Traffic Data Imputation Algorithm Based on Improved Low-Rank Matrix Decomposition

Xianglong Luo,1,2 Xue Meng,1 Wenjuan Gan,1 and Yonghong Chen1

1School of Information Engineering, Chang’an University, Xi’an 710064, China
2School of Highway, Chang’an University, Xi’an 710064, China

Correspondence should be addressed to Xue Meng; 2017124037@chd.edu.cn

Received 16 January 2019; Revised 8 May 2019; Accepted 11 June 2019; Published 1 July 2019

Abstract: Traffic data plays a very important role in Intelligent Transportation Systems (ITS). ITS requires complete traffic data in transportation control, management, guidance, and evaluation. However, the traffic data collected from many different types of sensors often includes missing data due to sensor damage or data transmission error, which affects the effectiveness and reliability of ITS. In order to ensure the quality and integrity of traffic flow data, it is very important to propose a satisfying data imputation method. However, most of the existing imputation methods cannot fully consider the impact of sensor data with data missing and the spatiotemporal correlation characteristics of traffic flow on imputation results. In this paper, a traffic data imputation method is proposed based on improved low-rank matrix decomposition (ILRMD), which fully considers the influence of missing data and effectively utilizes the spatiotemporal correlation characteristics among traffic data. The proposed method uses not only the traffic data around the sensor including missing data, but also the sensor data with data missing. The information of missing data is reflected into the coefficient matrix, and the spatiotemporal correlation characteristics are applied in order to obtain more accurate imputation results. The real traffic data collected from the Caltrans Performance Measurement System (PeMS) are used to evaluate the imputation performance of the proposed method. Experiment results show that the average imputation accuracy with proposed method can be improved 87.07% compared with the SVR, ARIMA, KNN, DBN-SVR, WNN, and traditional MC methods, and it is an effective method for data imputation.

1. Introduction

With the rapid development of the social economy, many kinds of the massive road infrastructure are implemented [1–4] but traffic congestion still exists in the highway. Therefore, it is necessary to collect the information of the highway for the convenience of people’s travelling demand. With the development of information technology, the collection of highway information becomes possible, and the collection equipment used for highways includes Bluetooth sensor, remote traffic microwave sensor, video sensors, and loop detectors. However, traffic flow data are lost in different degrees due to sensor damage, malfunction, or transmission errors, etc. Missing data makes it difficult to extract valid information from traffic data. Meanwhile, the missing data is also an obstacle in the traffic and the travel time prediction field [5–8], and the integrity of traffic flow data is the premise of data analysis in ITS. Therefore, it is very important to put forward an effective traffic data imputation method. At present, various methods have emerged in the field of traffic flow data imputation. These imputation methods can be roughly divided into three categories: prediction methods, interpolation methods, and statistical learning methods [9].

Traffic flow prediction models [10–12] are critical for road traffic management in complex road networks. Prediction methods usually build predictive models with historical data and treat missing data as values to be predicted. There are many ways to build traffic flow prediction models, from a simple null value imputation to complex spatiotemporal imputation models [13]. The representative prediction methods include Autoregressive Integrated Moving Average Model (ARIMA) [14–16], Bayesian networks (BNs) [17–19], and support vector regression (SVR) [20, 21]. Elshenawy et al. [22] proposed an intelligent data imputation method...
with ARIMA model and presented a mechanism based
on Hyndman-Khandakar algorithm to determine ARIMA
parameters. Sun et al. [23] partitioned a day into different
time section and used SVR to forecast traffic flow data. Chen
et al. [24] proposed an Autoregressive Integrated Moving
Average with Generalized Autoregressive Conditional Het-
eroscedasticity (ARIMA-GARCH) model for traffic flow pre-
diction. However, these prediction methods failed to utilize
the sensor information with missing data, which would affect
data imputation accuracy.

Interpolation methods are divided into temporal-
neighboring and pattern-neighboring [25]. Temporal-
neighboring methods fill up the missing data by the known
data from the same sensors at the same daily time but
on some neighboring days [20, 26]. Pattern-neighboring
methods use the similarity characteristics of the daily traffic
flow data [27] and estimate missing data using historical data
collected from the same sensors on different days [17, 20].
The typical pattern-neighboring methods include K-nearest
neighbors (KNN) model [28, 29] and Local Least Squares
(LLL) [30, 31] model, and the key difficulty of these methods
is to determine the neighbors by an appropriate distance
metric [32, 33]. Nguyen et al. [34] used the mean value of
the historical data to estimate missing data. Smith et al. [35]
used historical data or the data from surrounding periods
and locations to impute the missing data. The interpolation
model assumes that the daily traffic flow data are similar, but
the actual traffic flow data fluctuates and changes with time.
Therefore, it is impossible to obtain satisfactory imputation
performances.

The method based on statistical learning has been devel-
op en in recent years. This method primarily assumed the
probability distribution model of traffic data and used iter-
ative methods to estimate the parameters of the probability
distribution. Then the observed data was used to impute
the missing data. The statistical learning methods include
Probabilistic Principal Component Analysis (PPCA) [6, 9],
Bayesian Principal Component Analysis (BPCA) [26], neu-
ral network method [36], and Markov Chain Monte Carlo
(MCMC) [37]. The MCMC is a typical imputation method
based on statistical learning. The basic idea of the MCMC
method regards the missing data as the target parameter
and estimate the parameter by the sample values of the
parameter. Y Higashijima et al. [38] proposed a regression
tree imputation method and used a preprocessing method to
improve imputation accuracy. Wei et al. [39] proposed a data-
driven imputation method and used k-means clustering to
group the most correlated road segments; the trained model
is able to estimate the missing data at multiple locations
under a unified framework. Although the methods based
on statistical learning have strong hypothesis about traffic
data, their performance is superior to traditional imputation
methods [40] because the assumed probability distribution
captures the essentials of traffic flow.

The methods based on prediction and interpolation sim-
ply impute the data with the temporal or spatial correlation
characteristic and only consider the information of historical
data. The historical imputation methods fill the missing data
with the known data point collected on the same sensors at
the same daily time but from different days. These methods
require higher stability of historical data, but traffic flow
data is usually unstable and fluctuate to some extent in
practical applications. The traditional imputation method
sets all the missing data to zero and uses the data matrix
with zero-padding to participate the operation for the data
imputation, which cannot consider the impact of missing
sensor data into the imputation result. Generally, the sensors
including missing data have the highest correlation with
final imputation results. However, the missing data is set to
zero directly in the traditional imputation method, which
ignores the effect of the missing data on the imputation
results and reduces the accuracy of the imputation results. In
order to address the above problems, a traffic data imputation
method is proposed based on improved low rank matrix
decomposition (ILRMD). Compared with the traditional
imputation method, the ILRMD method fully considers the
impact of missing data in the imputation results. In the
process of data imputation, the ILRMD method does not
directly discard the information of missing data, and the
effect of missing data is reflected in the coefficient matrix.
The reconstructed data matrix multiplied by the coefficient
matrix, containing the missing data information, is the
imputation result. The ILRMD method uses not only the
traffic data around the sensor including missing data, but also
sensor data with data missing. The information contained in
the missing data is fully considered, and the spatiotemporal
correlation characteristics of the traffic flow are adequately
utilized. The tested results with traffic data collected from
the Caltrans Performance Measurement System (PeMS)
show that the proposed algorithm has superior imputation
accuracy.

The rest of this paper is organized as follows. Section 2
reviews the related work in traffic data imputation and gives
a brief introduction. The traditional imputation approach
is introduced in Section 3. Section 4 describes the ILRMD
method proposed in this paper. Section 5 discusses the
result analysis and method comparison. Section 6 makes the
conclusion of this paper and gives some recommendations.

2. Related Work

With the rapid development of machine learning, pattern
derognition, computer vision, and data mining, the process-
ing of big data is becoming more and more important. The
scale and growth rate of big data are continuously increasing,
but large-scale high-dimensional data is often correlative and
redundant. Therefore, it is necessary to perform reasonable
compression processing on large-scale data. In order to
reduce data redundancy, Candes [41] proposed the concept
of low rank sparse matrix decomposition in 2009, which
is also called Low-Rank Matrix Recovery (LRR), Low-
Rank Matrix Decomposition (LRMD), or Robust Principal
Component Analysis (RPCA).

2.1. Low-Rank Matrix Decomposition. For a given data matrix
$D \in R^{m \times n}$ distributed in a linear subspace with approximately
low dimension, it can be decomposed into a low-rank matrix
A and a sparse matrix $E$ [42].

$$\min_{A,E} \ \text{rank} (A) + \lambda \| E \|_0$$

s.t. $D = A + E$  \hspace{1cm} (1)

where $\| E \|_0$ represents the $L_0$ norm of the matrix $E$ and $\lambda$ represents the compromise factor of matrices $A$ and $E$.

Since the optimization problem of (1) is a NP-hard problem, it can be relaxed to the convex optimization problem [41–43], which is noted as follows:

$$\min_{A,E} \ \| A \|_* + \lambda \| E \|_{2,1}$$

s.t. $D = A + E$ \hspace{1cm} (2)

where $\| A \|_*$ represents the nuclear norm of matrix $A$; $\| E \|_{2,1} = \sum_{i=1}^m \sqrt{\sum_{j=1}^n e_{ij}^2}$ is the $L_{21}$ norm of the matrix $E$.

The low-rank characteristic of recovered matrix determines the matrix imputation performance. Therefore, choosing the suitable LRMD solution method is crucial. The main algorithms for solving LRMD problem include Iterative Threshold method [44, 45], the Dual Approach [46], Accelerated Proximal Gradient Algorithm [47], and Augmented Lagrange Multiplier method [48]. In this paper, Augmented Lagrange Multiplier method is used.

2.2. Matrix Imputation Based on Low-Rank Matrix Decomposition. Generally, we cannot recover all the data with partial sample data. But Candès [42] proved that the missing data can be recovered more accurately when data matrix is low or near low rank. From the Section 2.1, the low rank matrix $A$ is acquired based on LRMD, which can be used to impute the missing data.

The model of matrix imputation can be noted as follows:

$$\min_A \ \text{rank} (A)$$

s.t. $P_\Omega (A) = P_\Omega (D)$ \hspace{1cm} (3)

where $\Omega$ is the set of known element subscripts, and $\Omega \subseteq \mathbb{R}^{m \times n}$, $P_\Omega : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$ is a linear projection operator, which can be defined as follows:

$$P_\Omega (D_{ij}) = \begin{cases} D_{ij} & (i, j) \in \Omega \\ 0 & (i, j) \notin \Omega \end{cases} \hspace{1cm} (4)$$

The optimization problem of (3) is also a NP-hard problem, so it needs to be relaxed into a convex optimization problem:

$$\min_A \ \| A \|_*$$

s.t. $P_\Omega (A) = P_\Omega (D)$ \hspace{1cm} (5)

2.3. Matrix Imputation Based on Low-Rank Matrix Representation. The low-rank matrix imputation method mentioned above directly minimizes the rank of imputed data. In order to improve imputation efficiency, a self-expression is applied to LRMD, which is called the low-rank matrix representation [49, 50]. The data matrix $D$ is represented as a linear combination with a dictionary matrix $B$, that is, $D = BZ$. The matrix $Z$ is the coefficient matrix, and it is expected to be low rank. $Z$ can be obtained by solving the optimization problem in the following:

$$\min_Z \ \text{rank} (Z)$$

s.t. $D = BZ$ \hspace{1cm} (6)

Equation (6) can be convexly relaxed to obtain the following:

$$\min_Z \ \| Z \|_*$$

s.t. $D = BZ$ \hspace{1cm} (7)

If the data matrix $D$ is selected as the dictionary matrix, (7) can be noted as follows:

$$\min_Z \ \| Z \|_*$$

s.t. $D = DZ$ \hspace{1cm} (8)

In practical applications, the data matrix $D$ may be disturbed by noise. In order to enhance the robustness, (8) can be revised as follows:

$$\min_{Z,E} \ \| Z \|_* + \lambda \| E \|_{2,1}$$

s.t. $D = DZ + E$ \hspace{1cm} (9)

A data matrix $D$ is represented by a data dictionary $B$, and the coefficient matrix $Z$ is sparser when $D$ has higher similarity with $B$. But the stochastic noise is usually appended in data matrix $D$, which will influence the correlation within the data matrix. When the stochastic noise $E$ is removed, the correlation of data matrix can be enhanced. $D$ is selected as a dictionary, and its essence is to reveal the correlation within the data matrix. When the coefficient matrix $Z$ is sparse, data columns in data matrix $D$ are represented by each other’s columns with few coefficients as possible. For the traffic flow data, it has high spatiotemporal correlation characteristics, but it is affected by the weather, holidays, and other factors, which makes the traffic flow data have stochastic volatility. Therefore, if the influence of this stochastic volatility on the traffic data is removed, the correlation between the traffic data will be enhanced. After removing the influence of stochastic noise, the correlation between the data itself is further explored, and the similarity between the data is expressed with as little information as possible. Then the internal correlation of traffic flow data is used to impute the data.

2.4. The Solution of the Coefficient Matrix. In order to obtain the solution of (9), a variable $J$ is introduced and let $J = Z$ to separate the variable $Z$. The coefficient matrix $Z$ can be
calculated with the Augmented Lagrange Multiplier method, and the optimization model becomes the following:

$$
\begin{align*}
\min_z & \quad \|Y\|_2 + \lambda \|D - DZ\|_2, \\
\text{s.t.} & \quad Z = J
\end{align*}
$$

(10)

Construct an Augmented Lagrange function as (11), where $Y$ is a Lagrange Multiplier, $\| \cdot \|_2$ is Fibonacci norm, which represents the sum of the absolute squares of elements, and $\mu$ is a weight to tune the error term $\|Z - J\|_F$.

$$
L(Z, J) = \|Y\|_2 + \frac{1}{2} \|D - DZ\|_F^2 + \langle Y, Z - J \rangle + \frac{\mu}{2} \|Z - J\|_F^2
$$

(11)

The Exact Augmented Lagrange Multiplier (EALM) method is used to solve the matrices $J$ and $Z$ according to the following:

$$
J^{k+1} = \arg \min_J L(Z^k, J)
$$

(12)

$$
Z^{k+1} = \arg \min_Z L(Z, J^{k+1})
$$

(13)

The updating of the coefficient matrix $Z$ is as follows. Firstly, a projection matrix $D$ is used to express the unmissing position of the matrix $D$, and $D = W \cdot (D)$. For convenience, set $G = J - Y / \mu$ and (13) can be expressed as follows:

$$
\begin{align*}
\min_z & \quad \frac{1}{2} \|W \cdot (D) - W \cdot (DZ)\|_F^2 + \frac{\mu}{2} \|Z - G\|_F^2 \\
\text{s.t.} & \quad \|Z - J\|_F^2
\end{align*}
$$

(14)

In order to get a derivative about $Z$ in (14), the cross product should be changed to inner product. The matrices of (14) are spread in column as follows:

$$
\begin{align*}
\min_z & \quad \frac{1}{2} \sum_{i=1}^n \|W_i \cdot (d_i - w_i \cdot (Dz_i))\|_2^2 \\
& \quad + \frac{\mu}{2} \sum_{i=1}^n \|z_i - g_i\|_2^2
\end{align*}
$$

(15)

where $W_i$, $d_i$, and $z_i$ are, respectively, the $i^{th}$ column of matrices $W$, $D$, and $Z$.

Change vector $w_i$ to a diagonal matrix, i.e., $\tilde{W}_i = \text{diag}(w_i)$ and $w_i \cdot d_i = \tilde{W}_i \cdot d_i$. Therefore, (15) can be expressed as follows:

$$
\begin{align*}
\min_z & \quad \frac{1}{2} \sum_{i=1}^n \|\tilde{W}_i \cdot (d_i - \tilde{W}_i \cdot (Dz_i))\|_2^2 + \frac{\mu}{2} \sum_{i=1}^n \|z_i - g_i\|_2^2 \\
\text{s.t.} & \quad \|Z - J\|_F^2
\end{align*}
$$

(16)

For simplifying (16), $\tilde{W}_i \cdot d_i$ is denoted as $K_i$, and $\tilde{W}_i \cdot D$ is denoted as $H_i$. Then (16) can be simplified as follows:

$$
\begin{align*}
\min_z & \quad \frac{1}{2} \sum_{i=1}^n \|K_i - H_i z_i\|_F^2 + \mu \|z_i - g_i\|_2^2 \\
\text{s.t.} & \quad \|Z - J\|_F^2
\end{align*}
$$

(17)

For (17), $z_i$ can be updated by the following:

$$
\begin{align*}
z_i = \left( H_i^T H_i - \mu I \right)^{-1} (H_i^T K_i - \mu g_i)
\end{align*}
$$

(18)

Then repeat the above process until the objective function convergence. The coefficient matrix $Z$ can be obtained when the termination condition is met, and it is expressed as follows:

$$
Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\
z_{21} & z_{22} & \cdots & z_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
z_{n1} & z_{n2} & \cdots & z_{nn} \end{bmatrix} \in \mathbb{R}^{n \times n}
$$

(19)

3. Traditional Imputation Method with LRMD

The traditional method imputed the missing data by zero-padding operation. For an original matrix $D_0 = [d_1, d_2, \ldots, d_n] \in \mathbb{R}^{m \times n}$, suppose that $d_p$ is missing, where $d_p$ represent $p^{th}$ column in $D_0$. The missing column of the matrix $D_0$ is imputed by 0, which can be represented as a matrix $\tilde{D}_1$.

$$
\tilde{D}_1 = [\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_{p-1}, 0, \tilde{d}_{p+1}, \ldots, \tilde{d}_n] \in \mathbb{R}^{m \times n}
$$

(20)

where $\tilde{d}_i = \begin{cases} 0 & \text{if } i = 1, 2, \ldots, m \text{ is the specific elements in the matrix } \tilde{D}_1. \\
d_j & \text{if } i = j \text{ and } j \neq p \
\end{cases}$

Multiplying $\tilde{D}_1$ by the $p^{th}$ column of coefficient matrix $Z$, $\tilde{d}_p = [\tilde{d}_{1p}, \tilde{d}_{2p}, \ldots, \tilde{d}_{np}]^T$ can be recovered by the following:

$$
\tilde{d}_{ip} = \sum_{k=1}^n d_{ik} \cdot z_{kp}
$$

(21)

The zero-padding operation is used for the traditional matrix imputation method to filling the missing column. Then the reconstructed matrix is multiplied by the corresponding column of the coefficient matrix $Z$; the imputed data of the missing column is obtained. This method only uses the data around the missing column to impute the
missing data; that is to say, the missing column does not contribute to the imputation result. Generally, the sensors including missing data have the highest correlation with final imputation results. However, the missing data is set to zero directly in the traditional imputation method, which ignores the effect of the missing data on the imputation results and reduces the accuracy of the imputation results.

4. Traffic Data Imputation with ILRMD

The missing data generally can be divided into three different types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing at Determinate (MAD). This paper mainly deals with the problem of determinate missing. In road networks, traffic data was collected by various types of sensors, which usually demonstrated high temporal-spatial correlation characteristics; that is, traffic data have low-rank characteristic.

In a road network, suppose that there are $m$ sensors and each sensor has $n$ data samples, which can be denoted as a data matrix $D_{n,m}$. This paper assumed that the data in the $p^{th}$ column is missing. The traditional imputation method based on LRMD failed to consider the impact of missing data columns on imputation results. In order to address this shortcoming and combine the temporal-spatial correlation characteristics of traffic flow, this paper proposes a data imputation method based on ILRMD.

4.1. The Proposed ILRMD Model.

In (9), it is assumed that $d_{ij}$, $e_{ij}$ are the elements of the $j^{th}$ ($1 < j < n$) observed sensor at the $i^{th}$ ($1 < i < m$) time, respectively, existing in the observed matrix $D$ is $\mathbb{R}^{n,m}$ and the noise matrix $E$ is $\mathbb{R}^{m,n}$. $z_{ij}$ is the element of the coefficient matrix $Z$, and the coefficient matrix $Z$ is $\mathbb{R}^{n,m}$. According to the multiplication rule, the following is obtained:

$$d_{ij} = \left( \sum_{k=1}^{n} d_{ik}z_{kj} \right) + e_{ij}$$

(22)

Then, (22) can be transformed into the following:

$$d_{ij} = \frac{d_{i1}z_{1j} + \cdots + d_{(i-1)}z_{(i-1)j} + \cdots + d_{in}z_{nj}}{1 - z_{jj}} + \frac{e_{ij}}{1 - z_{jj}}$$

(23)

The coefficient matrix of the $j^{th}$ observed sensor can be expressed as follows:

$$z_{ij}$$

$$1 - z_{jj}$$

$$1 - z_{jj}$$

$$1 - z_{jj}$$

(24)

The final coefficient matrix $Z_0$ of all observed sensors is described as follows:

$$Z = [Z_{01}, Z_{02}, \ldots, Z_{0n}]$$

$$[\begin{array}{cccc}
  z_{11} & z_{12} & \cdots & z_{1j} & \cdots & z_{1n} \\
  1 - z_{11} & 1 - z_{22} & \cdots & 1 - z_{jj} & \cdots & 1 - z_{nn} \\
  z_{31} & z_{32} & \cdots & z_{3j} & \cdots & z_{3n} \\
  1 - z_{11} & 1 - z_{22} & \cdots & 1 - z_{jj} & \cdots & 1 - z_{nn} \\
  \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
  z_{n1} & z_{n2} & \cdots & z_{nj} & \cdots & z_{nn} \\
  1 - z_{11} & 1 - z_{22} & \cdots & 1 - z_{jj} & \cdots & 1 - z_{nn} \\
  \end{array}]_{(n-1)\times n}$$

(25)

Assuming that $D_3$ represents the matrix $D$ that removes the $p^{th}$ column. According to the matrix multiplication rule, the matrix $D_2 = [d_{11}, d_{12}, \ldots, d_{p-1}, d_{p+1}, \ldots, d_{n}] \in \mathbb{R}^{m\times(n-1)}$ is multiplied by the $p^{th}$ column of the coefficient matrix $Z_0$. The value $\bar{d}_p = [\bar{d}_{1p}, \bar{d}_{2p}, \ldots, \bar{d}_{mp}]^T$ is obtained and can be noted as follows:

$$\bar{d}_p = \sum_{k=1}^{p-1} d_{ik} \cdot \frac{z_{kp}}{1 - z_{pp}} + \sum_{k=p+1}^{n} d_{ik} \cdot \frac{z_{kp}}{1 - z_{pp}}$$

(26)

$$\bar{d}_p = d_{1p}z_{1p} + \cdots + d_{(p-1)}z_{(p-1)p} + d_{(p+1)}z_{(p+1)p} + \cdots + d_{np}z_{np}$$

(27)

The ILRMD method proposed in this paper assumes that a certain column of data in the matrix is lost and then multiplies the matrix by the coefficient matrix $Z_0$ to recover the missing data. The influence of all observed sensors is considered including the sensor with missing data. In (24), if the value $Z_{jj}$ is zero, the data of the surrounding sensors is used for imputation. If the value $Z_{jj}$ is not zero, both the data of the surrounding sensors and the sensor including missing data are used.

The differences between the ILRMD method and the traditional imputation method are discussed as follows. The traditional imputation method performs the zero-padding operation on the missing column and then is directly multiplied by the corresponding column of the coefficient matrix $Z$. The traditional imputation method utilizes the data collected from the surrounding sensors to recover the matrix and ignores the effect of the sensors including missing data. The ILRMD method assumes that the $p^{th}$ column of the data is completely missing and the matrix $D_2$ represents the matrix $D$ after removing the $p^{th}$ column data. Then after the conversion, the weight that is most relevant to each sensor itself is expressed in another form, in order to reduce the effect of the most relevant weight to the imputation result. From (22)-(24), a coefficient matrix $Z_0$ is obtained. The coefficient matrix $Z_0$ considers not only the surrounding sensors, but also the influence of the sensor including missing
data. Ultimately the matrix $D_3$ is multiplied by the coefficient matrix $Z_0$ for obtaining imputation result. The main steps of the proposed imputation method are as follows.

**Step 1.** The traffic flow data is preprocessed by smoothing and filtering, and the complete traffic flow data of one day is randomly selected to construct the training matrix $D$.

**Step 2.** The preprocessed matrix $D$ is decomposed into the low-rank matrix $A$ and the sparse matrix $E$ according to (1).

**Step 3.** According to (9), matrix $A$ is decomposed into $D$ and $Z$, and, from (10) to (20), the coefficient matrix $Z$ is solved.

**Step 4.** Construct test matrix $D$ and set matrix $D_2$ as the dictionary matrix. $D_2$ represents the matrix $D$ that removes the $p^{th}$ column.

**Step 5.** The coefficient matrix $Z_0$ is obtained according to (25) and the missing data which need to be imputed is obtained by (26).

4.2. **Performance Evaluation Criteria.** The evaluation criteria to measure the error of the imputed data included root mean square error (RMSE), mean absolute error (MAE), mean squared percentage error (MSPE), and mean absolute percentage error (MAPE). The RMSE and MAPE are selected in this paper. The formulas are as follows:

$$
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}
$$

$$
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\%
$$

where $N$ is the total number of the missing data, $y_i$ is the actual value of the $i^{th}$ missing data point, and $\hat{y}_i$ is the corresponding estimated value.

5. **Experiment Results**

5.1. **Data Description.** The data used to evaluate the performance of the proposed model was collected in mainstream detectors provided by the PeMS database, which includes more than 39,000 individual sensors that span the highway system in all major metropolitan areas of California. In this paper, 46 mainline sensors numbered from 1108512 to 1221232 are selected to perform data imputation test from April 1st, 2018, to April 30th, 2018. The traffic flow data is aggregated at 5-minute intervals and generate 288 data points for the daily flow. The data of 1 day, 7 days, and 14 days are, respectively, selected to construct the training matrix; however, the experimental results show that the improvement of the imputation accuracy is not obvious when the training samples become larger and larger. Therefore, the traffic flow data on April 23th, 2018, is used as training data, and the data on April 30th, 2018, is used as test data. The data in sensor numbered 1108512 is assumed to be missing, which needs to be imputed. According to the analysis of the spatial-temporal correlation characteristics of traffic flow, the traffic flow data on the same day in different consecutive weeks have high regularity and relevancy. Therefore, this paper selects traffic flow data from the same day on consecutive weeks (two Mondays) to perform the experiment. The traffic flow data of 46 observed sensors on April 23th, 2018, are selected as training matrix, and the data in sensor numbered 1108512 on April 30th, 2018, is assumed to be missing, which needs to be imputed.

Due to the influence of people’s willing for a trip, weather, and other factors, the traffic flow data presents certain stochastic fluctuation and abrupt. In order to reduce the impact of stochastic fluctuation of traffic flow data on imputation results, a five-point smoothing filtering method was used to preprocess the data. The original and filtered data, in the sensor numbered 1108512 on April 8, 2018, are shown in Figure 1.

From Figure 1, it can be seen that the filtered data intuitively reflects the regularity of the traffic data, and the abrupt points are effectively filtered out in the original traffic flow data.

In this paper, the training data and the test data are all preprocessed with a smoothing filtering method at first, which can remove the abnormal points in the sensor data. Then we randomly assume that a sensor data is missing and then impute the missing sensor data with the proposed model.

5.2. **Results and Performances Analysis**

5.2.1. **Influence of Parameter $\lambda$.** The compromise factor $\lambda$ is an important parameter of low rank matrix decomposition, and the different $\lambda$ values have an important impact on the performance of data imputation. In order to verify the effectiveness of ILRMD method, the influence of parameter $\lambda$ is analyzed. The RMSE and MAPE of imputation results changes with the compromise factor $\lambda$ are, respectively, shown in Figures 2(a) and 2(b).
From Figure 2, we can see that, for the traditional MC method, both RMSE and MAPE gradually decrease with the increase of the compromise factor $\lambda$. After RMSE and MAPE reach the minimum value ($\lambda = 0.08$), which increase again. For the ILRMD method, RMSE and MAPE all decrease with the change of $\lambda$. When $\lambda = 0.15$, they reach the minimum and then increase slowly. In any case, the traditional MC method is far less effective than ILRMD method. Therefore, in order to compare the imputation results of the two methods in the best state, $\lambda$ is set as 0.08 for traditional MC method and 0.15 for ILRMD method in this paper.

### 5.2.2. The Selection of the Training Data.

Due to traffic flow has high spatial-temporal correlation characteristics, it is necessary to analyze the effect of different training data to imputation results. However, the selection of training data has little influence on the performance of the proposed ILRMD method. In order to show that the performance of the proposed method is not sensitive to the time, the traffic flow data of four days (April 21th, 2018, April 22th, 2018, April 23th, 2018, and April 24th, 2018) are randomly selected as training data to impute the data of April 30th, 2018. The experimental results are shown in Figures 3(a), 3(b), 3(c), and 3(d).

It can be seen from Figure 3 that the proposed ILRMD method always has good performance and is not sensitive to the selection of training data. And the imputation performance of different training data is shown in Table 1.

<table>
<thead>
<tr>
<th>Training data</th>
<th>MAPE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 21th, 2018</td>
<td>0.0294</td>
<td>0.0454</td>
</tr>
<tr>
<td>April 22th, 2018</td>
<td>0.0364</td>
<td>0.0588</td>
</tr>
<tr>
<td>April 23th, 2018</td>
<td>0.0260</td>
<td>0.0453</td>
</tr>
<tr>
<td>April 24th, 2018</td>
<td>0.0207</td>
<td>0.0409</td>
</tr>
</tbody>
</table>

It can be seen from Table 1 that the proposed method always has good performance although the different training data is used. The results indicate that the selection of time has little influence on the proposed ILRMD method. Therefore, we only select the traffic flow data of one day (April 23th, 2018) to verify the proposed model in the paper.

### 5.2.3. Comparison of Imputation Results.

For the purpose of verifying the performances of ILRMD method, the proposed method is compared with the traditional method. The imputation results of the ILRMD method under the best condition ($\lambda = 0.15$) and the traditional method under the best condition ($\lambda = 0.08$) are shown in Figures 4(a) and 4(b).

From Figure 4, it can be seen that the imputation results of traffic flow data through the ILRMD are more accurate than the traditional MC method. Although the imputation result is obtained in the optimal compromise factor $\lambda$ with the traditional MC method, there is a big deviation between the imputation result and the real data, and the ILRMD method still recovers the missing traffic data more accurately. When compromise factor $\lambda$ is set as the optimal value for the ILRMD method, the imputation result is almost identical with the real value, but there are more deviations in traditional methods. It is observed that the imputation results of the proposed ILRMD method have similar traffic patterns with the real traffic flow, especially in morning and evening peak hours.

### 5.2.4. The Comparison of ILRMD and Other Imputation Methods.

In order to evaluate the advantages of our proposed approach, the ARIMA, SVR, DBN-SVR, WNN, KNN, and Traditional MC imputation methods are selected under the premise of testing with the same experimental data. In the ARIMA model, the orders of autoregressive $p$, moving average $q$, and difference $d$ are, respectively, set as 5, 5, and 1. In the SVR model, the nuclear function is configured as “rbf”, the number of iterations is 10,000, and the penalty factor is taken as 0.01. In the WNN model, the number of iterations is 1000, the number of the hidden layer nodes is 3. In the DBN-SVR model, the number of network layers in the DBN model is set as 3 and the number of iterations is 200. The ILRMD model proposed in this paper is compared with
Figure 3: The imputation results of traditional MC and ILRMD methods.

Figure 4: The imputation results of two imputation methods (\(\lambda = 0.08\))

Figure 4: The imputation results of two imputation methods (\(\lambda = 0.15\))
Table 2: The performance comparison of data imputation models.

<table>
<thead>
<tr>
<th>Missing sensor ID</th>
<th>1108512</th>
<th>1119921</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>MAPE</td>
<td>RMSE</td>
</tr>
<tr>
<td>KNN</td>
<td>0.1338</td>
<td>0.5022</td>
</tr>
<tr>
<td>SVR</td>
<td>0.1024</td>
<td>0.5396</td>
</tr>
<tr>
<td>WNN</td>
<td>0.1442</td>
<td>0.6660</td>
</tr>
<tr>
<td>DBN-SVR</td>
<td>0.1129</td>
<td>0.6402</td>
</tr>
<tr>
<td>ARIMA</td>
<td>0.2078</td>
<td>0.6065</td>
</tr>
<tr>
<td>Traditional MC</td>
<td>0.3717</td>
<td>0.1830</td>
</tr>
<tr>
<td>ILRMD</td>
<td>0.0260</td>
<td>0.0453</td>
</tr>
</tbody>
</table>

Figure 5: The imputation results of different models.

The imputation results of different models and real traffic flow are shown within one day in Figure 5.

It can be seen from Figure 5, the imputation traffic flow has similar traffic patterns with the real traffic flow. The DBN-SVR model has the worst imputation performance; the ARIMA, SVR, KNN, and WNN are better than the DBN-SVR, while they show weakness compared with the ILRMD method. The imputation value of the proposed ILRMD model is almost coincided with the measured data. It is observed that the proposed ILRMD model has better imputation performance.

The error analysis test is conducted using two error evaluation criteria, which is expressed in Table 2. In order to more precisely verify the performance of the proposed model, another sensor numbered 1119921 is randomly selected to perform the test. In Table 2, the sensors numbered 1108512 and 1119921 are, respectively, assumed to be imputed to verify the performance of the proposed model. It can be seen from Table 2, when the sensors numbered 1108512 and 1119921 are assumed to be imputed, the proposed ILRMD models all have the best performance compared with other approaches. These experiments can verify that the ILRMD model proposed in this paper is an effective method for data imputation.

From Table 2 of the first condition (1108512 sensor), it can be seen that the imputation accuracy of the ILRMD model, respectively, improves 93.01%, 74.61%, 95.96%, 80.57%, 96.30%, and 81.97% compared with the traditional MC, SVR, ARIMA, KNN, DBN-SVR, and WNN methods. The average imputation accuracy is 87.07% higher than other imputation methods. Results demonstrate that the proposed ILRMD model has the best performance compared with other approaches, and it is an effective method for data imputation.

6. Conclusions and Recommendations

In the paper, a data imputation method is proposed to impute the missing traffic flow data. Different from the most known traffic flow data imputation methods, the ILRMD model makes an effective use of the information of missing sensors and takes full advantage of the high spatiotemporal correlation characteristics of traffic flow data. The experiment result shows that the proposed imputation method is superior to other methods. However, this paper focuses on dealing with the missing traffic data at a single sensor; we only considered one observed sensor with missing data. In practical terms, the missing traffic data is always distributed on multisensors.

In our future research, the missing data analysis on multisensors is being studied. The concept of missing rate can be introduced, and the more effective data imputation
method can be proposed for different degrees of missing data in order to improve the imputation accuracy.

Data Availability

The data used in this paper are collected from the Caltrans Performance Measurement System (PeMS) in 46 sensors numbered from 1108512 to 1221232 on 04/01/2018–04/27/2018. If any researcher requests for these data, he can log into the website: http://pems.dot.ca.gov/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was partly supported by the National Key R&D Program of China (2018YFC0808706) and the National Natural Science Foundation of China (Grant no. 5157081053). The authors are also grateful to the PeMS for providing the data.

References


