Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 4840021, 16 pages
https://doi.org/10.1155/2022/4840021

Research Article

Data Imputation for Detected Traffic Volume of Freeway Using Regression of Multilayer Perceptron

Xiang Wang, Yingying Ma, Shengwen Huang, and Yu Xu

Department of Transportation Engineering, South China University of Technology, Guangzhou, Guangdong 510640, China

Correspondence should be addressed to Yingying Ma; mayingying@scut.edu.cn

Received 17 January 2022; Revised 24 March 2022; Accepted 7 April 2022; Published 5 May 2022

Academic Editor: Yu-Sheng Ci

Traffic volume data are the important part of research and application of intelligent transportation systems (ITS). However, data loss often happens due to various factors in the real world, which may cause large deviations in prediction or bad accuracy of optimizations. Imputation is a valid way to handle missing values. A multilayer perceptron-multivariate imputation of chain equation (MLP-MICE) regression imputation method optimized by the limit-memory-BFGS algorithm is proposed, considering the temporal and spatial characteristics of traffic volume. Also, 32 groups of simulated imputation experiments based on the detected traffic volume of road sections in the Guangdong freeway system are conducted, which take the scenarios of continuous missing and jumped missing into account. The results of the experiments show that the MLP-MICE can optimize the imputation performance in the missing value of traffic volume with the MAPE of imputation results from 6.38% to 30%. Meanwhile, the proposed model has higher imputation accuracy for the traffic volume data with a lower degree of mutation. Lastly, the performance of the proposed model of imputation in short-term traffic volume prediction is discussed using the support vector machine. The results of it show that the MAPE of prediction under the proposed model is much lower than all-zero imputation. Therefore, the proposed model in this study is positive on improving the accuracy of traffic volume prediction and intelligent traffic control and management.

1. Introduction

Traffic volume data are an important part of Intelligent Transportation Systems (ITSs); however, missing data are widespread and inevitable problem due to the failure of detectors and information processing errors, which certainly have negative effects on the application of ITS due to the temporal and spatial characteristics of traffic flow [1, 2]. Moreover, the data failure is likely to be culled in the actual study because the size of overall data is large [3], which will reduce the accuracy of prediction and optimization of the road management system. Therefore, effective imputation methods for missing traffic volume data are necessary.

Imputation of missing data means to replace the missing values with estimated values [4]. The imputation methods can be divided into single imputation and multiple imputation according to the times of imputation [5]. Single imputation includes statistical and machine learning-based methods [6]. Mean imputation and regression imputation are mainly statistical-based methods. Mean imputation is easy to operate, but its performance is limited because of the underestimation of imputation results and the neglect of the relations with other variables [7], while regression imputation is realized by obtaining alternative values by regression, such as spatial autoregression models [8] and logistic regression [9]. Machine learning-based imputation methods have been proposed with the development of data information technology in recent years [10], such as support vector regression [11], residual learning networks [12], and semisupervised regression [13]. Hot deck imputation [14, 15] and cluster imputation [16] are also widely used in the imputation of missing data. Previous studies show that the combination of single imputation methods can improve the performance of imputation, for example, general regression auto associative neural network (GRAANN) [17] is a hybrid of mean imputation and machine learning, and fuzzy c-means support vector regression genetic algorithm

(Fuzzy C-means SVRGA) [18] is combined by clustering imputation and machine learning.

Generally, multiple imputation (MI) requires at least two times of imputation [19]. The basic idea of MI is to create multiple copies of data containing missing values, perform the imputation of each copy independently, and then select the imputation results according to certain evaluation criteria [4, 20]. MI is efficient to improve the precision of imputation [21–24] and Multivariate Imputation by Chained Equations (MICE) [25] is a commonly used method for MI. It is verified that good performance can be obtained from the hybrid of MICE and regression imputation [26].

The imputation methods are widely used in the fields of economy [8], computer [9, 13], biology [11], environment [12, 27], medicine [21, 28], and so on. Most of them are used to resolve the nonresponse of the questionnaire survey, which has weak temporal and spatial correlation. Traffic volume is a kind of data with strong temporal continuity and spatial correlation, which should be considered in the imputation process of accuracy-improving [29]. Previous studies show that traffic volume imputation methods are also mainly based on statistical learning and machine learning [30]. Statistical learning methods take full advantage of the statistical feature of traffic volume [31], and it mainly includes improved principal components analysis (PCA) methods such as PPCA [32], KPPCA [31], fuzzy theory [33, 34], and tensor completion [35–38]. Machine learning methods estimate the missing value by machine learning [30] or deep learning algorithms [39], for example, SVR [40, 41], DSAE [42], KNN [43], CNN [44, 45], LSTM [46, 47], GAN [48], and DEB [49]. Comparing with the statistical learning method, the machine learning method makes use of more characteristics of traffic volume, especially the temporal features [30]. And the spatiotemporal (ST) features are considered such as ST-BiRT [38] and ST-PTD [50]. The details of the imputation methods used in dealing with traffic volume data are shown in Table 1.

Most imputation methods for traffic volume are single imputation methods, the results of which are generally underestimated [18], and they are more suitable for complex and large traffic systems with a large scale of traffic volume data, where a combination with other kinds of methods is not recommended due to the increasing calculation. However, it is different for smaller systems with a small scale of data; the imputation accuracy of it can be improved by using a combination of multiple methods, while MI is an effective one. In addition, such a combination can effectively reduce the influence of abnormal fluctuations of traffic data caused by external random factors on the final output result.

Neural network model is promising to obtain accurate imputation results [39]. Multilayer perceptron (MLP), a deep learning model, finds application in representing the nonlinear features in traffic prediction, and missing value imputation [41] has been chosen in this study. Therefore, to study a simple traffic system and to fully use the performance of MLP with MI to improve the precision of imputation, a hybrid model of deep learning and multiple imputation called MLP-MICE is proposed to impute the missing data of detected traffic volume. Comparing with the unprocessed data, inputting the data fixed by the proposed MLP-MICE into the prediction model improves the accuracy effectively, which benefit the intelligent traffic control and management. The rest of this study is organized as follows. Section 2 introduces the methodology, including the framework of MLP-MICE and each part of it. Section 3 verifies the proposed model using freeway detected data and analyses the performance. Section 4 concludes this study with some remarks.

2. Methodology

2.1. Model Framework. The structure of the model is shown in Figure 1. Considering the temporal continuity and spatial correlation of freeway traffic volume, MLP was used to impute the missing traffic volume data, and the L-BFGS algorithm was used to optimize the parameters of MLP. The model is divided into three parts, which are data preparation, MLP imputation, and MICE process.

2.2. Data Preparation. The input of the model is defined as an $M \times q$ matrix denoted as D:

$$\text{input} = D = \begin{pmatrix} d_{t+1,1} & \cdots\cdots & d_{t+1,q} \\ d_{t+2,1} & \cdots\cdots & d_{t+2,q} \\ \vdots & & \vdots \\ d_{t+M,1} & \cdots\cdots & d_{t+M,q} \end{pmatrix}, \tag{1}$$

where $d_{t+i,j}$ is the traffic volume of time interval i at location q, M represents the number of time intervals for the imputation model, q is the number of detecting locations, and t is the started time of detectable data for imputation. Suppose N elements are missing from $d_{t+m,n}$, where m represents the starting time interval of the missing data, n represents the location number, and $M = m + N$, which means the data of all the locations during and before the missing data are used to input into the model. M needs to be determined by the data distribution, and equation (2) can be used when the model has been tested to have excellent imputation performance on the missing rate α, in which $M_{\min}$ is the minimum data sample size to get relatively accurate results:

$$M = \max\left(M_{\min}, \frac{N}{\alpha}\right). \tag{2}$$

Finally, the following equation normalizes the elements of nonmissing values of the matrix to speed up the convergence of the model as $d'_{t,q}$:

$$d'_{t,q} = \frac{d_{t,q}}{\max\left(d_{t,q}\right) - \min\left(d_{t,q}\right)} \quad t \in [t+1, M+1], \quad q \in [1, q]. \tag{3}$$

2.3. Multilayer Perceptron Imputation. The multilayer perceptron (MLP) neural network structure is interconnected by many nodes and contains four layers.

Table 1: Measurement conversions.

Imputation method	MAPE or RMSE	Urban road or freeway	Data sources
Principal component analysis (PCA)	76~84 [32]	Urban road	China
	14~24 [31]	Freeway	America
Fuzzy rough set (FRS)	4.768~6.533 [33]	Freeway	China
	6.9766~10.4998 [34]	Freeway	China
Tensor completion	4.0893~5.3544 [35]	Urban road	China
	10.3%~12.71% [36]	Freeway	America
	7.3%~19.98% [37]	Urban road	China
	0.91%~64.95% [38]	Urban road	China
Support vector machine (SVR)	4~14 [40]	Both	China
	5.7232% [41]	Urban road	China
Denoising stacked autoencoders (DSAE)	13.9~20.9 [42]	Freeway	America
Convolutional neural network (CNN)	≤24 [44]	Freeway	America
Long short-term memory (LSTM)	9.63~17.54 [46]	Urban road	America
	1.927~9.192 [47]	Freeway	America
Generative adversarial networks (GAN)	3.66~10.86 [48]	Urban road	China and America
Dual-stage error-corrected boosting regressor (GBR)	1.39%~6.08% [49]	Freeway	America
Spatiotemporal-PTD	3.45~8.35 [50]	Urban road	China

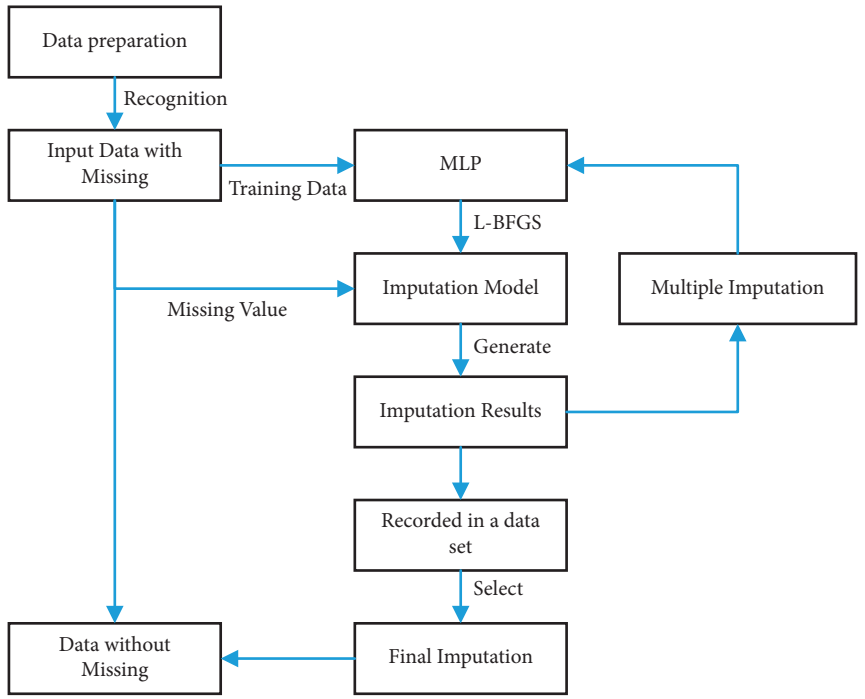

Figure 1: Overall structure of the model.

The nodes belonging to different layers process the input data through the activation function. Then, the results are transmitted from layer to layer [51]. The MLP regression imputation model we use is illustrated in Figure 2. Define the data $a_{1\sim m}^{[l-1]}$, the weight matrix W, and the bias term matrix B, where $a_{1\sim m}^{[l-1]}$ is the data that input to layer l and $w_{ij}^{(l)}$ is the element of W, which denotes the weight of the i^{th} node of the layer $l-1$ to the j^{th} node of the layer l, and $b_j^{(l)}$ is the element of B, which represents the numerical deviation of the input to the j^{th} node of the layer l. $z^{(l)}$ represents the output dataset of each node in layer l:

$$z^{[l]} = \begin{bmatrix} w_{11}^{[l]} & w_{12}^{[l]} & \cdots & w_{1m}^{[l]} \\ w_{21}^{[l]} & w_{22}^{[l]} & & w_{2m}^{[l]} \\ \vdots & & \ddots & \vdots \\ w_{n1}^{[l]} & w_{n2}^{[l]} & \cdots & w_{nm}^{[l]} \end{bmatrix} \cdot \begin{bmatrix} a_1^{[l-1]} \\ a_2^{[l-1]} \\ \vdots \\ a_m^{[l-1]} \end{bmatrix} + \begin{pmatrix} b_1^{[l]} \\ b_2^{[l]} \\ \vdots \\ b_n^{[l]} \end{pmatrix}. \tag{4}$$

Choose tanh as the activation function and the following expression can be obtained:

$$\begin{aligned} a^{[l]} &= \tan h\left(z^{[l]}\right) \\ &= \tan h\left(w^{[l]}a^{[l-1]} + b^{[l]}\right). \end{aligned} \tag{5}$$

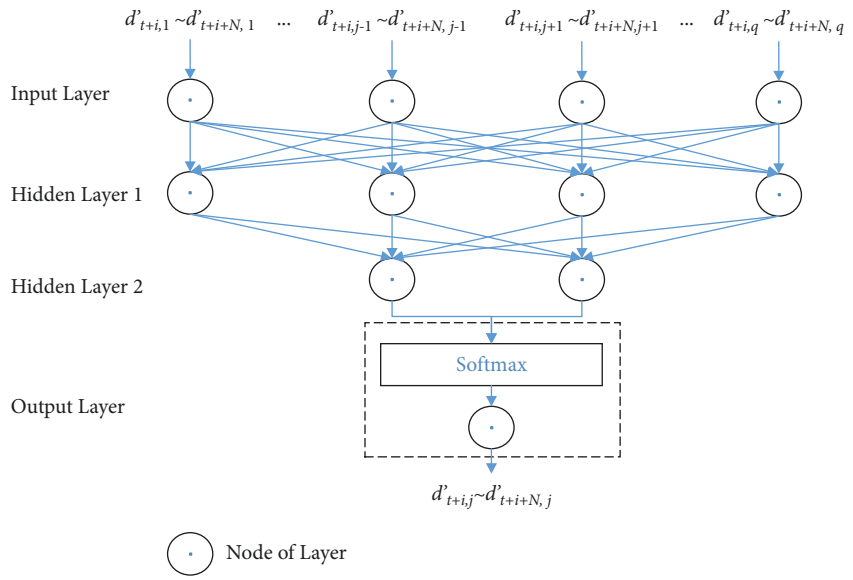

Figure 2: Working principle of MLP.

Definition:
θ_0: the first run parameter of MLP
μ: the tolerable maximum positive number
iter: iteration of L-BFGS, $iter = k(k = 0,1,2, \ldots)$
m: latest m groups of iteration results are used for calculation
ε: optimized step length
Procedure L-BFGS
Calculate B_0 and $a\nabla h(\theta_k)$ according to θ_0 and the value of the loss function
While $|\nabla h'(\theta_k)| > \mu$ do
get $(\theta_k + \varepsilon r) = \min(f(\theta k + \varepsilon_i r))$ from:
Let: $\Delta g = \Delta g_k$, $B = B_k$
for $i = k-1, k-2, \ldots, k-m$:
$\varepsilon_i = (\Delta\theta_i^T \cdot \Delta g / \Delta g_i^T \cdot \Delta\theta_i)$
$\Delta g = \Delta g - \varepsilon_i \Delta g_i$
end for
$r = B \cdot \Delta g$
for $i = k-m, k-m+1, \ldots, k-1$:
$\beta = (\Delta g_i^T \cdot r / \Delta g_i^T \cdot \Delta\theta_i)$
$r = r + \Delta\theta_i(\varepsilon_i - \beta)$
end for
stop with result $B_k \nabla h(\theta) = -r$
$\theta_{k-\text{renew}} = \theta_k + \varepsilon r$, $\theta \in \{w, b\}$
$k = k + 1$
$|h'(\theta_k)| = |h'(\theta_{k-\text{renew}})|$
end while
end procedure

Algorithm 1: L-BFGS.

When l is the output layer, it outputs the normalizing estimation of the missing values. The activation method in the output layer is softmax regression.

2.4. Parameters Optimization Based on L-BFGS. L-BFGS algorithm (limit-memory-BFGS) is used to optimize the parameters of MLP. L-BFGS is a kind of approximate quasi-Newton method. It is commonly used to solve unconstrained nonlinear programming problems with the advantages of fast convergence and low memory overhead (Algorithm 1).

Define $x^{(i)}$ is the i^{th} element of set X, $R_{WB}(x^{(i)})$ is the result of imputation for the value of the i^{th} time interval with weight W and the numerical deviation B for a certain location, $y^{(i)}$ is the observed value of the i^{th} time interval for this location, and $w^{(ij)}$ are elements of W. Then, the loss function of the set X, denoted as $f_{\text{loss}}(X)$, can be described as follows:

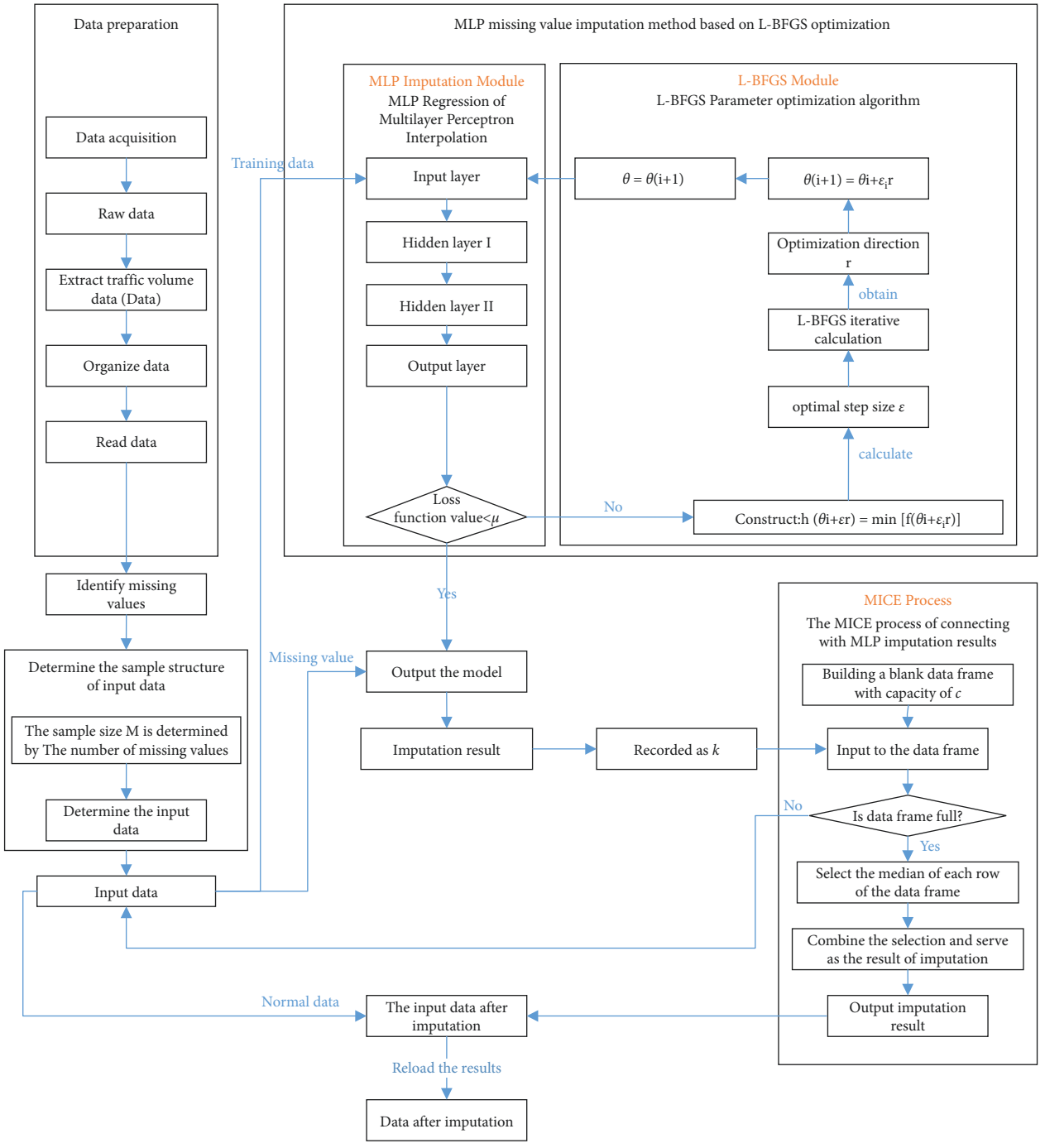

Figure 3: Flowchart of missing value imputation of the proposed model.

$$f_{\text{loss}}(X) = \frac{1}{2M}\sum_{i=1}^{m}\left[R_{WB}\left(x^{(i)}\right) - y^{(i)}\right]^2 + \delta L_2,$$
$$L_2 = \sqrt{\sum\left(w^{(ij)}\right)^2}, \quad w^{(ij)} \in W. \tag{6}$$

L_2 is added in equation (6) to avoid the overfitting phenomenon and local optimization, and δ is the coefficient of L_2. The value of the loss function is related to the parameters W and B of the MLP. Let θ denote the set of W and B, and then, define $h(\theta)$ as follows:

$$h(\theta) = f_{\text{loss}}(X), \quad \text{in which } \theta = \{W, B\}. \tag{7}$$

θ needs to be generated by iteration to meet the near minimum of $f_{\text{loss}}(X)$ [52]. Let k denote the iteration times. When the formula is expanded by Taylor expansion at θ_k after k iterations, equation (8) can be obtained:

$$h(\theta) = h(\theta_k) + \nabla h(\theta_k)(\theta - \theta_k) + \frac{1}{2}(\theta - \theta_k)^T \nabla^2 h(\theta_k)(\theta - \theta_k), \tag{8}$$

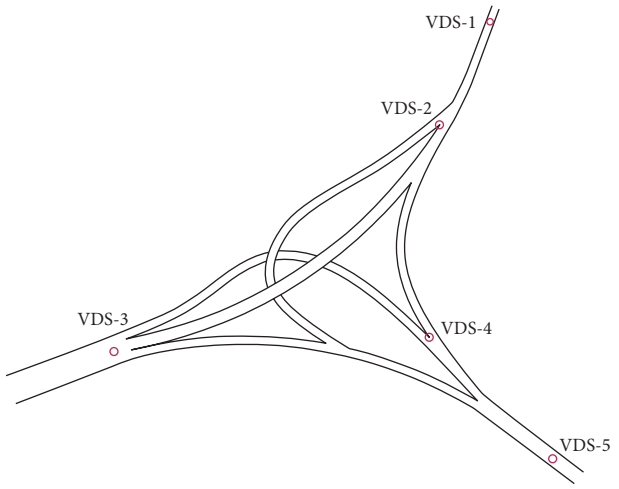

Figure 4: Locations of the five VDS.

where $\nabla h(\theta_k)$ is the gradient vector of $h(\theta)$, $\nabla^2 h(\theta_k)$ represents the Hessian matrix for $h(\theta)$, and R parameters are in θ_k, and θ_{k+1} can be obtained by the following equation :

$$g_k = \nabla h(\theta_k) = \begin{bmatrix} \frac{\partial h}{\partial \theta_{k1}} \\ \frac{\partial h}{\partial \theta_{k2}} \\ \vdots \\ \frac{\partial h}{\partial \theta_{kR}} \end{bmatrix},$$

$$H_k = \nabla^2 h(\theta_k) = \begin{bmatrix} \frac{\partial^2 h}{\partial \theta_{k1}^2} & \frac{\partial^2 h}{\partial \theta_{k1}\partial \theta_{k2}} & \cdots & \frac{\partial^2 h}{\partial \theta_{k1}\partial \theta_{kR}} \\ \frac{\partial^2 h}{\partial \theta_{k2}\partial \theta_{k1}} & \frac{\partial^2 h}{\partial \theta_{k2}^2} & \cdots & \frac{\partial^2 h}{\partial \theta_{k2}\partial \theta_{kR}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 h}{\partial \theta_{kR}\partial \theta_{k1}} & \frac{\partial^2 h}{\partial \theta_{kR}\partial \theta_{k2}} & \cdots & \frac{\partial^2 h}{\partial \theta_{kR}^2} \end{bmatrix}, \tag{9}$$

$$\theta_{k+1} = \theta_k - H_k^{-1} g_k.$$

To simplify the calculation, an approximation is made to equation (9). Take the derivative at $k+1$ of $h(\theta)$, and express it in the form of a gradient operator as in the following equation:

Table 2: Statistical property of 15-minute traffic flow of each VDS (unit: Veh).

VDS	VDS-1	VDS-2	VDS-3	VDS-4	VDS-5
Average	138.5	317.8	185	152.6	146.3
Median	119	287	167	132	127
Maximum	667	814	559	484	665

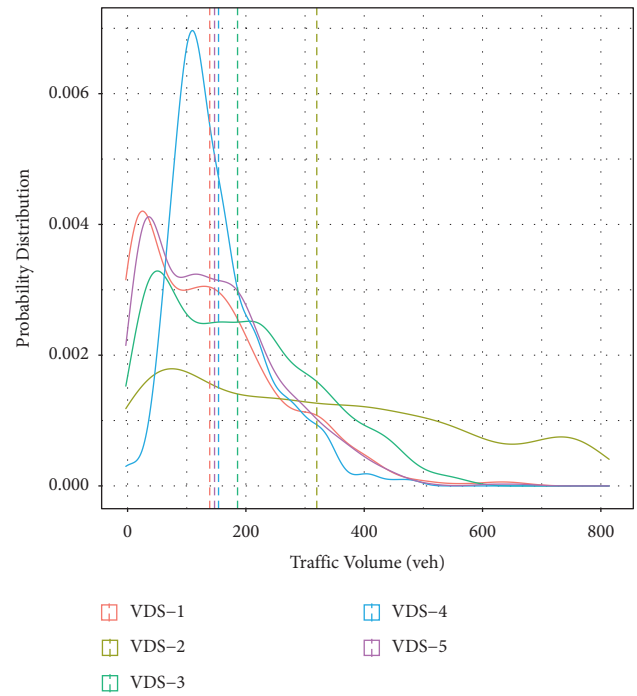

Figure 5: Kernel density estimation (KDE) probability distribution curve of traffic flow.

$$\begin{cases} \nabla h(\theta_k) \approx \nabla h(\theta_{k+1}) + H_{k+1}(\theta_k - \theta_{k+1}) \\ \Leftrightarrow g_{k+1} - g_k = H_{k+1}(\theta_{k+1} - \theta_k), \\ \Delta g_k = g_{k+1} - g_k, \\ \Delta \theta_k = \theta_{k+1} - \theta_k, \\ H_{k+1} = B_{k+1}, \end{cases} \tag{10}$$

$$\Rightarrow \begin{cases} \Delta g_k = B_{k+1} \cdot \Delta \theta_k, \\ \Delta \theta_k = B_{k+1}^{-1} \cdot \Delta g_k. \end{cases}$$

Record B_{k+1} for the first approximation:

$$B_{k+1} = B_k + \Delta B_k. \tag{11}$$

Therefore,

$$\Delta B_k = \Delta g_k \cdot (\Delta \theta_k)^{-1} - B_k \cdot \Delta \theta_k \cdot (\Delta \theta_k)^{-1},$$

$$B_{k+1} = B_k + \Delta B_k = B_k + \frac{\Delta g_k \cdot \Delta g_k^T}{\Delta g_k^T \cdot \Delta \theta_k} - \frac{B_k \cdot \Delta \theta_k \cdot \Delta \theta_k^T \cdot B_k}{\Delta \theta_k^T \cdot B_k \cdot \Delta \theta_k}. \tag{12}$$

Then, B_{k+1}^{-1} can be formulated as follows, where I is a unit matrix:

Table 3: Autocorrelation coefficient and correlation coefficient of VDS-1 to VDS-5.

VDS	Autocorrelation coefficient		Correlation coefficient				
	lag = 1	lag = 2	VDS-1	VDS-2	VDS-3	VDS-4	VDS-5
VDS-1	0.903	0.861	1	0.798	0.759	0.496	0.882
VDS-2	0.950	0.923		1	0.834	0.593	0.791
VDS-3	0.940	0.908			1	0.674	0.836
VDS-4	0.773	0.702				1	0.582
VDS-5	0.925	0.892					1

Table 4: Data fragmentation.

VDS	Missing data	Number of missing data	Missing rate (%)
VDS-1	010709-010723	15	2.3
VDS-2	020409-020446,020448	38	5.7
VDS-3	—	0	0
VDS-4	040357,040392,040507	3	0.4
VDS-5	050670-050673	4	0.6

$$B_{k+1}^{-1}=\left(I-\frac{\Delta\theta_k\cdot\Delta g_k^T}{\Delta g_k^T\cdot\Delta\theta_k}\right)B_k^{-1}\left(I-\frac{\Delta g_k\cdot\Delta\theta_k^T}{\Delta g_k^T\cdot\Delta\theta_k}\right)+\frac{\Delta\theta_k\cdot\Delta\theta_k^T}{\Delta g_k^T\cdot\Delta\theta_k}, \quad (13)$$

while

$$V_k=I-\frac{\Delta\theta_k\cdot\Delta g_k^T}{\Delta g_k^T\cdot\Delta\theta_k}. \quad (14)$$

Therefore,

$$B_{k+1}=\left(V_k^TV_{k-1}^T\ldots V_1^TV_0^T\right)B_0\left(V_0V_1\ldots V_{k-1}V_k\right)+\sum_{i=0}^{k-1}\left[\left(V_k^TV_{k-1}^T\ldots V_{i+2}^TV_{i+1}^T\right)\cdot\frac{\Delta\theta_i\cdot\Delta\theta_i^T}{\Delta g_i^T\cdot\Delta\theta_i}\cdot\left(V_{i+1}V_{i+2}\ldots V_{k-1}V_k\right)\right]+\frac{\Delta\theta_k\cdot\Delta\theta_k^T}{\Delta g_k^T\cdot\Delta\theta_k}. \quad (15)$$

In order to simplify the calculation and reduce the required memory, the latest m groups of θk are used in the calculation, and B_{k+1} is approximated for the second time. Then, the following equation can be expressed as

$$B_k=\left(V_{k-1}^T\ldots V_{k-m}^T\right)B_{k-m}\left(V_{k-m}\ldots V_{k-1}\right)+\sum_{i=k-m}^{k-2}\left[\left(V_{k-1}^T\ldots V_{i+1}^T\right)\cdot\frac{\Delta\theta_i\cdot\Delta\theta_i^T}{\Delta\theta_i^T\cdot\Delta g_i}\cdot\left(V_{i+1}\ldots V_{k-1}\right)\right]+\frac{\Delta\theta_{k-1}\cdot\Delta\theta_{k-1}^T}{\Delta\theta_{k-1}^T\cdot\Delta g_{k-1}}. \quad (16)$$

Equation (16) has two purposes: one is to find the feasible direction of iterations, and the other is to determine the specific calculation method of iterative optimization. The expression of the direction r is

$$r=-B_k\cdot\nabla h\left(\theta_k\right). \quad (17)$$

Table 5: Imputation results of continuous missing experiments (MAPE).

VDS missing rate (%)	VDS-1	VDS-2	VDS-3	VDS-4	VDS-5
	Continuous missing: subset I				
10	8.21	5.01	12.38	13.75	12.37
20	7.34	4.86	9.84	12.5	9.48
30	8.25	4.44	8.88	11.65	7.92
40	8.41	8.44	9.77	18.41	6.38
50	20.66	8.42	8.16	23.92	16.19
60	48.02	8.82	22.55	76.63	25.97
70	57.26	9.29	36.7	72.46	20.22
80	57.99	17.78	36.67	70.12	29.47
	Continuous missing: subset II				
10	28.79	13.04	24.96	11.19	48.01
20	56.16	14.63	25.29	11.75	48.48
30	43.54	12.4	34.99	11.46	65.95
40	122.45	49.97	40.58	28.83	58.48
50	243.51	172.79	14.38	13.99	162.64
60	182.74	158.25	52.27	11.86	149.89
70	402.28	130.66	44.61	20.10	214.85
80	437.74	143.48	258.09	121.44	170.83
	Jumped missing: subset I				
10	19.36	14.34	10.38	18.42	20.45
20	20.48	13.23	7.60	15.33	24.84
30	19.58	12.69	6.46	17.93	25.10
40	17.01	11.70	7.79	15.23	23.38
50	11.23	15.11	10.82	19.74	14.64
60	19.60	17.65	10.92	30.99	23.51
70	68.41	39.38	39.36	52.96	59.41
80	80.42	64.31	69.68	57.63	75.78
	Jumped missing: subset II				
10	81.73	34.40	37.06	20.89	28.68
20	63.29	29.21	22.40	33.24	26.11
30	121.62	50.11	27.97	48.75	47.77
40	138.06	55.89	30.76	53.66	54.15
50	199.04	86.17	25.27	50.43	98.64
60	138.16	91.35	29.80	52.79	89.44
70	303.07	86.42	99.19	55.77	91.69
80	291.91	72.55	140.31	81.36	107.50

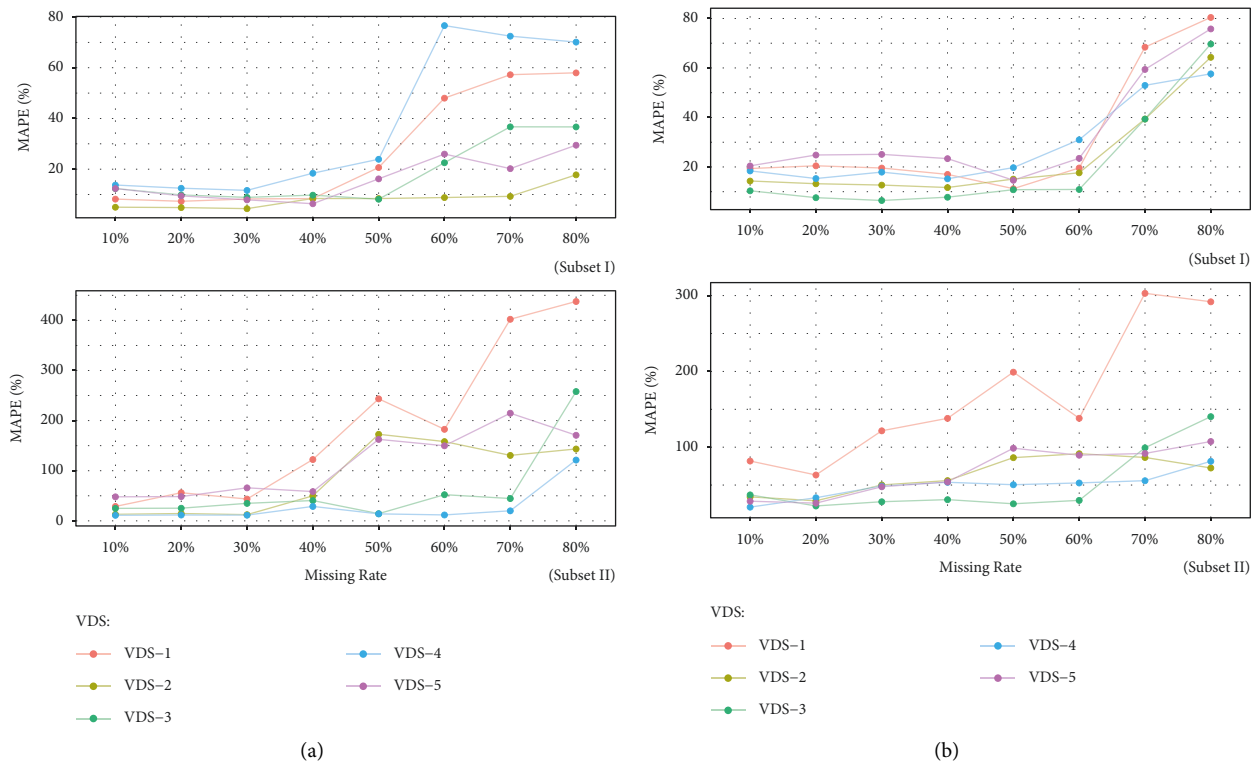

Figure 6: Imputation experimental results of Subset I and Subset II: cross analysis of (a) continuous missing scenarios and (b) jumped missing scenarios.

A two-loop recursion algorithm can be used to obtain the specific direction of parameter optimization and the optimized step length ε according to equation (17).

The L-BFGS algorithm is shown as follows:

2.5. Multivariate Imputation by Chain Equation. The parameters of MLP are updated when the algorithm ends. Multivariate imputation by chained equations (MICE) is a multiple imputation method that can realize flexible imputation of missing values as shown in the following four steps [53]. The process of MICE is shown as follows.

Step 1: construct a data frame with a capacity.

Step 2: fill the data frame with the imputation results from MLP and the evaluation of each result. Different types of missing should have different filling functions.

Step 3: repeat step 2. If the data frame is filled, go to step 4.

Step 4: select the final imputation result of the missing values from the data frame according to the evaluation.

Through the above process, the MLP-MICE regression imputation method optimized by the L-BFGS algorithm is proposed. The complete process of the imputation model proposed in this study is shown in Figure 3.

3. Results and Discussion

3.1. Empirical Analysis. This study takes the detected traffic volume data of VDS (Video Detection Systems) in 5 locations around an interchange of the freeways in Guangdong Province as an example, which is shown in Figure 4, to conduct an empirical analysis of the proposed model. Traffic volume is extracted by image recognition from each VDS. The data used in the study are collected from 0:00 on May 1, 2020, to 24:00 on May 7, 2020, including two workdays and five holidays, which cover many various scenarios. Name each VDS of a different location, from VDS-1 to VDS-5. The time interval of data collection is 15 minutes, and 672 pieces of data are collected in total. The statistical property of the 15-minute traffic volume of each VDS is shown in Table 2.

The probability distribution of the collected data is drawn by kernel density estimation (KDE), as shown in Figure 5. The higher the peak of probability curve, the more concentrated the traffic volume. The further to left the area enclosed by the coordinate axis and curve, the lower the overall traffic volume.

In Table 3, lag represents the different order of autocorrelation. A high correlation is considered when the correlation coefficient is greater than 0.5. Table 3 illustrates that the adjacent VDS has a significant correlation, which means the detected traffic volume has not only temporal relation but also spatial relation among VDSs near each other.

In the process of data collection, data loss is inevitable due to the breakdown of electronic equipment or other environmental factors. The data are sorted and labeled in the form of a VDS-days-time series. For example, 010236 represents the 36^{th} data collected on the second day of VDS-1. The specific missing data from real-world VDS are shown in Table 4.

Table 6: Experimental results of simulated imputation.

VDS		VDS-1	VDS-2	VDS-3	VDS-4	VDS-5
Variance of adjacent difference		44.42	59.08	34.90	40.25	36.22
			VDS imputation result (best MAPE)			
Continuous missing	Subset I	8.25	4.44	8.88	11.65	7.92
	Subset II	28.79	13.04	24.96	11.19	48.01
Jumped missing	Subset I	11.23	15.11	10.82	19.74	14.64
	Subset II	63.29	29.21	22.40	33.24	26.11

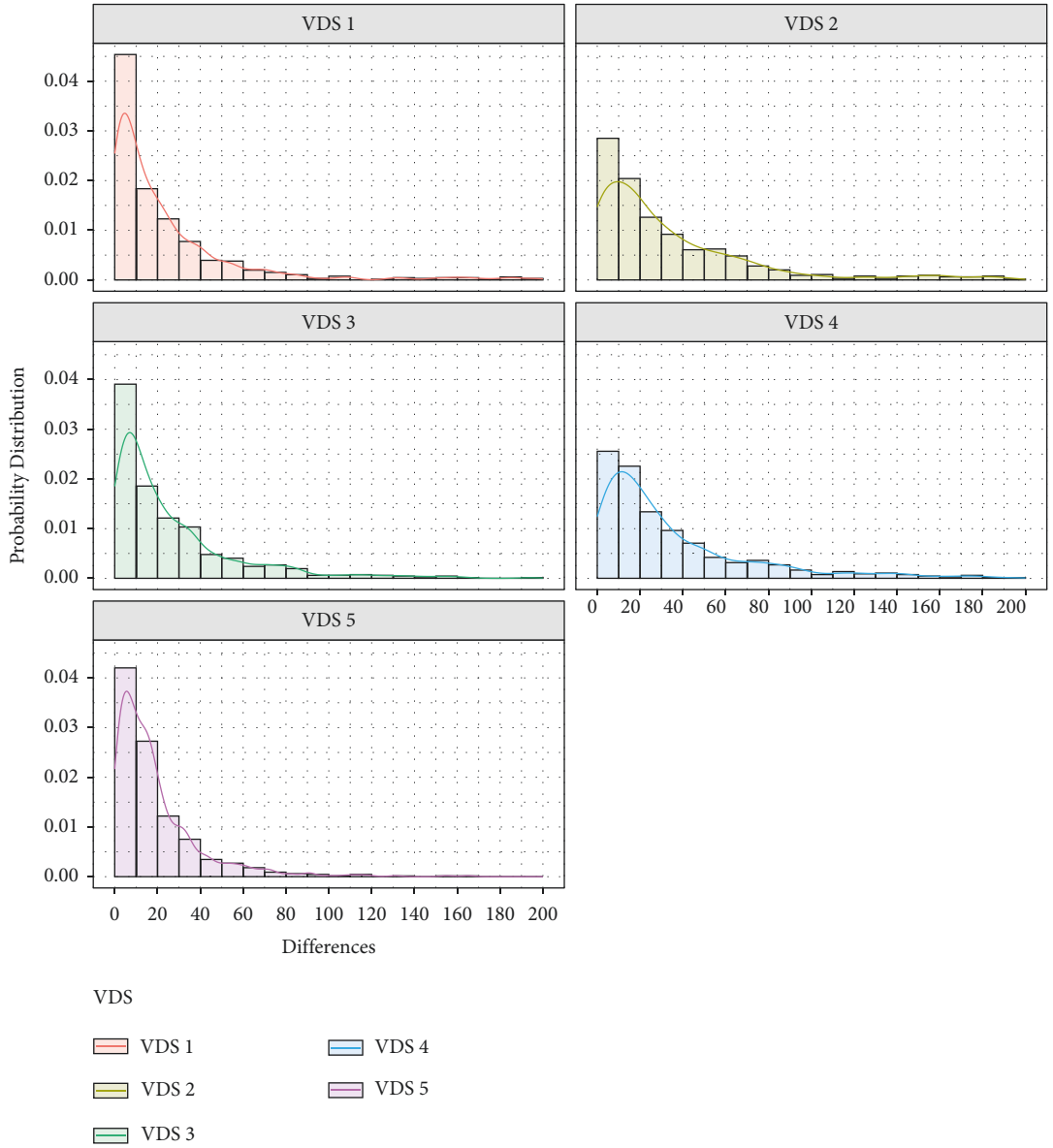

Figure 7: Probability distribution of the adjacent difference of traffic volume among VDS.

According to the data of case study, we can see that data missing happens. Although the data missing rate is not very high in this case, it will still have some impact on data applications. On the other hand, for the imputation model in this study, the small data missing rate ensures that there are sufficient observed values for the model test and performance analysis.

Imputation experiments of different scenarios are carried out and comparisons are made with other methods. The missing rate denotes the degree of missing in the dataset, which affects the output of the imputation model. There are two patterns of missing value, which are continuous missing and jumped missing. Continuous missing happens when the VDS cannot work for a long

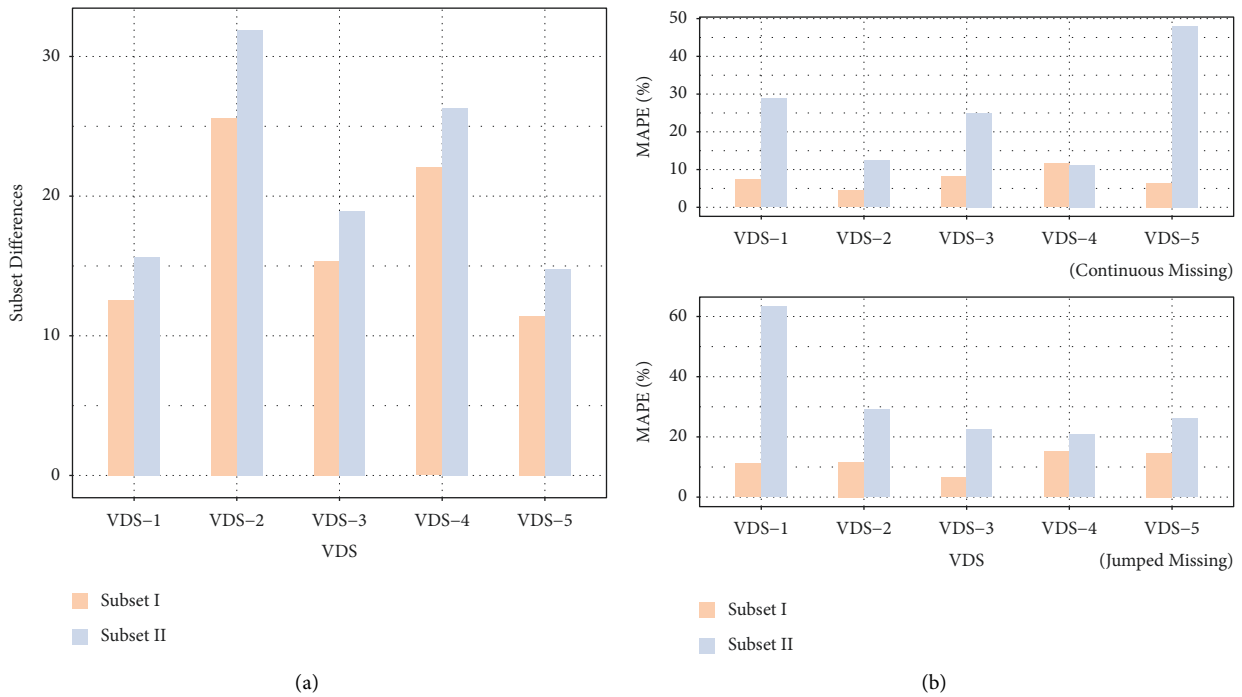

Figure 8: Comparison of the subsets' differences of the VDS. (a) Subset differences' comparison and (b) imputation performance comparison.

Table 7: Comparison of imputation results of different models on continuous missing (MAPE).

Imputation method	Subset I ($\alpha = 10\%$)	Subset II ($\alpha = 30\%$)	Imputation method	Subset I ($\alpha = 10\%$)	Subset II ($\alpha = 30\%$)
MLP-MICE	4.42	14.88	Random Forest	5.72	17.11
MLP	5.75	16.17	Ababoost	6.77	29.71
KNN	7.33	25.51	Gradient rise	5.71	17.37
Decision tree	8.68	19.77	Bagging	7.51	21.44
SVR	8.38	91.97	Extremely random tree	7.69	24.15

time for some reason, while jumped missing happens when the VDS temporarily breaks down. To simulate the experiment of different missing scenarios, set the missing rate from 10% to 80% for both continuous missing and jumped missing, and 10% is used as the span. The MLP used in this case has two hidden layers which consist of 4 nodes and 2 nodes separately, and let M be 60 according to the analysis of the detected data.

Select data of the 1^{st}–120^{th} time intervals of all the locations as the dataset of the simulated imputation experiment. Define the 1^{st}–60^{th} elements as subset I and the 61^{th}–120^{th} as subset II. During the experiment, a piece of continuous data with a length of ten was randomly removed to simulate the continuous missing scenario. Ten discontinuous data were randomly removed to simulate a jumped missing scenario. Meanwhile, to verify the superiority of the imputation model, MLP, random forest, and decision tree were selected as the control groups of the experiment.

The experiment uses mean absolute percentage error (MAPE) to evaluate the imputation performance:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{\ddot{y}_i - y_i}{y_i} \right|. \quad (18)$$

In equation (18), y_i represents the i^{th} observed value and $\ddot{y}_i$ represents the i^{th} imputation result. The smaller the MAPE, the better the imputation performance.

3.2. Results' Analysis. The results of experiments of MLP-MICE are shown in Table 5.

Table 5 and Figure 6(a) show that the model has the best imputation performance for continuous missing-subset I when the missing rate is 30%. The imputation accuracy of the model changes abruptly when the missing rate is 60%. And the model has the best imputation performance for continuous missing-subset II when the missing rate is 10%. The imputation accuracy of the model changes abruptly when the missing rate is 40%.

Table 5 and Figure 6(b) show that the model has the best imputation performance for jumped missing-subset I when the missing rate is 10%. The imputation accuracy of the model changes abruptly when the missing rate is 60%.

When the missing rate is between 10% and 30%, the MAPE of the most imputation result is between 6.38% and

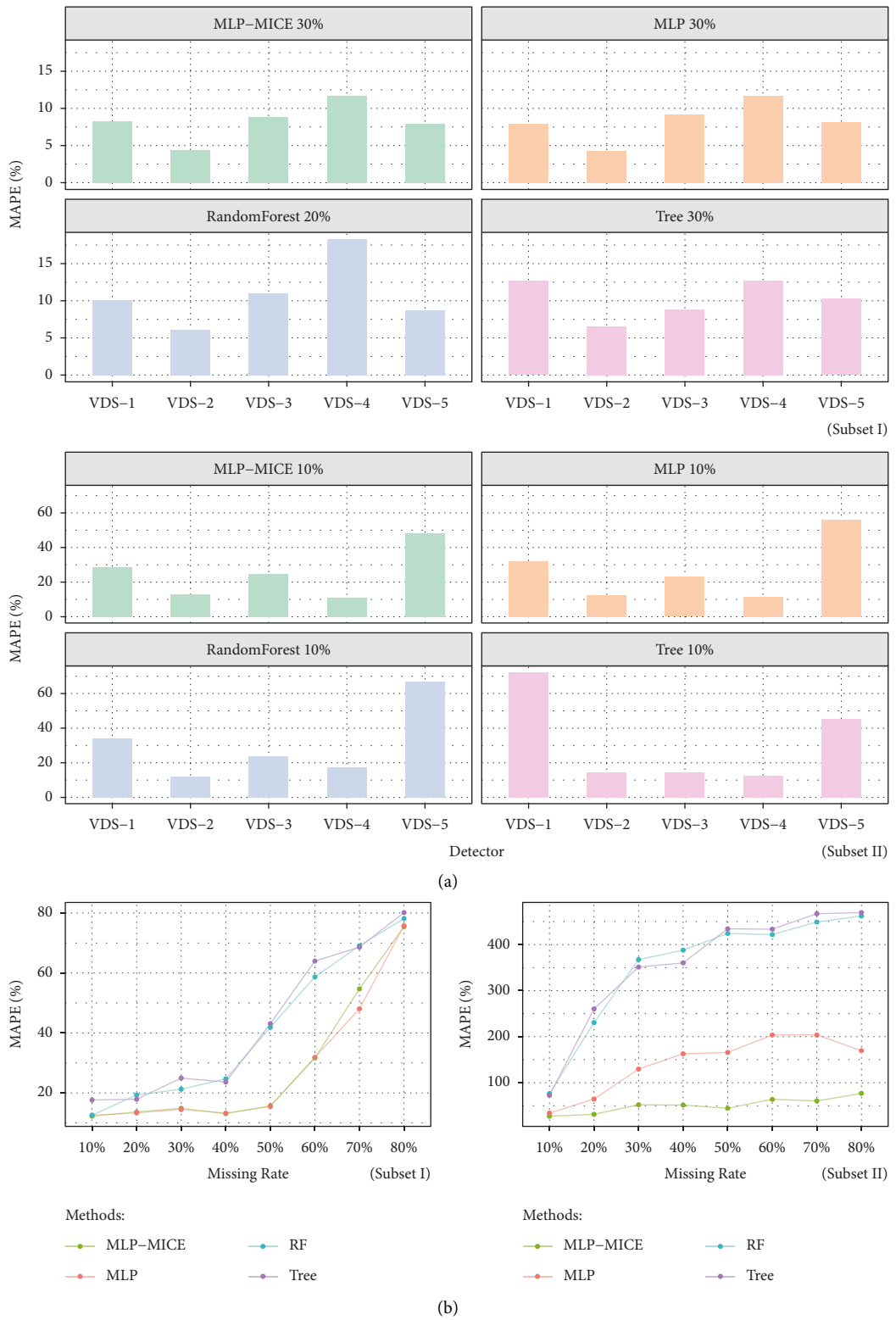

FIGURE 9: Imputation results and comparison of the proposed model with other models on the same data subset. (a) Continuous missing and (b) jumped missing.

30%, which illustrates that the proposed model has good imputation performance [31–50].

In this case, the adjacent difference is defined as the absolute value of the difference between the adjacent traffic volumes in the same VDS during a certain period. Total difference in a data subset is called subset differences, which measures the degree of the mutation of data subsets. The less fluctuation of the traffic volume in a

Table 8: Prediction experiment results of other VDS.

VDS no.	Imputation method	SVR-MAPE (%)	VDS no.	Imputation method	SVR-MAPE (%)
VDS-1	MLP-MICE	46.12	VDS-4	MLP-MICE	19.71
	All zero	128.40		All zero	19.75
VDS-2	MLP-MICE	26.13	VDS-5	MLP-MICE	16.99
	All zero	14.78		All zero	18.16
VDS-3	MLP-MICE	24.19			
	All zero	44.97			

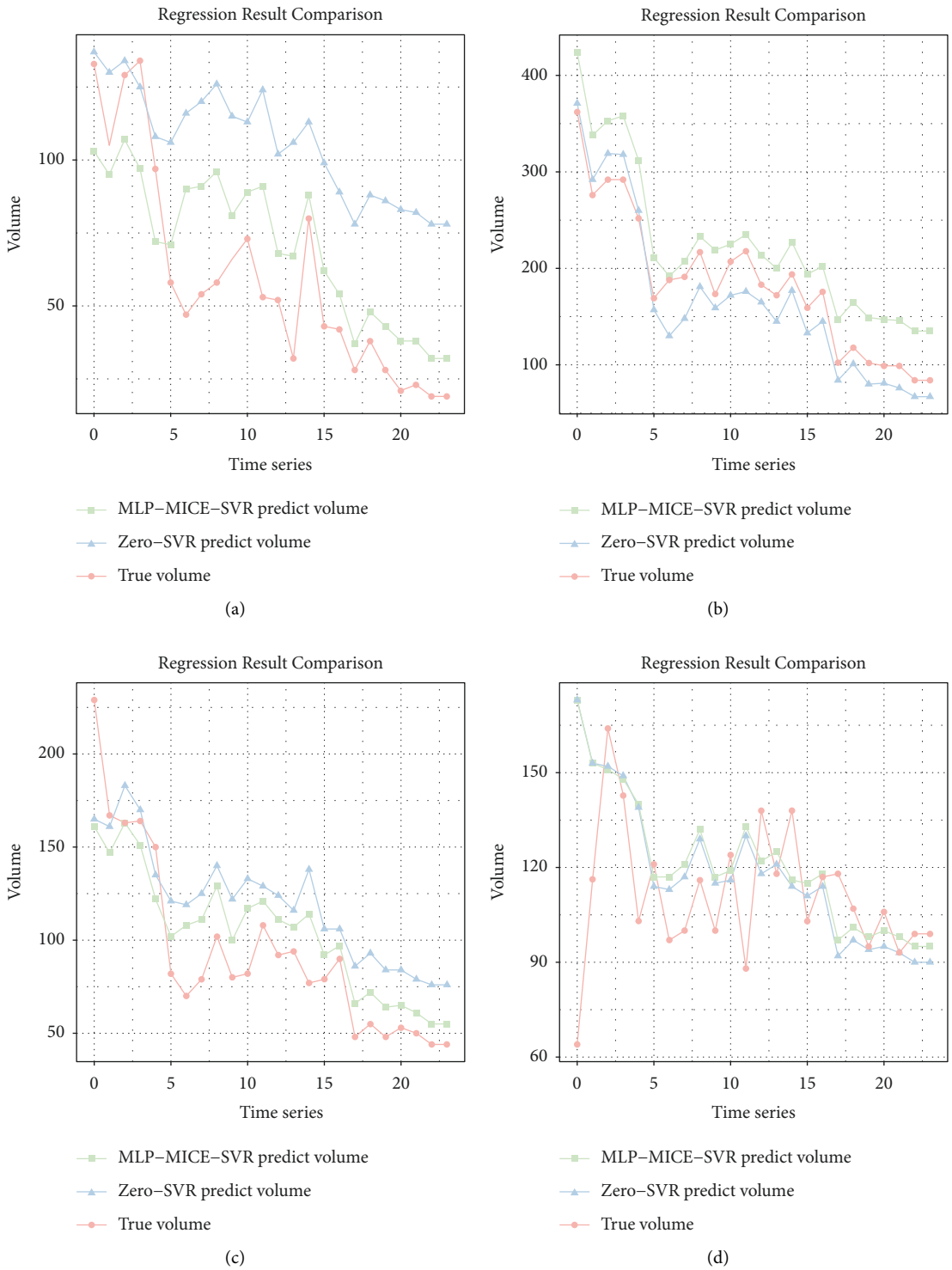

Figure 10: Continued.

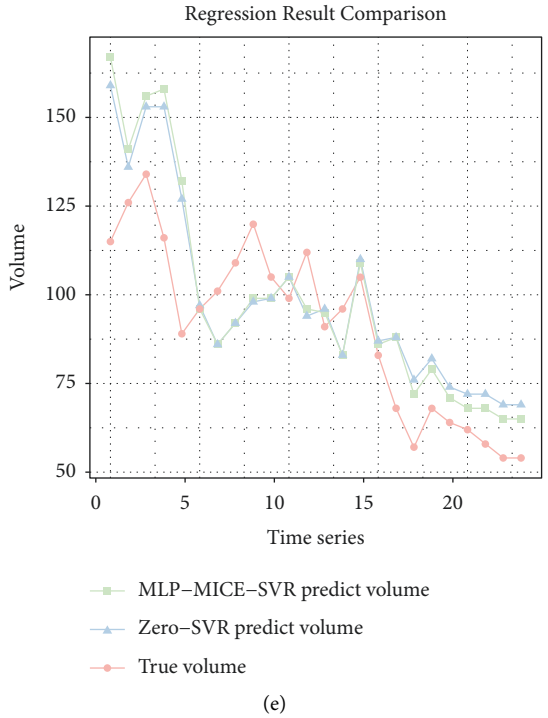

Figure 10: MLP-MICE imputation application analysis results of VDS-1 to VDS-5. (a) VDS-1, (b) VDS-2, (c) VDS-3, (d) VDS-4, and (e) VDS-5.

certain period happens, the smaller the value of subset differences is.

Table 6 and Figure 7 show the adjacent difference probability distribution among VDS-1, VDS-2, and VDS-4 is relatively uniform, while the variances are larger among VDS-1 to VDS-5. Comparing with jumped missing, the imputation performance of continuous missing of these three detectors is better, and VDS-3 and VDS-5 with relatively concentrated adjacent differences have better imputation results in jumped missing. The curve concentrates to the left, which means stable data because the adjacent difference probability distribution describes the degree of the mutation of the traffic volume. Therefore, when the traffic volume changes significantly, the continuous missing imputation performance is better. When the traffic volume changes stably, the imputation performance for jumped missing is better.

The main difference between subset I and subset II is the value of subset differences, as shown in Figure 8. The results of Table 5 and Figure 8(a) show that the imputation model has better performance for subset I, which means that MLP-MICE has better imputation performance for data subsets with smaller "subset differences."

From Figure 8(b) and Table 5, we can see that, for the two subsets, the higher the subset differences, the lower the imputation performance of MLP-MICE. And for continuous missing, the model has higher imputation performance for data subsets with small subset differences. While for jumped missing, the imputation performance is not strictly related to the subset differences. However, in general, the imputation performance for data subsets is better when the subset difference is smaller. We can draw the conclusion that MLP-MICE has high imputation performance for the data subsets where the mutation degree is low.

To verify the performance of the proposed model, regression imputation methods such as MLP, KNN, decision tree, SVR, and random forests are chosen for comparison experiments. Input subsets I and II into separate models and set the missing rate to 30% and 10%, respectively. The results are shown in Table 7.

Table 7 shows that the MAPE of the proposed model is lower than the other methods in each test. To verify the performance of existing missing imputation models, MLP, decision tree, and random forest are chosen by the average MAPE of imputation on both data subsets to do more tests on subsets I and II on continuous missing and jumped missing. Set the missing rate to 10–80% with the gradient at 10% and compare them with the imputation results of MLP-MICE, respectively. The results in Figure 9 show that MLP-MICE has good imputation performance and also proves the superiority of the proposed imputation model.

Finally, the proposed model is tested in the short-term prediction of traffic volume. The MLP-MICE with a missing rate of 20% is constructed to perform regression imputation on jumped missing and continuous missing of the collected dataset in Section 3. Meanwhile, all-zero imputation was taken as the comparison. Short-term prediction of traffic volume is carried out with the support vector machine model.

The dataset that complete the missing value by different imputation methods was taken as the input data, and the short-term traffic volume prediction on the last six hours (24

data volumes) was taken as an example. Comparing with the datasets whose missing values are all replaced by 0, the MAPE calculated from the prediction of the dataset with MLP-MICE repairing is significantly better. The results are shown in Table 8 and Figure 10.

Figure 10 shows that the prediction accuracy of VDS-1, VDS-3, VDS-4, and VDS-5 increases compared to all-zero imputation, while VDS-2 decreases because continuous zeros are recognized as outliers by SVR, which makes the all-zero imputation dataset predicted by SVR better. Despite the above defects, the average of the MLP-MICE imputation MAPE is 26.63%, while the MAPE of all-zero imputation is 45.21%. Therefore, the MLP-MICE can effectively improve the accuracy of short-term prediction of traffic volume.

4. Conclusions

This study proposes the MLP-MICE imputation model optimized by the L-BFGS algorithm, in which temporal and spatial characteristics of freeway traffic volume have been considered. According to the experiments and application analysis of the real-world data, the following conclusions can be drawn. (i) The proposed MLP-MICE in this study has better imputation performance and a strong superiority compared with other models. (ii) The imputation performance of the proposed model is better for continuous missing than for jumped missing. In the imputation process of the missing value of traffic volume data, the more smoothly the data change, the better the imputation performance of MLP-MICE in jumped missing is. When the traffic volume changes significantly, the imputation performance of MLP-MICE for continuous missing is improved. (iii) Whether continuous missing or jumped missing, there is always a gap of imputation performance among different data subsets that are from the same dataset but have a diverse degree of mutation. The smaller the degree is, the better the imputation performance of the missing value is. The gap between the imputation performances widens with the concentration of datasets and narrows with the divergence of datasets. (iv) For freeway traffic volume data, the proposed model is applied to conduct a short-time traffic prediction can get a more accurate result than only filling the missing data with zero. However, spatial and temporal characteristics of traffic flow are mainly considered for the imputation model in this study, but features such as weather, road conditions, and travel demand may also have an influence on the imputation performance, which can be considered in further study.

Data Availability

The data used to support the findings of this study were supplied under license and so cannot be made freely available. The data can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research and publication of this work was funded by the National Natural Science Foundation of China (project no. 52072129). The authors would like to express appreciation to Tianjiao Wang, Lingbin Kong, and Yuxin Guo for their help on this paper.

References

[1] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[2] M. T. Asif, N. Mitrovic, L. Garg, J. Dauwels, and P. Jaillet, "Low-dimensional models for missing data imputation in road networks," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3527–3531, Vancouver, BC, Canada, May 2013.

[3] Y. Yang, K. He, Y.-p. Wang, Z. Z. Yuan, Y. H. Yin, and M. Z. Guo, "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods," *Physica A: Statistical Mechanics and Its Applications*, vol. 595, Article ID 127083, 2022.

[4] A. R. T. Donders, G. J. .M. G. van der Heijden, S. Theo, and G. M. M. Karel, "A gentle introduction to imputation of missing values," *Journal of Clinical Epidemiology*, vol. 59, pp. 1087–1091, 2006.

[5] J. Deng, L. Shan, D. He, and R. Tang, "The imputation method of missing data and its development trend," *Statistics & Decisions*, vol. 35, pp. 28–34, 2019.

[6] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1487–1509, 2020.

[7] Z. Zhang, "Missing value imputation: focusing on single imputation," *Annals of Translational Medicine*, vol. 4, no. 9, 2016.

[8] X. Wang, A. Li, Z. Jiang, and H. Feng, "Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme," *BMC Bioinformatics*, vol. 7, p. 32, 2006.

[9] X. Li, "Imputation method for regional missing data using spatial autoregression models," *Journal of Applied Sport Management*, vol. 3, pp. 45–50, 2005.

[10] Y. Ci, H. Wu, Y. Sun, and W. Lina, "A prediction model with wavelet neural network optimized by the chicken swarm optimization for on-ramps metering of the urban expressway," *Journal of Intelligent Transportation Systems*, vol. 1-18, 2021.

[11] P. Sentas and L. Angelis, "Categorical missing value imputation for software cost estimation by multinomial logistic regression," *Journal of Systems and Software*, vol. 79, pp. 404–414, 2006.

[12] B. Wang, N. Zhang, W. Lu, and J. Wang, "Deep-learning-based seismic data interpolation," *A preliminary result*, vol. 84, pp. 1–73, 2018.

[13] X. Jing, F. Qi, F. Wu, and B. Xu, "Missing Data Imputation Based on Low-Rank Recovery and Semi-supervised Regression for Software Effort Estimation," in *Proceedings of the 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*, pp. 607–618, Austin, TX, USA, May 2016.

[14] J. Chen and J. Shao, "Jackknife variance estimation for nearest-neighbor imputation," *Journal of the American Statistical Association*, vol. 96, p. 260, 2001.

[15] L. Yu, Y. Jin, and J. Wang, "The research of missing data imputation method: based on nearest neighbor imputation and association rules," *Statistics & Information Forum*, vol. 30, pp. 35–40, 2015.

[16] D. Li, J. Deogun, W. Soaulding, and B. Shuart, "Towards missing data imputation: a study of fuzzy K-means clustering method," in *Proceedings of the Rough Sets and Current Trends in Computing*, pp. 573–579, Berlin, Heidelberg, 2004.

[17] V. Ravi and M. Krishna, "A new online data imputation method based on general regression auto associative neural network," *Neurocomputing*, vol. 138, pp. 106–113, 2014.

[18] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Information Sciences*, vol. 233, pp. 25–35, 2013.

[19] R. De jong, S. Van Buuren, and M. Spiess, "Multiple imputation of predictor variables using generalized additive models," *Communications in Statistics - Simulation and Computation*, vol. 45, pp. 968–985, 2016.

[20] P. Royston, "Multiple imputation of missing values," *STATA Journal*, vol. 4, pp. 227–241, 2004.

[21] J. W. Graham, A. E. Olchowski, and T. D. Gilreath, "How many imputations are really needed? Some practical clarifications of multiple imputation theory," *Prevention Science*, vol. 8, pp. 206–213, 2007.

[22] L. L. Doove, S. Van Buuren, and E. Dusseldorp, "Recursive partitioning for missing value imputation in the presence of interaction effects," *Computational Statistics & Data Analysis*, vol. 72, pp. 92–104, 2014.

[23] K. J. Lee and J. A. Simpson, "Introduction to multiple imputation for dealing with missing value," *Respirology*, vol. 19, pp. 162–167, 2014.

[24] L. F. Burgette and J. P. Reiter, "Multiple imputation for missing value via sequential regression trees," *American Journal of Epidemiology*, vol. 172, pp. 1070–1076, 2010.

[25] J. Zhu and T. E. Raghunathan, "Convergence properties of a sequential regression multiple imputation algorithm," *Journal of the American Statistical Association*, vol. 110, pp. 1112–1124, 2015.

[26] T. E. Ranghuanthan, J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger, "A multivariate technique for multiply imputing missing values using a sequence of regression models," *Survey Methodology*, vol. 27, pp. 85–96, 2001.

[27] M. Esmaeilbeigi, O. Chatrabgoun, A. Hosseinian-Far, R. Montasari, and A. Daneshkhah, "A low cost and highly accurate technique for big data spatial-temporal imputation," *Applied Numerical Mathematics*, vol. 153, pp. 492–502, 2020.

[28] W. Xiong, H. Pan, and L. Liu, "Robust efficient imputation of rounded zeros in high-demensional compositional data and its applications," *Statistical Research*, vol. 37, pp. 104–116, 2020.

[29] W. Han, J. Wang, and J. Hu, "Imputation methods for missing values in traffic flow data," *Computer and Communications*, vol. 1, pp. 39–42, 2005.

[30] L. Li, J. Zhang, Y. Wang, and B. Ran, "Missing value imputation for traffic-related time series data based on a multi-view learning method," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, pp. 2933–2943, 2019.

[31] L. Li, Y. Li, and Z. Li, "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence," *Transportation Research Part C: Emerging Technologies*, vol. 34, pp. 108–120, 2013.

[32] L. Qu, L. Li, Y. Zhang, and J. Hu, "PPCA-based missing value imputation for traffic flow volume: a systematical approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, pp. 512–522, 2009.

[33] J. Tang, X. Zhang, T. Yu, and F. Liu, "Missing traffic data imputation considering approximate intervals: a hybrid structure integrating adaptive network-based inference and fuzzy rough set," vol. 573, Article ID 125776, 2009.

[34] J. Tang, X. Zhang, W. Yin, Z. Yajie, and W. Yinhai, "Missing data imputation for traffic flow based on combination of fuzzy neural network and rough set theory," *Journal of Intelligent Transportation Systems*, vol. 1-16, 2020.

[35] C. Gong and Y. Zhang, "Urban traffic data imputation with detrending and tensor decomposition," *IEEE Access*, vol. 8, Article ID 11124, 2020.

[36] Q. Li, H. Tan, Y. Wu, L. Ye, and F. Ding, "Traffic flow prediction with missing data imputed by tensor completion methods," *IEEE Access*, vol. 8, Article ID 63188, 2020.

[37] H. Zhang, P. Chen, J. Zheng et al., "Missing data detection and imputation for urban ANPR system using an iterative tensor decomposition approach," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 337–355, 2019.

[38] J. Li, L. Xu, R. Li, W. Pan, and H. Zilin, "Deep spatial-temporal bi-directional residual optimisation based on tensor decomposition for traffic data imputation on urban road network," *Information Sciences*, vol. 586, pp. 344–373, 2022.

[39] Y. Duan, Y. Lv, W. Kang, and Y. Zhao, "A deep learning based approach for traffic data imputation," in *Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems*, pp. 912–917, Qingdao, China, October 2014.

[40] Q. Shang, Z. Yang, S. Gao, and D. Tan, "An imputation method for missing traffic data based on FCM optimized by PSO-SVR," *Journal of Advanced Transportation*, vol. 2018, Article ID 2935248, 2018.

[41] C. Fu, S. Yang, and Y. Zhang, "Promoted short-term traffic flow prediction model based on deep learning and support vector regression," *Journal of Transportation Systems Engineering and Information Technology*, vol. 19, pp. 130–134+148, 2019.

[42] Y. Duan, Y. Lv, Y. Liu, and W. Fei-Yue, "An efficient realization of deep learning for traffic data imputation," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 168–181, 2016.

[43] B. Sun, L. Ma, W. Cheng, W. Wen, P. Goswami, and G. Bai, "An improved k-nearest neighbours method for traffic time series imputation," *Chinese Automation Congress*, pp. 7346–7351, 2017.

[44] Y. Zhuang, R. Ke, and Y. Wang, "Innovative method for traffic data imputation based on convolutional neural network," *IET Intelligent Transport Systems*, vol. 13, pp. 605–613, 2019.

[45] O. Benkraouda, B. T. Thodi, H. Yeo, M. Menéndez, and S. E. Jabari, "Traffic data imputation using deep convolutional neural networks," *IEEE Access*, vol. 8, Article ID 104740, 2020.

[46] A. J. Saroj, A. Guin, and M. Hunter, "Deep LSTM recurrent neural networks for arterial traffic volume data imputation," *Journal of Big Data Analytics in Transportation*, vol. 3, 2021.

[47] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values," *Transportation Research Part C: Emerging Technologies*, vol. 118, Article ID 102674, 2020.

[48] B. Yang, Y. Kang, Y. Yuan, X. Huang, and H. Li, "St-Lbagan: Spatio-temporal learnable bidirectional attention generative adversarial networks for missing traffic data imputation," *Knowledge Based System*, vol. 215, Article ID 106705, 2021.

[49] M. Kaur, S. Singh, and N. Aggarwal, "Missing traffic data imputation using a dual-stage error-corrected boosting regressor with uncertainty estimation," *Information Sciences*, vol. 586, pp. 344–373, 2022.

[50] P. Wang, T. Hu, F. Gao, R. Wu, W. Guo, and X. Zhu, "A hybrid data-driven framework for spatiotemporal traffic flow data imputation," *IEEE Internet of Things Journal*, 2022, In press.

[51] H. K. Ghritlahre and R. K. Prasad, "Exergetic performance prediction of solar air heater using MLP, GRNN and RBF models of artificial neural network technique," vol. 223, pp. 566–575, 2018.

[52] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Mathematics of Computation*, vol. 35, pp. 773–782, 1980.

[53] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice," *Statistics in Medicine*, vol. 30, pp. 377–399, 2011.