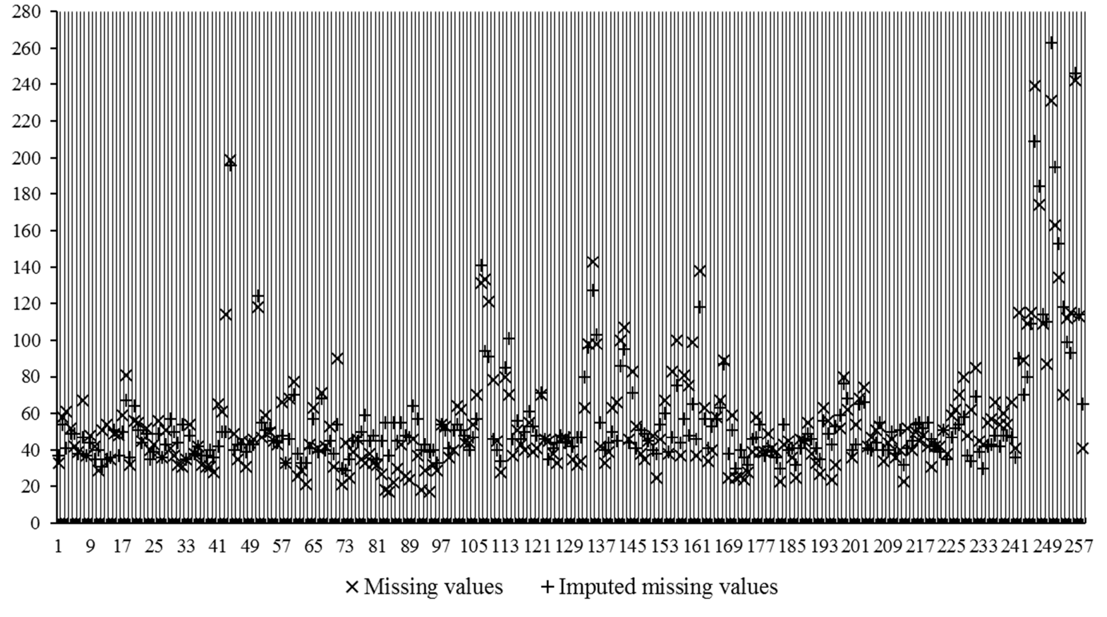

Fig. 13 The fitness between the missing values in PM_{10} and the corresponding imputed values by using MLR-RNN method where the missing ratio is (25%) and the missing distribution is discretely.

Figure (13) proved that the fitness between the missing values in PM_{10} and the imputed values where the missing ratio is (25%) and the missing distribution is discretely by using the hybrid MLR-RNN method. It iss better than the corresponding fitness by using MLR model because the missing values are more fitted with the corresponding imputed values in the hybrid MLR-RNN method's fitness. Figure (14) shows the fitness between the missing values and the corresponding imputed values by using the MLR-RNN method where the missing ratio is (25%) and the missing distribution is successive.

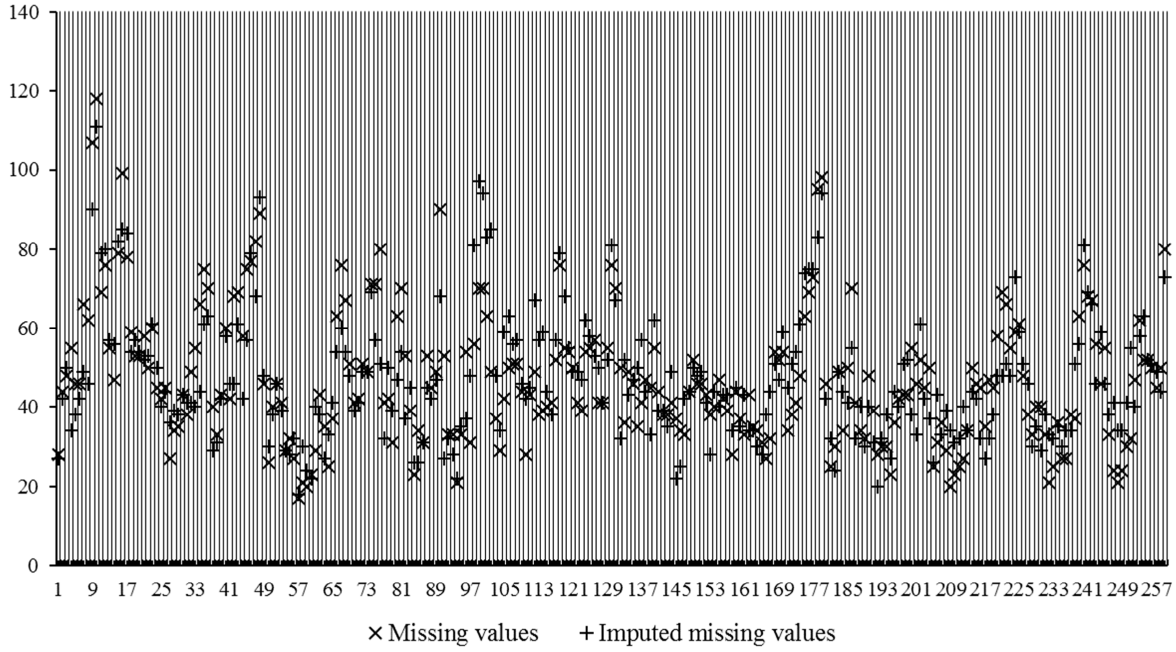

Fig. 14 The fitness between the missing values in PM_{10} and the corresponding imputed values by using MLR-RNN method where the missing ratio is (25%) and the missing distribution is successive.

Figure (14) proved that the fitness between the missing values in PM_{10} and the imputed values where the missing ratio is (25%) and the missing distribution is successively using the hybrid MLR-RNN method. It is better than the corresponding fitness by using MLR model because the missing values are more fitted with the corresponding imputed values in the hybrid MLR-RNN method's fitness. RMSE and MAE of the results of imputing the missing values by using the hybrid MLR-RNN method for different missing rates in two distribution methods (discretely and successively) are summarized as in Table 8.

TABLE VIII
SUMMARY OF THE ERROR MEASUREMENTS OF MISSING VALUES IMPUTATION BY USING THE HYBRID MLR-RNN METHOD.

Missing Values	5%		10%		25%	
Distribution	RMSE	MAE	RMSE	MAE	RMSE	MAE
Discretely	14.0	10.7	13.4	10.3	14.0	10.6
Successively	22.7	14.4	17.8	12.1	10.0	7.8

Table 8 shows that the results of the hybrid MLR-RNN method outperformed MLR model results because of RMSE and MAE values of the hybrid MLR-RNN method were less than the corresponding RMSE and MAE values of MLR model in Table 8.

IV. CONCLUSION

The imputation of missing values is essential before analyzing time series. The comparison between the proposed method and the traditional imputation methods had been shown that hybrid MLR-RNN outperformed the traditional MLR method. In conclusion, the missing values in PM_{10} data based on several climate independent variables with the nonlinearity problem can be imputed more accurately using the proposed method. Therefore, imputing the missing values of PM_{10} using the proposed method leads to more accurate performance.

REFERENCES

[1] Hardle W., Simar L., " Applied multivariate statistical analysis ", Berlin and Louvain-la-Neuve, Germany, 2003.5-Neil H.Timm," Applied multivariate analysis ",Springer verlag New York, Inc, 2002.

[2] Dubrov A., "Applied multivariate data analysis ", Statistica, Moscow, 1992.

[3] GBD Factors Collaborators. Global, regional, and national comparative risk assessment of 79 behavioral, environmental and occupational and metabolic risks or clusters of risks, 1990-2015 a systematic analysis for the Global Burden of Disease Study 2015.Lancet.2016 oct, 388(10053):1659-1724.

[4] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., & Yidris, N. (2020). Conceptual design of the cross-arm for the application in the transmission towers by using TRIZ–morphological chart–ANP methods. Journal of Materials Research and Technology, 9(4), 9182-9188.

[5] Luo, Y., Cai, X., Zhang, Y., & Xu, J. (2018). Multivariate time series imputation with generative adversarial networks. In Advances in Neural Information Processing Systems (pp. 1596-1607).

[6] Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). Brits: Bidirectional recurrent imputation for time series. Advances in Neural Information Processing Systems, 31, 6775-6785.

[7] Suo, Q., Yao, L., Xun, G., Sun, J., & Zhang, A. (2019, June). Recurrent Imputation for Multivariate Time Series with Missing Values. In 2019 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 1-3). IEEE.

[8] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., Yidris, N., & Fattahi, A. (2020). Experimental and numerical investigation of the mechanical behavior of full-scale wooden cross arm in the transmission towers in terms of load-deflection test. Journal of Materials Research and Technology, 9(4), 7937-7946.

[9] Nassar, L., Saad, M., Okwuchi, I. E., Chaudhary, M., Karray, F., & Ponnambalam, K. (2020, October). Imputation impact on strawberry yield and farm price prediction using deep learning. In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3599-3605). IEEE.

[10] Saad, M., Nassar, L., Karray, F., & Gaudet, V. (2020, October). Tackling Imputation Across Time Series Models Using Deep Learning and Ensemble Learning. In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 3084-3090). IEEE.

[11] Kim, C., Son, Y., & Youm, S. (2019). Chronic disease prediction using character-recurrent neural network in the presence of missing information. Applied Sciences, 9(10), 2170.

[12] Yoon, J., Zame, W. R., & van der Schaar, M. (2018). Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Transactions on Biomedical Engineering, 66(5), 1477-1490.

[13] Sangeetha, M., & Kumaran, M. S. (2020). Deep learning-based data imputation on time-variant data using recurrent neural network. Soft Computing, 1-12.

[14] Khan, Z., Khan, S. M., Dey, K., & Chowdhury, M. (2019). Development and evaluation of recurrent neural network-based models for hourly traffic volume and annual average daily traffic prediction. Transportation Research Record, 2673(7), 489-503.