Hindawi
Journal of Advanced Transportation
Volume 2018, Article ID 2935248, 21 pages
https://doi.org/10.1155/2018/2935248

Research Article

An Imputation Method for Missing Traffic Data Based on FCM Optimized by PSO-SVR

Qiang Shang[1], **Zhaosheng Yang**[2], **Song Gao**[1] **and Derong Tan**[1]

[1]*School of Transportation and Vehicle Engineering, Shandong University of Technology, Zibo, Shandong 255049, China*
[2]*College of Transportation, Jilin University, Changchun 130022, China*

Correspondence should be addressed to Qiang Shang; shangqiangv587@163.com

Received 26 August 2017; Revised 3 December 2017; Accepted 14 December 2017; Published 8 January 2018

Academic Editor: Taha H. Rashidi

Missing traffic data are inevitable due to detector failure or communication failure. Currently, most of imputation methods estimated the missing traffic values by using spatial-temporal information as much as possible. However, it ignores an important fact that spatial-temporal information of the traffic missing data is often incomplete and unavailable. Moreover, most of the existing methods are verified by traffic data from freeway, and their applicability to urban road data needs to be further verified. In this paper, a hybrid method for missing traffic data imputation is proposed using FCM optimized by a combination of PSO algorithm and SVR. In this method, FCM is the basic algorithm and the parameters of FCM are optimized. Firstly, the patterns of missing traffic data are analyzed and the representation of missing traffic data is given using matrix-based data structure. Then, traffic data from urban expressway and urban arterial road are used to analyze spatial-temporal correlation of the traffic data for the determination of the proposed method input. Finally, numerical experiment is designed from three perspectives to test the performance of the proposed method. The experimental results demonstrate that the novel method not only has high imputation precision, but also exhibits good robustness.

1. Introduction

With the continuous increase in travel demand, the urban road traffic congestion is becoming ever more serious. However, it is not sufficient to solve the problem of traffic congestion only by building new roads and other infrastructures, because of some economic and environmental reasons. Therefore, more and more researchers are focusing on the optimization of the existing traffic network in order to avoid or mitigate traffic congestion. Intelligent transportation systems (ITS) play an important role in the optimization of the existing traffic network. ITS applications such as intelligent traffic control and dynamic traffic guidance are able to improve the actual capacity of existing roads based on traffic data collected in real time. Nowadays, traffic data collection methods such as loop detectors, microwave detectors, video sensors, and GPS become more and more widely in the ITS. Unfortunately, missing traffic data are inevitable because of detector faults or transmission distortion [1], which severely limits the application and generalization of the ITS. For example, advanced traffic control system (a subsystem of ITS) requires sufficient and complete traffic flow data (including but not limited to traffic volume, speed, and occupancy) to generate appropriate traffic control strategies [2].

Many research efforts have been undertaken to estimate missing traffic data and many excellent imputation methods have been proposed. From the viewpoint of modeling philosophy, these methods are roughly divided into three categories: prediction based methods, interpolation based methods, and statistical learning based methods [3, 4].

Prediction based methods directly used existing traffic flow prediction methods, including Auto-Regressive Integrated Moving Average (ARIMA) model [5] and Feed-Forward Neural Network (FFNN) [2]. In these methods, a missing data point is regarded as a value to be predicted, and then the value is predicted using the relationship extracted from historical past-to-future data pairs [6]. However, two major differences between missing traffic data imputation and traffic flow prediction had not been fully considered in these methods. On the one hand, most of the prediction

methods do not use the traffic data collected after the missing value, which may reduce performance of missing traffic data imputation. On the other hand, if a consecutive series of traffic data are all missed, the prediction based methods are not able to provide satisfactory results for missing traffic data imputation.

Interpolation based methods estimate the missing data according to an average or weighted average of known data which is neighboring the missing data. There are two types of neighboring data, one is temporal-neighboring (collected from the same detector in the same time period but in neighboring days) [7], and the other is pattern-neighboring (collected from the same detector at the same time period but in other days with similar daily flow variation patterns) [8]. The historical average model is a typical temporal-neighboring interpolation method that completes the missing data using the average of the known historical data collected from the same location or in the same time period but in the previous few days [9]. Pattern-neighboring interpolation methods often estimate the missing data using the average or weighted average of known data of neighboring detectors which is a typical pattern-neighboring [10]. The K-Nearest Neighbor (KNN) model is a typical pattern-neighboring interpolation methods [11], whose key work is to determine the neighboring points by appropriate distance metrics. Interpolation based methods are highly dependent on the assumption that neighboring traffic flows are strongly similar to each other. However, this assumption sometimes fails in practice. In addition, when the neighboring traffic data are also missing, the performance of the interpolation based methods will be degraded severely.

Statistical learning based methods often use the observed data to learn a scheme and then inference the corrupted or missing data points in an iterated fashion, which try to take advantages of the statistical feature of traffic flow [12]. These methods usually include two steps: first, assuming a special probability distribution that is followed by the observed data; secondly, the values which best fit the assumed probability distribution which will be used to fill the missing values. Markov Chain Monte Carlo (MCMC) imputation method [13] and Probabilistic Principal Component Analysis (PPCA) imputation method [3] are two classical statistical learning based methods. Moreover, Kernel Probabilistic Principal Component Analysis (KPPCA) [14] and Bayesian Principal Component Analysis (BPCA) method [15] have also been used for missing traffic data imputation. In 2014, Chiou et al. propose an imputation method for missing traffic values by using the conditional expectation approach to functional principal component analysis (FPCA) [16], and their simulation study shows that the FPCA method performs better than the PPCA and BPCA. Though these methods have a strong assumption usually, their imputation performance is often better than that of conventional methods. This is mainly because the assumed probability distribution almost captures the essentials of traffic flow variations.

In recent years, a number of new methods have been proposed to impute missing traffic data. In 2013, Tan et al. [4] proposed a missing traffic data imputation method based on tensor for the first time, and the experimental results show that this tensor-based method has achieved a better performance, especially in the case with a high missing ratio. As multiway matrices, tensor can take full advantage of the temporal and spatial information of traffic flow data to impute the missing data with a higher precision. Subsequently, Tan et al. [17, 18] proposed several other tensor-based imputation methods according to different perspectives and tested the proposed methods using the traffic data from PeMS, and the results show that these methods have good performance under extreme conditions. In 2015, Tang et al. [19] proposed a missing traffic imputation method based on the fuzzy C-means (FCM) optimized by the genetic algorithm. In this method, the matrix form is used to express missing traffic data and the genetic algorithm is employed to optimize the FCM parameters. The empirical results demonstrate that the optimized FCM method has good imputation performance for the missing traffic data with different missing ratios and different sampling intervals. In 2016, Asif et al. [20] proposed a matrix and tensor-based method to estimate the missing traffic data of road network and verified the performance of this method from three aspects: estimation accuracy, data variance, and estimation bias. In 2016, Duan et al. [21] proposed a deep learning method to impute the missing traffic data, and the empirical results show that the average imputation error of this method is below 10 veh/5 min, and the imputation performance is better than the historical average model, ARIMA model, and BP neural network model. More recently, Chen et al. [22] proposed an ensemble correlation-based low-rank matrix completion method which achieved better imputation performance than competing methods (including temporal average imputation and the PPCA imputation). In this method, low-rank matrix is used to represent traffic data and ensemble KNN learning is used to explore the relationship between the missing data and the complete data.

By reviewing the existing methods (especially in recent years) for missing traffic data imputation, it can be found that there are two important research directions in this field. One is to study how to use more spatial-temporal correlation data (which contain more spatial-temporal information) to impute missing data. The other is to study how to use advanced algorithms or improved methods for more fully mining spatial-temporal information contained in spatial-temporal correlation data. However, it ignores an important fact that spatial-temporal information of the missing traffic data is often incomplete and unavailable. Moreover, most of the existing methods are validated using traffic data from freeway (such as [16–19, 21, 22]), and their applicability to urban road data needs to be further validated. As we all know, the freeway is relatively closed and there are no intersections to hinder the operation of the traffic flow; thus the continuity of traffic flow is better and the traffic flow data show a strong temporal-spatial correlation characteristic. Obviously, the traffic data spatial-temporal information of urban road is less than that of the freeway due to its own structural characteristics.

To tackle the shortcoming as mentioned above, a novel method is proposed to improve the performance of missing traffic data imputation. In order to fully utilize the available

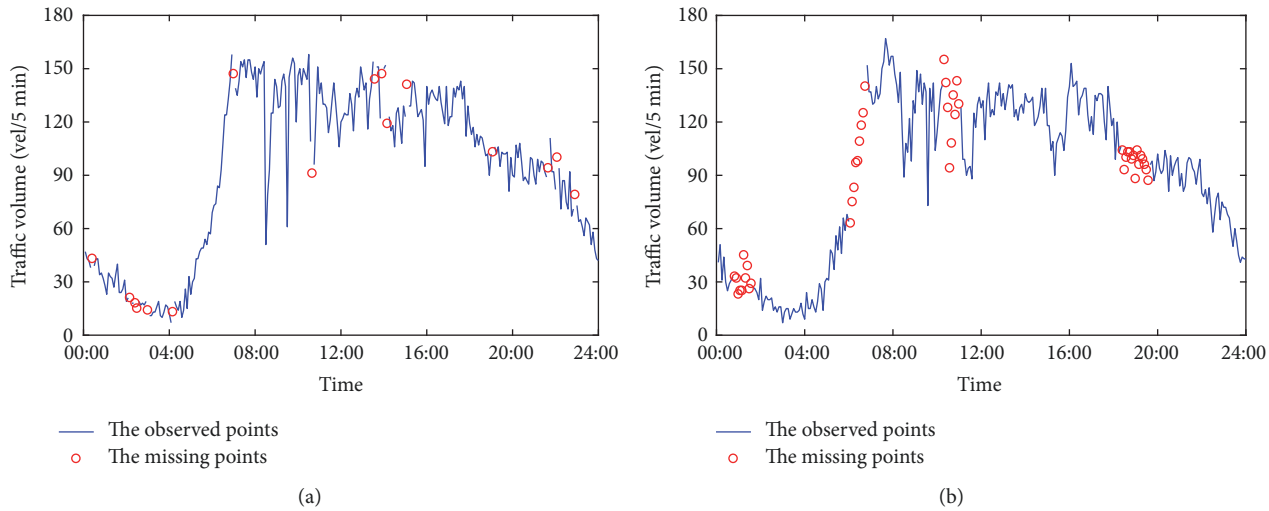

Figure 1: Typical missing patterns of traffic flow data: (a) MCR and (b) MR.

spatial-temporal correlation data, fuzzy C-means (FCM) is chosen as the basic algorithm, because of its excellent performance to analyze the data with multiple attributes, and the spatial-temporal correlation data can be expressed as multiple attributes. In addition, FCM not only has been successfully applied to address the clustering with incomplete data, but also has been applied to the estimation of missing data [23–26], including estimation of missing traffic volume data [19]. In this study, we draw on the principle of a SVR-based method for missing data imputation and taking into account the advantages of particle swarm optimization (PSO) algorithm [27] and then propose a new optimization algorithm that combines PSO and SVR to optimize the parameters of FCM. In order to fully test the imputation performance of the proposed method, the experiment is designed from three perspectives.

The main contributions of this paper are summarized as follows. (a) The combination of PSO and SVR is employed to optimize FCM parameters for the first time. (b) The urban road data are used to test imputation performance of the proposed method, including urban expressway data and urban data arterial. (c) Choose the available data as input of the method, according to correlation analysis of experimental data, rather than taking all possible spatial-temporal correlation data into account. (d) The imputation performance of the proposed method is tested using the traffic data without complete spatial-temporal correlation data (i.e., some of the spatial-temporal correlation data are unavailable or inaccessible).

To give an explanation of the proposed imputation method in detail, the rest of this paper is organized as follows: Section 2 introduces the missing traffic data, including the patterns of the missing data points and the matrix-based missing traffic data representation. Section 3 presents the experimental result and discussion. Finally, the conclusions and future work are outlined in Section 4.

2. Missing Traffic Data

2.1. Patterns of Missing Traffic Data. Reasons for missing data are uncertain and beyond our control. Therefore, it is necessary to investigate the process of missing data generation. In many studies, missing data are regarded as a probabilistic phenomenon and missing data points present one or more probability distributions. In general, there are three patterns of missing traffic data as follows [3, 16]:

(1) Missing Completely at Random (MCR), where the missing data points are completely independent of each other. Therefore, the missing data usually appear as some isolated points distributed randomly (see Figure 1(a)).

(2) Missing at Random (MR), where the missing data points are associated with their neighboring points. Therefore, missing data usually appear as a small group of consecutive points lost at the same time, but these groups are random distribution (see Figure 1(b)).

(3) Not Missing at Random (NMR), where the generation of missing data points have certain patterns. In Intelligent Transportation Systems (ITS), NMR is usually caused by a long time failure of detectors, which results in poor availability of collected traffic data. In this study, we assume that unexpected NMR data points of traffic flow time series have already been found and discarded.

In view of the above analysis, we focus on missing data imputation under three kinds of missing patterns including MCR, MR, and mixed MCR/MR which is a combination of MCR and MR.

2.2. Matrix-Based Missing Traffic Data Representation. At present, matrix-based structure is one of the most widely used and most effective forms for missing traffic data representation. Compared with the traditional vector-based data structure, the matrix-based data structure can make more full use of spatial-temporal correlation information which is usually contained in similar traffic patterns, including

Table 1: A matrix-based missing data representation for "day pattern."

	Monday	Monday	Monday	Monday	Monday
00:00–00:05	54	?	47	?	41
00:05–00:10	44	43	51	?	36
00:10–00:15	?	42	31	34	43
00:15–00:20	28	38	39	36	44
...	...	...	...	...	...
23:40–23:45	67	56	50	45	54
23:45–23:50	57	50	?	45	55
23:50–23:55	71	45	56	48	?
23:55–00:00	59	?	49	48	46

Note. "?" is the missing values in traffic volume dataset.

Table 2: A matrix-based missing data representation for "week pattern."

	Monday	Tuesday	Wednesday	Thursday	Friday
00:00–00:05	58	49	62	57	54
00:05–00:10	?	58	57	56	57
00:10–00:15	?	57	44	52	53
00:15–00:20	44	58	42	32	37
...	...	...	...	...	...
23:40–23:45	72	?	91	?	77
23:45–23:50	85	63	65	66	?
23:50–23:55	75	?	61	76	79
23:55–00:00	80	77	86	83	75

Note. "?" is the missing values in traffic volume dataset.

temporal patterns (such as "day pattern" and "week pattern") and spatial patterns (such as "link pattern" and "section pattern"). The "day pattern" is the traffic flow data collected in the same day but in the neighboring weeks, such as several consecutive Monday. The "week pattern" is the traffic flow data collected in the neighboring days of a week, such as several consecutive weekdays or weekends. The "link pattern" is the traffic flow data collected in the same link (lane) but in different sections. The "section pattern" is the traffic flow data collected in the same section but in different links (lanes).

In this study, 5 min-interval traffic volume data are taken as an example. A matrix-based structure of "day pattern" is presented in Table 1. A matrix-based structure of "week pattern" is given in Table 2. Matrix-based structures of "link pattern" and "section pattern" are shown in Table 3. As we can see, with a matrix-based data structure for missing data representation, the FCM method can make better use of the explicit topological around the missing data to improve the imputation performance.

3. Methodology

In this section, a brief summary of the relevant methods of this study is given firstly, which include support vector regression imputation, fuzzy C-means imputation, particle swarm optimization (PSO), and PSO-based FCM imputation. Then, a novel imputation method named PSO-SVR-FCM is proposed using FCM optimized by a combination of PSO and SVR.

3.1. FCM-Based Imputation Method. Clustering algorithms can be divided into two categories including hard clustering and soft (fuzzy) clustering. For hard clustering, a record of a dataset belongs to one and only one cluster, in which the record is the most similar to other records. However, for soft clustering, each record has a certain probability (known as the membership degree) that belongs to each of the clusters. As a typical hard clustering algorithm, the C-means clustering is a powerful technique for data clustering in many fields. As the most famous soft clustering algorithm, the FCM is an improved algorithm of traditional C-means clustering, which can overcome the limitations of local optimum in traditional C-means clustering and also make a better clustering performance when the clusters are not well separated [28].

For a dataset $X = \{x_1, x_2, \ldots, x_n\}$, each x_i $(1 \le i \le n)$ has l attributes. And then X can be expressed as (1) and transform into a matrix-based data structure.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1l} \\ x_{21} & x_{22} & \cdots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nl} \end{bmatrix}, \quad (1)$$

where x_{ij} represents the jth attribute value collected at the ith time interval.

$\forall x_{ij} \neq \Phi$, $x_i = \{x_{i1}, x_{i2}, \ldots, x_{il}\}$ can be regarded as a complete data vector, where Φ is an empty dataset. For all complete data vector x_i, $R = \{x_{ij} \neq \Phi,\ 1 \le j \le l,\ 1 \le i \le n\}$

Table 3: A matrix-based missing data representation for "section pattern" and "lane pattern."

	At the same lane but the adjacent sections				At the same section but the adjacent lanes			
	Detector 1	Detector 2	⋯	Detector n	Detector 1	Detector 2	⋯	Detector n
00:00–00:05	15	10	⋯	12	34	47	⋯	?
00:05–00:10	?	23	⋯	21	?	36	⋯	44
00:10–00:15	13	?	⋯	?	26	32	⋯	29
00:15-00:20	13	10	⋯	?	12	32	⋯	?
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
23:40–23:45	45	43	⋯	44	?	53	⋯	46
23:45–23:50	46	39	⋯	37	43	52	⋯	48
23:50–23:55	?	29	⋯	34	48	?	⋯	56
23:55–00:00	44	33	⋯	35	36	?	⋯	54

Note. "?" is the missing values in traffic volume dataset.

indicates the set of available attributes, which can be used to estimate missing values.

The main steps of FCM-based imputation method are as follows.

Step 1. Set the values of parameters including cluster size K and weighting factor m, and initialize the value of membership function U.

Step 2. Calculate the clusters centroids $C = \{c_1, c_2, \ldots, c_K\}$ by

$$c_k = \frac{\sum_{i=1}^{n} U\left(x_i, c_k\right)^m \cdot x_i}{\sum_{i=1}^{n} U\left(x_i, c_k\right)^m}, \tag{2}$$

where c_k $(1 \le k \le K)$ is the centroid of the kth cluster, the parameter m $(1 < m < +\infty)$ is weighting factor to quantify the fuzzy degree for clustering, membership function $U(x_i, c_k)$ is a $n \times K$ matrix and means the degree that x_i belongs to c_k, which can be calculated by (3). For all x_i, there is $\sum_{j=1}^{K} U(x_i, c_j) = 1$.

$$U\left(x_i, c_k\right) = \frac{d\left(x_i, c_k\right)^{-2/(m-1)}}{\sum_{j=1}^{k} d\left(x_i, c_j\right)^{-2/(m-1)}}, \tag{3}$$

where $d(x_i, c_k)$ is a generalized norm distance that between the specific data x_i and the centroid c_k, which can be calculated by (4). When $p = 1$, (4) indicates the Manhattan distance, and when $p = 1$, (4) indicates the Euclidean distance.

$$d\left(x_i, c_k\right) = \left(\sum_{j=1}^{l} \left|x_{ij} - c_k\right|^p\right)^{1/p}. \tag{4}$$

Step 3. Minimize the objective function defined as follows and search the optimal values of U and C.

$$J\left(U, C\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} U\left(x_i, c_k\right)^m \cdot d\left(x_i, c_k\right) \tag{5}$$

Step 4. Whether the termination condition is met, if the objective function value is less than a preset threshold, the difference between objective function values of two consecutive iterations is less than a preset threshold, or the number of iterations reaches its preset maximum number, then the termination condition is met and go to the next step; otherwise update the U according to (6) and return to Step 2.

$$U\left(x_i, c_k\right) = \frac{d\left(x_i, c_k\right)^{-2/(m-1)}}{\sum_{j=1}^{k} d\left(x_i, c_j\right)^{-2/(m-1)}}. \tag{6}$$

Step 5. Obtain the optimal values of U and C to estimate the missing attribute values of x_i based on

$$\widehat{x}_{ij} = \sum_{k=1}^{K} U\left(x_i, c_k\right) \cdot c_k, \tag{7}$$

where $\widehat{x}_{ij}$ presents missing values regarded as the nonreference attributes.

Figure 2 illustrates the process of FCM-based method for missing traffic data imputation, where "?" is supposed to be a sample missing value in the dataset. In this example, the complete data are clustered into 3 clusters with a weighting factor value of 2, which means that the parameters $K = 3$ and $m = 2$. As shown in Figure 2, each data vector contains two attributes which, respectively, correspond to the abscissa and ordinate values. The membership values of missing value "?" are estimated as (0.1, 0.3), (0.2, 0.5), and (0.7, 0.2), and the clustering centroids are estimated to be (120, 119), (87, 85), and (25, 26). Therefore, if the abscissa value is missing, the missing value is calculated as "?" $= 0.1 \times 120 + 0.2 \times 87 + 0.7 \times 25 = 46.9$. Similarly, if the ordinate value is missing, the missing value is calculated as "?" $= 0.5 \times 119 + 0.3 \times 85 + 0.2 \times 26 = 90.2$. Here, only two-dimensional data (two attributes) is used as an example, and the FCM-based method is also applicable to multidimensional data.

3.2. Support Vector Regression. The Support Vector Machine (SVM) [29] is a popular machine learning method based on statistical learning theory. In general, the SVM can be divided into two categories according to their uses. The SVM is originally developed for the classification problem using

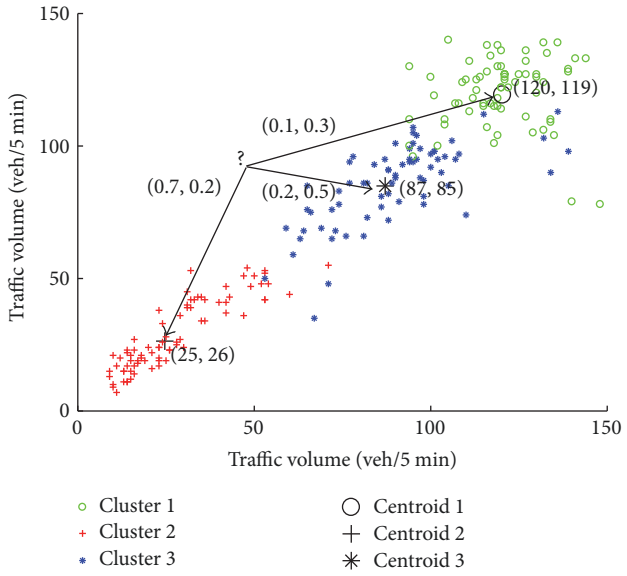

Figure 2: The process of FCM-based method for missing traffic data imputation.

structural risk minimization principle, which is also called support vector classification (SVC). Then, the SVM is modified to solve nonlinear regression problems by incorporating the ε-intensive loss function and is also called support vector regression (SVR). In SVR, the input data is mapped to a high-dimensional feature space using a nonlinear function (known as the kernel function that is satisfied with Mercer's condition) and then a linear regression function is computed in the mapped feature space [30]. A significant advantage of SVR is that mathematical calculations are relatively simple because nonlinear problems of input space are transformed into linear problems of high-dimension feature space. SVR only needs to select the appropriate kernel function without the need for knowledge about the specific form of nonlinear mapping. Then, the high-dimensional feature space can be transformed into a low dimensional space via the selected kernel function; thus SVR avoids the "dimension disaster." However, there is no mature and solid theory for the kernel function selection and parameters optimization. In this study, Gaussian radial basis function (RBF) is selected as the kernel function because of its excellent performance [31, 32], which is defined as follows:

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2}\right), \quad \sigma > 0, \tag{8}$$

where σ is the kernel parameter and also called kernel width. It is well known that the regression accuracy of the SVR with RBF kernel function is closely related to the settings of penalty factor (regularization parameter) C, loss function parameter ε, and kernel parameter σ. In our study, the PSO algorithm is also employed to optimize the three parameters of SVR. The basic principle of the SVR algorithm and its solution are described in detail in [29, 30], which are not described here because of the limited length.

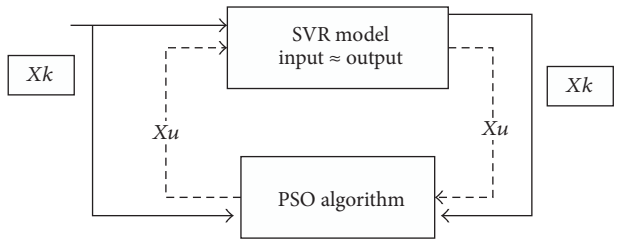

Figure 3: The process of the SVR-based imputation method with PSO.

3.3. SVR-Based Imputation Method. SVR is widely used for missing data imputation. In some methods, SVR is used to predict the missing values directly. In other methods, SVR is used to evaluate the accuracy of the estimated values, rather than estimate missing values directly. In these methods, the optimal estimated values can be found by intelligent optimization algorithm such as genetic algorithm and PSO algorithm. The SVR-based method with PSO is selected as an example and the main steps of this method are as follows.

Step 1. Select samples without any missing attribute values.

Step 2. Set one of the condition attributes (input attribute), some of whose values are missing, as the decision attributes (output attribute), and on the contrary, set the decision attributes as the condition attributes.

Step 3. SVR is used to predict the decision attribute values [33].

The above three steps are used for each attribute one by one, and then all attribute outputs are combined into the model output which corresponds to the model input. In this way, the model is trained by recall itself. The process of the SVR-based imputation method with PSO is illustrated in Figure 3. First, SVR model needs to be trained with complete records before it can be used to estimate missing data. When use the trained SVR, the input will be recalled on the output. Xu is the unknown data attribute (the missing data), which is approximated by PSO. Xk is the known attribute values (the complete data). The model input is shown in (9), the model output is shown in (10), where f (function) represents the mapping between the model input and output. The input data are recalled in the model, and the difference is known as an error and shown in (10). The PSO is used to reduce the error between the input and the output SVR model. Thus, the fitness function (objective function) should be nonnegative to minimize the error, which results in most approximate value for the missing value. Equation (12) gives a commonly used the fitness function and all outputs are used to reduce fitness function value for completeness.

$$\text{SVR input} = \begin{pmatrix} Xk \\ Xu \end{pmatrix} \tag{9}$$

$$\text{SVR output} = f\begin{pmatrix} Xk \\ Xu \end{pmatrix} \tag{10}$$

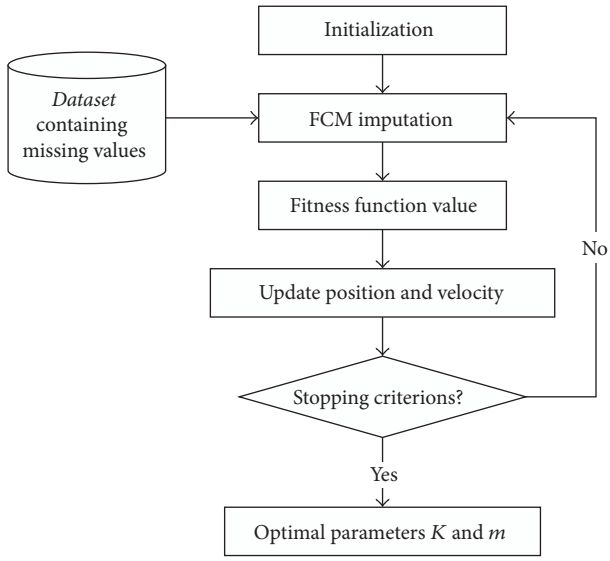

Figure 4: The flowchart of PSO-FCM-based imputation method.

$$\text{error} = \begin{pmatrix} Xk \\ Xu \end{pmatrix} - f\begin{pmatrix} Xk \\ Xu \end{pmatrix} \tag{11}$$

$$\text{PSO fitness function} = \left(\begin{pmatrix} Xk \\ Xu \end{pmatrix} - f\begin{pmatrix} Xk \\ Xu \end{pmatrix}\right)^2. \tag{12}$$

3.4. PSO-FCM-Based Imputation Method. As mentioned in Section 3.1, there are two important parameters of FCM that need to be determined, one is cluster sizes K and the other is weighting factor m. However, there is no definite theory to determine the optimal values of these two parameters. The choice of parameters K and m depends on characteristics of the dataset and the relationship between each attributes. In recent years, intelligent optimization algorithm, such as PSO algorithm and genetic algorithm, is employed to optimize FCM parameters with a good performance [19, 26]. Figure 4 is the flowchart of PSO-FCM-based imputation method. In this method, some values of complete attributes are artificially deleted to simulate missing data, according to the patterns of missing data; then PSO algorithm is used to search optimal parameters K and m for the best imputation accuracy of the missing data.

The main steps of the PSO-FCM-based imputation method are as follows.

Step 1. Initialize PSO algorithm and FCM parameters.

Step 2. Missing data are estimated by FCM method.

Step 3. Calculate the fitness function value, and the fitness function is shown as (12), that is, the mean square error between estimated data and actual data.

Step 4. Update the velocity and position of particles, according to their respective update rules.

Step 5. Judge whether the stopping criteria (usually defaulted to a certain calculation accuracy or maximum number of iterations) are reached; if it is reached, output optimal parameters K and m, otherwise, return to Step 2.

3.5. PSO-SVR-FCM-Based Imputation Method (the Proposed Method). In this study, a novel imputation method named PSO-SVR-FCM is proposed. Similarly, basic algorithms of the PSO-SVR-FCM method and the PSO-FCM method are both FCM-based imputation method. The difference between these two methods is optimization for FCM. In the PSO-SVR-FCM method, FCM is optimized via a combination of PSO algorithm and SVR, while only PSO is used to optimize the FCM in the PSO-FCM method.

In this paragraph, motivation and the unique features for proposed methods are given. Owing to a variety of reasons, the traffic data of urban road contain more noise data and outliers than that of freeway. Unfortunately, some noise data cannot be identified and processed, and these low-quality data are mixed in the complete data [3]. For the proposed method, FCM is selected as the basic algorithms, which is a strong tool for the identification of changing class structures and flexible, moveable, and creatable for uncertain data (i.e., noise and outliers) [34] to improve the imputation accuracy. However, FCM with constant parameters is difficult to apply for the missing values imputation of complex and diverse traffic datasets. Currently, when most heuristic optimization algorithms (e.g., PSO and GA) are used, the optimization objective function (fitness function) value is set as the observed value, and then FCM is trained using complete data with noise and outliers, which may lead to overfitting. Therefore, the optimized parameters are not optimal and need to be further optimized. For the proposed method, FCM parameters are optimized by a combination of PSO algorithm and SVR innovatively. SVR yields more sensible results for outliers, which is robust against the noise [35]. For the proposed method, SVR is introduced to build fitness function for FCM optimization, and the combination of PSO and SVR is used to optimize FCM parameters. In theory, the proposed method is likely to achieve better results for missing traffic data imputation.

Figure 5 illustrates the flowchart of PSO-SVR-FCM-based imputation method. A dataset with missing values can be divided into two categories: complete dataset and incomplete dataset. The dataset consists of a series of data records and each record is obtained at each sampling interval. Any record in an incomplete dataset has one missing value (attribute) at least, while that in complete dataset has no missing value(s). As the basic algorithm, FCM is used to estimate missing data. The parameters K and m are optimized by a combination of PSO and SVR for the best imputation accuracy. As shown in Figure 5, the purpose of the PSO algorithm combined with SVR is to minimize error. The fitness (objective) function is minimized error that is error = $(X - Y)^2$, where X is the output of FCM imputation and Y is the output of SVR prediction. Before imputation for the missing values, the SVR must be trained using the complete records to estimate the output values that closely correspond

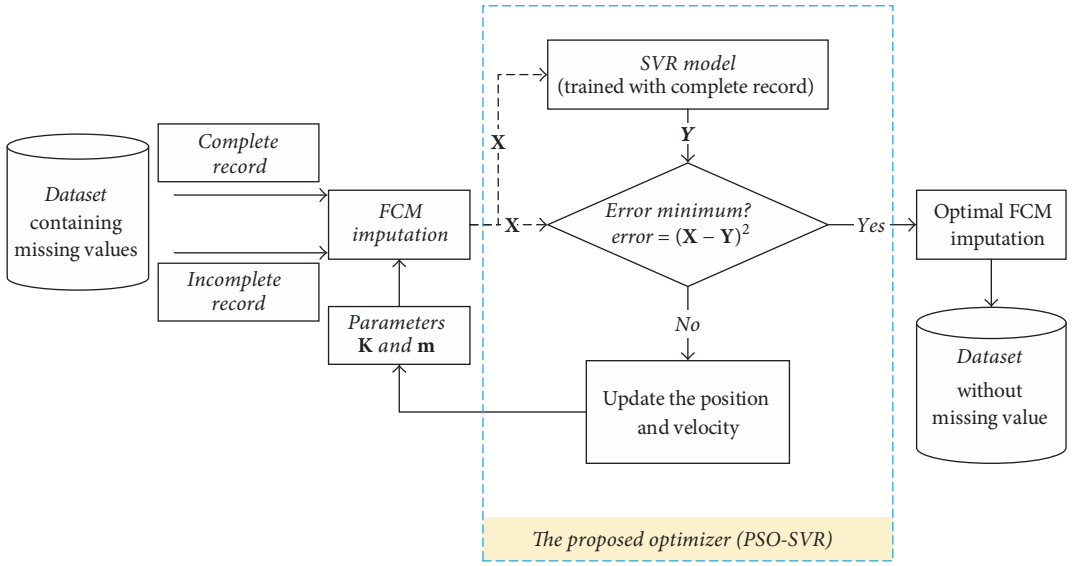

Figure 5: The flowchart of PSO-SVR-FCM-based imputation method.

to the input. Therefore, the PSO algorithm searches optimal parameters K and m to minimize the fitness function value.

The main steps of the PSO-SVR-FCM-based imputation method are as follows.

Step 1. SVR model is trained with complete records, for which Input (X) ≈ Output (Y).

Step 2. FCM is used to impute the incomplete record and compare the FCM output with the SVR output, that is, calculate the fitness function value.

Step 3. The PSO algorithm is used to optimize parameters K and m to minimize the fitness function value.

Step 4. FCM with optimal parameters is used for missing data imputation.

4. Numerical Experiment

In this section, numerical experiments are conducted to test the performance of the proposed imputation method. First, experimental data are described, which include urban expressway data and urban arterial data. Then, analyze spatial-temporal correlation of these two types of traffic data to determine the input of the proposed method. Finally, the experiment is designed from three perspectives to test the performance of the proposed method. In addition, three state-of-the-art imputation methods are introduced for comparison.

4.1. Description of Experimental Data. Two types of traffic data collected from urban expressway and urban arterial are both used to verify the proposed imputation method. Urban expressway traffic data are collected by some loop detectors located in North and South Elevated Expressway, Shanghai, China; and the urban arterial traffic data are collected by some loop detectors located in Lianqian West Road, Xiamen, China. Loop detectors can record traffic data (including traffic volume, average speed, and average occupancy) as time series at a certain time interval (e.g., 5 minutes). Although three types of traffic data can be obtained at the same time, this study only uses the traffic volume data as an example. According to the preliminary statistics, the missing data rate of traffic volume collected from urban expressway is 5.10%, and the missing data rate of traffic volume collected from the arterial road is 3.46%. It is worth noting that there are some loop detectors with a high missing ratio at certain sampling times.

4.2. Spatial-Temporal Correlation Analysis for Traffic Data. The key for missing traffic data imputation is to make full use of spatial-temporal information. At present, many studies have demonstrated the freeway traffic volume with a strong spatial-temporal correlation. However, due to the different structure and function of freeway and urban road, it is necessary to further explore whether there is spatial-temporal correlation for the urban road traffic volume. Moreover, in view of the differences between the urban expressway and the urban arterial road, the traffic volume data from these two roads should be analyzed to determine which spatial-temporal correlation data is available. In this way, available spatial-temporal correlation data can be used for imputation method.

4.2.1. Temporal Correlation Analysis. The temporal correlation of traffic volume data is mainly reflected in the "day pattern" and the "week pattern." The "day pattern" is the traffic volume collected in the same day but in the neighboring weeks. The "week pattern" is the traffic flow volume collected in the neighboring days of a week. In order to analyze the temporal correlation for urban road traffic volume, two loop detectors are randomly selected from the urban expressway

Table 4: Correlation coefficient matrices of two traffic volume series from the same detector located at urban expressway but in different Tuesday.

	09.02 (Tuesday)	09.09 (Tuesday)	09.16 (Tuesday)	09.23 (Tuesday)	09.30 (Tuesday)
09.02 (Tuesday)	1	0.9721	0.9735	0.9661	0.9683
09.09 (Tuesday)	0.9721	1	0.9704	0.9527	0.9654
09.16 (Tuesday)	0.9735	0.9704	1	0.9567	0.9685
09.23 (Tuesday)	0.9661	0.9527	0.9567	1	0.9571
09.30 (Tuesday)	0.9683	0.9654	0.9685	0.9571	1

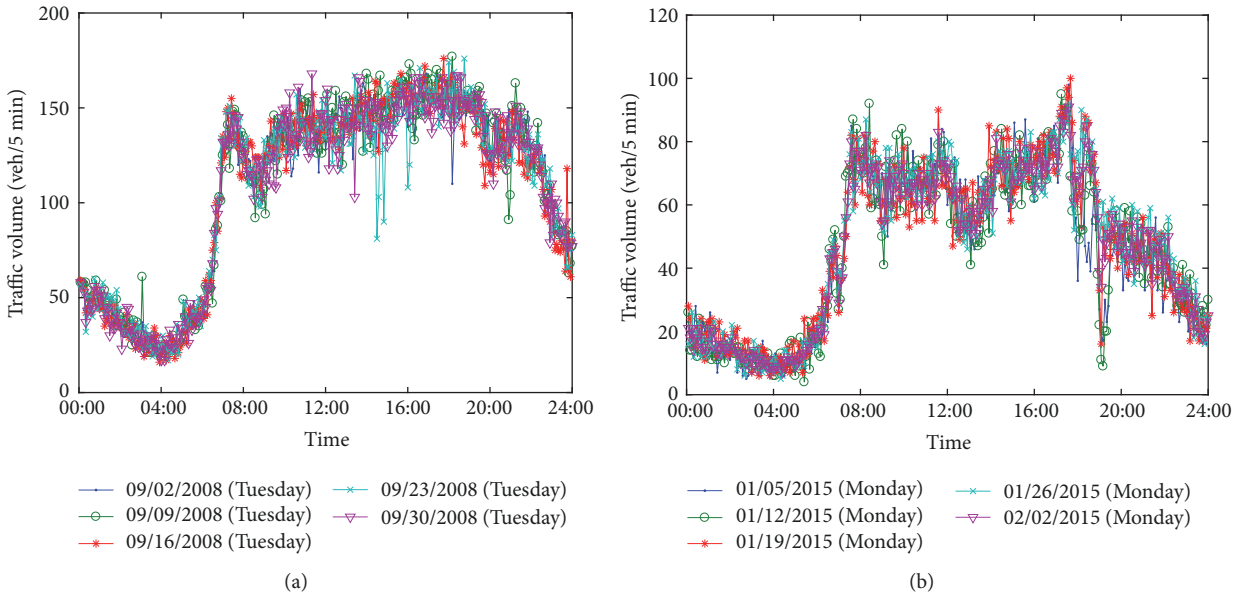

Figure 6: An example for illustrating the "day pattern": (a) urban expressway and (b) urban arterial road.

and the arterial road, respectively, whose traffic volume is used for temporal correlation analysis.

Figure 6 shows traffic volume data from five consecutive Mondays/Tuesdays to demonstrate the "day pattern." As can be seen from Figure 6, the traffic volume series of each day is similar to each other obviously, which not only has a similar trend in the whole, but also has a similar traffic volume value at the same sampling interval. To quantify the correlation between each traffic volume series, Pearson's correlation coefficient is applied to measure the data correlation, which is given as follows:

$$R = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{cov}(X, X) \cdot \operatorname{cov}(Y, Y)}}, \tag{13}$$

where X and Y represent two traffic volume time series, respectively, and cov is the covariance of two traffic volume time series. In particular, Tables 4 and 5 give the correlation coefficient matrices of the traffic volume data shown in Figures 6(a) and 6(b), respectively. All the correlation coefficients are greater than 0.9, which illustrates that the traffic volume time series of urban expressway and urban arterial road are both with strong daily correlation.

Figure 7 shows traffic volume data from five consecutive working days in a week to demonstrate the "week pattern." As can be seen from Figure 7, the traffic volume series of each day is similar to each other obviously, which not only has a similar trend in the whole, but also has a similar traffic volume value at the same sampling interval. In particular, Tables 6 and 7 give the correlation coefficient matrices of the traffic volume data shown in Figures 7(a) and 7(b), respectively. All the correlation coefficients are greater than 0.9 except the correlation coefficient (that is 0.8957 and very close to 0.9) between Thursday and Thursday in Table 7, which illustrates that the traffic volume time series of urban expressway and urban arterial road are both with strong week correlation.

According to Section 4.2.1, it can be found that the traffic flow data of urban expressway and urban arterial road are both with strong temporal correlation (daily correlation and week correlation). In theory, the temporal correlation can be used for missing traffic data imputation effectively.

4.2.2. Spatial Correlation Analysis. The spatial correlation of traffic volume data is mainly reflected in the "link pattern" and the "section pattern." The "link pattern" is the traffic flow data collected in the same link (lane) but in different sections. The "section pattern" is the traffic flow data collected in the same section but in different links (lanes). In order to analyze the spatial correlation for urban road traffic volume, several

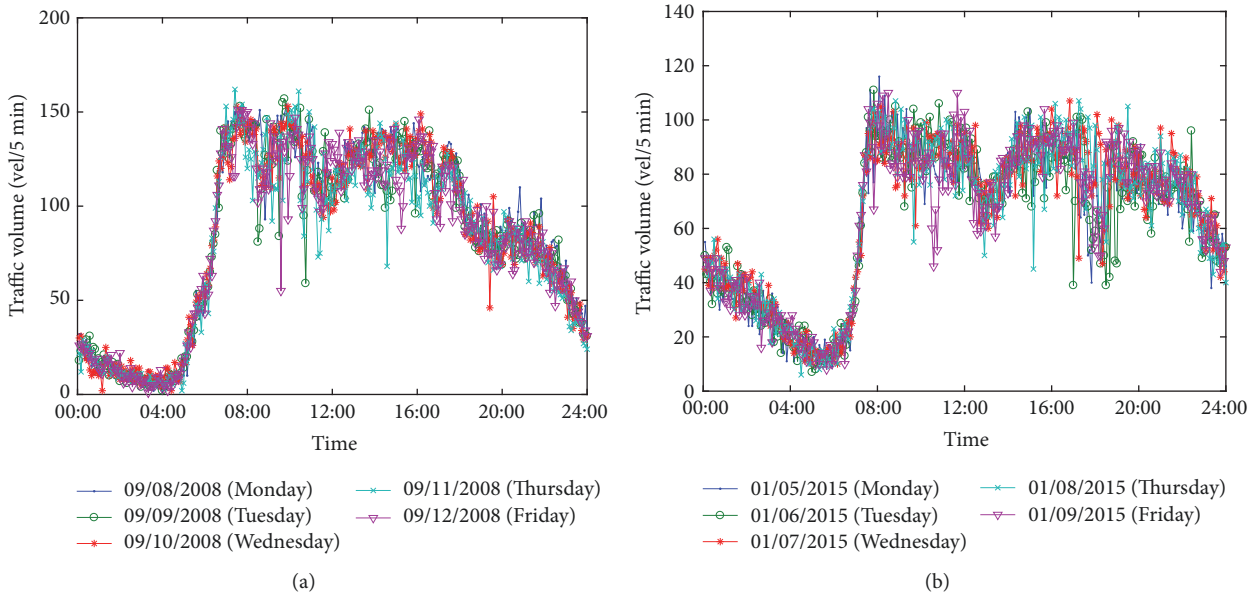

Figure 7: An example for illustrating the "week pattern": (a) urban expressway and (b) urban arterial road.

Table 5: Correlation coefficient matrices of two traffic volume series from the same detector located at urban arterial road but in different Monday.

	01.05 (Monday)	01.12 (Monday)	01.19 (Monday)	01.26 (Monday)	02.02 (Monday)
01.05 (Monday)	1	0.9411	0.9277	0.9224	0.9469
01.12 (Monday)	0.9411	1	0.9219	0.9254	0.9571
01.19 (Monday)	0.9277	0.9219	1	0.9373	0.9764
01.26 (Monday)	0.9224	0.9254	0.9373	1	0.9874
02.02 (Monday)	0.9469	0.9571	0.9764	0.9874	1

Table 6: Correlation coefficient matrices of two traffic volume series from the same detector located at urban expressway but in five consecutive working days.

	09.08 (Monday)	09.09 (Tuesday)	09.10 (Wednesday)	09.11 (Thursday)	09.12 (Friday)
09.08 (Monday)	1	0.9609	0.9694	0.9585	0.9565
09.09 (Tuesday)	0.9609	1	0.9701	0.9500	0.9554
09.10 (Wednesday)	0.9694	0.9701	1	0.9537	0.9625
09.11 (Thursday)	0.9585	0.9500	0.9537	1	0.9417
09.12 (Friday)	0.9565	0.9554	0.9625	0.9417	1

Table 7: Correlation coefficient matrices of two traffic volume series from the same detector located at urban arterial road but in five consecutive working days.

	01.05 (Monday)	01.06 (Tuesday)	01.07 (Wednesday)	01.08 (Thursday)	01.09 (Friday)
01.05 (Monday)	1	0.9051	0.9204	0.9178	0.9194
01.06 (Tuesday)	0.9051	1	0.9040	0.8957	0.9009
01.07 (Wednesday)	0.9204	0.9040	1	0.9097	0.9175
01.08 (Thursday)	0.9178	0.8957	0.9097	1	0.9058
01.09 (Friday)	0.9194	0.9009	0.9175	0.9058	1

Table 8: Correlation coefficient matrices of two traffic volume series from several adjacent sections in a lane of the urban expressway.

	Section 1	Section 2	Section 3	Section 4	Section 5
Section 1	1	0.9910	0.9711	0.9477	0.9592
Section 2	0.9910	1	0.9808	0.9510	0.9651
Section 3	0.9711	0.9808	1	0.9481	0.9695
Section 4	0.9477	0.9510	0.9481	1	0.9572
Section 5	0.9592	0.9651	0.9695	0.9572	1

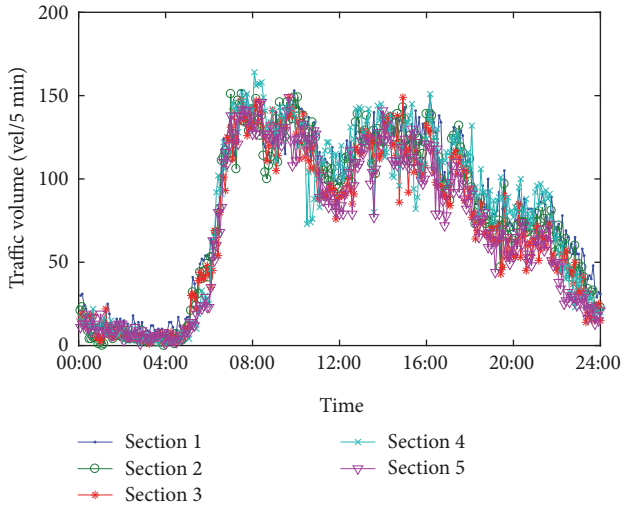

Figure 8: Traffic volume data from several adjacent sections in a lane of the urban expressway.

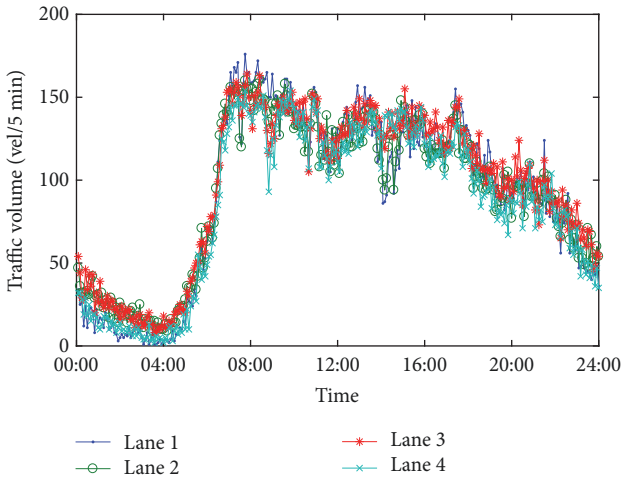

Figure 9: Traffic volume data from several adjacent lanes in a section of the urban expressway.

typical loop detectors are randomly selected from the urban expressway and the urban arterial road, respectively, whose traffic volume data are used for spatial correlation analysis.

The spatial correlation of traffic flow data is usually closely related to the structure of road where the traffic data are collected. Taking into account the different structures of urban expressway and urban arterial road, the traffic flow data from urban expressway and urban arterial roads are analyzed, respectively.

In urban expressway, there are usually no intersections to hinder the operation of the traffic flow. In general, traffic volume data of urban expressway have "link pattern" and the "section pattern" obviously. Figure 8 shows traffic volume data from several adjacent sections in a lane of the urban expressway to demonstrate the "link pattern." Figure 9 shows traffic volume data from several adjacent lanes in a section of the urban expressway to demonstrate the "section pattern." As can be seen from Figure 8, the traffic volume series of each section is similar to each other obviously. In Figure 9, the traffic volume series of each lane is also similar to each other obviously. Tables 8 and 9 give the correlation coefficient matrices of the traffic volume data shown in Figures 8 and 9, respectively. It can be seen in Tables 8 and 9 that all the correlation coefficients are greater than 0.9, which illustrates that the traffic volume time series of urban expressway show strong spatial correlation.

Table 9: Correlation coefficient matrices of two traffic volume series from several adjacent lanes in a section of the urban expressway.

	Lane 1	Lane 2	Lane 3	Lane 4
Lane 1	1	0.9843	0.9739	0.9642
Lane 2	0.9843	1	0.9761	0.9718
Lane 3	0.9739	0.9761	1	0.9843
Lane 4	0.9642	0.9718	0.9843	1

The structure of the urban arterial road is significantly different from that of the urban expressway. For the urban arterial road, signal intersections affect the continuity of traffic flow, because there are many confluences and separations of traffic flow at the intersection. In general, the traffic volume of urban arterial road is considered not to have "link pattern." Figure 10 shows traffic volume data from two adjacent sections in a lane of the urban arterial road. It can be seen from Figure 10 that traffic volume data of the two sections have a large difference, and the correlation coefficient is calculated as 0.7931, which indicate that the urban arterial road has a weak "link pattern." Next, we analyze whether there is "section pattern" for the urban arterial road data.

Figure 11 shows traffic volume data from three adjacent lanes in a section of the urban arterial road. It can be seen from Figure 11 that traffic volume of data the two straight

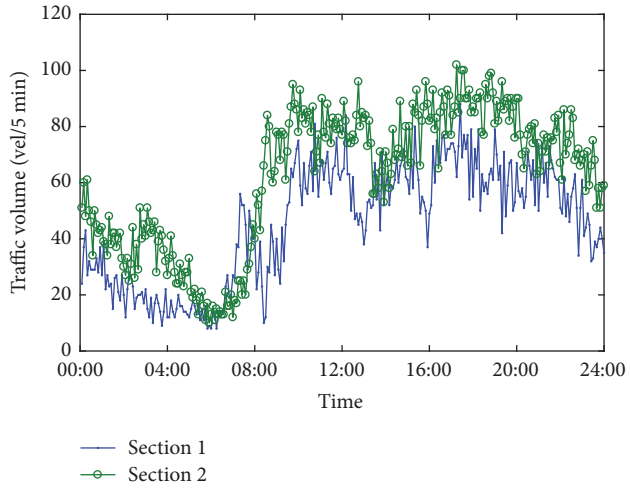

Figure 10: Traffic volume data from two adjacent sections in a lane of the urban arterial road.

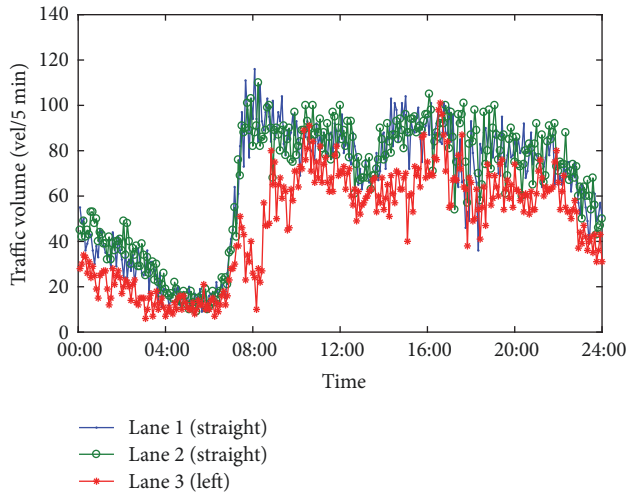

Figure 11: Traffic volume data from three adjacent lanes in a section of the urban arterial road.

lanes are very similar but are very different from the traffic volume data of the left lane. Table 10 gives the correlation coefficient matrices of the traffic volume data of these three lanes. The correlation coefficient of these two straight lanes is 0.9215, while the correlation coefficients between the straight lane and the left lane are less than 0.85. Therefore, traffic volume data from straight lanes of an urban arterial road have a strong "section pattern."

4.3. Experimental Scheme. In order to evaluate the performance of the proposed method, we select no missing data (or the data with a very small missing ratio) from the urban expressway and the urban arterial road mentioned above; then, according to several certain missing patterns, the missing data are generated artificially; finally, the PSO-SVR-FCM method is used to impute these missing data and the differences between the imputed values and the actual (observed) values are compared. To analyze the performance

Table 10: Correlation coefficient matrices of two traffic volume series from three adjacent lanes in a section of the urban arterial road.

	Lane 1 (straight)	Lane 2 (straight)	Lane 3 (left)
Lane 1 (straight)	1	0.9215	0.8048
Lane 2 (straight)	0.9215	1	0.8447
Lane 3 (left)	0.8048	0.8447	1

of the imputation method more comprehensively, experimental scheme is designed from three aspects as follows.

(1) Two kinds of data sources mentioned in Section 4.1 are selected, including urban expressway data and urban arterial road data. Both urban expressway data and urban arterial road data are collected in 5 min intervals so that a daily traffic volume time series contains 288 data points.

For the urban expressway, the "day pattern" data are collected from five consecutive Tuesdays (09/02/2008, 09/08/2008, 09/16/2008, 09/23/2008, and 09/30/2008); the "week pattern" data are collected from five consecutive working days (09/08/2008~09/12/2008). The "line pattern" data are collected from two sets of detectors, and one set of detectors contains five detectors numbered NBXX11(4), NBXX12(4), NBXX13(4), NBXX14(4), and NBXX15(4), and the other set of detectors contains five detectors numbered NBXX11(2), NBXX12(2), NBXX13(2), NBXX14(2), and NBXX15(2). The "section pattern" data are also collected from two sets of detectors, and one set of detectors contains five detectors numbered NBXX13(1), NBXX13(2), NBXX13(3), and NBXX13(4), and the other set of detectors contains five detectors numbered NBXX11(1), NBXX11(2), NBXX11(3), and NBXX11(4).

For the urban arterial road, the "day pattern" data are collected from five consecutive Mondays (01/05/2015, 01/12/2015, 01/19/2015, 01/26/2015, and 02/02/2015); the "week pattern" data are collected from five consecutive working days (01/05/2015~01/09/2015). The "section pattern" data are collected from the detectors in the same section but in different straight lanes. Here, two sets of detectors are used for test the proposed method, and the first set contains three detectors numbered DC00004964(E1), DC00004964(E2 and DC00004964(E3), and the second set contains two detectors numbered DC00004963(W2) and DC00004963(W3). According to the spatiotemporal correlation analysis of traffic data from an urban arterial road (see Section 4.2), we can see that urban arterial road traffic data show weak "link pattern." Therefore, "link pattern" is not taken into account for imputation of the missing traffic data from an urban arterial road.

(2) Whether spatial-temporal correlation data for the missing traffic data is complete and available, in the process of traffic data collection, spatial-temporal correlation data for the missing traffic data are often incomplete or unavailable data, especially spatial correlation data. Even for urban expressway with better continuous traffic flow, this problem still exists. For example, the detector of the most

upstream/downstream detection section or the detector of the most edge lane has less adjacent detectors, so that the available spatial-temporal correlation data are also less. In addition, spatial-temporal correlation data of other detectors may also be unavailable due to long time failures of adjacent detectors. Therefore, it is necessary to verify the method performance for missing traffic data imputation using incomplete spatial-temporal correlation data.

In this paper, the detector, whose missing data need to be estimated, is defined as the target detector. For urban expressway, the detector NBXX13(2) and the detector NBXX11(4) are used as target detectors respectively. Since the detector NBXX11(4) is located on the outermost lane of the most upstream section, its spatial correlation data is not comprehensive. In contrast, the detector NBXX13(2) is located in the middle lane of the midsection, so that its spatial correlation data is relatively comprehensive. Similarly, two detectors DC00004964(E2) and DC00004963(W2) that located on urban arterial road are selected as the target detectors. The spatial correlation data of detector DC00004964 (E2) is relatively comprehensive, and the spatial correlation data of the detector DC00004963 (W2) is not comprehensive.

(3) Different patterns of missing traffic data are taken into account, including Missing Completely at Random (MCR), Missing at Random (MR), and a combination of these two patterns (mixed MCR/MR). For each pattern of missing data, missing traffic volume data points are simulated by setting different missing ratios: 1%, 5%, 10%, 15%, 20%, 25%, and 30%.

4.4. Results and Discussion. In order to analyze the performance of the proposed method more clearly, several typical missing traffic data imputation methods are introduced for comparison, including the imputation method based on genetic algorithm and fuzzy C-means (GA-FCM) [19], the imputation method based on K-Nearest Neighbor and Non-Parametric Regression (KNN-NPR) [36], the imputation method based on ARIMA model (ARIMA) [5], and the SVR-based imputation method optimized by PSO (PSO-SVR) [33]. The GA-FCM and PSO-SVR belong to machine learning methods, the KNN-NPR belongs to interpolation methods, and ARIMA belongs to prediction methods. In order to ensure the performance of the comparison methods (GA-FCM, KNN-NPR, ARIMA, and PSO-SVR), their parameters are set and optimized according to the corresponding literatures.

In addition, two evaluation criteria are selected to measure the imputation accuracy of the methods: Root Mean Square Error (RMSE) and Relative Accuracy (RA). RMSE measures the error between the actual values and the estimated values and can be calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \tilde{y}_i\right)^2}, \tag{14}$$

where y_i is the actual value of the ith missing data point, $\tilde{y}_i$ is the estimated value of the ith missing data point, and n is the number of missing data points.

RA is a measure of how many estimations are made within a certain tolerance. In this study, the tolerance is set to 10% as performed in [19, 23]. RA is calculated as (15).

$$\mathrm{RA} = \frac{n_p}{n} \times 100\% \tag{15}$$

$$\mathrm{PAE} = \frac{\left|y_i - \tilde{y}_i\right|}{y_i} \times 100\%, \tag{16}$$

where n_p is the number of correct estimations within a certain tolerance (here is PAE <= 10%). PAE means Percentage Absolute Error and is calculated as (16).

4.4.1. Optimization of the Proposed Method Parameters. It is known from the experimental scheme that many groups of traffic data are used to test the PSO-SVR-FCM method. Next, the traffic data from NBXX13(2) detector is taken as an example to demonstrate the process of parameters optimization of the PSO-SVR-FCM method. The missing data of NBXX13(2) detector is generated with a missing ratio of 15% in the mixed MCR/MR pattern. Randomly select 75% of the data as a training set and 25% of the data as a test set. The k-fold cross-validation method [37] is used for training, which can make full use of the information in the sample and avoid overfitting and underfitting. In other words, it can improve the generalization ability of the model under the premise of ensuring good imputation accuracy. In k-fold cross-validation, the training dataset is randomly divided into k subsets. The $k-1$ subsets are used as training set for building the model and the kth subset is used as a validation set for verifying model performance. Each subset is used as a validation set and the verifying is repeated k times in total. The average value of the results of k times verifying is used to evaluate the model performance. In this study, 5-fold cross-validation is selected.

For the proposed method, FCM is the basic algorithm and the FCM parameters are optimized via a combination of PSO algorithm and SVR. Firstly, the parameters of the PSO algorithm are set as follows: population number is 20, acceleration coefficient $c_1 = c_2 = 2$, maximum iteration number $t_{\max} = 100$, starting inertia factor $w_{\mathrm{start}} = 0.9$, and termination inertia factor $w_{\mathrm{end}} = 0.4$. Secondly, the parameters of SVR should be set before the SVR is trained using complete data. In this study, the parameters of SVR are also optimized using the PSO algorithm, and the optimized parameters of SVR are obtained as follows: penalty factor $C = 15.48$, loss function parameter $\varepsilon = 0.01$, and kernel parameter $\sigma = 0.53$. Thirdly, the iteration termination condition for training FCM is that the reduction between objective function values in two iterations is less than 0.0001, or the maximum iterations number 100 is reached. It is necessary to determine the range of FCM parameters (weighting factor m and cluster size K) before optimizing the FCM parameters. Here, the range of FCM parameters are set as $1 < m \leq 5$ and $K \leq \sqrt{n}$, where is the n sample size. In addition, the similarity between the data and the cluster centroid is measured using the Euclidean distance [19]. It is worth noting that all data need to be normalized at first, and the imputation data are

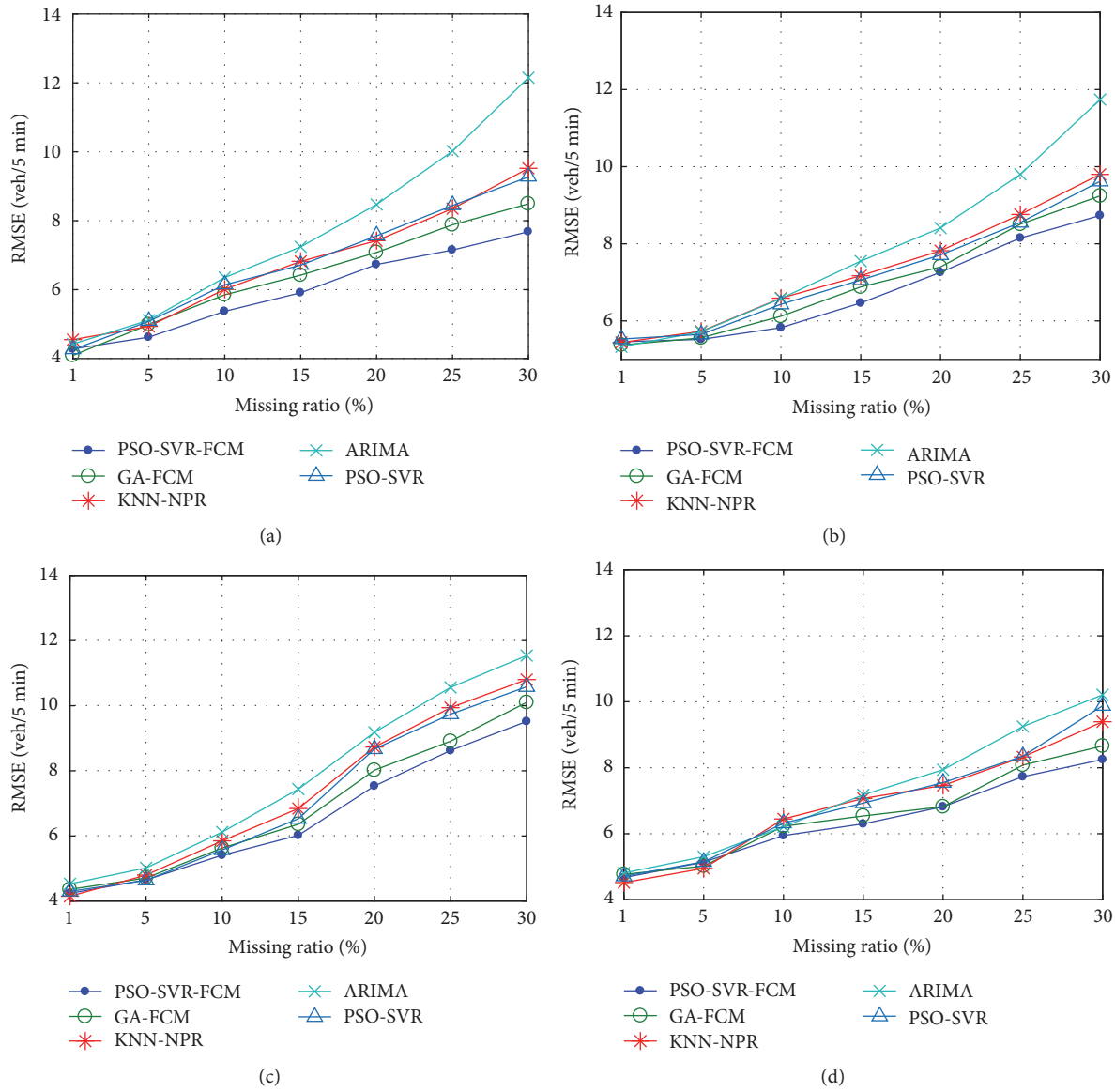

Figure 12: The RMSE curve of each method for comparison in MCR pattern, using the traffic data collected from the detectors (a) NBXX13(2), (b) NBXX11(4), (c) DC00004964(E2), and (d) DC00004963(W2).

obtained by the antinormalization. The normalized formula is as follows:

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{17}$$

where x_i is the ith raw data, y_i is the ith normalized data, $x_{\max}$ is the maximum value of all the raw data, and $x_{\min}$ is the minimum of all the raw data.

4.4.2. Test Results in MCR Pattern. Figure 12 shows the RMSE curve of each method for comparison in MCR pattern. As can be seen from Figure 12(a), the RMSE of each method is similar and smaller, when the missing ratio is lower. With the increase of the missing ratio, the RMSE of each method is gradually increasing, among which the increase rate of PSO-SVR-FCM method RMSE is the slowest. For almost any the missing ratios, RMSE of PSO-SVR-FCM method is less than the other three methods.

Figure 13 gives the RA curve of each method for comparison in MCR pattern. As can be seen from Figure 13(a), when the missing ratio is 1%, the RA of each method is 100%, which is due to the fact that the missing ratio of 1% corresponds to a small amount of missing data. With the increase of missing ratio, the RA of each method decreases gradually and the RA decrease rate of PSO-SVR-FCM method is the slowest, which shows that this method has better performance.

The RMSE curves shown in Figure 12(b) are similar to those in Figure 12(a), as can be seen from Figure 12(b), the RMSE of each method is close when the missing ratio is 1% and 5%, for other missing ratios, the RMSE of the PSO-SVR-FCM method is smaller than that of the other comparison methods, which shows that the PSO-SVR-FCM method

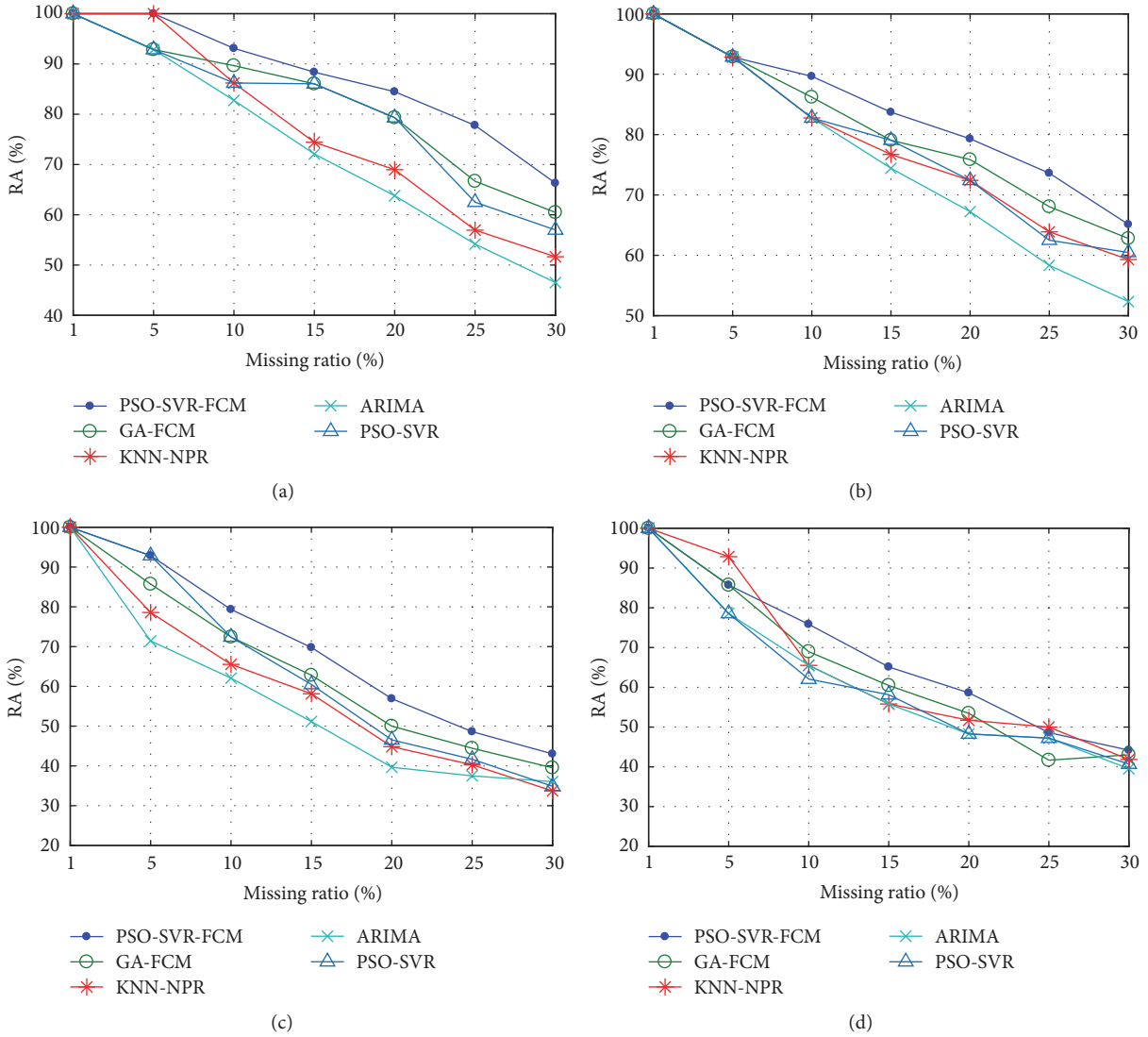

Figure 13: The RA curve of each method for comparison in MCR pattern, using the traffic data collected from the detectors (a) NBXX13(2), (b) NBXX11(4), (c) DC00004964(E2), and (d) DC00004963(W2).

can impute the missing values by using the incomplete spatial-temporal information and also has a relatively good performance. However, an important difference between Figure 12(b) and Figure 12(a) can be found that, for the urban expressway data with incomplete spatial-temporal information (shown in Figure 12(b)), the difference between the RMSE of the PSO-SVR-FCM method and the RMSE of the other methods is small, indicating that the PSO-SVR-FCM method has a smaller advantage than the other four methods. The reason is that incomplete spatial-temporal information makes the performance of the PSO-SVR-FCM method declined in a certain extent.

The RA curves shown in Figure 13(b) are similar to that in Figure 13(a). In Figure 13(b), the RA of the PSO-SVR-FCM method is higher than the other four methods except for the missing ratios of 1% and 5%, which indicates that performance of the PSO-SVR-FCM method is better. The difference is that the gap between the RA of the PSO-SVR-FCM method and the RA of other methods is reduced for urban expressway data without complete spatial-temporal information (see Figure 13(b)), which shows that incomplete spatial-temporal information makes the performance of the PSO-SVR-FCM method declined in a certain extent.

It can be seen from Figure 12(c) that the RMSE of the PSO-SVR-FCM method is very close to the RMSE of the GA-FCM method, when the missing ratio is 1%; and for other missing ratios, the RMSE of the PSO-SVR-FCM method is lower than the RMSE of the four comparison methods. It can be seen from Figure 13(c) that the RA of the PSO-SVR-FCM method is very close to the RA of the GA-FCM method, when the missing ratios are 1% and 5%; and for other missing ratios, the RA of the PSO-SVR-FCM method is higher than the RA of the four comparison methods. The comparison results show that the PSO-SVR-FCM method also has a superior

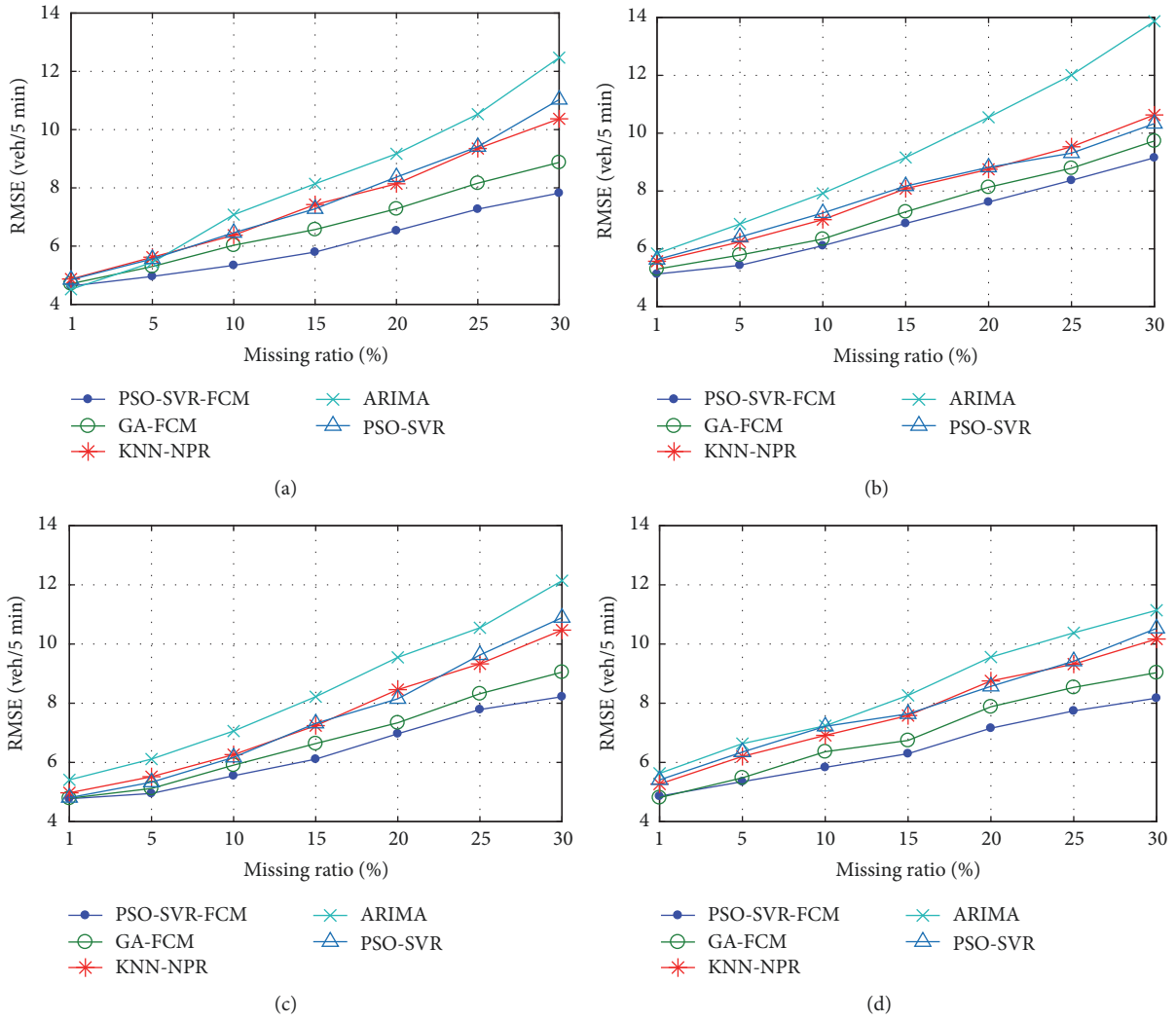

Figure 14: The RMSE curve of each method for comparison in MR pattern, using the traffic data collected from the detectors (a) NBXX13(2), (b) NBXX11(4), (c) DC00004964(E2), and (d) DC00004963(W2).

performance to estimate missing traffic volume of urban arterial road using complete spatial-temporal information.

It can be seen from Figure 12(d) that the RMSE of the PSO-SVR-FCM method is slightly lower than the RMSE of the KNN-NPR method when the missing ratios are 1% and 5%. However, with the increase of the missing ratio, the RMSE growth rate of the PSO-SVR-FCM method is the slowest. With the missing ratio of 10%, the RMSE of the PSO-SVR-FCM method is less than other methods, except for the missing ratio of 20%. It can be seen from Figure 13(d) that the RA of the PSO-SVR-FCM method is lower than the RA of the KNN-NPR method when the missing ratio is 5%; for other missing ratios, the RA of the PSO-SVR-FCM method is higher than or very close to the RA of other methods. The comparison results show that the PSO-SVR-FCM method also has relatively good performance to estimate missing traffic volume of urban arterial road without complete spatial-temporal information.

To sum up, in the MCR pattern, the PSO-SVR-FCM method shows better performance for missing traffic volume imputation, especially in the case of high missing ratio. The PSO-SVR-FCM method has a more significant advantage when the spatial-temporal information is complete and available.

4.4.3. Test Results in MR Pattern. Figure 14 shows the RMSE curve of each method for comparison in MR pattern. The results indicate that the RMSEs of all methods are similar to each other, when the missing ratio is 1%. Except for the missing ratio of 1%, the RMSE of the PSO-SVR-FCM method is less than comparison methods. Compared with the MCR pattern, the performance of PSO-SVR-FCM method is better in MR pattern, which is closely related to the data missing pattern. In the MR pattern, a large number of consecutive missing values are generated, especially with a high missing ratio. Therefore, it is more difficult to estimate the missing traffic values in the MR pattern using only

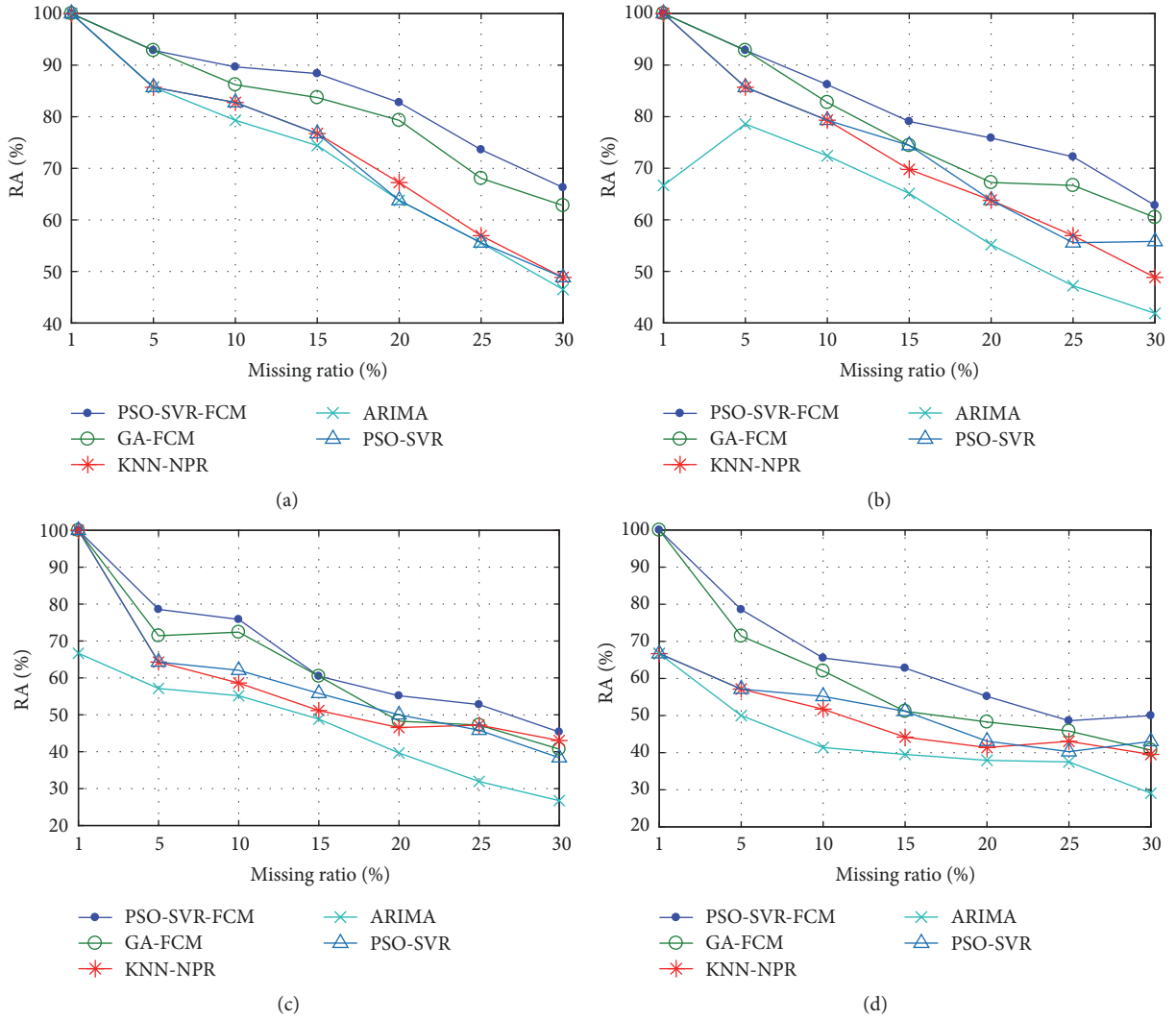

Figure 15: The RA curve of each method for comparison in MR pattern, using the traffic data collected from the detectors (a) NBXX13(2), (b) NBXX11(4), (c) DC00004964(E2), and (d) DC00004963(W2).

temporal correlation information. As a typical prediction based imputation method, ARIMA has a poor performance in the MR pattern.

Figure 15 shows the RA curve of each method for comparison in MR pattern. Clearly, the RA of the PSO-SVR-FCM method is greater than or equal to the RAs of other methods in any missing data. Therefore, for the MR pattern, the performance of the PSO-SVR-FCM method is superior to other comparison methods in terms of RA. And considering performance in terms of RMSE, the PSO-SVR-FCM method has better ability for missing traffic data imputation in MR pattern.

4.4.4. Test Results in Mixed MCR/MR Pattern. Figure 15 shows the RMSE curve of each method for comparison in mixed MCR/MR pattern. As can be seen from Figure 15, no matter what kind of data, the RMSE of each method is similar and smaller, when the missing ratio is 1%. With increase of the missing ratio, the RMSE of the PSO-SVR-FCM method is less than or equal to the RMSE of other methods. Therefore, in terms of RMSE, the performance of the PSO-SVR-FCM method is preferable to other comparison methods obviously.

Figure 17 shows RA curve of each method for comparison in mixed MCR/MR pattern. It can be seen from Figures 17(a)–17(c) that the RA of the PSO-SVR-FCM method is greater than or equal to RAs of the methods with any missing ratio. However, the RA curve given in Figure 17(d) is slightly different, and when the missing ratio is 30%, the RA of the GA-FCM method is not only larger than the RA of the PSO-SVR-FCM method, but also larger than the RA of the GA-FCM method with the missing ratio of 25%. Not as we expected, the higher the missing ratio is, the lower RA is. But, there is no such phenomenon for RMSE of GA-FCM method based on the same test data (see Figure 16(d)). Because RMSE is an absolute criterion, and RA is a relative criterion, they may present different characteristics, which are why we select RMSE and RA as the evaluation criteria. It can be seen from Figure 17(d) that the RA of the PSO-SVR-FCM method is

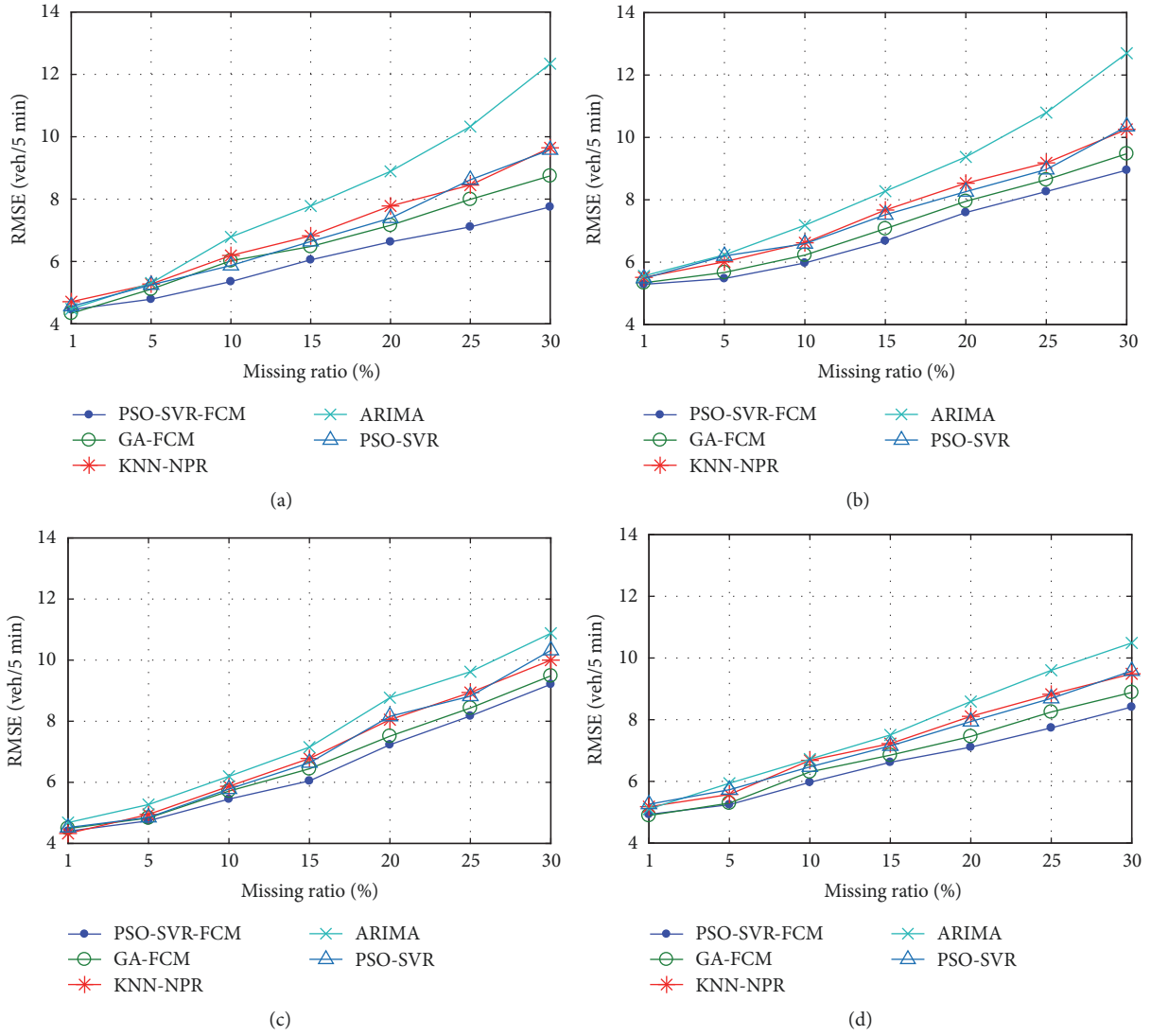

Figure 16: RMSE curve of each method for comparison in mixed MCR/MR pattern, using the traffic data collected from the detectors (a) NBXX13(2), (b) NBXX11(4), (c) DC00004964(E2), and (d) DC00004963(W2).

greater than or equal to RAs of other methods except for missing ratio of 30%.

As described above, the PSO-SVR-FCM method has a better capacity to impute missing traffic volume in the mixed MCR/MR pattern.

4.4.5. Statistical Analysis. The purpose of this section is to conduct a statistical analysis of the entire experimental results. The RMSE and RA of each method with different conditions have been given in above three sections. In this study, box-whisker plots are employed to illustrate the statistical results. Figure 18 shows box-whisker plots for RMSEs of different imputation methods. On each box, the central mark (red line) is the median; the edges of boxes are the 25th (Q1) and 75th (Q3) percentiles, and the interquartile range (IQR = Q3 − Q1) is used for evaluating the degree of concentration to median; the whiskers extend to the most extreme data points are not considered outliers (abnormal data points), and the outliers are the data points beyond the range of [Q3 + 1.5 IQR, Q1 − 1.5 IQR], which are usually plotted with symbol (+) individually.

It can be seen from Figure 18 that the median, 25th, and 75th percentiles of the PSO-SVR-FCM method RMSE are less than those of the other four methods RMSE, which indicates that the PSO-SVR-FCM method has less error for missing data imputation. In practice, an imputation method with a good statistical property is that its RMSE/RA is smaller/greater and has a fewer outliers. Fortunately, there is no outlier of RMSE obtained by each imputation method. However, the distance between the 25th and 75th percentiles of the PSO-SVR-FCM method RMSE is the smallest one and the normal range of the PSO-SVR-FCM method RMSE is also the smallest one, which shows that the PSO-SVR-FCM method can provide stable imputation results for different test data.

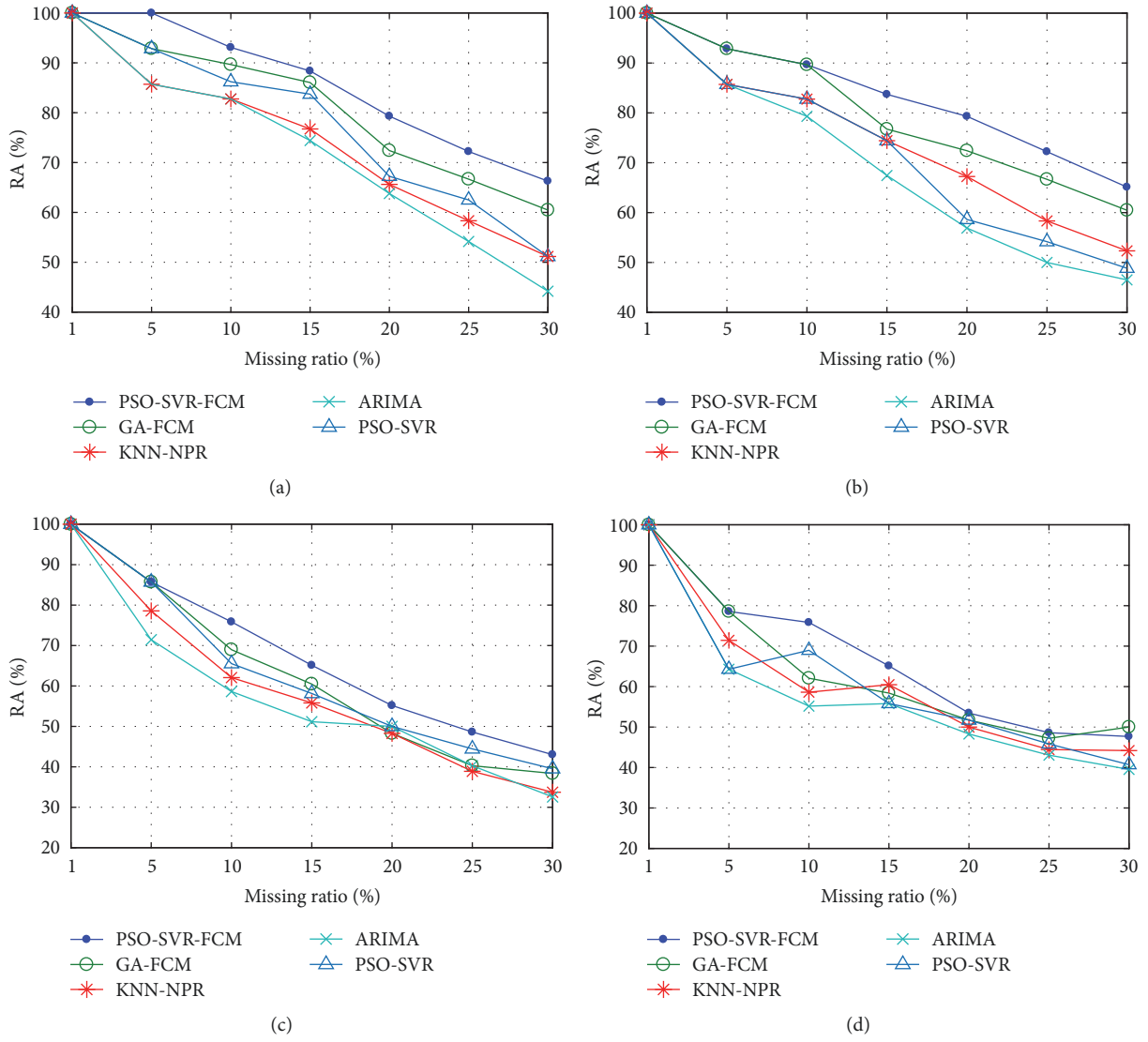

Figure 17: RA curve of each method for comparison in mixed MCR/MR pattern, using the traffic data collected from the detectors (a) NBXX13(2), (b) NBXX11(4), (c) DC00004964(E2), and (d) DC00004963(W2).

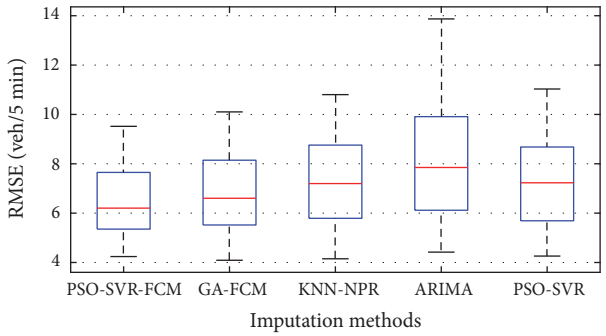

Figure 18: The box-whisker plots for RMSEs of different imputation methods.

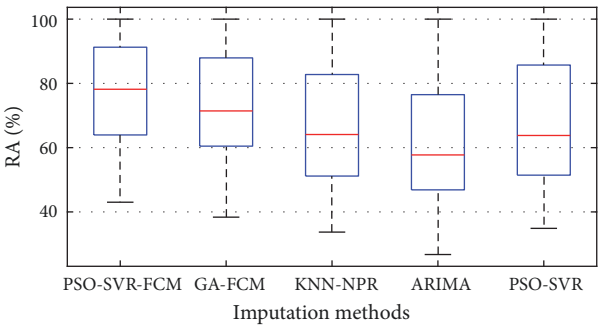

Figure 19: The box-whisker plots for RAs of different imputation methods.

Figure 19 illustrates the box-whisker plots for RAs of various imputation methods. It can be seen from Figure 19 that the median, 25th and 75th percentiles of the PSO-SVR-FCM method RA are greater than that of the other

four methods RA, which indicates that the PSO-SVR-FCM method has higher accuracy for missing data imputation. Similar to Figure 18, there is no outlier of RA obtained by each imputation method in Figure 19. However, the distance between the 25th and 75th percentiles of the PSO-SVR-FCM method RA is close to that of GA-FCM method RA and is one of the smallest RA, and the normal range of the PSO-SVR-FCM method RA is also the smallest, which indicate that the PSO-SVR-FCM method can provide accuracy imputation results with good robustness.

5. Conclusions

Missing data is a common problem which has to be faced in the transportation management and control systems. In this paper, we propose a hybrid method for missing traffic data imputation based on FCM optimized by a combination of PSO and SVR. The "day pattern," "week pattern," "link pattern," and "section pattern" of traffic flow data are taken into account, and matrix-based data structure is used to express the missing traffic data, so that the better imputation results can be achieved by making full use of the spatial-temporal correlation of traffic flow data. Then, the experiment is designed from three perspectives to test the proposed method and four typical imputation methods (GA-FCM, KNN-NPR, ARIMA, and PSO-SVR) are introduced for comparison. Based on the comparison and analysis of the experimental results, the conclusions can be drawn as follows:

(1) When the missing ratio is low, the four methods can provide good imputation accuracy. With the increase of the missing ratio, the imputation accuracy of the four methods gradually decreased. However, decreased rate of PSO-SVR-FCM method imputation accuracy is the slower than the compared methods, which indicate that PSO-SVR-FCM method performance is better.

(2) Compared with the MCR pattern, the imputation performances of the four methods are relatively poor in the MR pattern. Because the MR pattern has more continuous missing data points, it is more necessary to use spatial-temporal information for missing traffic data imputation.

(3) The available spatial-temporal information of the urban expressway traffic data is more than that of the urban arterial road traffic data. Therefore, an imputation method often achieves better performance to estimate the missing traffic data from the urban expressway.

(4) The PSO-SVR-FCM method is more sensitive to the spatial-temporal information of the urban expressway. When the spatial-temporal information of the urban expressway is incomplete, the performance of the PSO-SVR-FCM method is reduced. However, for urban arterial road traffic data, this phenomenon is not obvious; the reason may be that traffic data of urban arterial road has less spatial information originally.

Although the proposed method has achieved favorable results for missing traffic volume imputation, other traffic flow variables (such as traffic speed, travel time, and occupancy) data and the data collected under extreme conditions (such as accidents and extreme weather) will be used to test the proposed method in further study. Moreover, future work will also focus on the improvement of the proposed method. An important direction is the integration of more advanced and efficient intelligent optimization algorithm.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is supported by the National Natural Science Foundation of Shandong Province (Grant no. ZR2016EL19) and the Dr. Scientific Research Start Funding Projects of Shandong University of Technology (Grant no. 417006).

References

[1] D. Ni, J. D. Leonard II, A. Guin, and C. Feng, "Multiple imputation scheme for overcoming the missing values and variability issues in ITS data," *Journal of Transportation Engineering*, vol. 131, no. 12, pp. 931–938, 2005.

[2] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: differences, similarities and some insights," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 3, pp. 387–399, 2011.

[3] L. Qu, L. Li, Y. Zhang, and J. Hu, "PPCA-based missing data imputation for traffic flow volume: a systematical approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 3, pp. 512–522, 2009.

[4] H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li, "A tensor-based method for missing traffic data completion," *Transportation Research Part C: Emerging Technologies*, vol. 28, pp. 15–27, 2013.

[5] M. Zhong, S. Sharma, and P. Lingras, "Genetically designed models for accurate imputation of missing traffic counts," *Transportation Research Record*, no. 1879, pp. 71–79, 2004.

[6] M. Zhong, P. Lingras, and S. Sharma, "Estimation of missing traffic counts using factor, genetic, neural, and regression techniques," *Transportation Research Part C: Emerging Technologies*, vol. 12, no. 2, pp. 139–166, 2004.

[7] M. Zhong, S. Sharma, and Z. Liu, "Assessing robustness of imputation models based on data from different jurisdictions: Examples of alberta and saskatchewan, Canada," *Transportation Research Record*, no. 1917, pp. 116–125, 2005.

[8] W. Yin, P. Murray-Tuite, and H. Rakha, "Imputing erroneous data of single-station loop detectors for nonincident conditions: Comparison between temporal and spatial methods," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 16, no. 3, pp. 159–176, 2012.

[9] B. M. Williams, "Multivariate vehicular traffic flow prediction: evaluation of ARIMAX modeling," *Transportation Research Record*, no. 1776, pp. 194–200, 2001.

[10] J. W. C. van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Accurate freeway travel time prediction with state-space neural networks under missing data," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 5-6, pp. 347–369, 2005.

[11] Z. Liu, S. Sharma, and S. Datla, "Imputation of missing traffic data during holiday periods," *Transportation Planning and Technology*, vol. 31, no. 5, pp. 525–544, 2008.

[12] J. Wang, N. Zou, and G.-L. Chang, "Travel time prediction: Empirical analysis of missing data issues for advanced traveler information system applications," *Transportation Research Record*, no. 2049, pp. 81–91, 2008.

[13] D. Ni and J. D. Leonard II, "Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data," *Transportation Research Record*, no. 1935, pp. 57–67, 2005.

[14] 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.

[15] L. Qu, Y. Zhang, J. Hu, L. Jia, and L. Li, "A BPCA based missing value imputing method for traffic flow volume data," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV '08)*, pp. 985–990, June 2008.

[16] J.-M. Chiou, Y.-C. Zhang, W.-H. Chen, and C.-W. Chang, "A functional data approach to missing value imputation and outlier detection for traffic flow data," *Transportmetrica B*, vol. 2, no. 2, pp. 106–129, 2014.

[17] H. Tan, J. Feng, Z. Chen, F. Yang, and W. Wang, "Low multilinear rank approximation of tensors and application in missing traffic data," *Advances in Mechanical Engineering*, vol. 2014, Article ID 157597, 2014.

[18] B. Ran, H. Tan, Y. Wu, and P. J. Jin, "Tensor based missing traffic data completion with spatial-temporal correlation," *Physica A: Statistical Mechanics and its Applications*, vol. 446, pp. 54–63, 2016.

[19] J. Tang, G. Zhang, Y. Wang, H. Wang, and F. Liu, "A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation," *Transportation Research Part C: Emerging Technologies*, vol. 51, pp. 29–40, 2015.

[20] M. T. Asif, N. Mitrovic, J. Dauwels, and P. Jaillet, "Matrix and Tensor Based Methods for Missing Data Estimation in Large Traffic Networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 1816–1825, 2016.

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

[22] X. Chen, Z. Wei, Z. Li, J. Liang, Y. Cai, and B. Zhang, "Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation," *Knowledge-Based Systems*, vol. 132, pp. 249–262, 2017.

[23] 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.

[24] A. G. Di Nuovo, "Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario," *Expert Systems with Applications*, vol. 38, no. 6, pp. 6793–6797, 2011.

[25] J. Luengo, J. A. Sáez, and F. Herrera, "Missing data imputation for fuzzy rule-based classification systems," *Soft Computing*, vol. 16, no. 5, pp. 863–881, 2012.

[26] N. A. Samat and M. N. M. Salleh, "A study of data imputation using fuzzy c-means with particle swarm optimization," *Advances in Intelligent Systems and Computing*, vol. 549, pp. 91–100, 2017.

[27] J. Kennedy, "Particle swarm optimization," in *Encyclopedia of Machine Learning*, pp. 760–766, Springer US, Boston, MA, USA, 2011.

[28] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: a study of fuzzy K-means clustering method," in *Rough sets and current trends in computing*, vol. 3066 of *Lecture Notes in Comput. Sci.*, pp. 573–579, Springer, Berlin, 2004.

[29] V. N. Vapnik and V. Vapnik, *Statistical Learning Theory*, Wiley, New York, NY, USA, 1998.

[30] G. Sermpinis, C. Stasinakis, R. Rosillo, and D. de la Fuente, "European exchange trading funds trading with locally weighted support vector regression," *European Journal of Operational Research*, vol. 258, no. 1, pp. 372–384, 2017.

[31] J. Zhao, H. Wu, and L. Chen, "Road Surface State Recognition Based on SVM Optimization and Image Segmentation Processing," *Journal of Advanced Transportation*, vol. 2017, pp. 1–21, 2017.

[32] M. Castro-Neto, Y.-S. Jeong, M.-K. Jeong, and L. D. Han, "Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6164–6173, 2009.

[33] F. Honghai, C. Guoshun, Y. Cheng, Y. Bingru, and C. Yumei, "A SVM regression based approach to filling in missing values," in *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 3683 of *Lecture Notes in Computer Science*, pp. 581–587, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[34] A. Shahi, R. B. Atan, and M. N. Sulaiman, "Detecting effectiveness of outliers and noisy data on fuzzy system using FCM," *European Journal of Scientific Research*, vol. 36, no. 4, pp. 627–638, 2009.

[35] 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, article 32, no. 1, 2006.

[36] H. Chang, D. Park, Y. Lee, and B. Yoon, "Multiple time period imputation technique for multiple missing traffic variables: Nonparametric regression approach," *Canadian Journal of Civil Engineering*, vol. 39, no. 4, pp. 448–459, 2012.

[37] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *Journal of Econometrics*, vol. 187, no. 1, pp. 95–112, 2015.