Estimating Missing Traffic Volume Using Low Multilinear Rank Tensor Completion

Bin Ran, Huachun Tan, Jianshuai Feng, Wuhong Wang, Yang Cheng & Peter Jin


To link to this article: http://dx.doi.org/10.1080/15472450.2015.1015721

Accepted author version posted online: 03 Apr 2015.
Published online: 03 Apr 2015.

Submit your article to this journal

Article views: 65

View related articles

View Crossmark data

Citing articles: 1 View citing articles
Estimating Missing Traffic Volume Using Low Multilinear Rank Tensor Completion

BIN RAN,1 HUACHUN TAN,2 JIANSHUAI FENG,2 WUHONG WANG,2 YANG CHENG,1 and PETER JIN3

1Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, Wisconsin, USA
2Department of Transportation Engineering, Beijing Institute of Technology, Beijing, P.R. China
3Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA

Traffic volume data have been collected and used for various purposes in some aspects of intelligent transportation systems (ITS) applications. However, the unavoidable detector malfunction can cause data to be missing. It is often necessary to develop an effective approach to recover the missing data. In most previous methods, temporal correlation is explored to reconstruct missing traffic volume. In this article, a new missing traffic volume estimation approach based on tensor completion is proposed by exploring traffic spatial–temporal information. The tensor model is utilized to represent traffic volume, which allows for exploring the multicorrelation of traffic volume in spatial and temporal information simultaneously. In order to estimate the missing traffic volume represented by the tensor model, a novel tensor completion algorithm, called low multilinear rank tensor completion, is proposed to reconstruct the missing entries. The proposed approach is evaluated on the PeMS database. Experimental results demonstrate that the proposed method is more effective than the state-of-art methods, especially when the ratio of missing data is high.

Keywords Missing Data; Spatial–Temporal Correlation; Tensor Completion; Tensor Model; Traffic Volume Data

INTRODUCTION

In recent decades, a large amount of traffic volume data have been collected to fulfill operational and management needs (Ran, Jin, Boyce, Qiu, & Cheng, 2012), such as real-time operations, traffic network planning, congestion management, etc. The real-time traffic state data can also be used in many ITS applications: intelligent vehicle, route planning, driver assistance system (Höltl & Trommer, 2013; Wang, Zhang, Zhang, & Li, 2013) and connected vehicle (Wang, Li, Zheng, & Lu, 2015). However, regardless of which collection device is used, the problem of missing data is common in traffic monitoring programs due to detector failures and communication and processing errors (American Association of State Highway and Transportation Officials [AASHTO], 2009; Zhang & Liu, 2009).

The presence of missing data would decrease the efficiency of traffic monitoring programs and other intelligent transportation systems (ITS) applications, such as the Vehicle Information and Communication System (VICS), for which most analysis functions work on complete data. For certain purposes such as traffic forecast and traffic information issue, traffic counts with missing values may be the only data available, which may affect their performance. Hence, it is necessary to recover the missing traffic data for subsequent analysis. In accordance with the principle of Truth-in-Data, AASHTO (2009) also recommends that highway agencies document the procedures for editing traffic data.

Many imputation methods have been proposed for solving the missing data problem. These methods can generally be classified as traditional methods and statistically principled methods. The traditional methods involve the techniques such as historical (neighboring) imputation methods (Chen & Shao, 2000)
and spline (including linear)/regression imputation methods (Allison, 2001). These methods model the traffic data as vector patterns, which can cover little spatial–temporal information. Then spatial/temporal correlations are used to impute missing traffic data when few data are missing. Recently, the development of imputation techniques is moving on a statistically principled track. The Bayesian principal component analysis (BPCA) algorithm (Qu, Zhang, Hu, Jia, & Li, 2008) and probabilistic principal component analysis (PPCA) (Qu, Li, Zhang, & Hu, 2009) model the traffic data as a matrix pattern, which can cover more spatial–temporal information than a vector pattern, and can impute the missing traffic data from the whole matrix based on spatial–temporal correlations. These methods have been proved to be more effective with higher accuracy than the traditional imputation methods for missing traffic volume data.

The spatial and temporal correlations of traffic volume data are critical for imputing the missing traffic volume data (Qu et al., 2008, 2009). As mentioned earlier, traditional methods mostly only exploit part of correlations, such as historical or temporal neighboring correlations. And the statistically principled methods usually utilize the temporal correlations of traffic data from day to day. In fact, there are many correlations in the traffic volume. For example, the traffic data temporal correlations contain the relations from day to day, hour to hour, and so on. In addition, the spatial correlations exist in the adjacent detectors data. However, these correlations are not explored fully or simultaneously in previous imputing methods. From the view of traffic physics, the traditional methods and statistically principled methods mainly choose one or several detectors’ data for constructing the model for recovering the missing traffic data. However, they only exploit a part of the useful information for recovering the missing data.

To tackle the shortcoming just mentioned, a new approach is proposed to improve the performance of missing traffic volume imputation. A tensor (multiway matrix) pattern is used to model the traffic volume data, which can preserve the spatial–temporal structure of the traffic volume, and its multicorrelations (spatial and different temporal correlations) can be utilized simultaneously through the multiple modes of the tensor. Then a new missing data estimation method, called low multilinear rank tensor completion (LMRTC), is developed by extracting the underlying multi-mode correlations among the tensor pattern. Experimental results on PeMS (California Performance Measurement System) database (PeMS, 2011) show that the proposed approach can explore the spatial and temporal correlations of traffic volume data and improve the imputing performance.

The rest of this article is organized as follows. The second section presents the related work. The third section provides the notation used in this article. In the fourth section, the multi-correlations of traffic volume are analyzed by spatial–temporal properties of traffic volume, followed by introducing a tensor pattern to model traffic data. Details of the proposed tensor completion, LMRTC, to solve the missing traffic volume issue are described in the fifth section. Experiments and discussion are reported in the sixth section, and we conclude this article in the seventh section.

**RELATED WORK**

In this section, the previous traffic data imputation approaches are briefly reviewed. Then a simple introduction of tensor completion theory for missing data estimation is presented.

Traditional methods and statistically principled methods are the most popular imputation methods. The traditional methods contain historical (neighboring) imputation methods (Zhang, Zhang, Li, & Hu, 2004) and spline (including linear)/regression imputation methods (Chen & Shao, 2000).

The historical imputation method aims to fill a missing data point with a known data point collected on the same site at the same daily time but from a neighboring day (Allison, 2001). A variation of this algorithm fills the missing data with the average values taken over the most recent days. The spline/regression imputation method recovers the missing values by applying mathematical interpolation algorithm according to the surrounding known data points collected during the same day (Qu et al., 2009). Pei and Ma (2003) first calculate the linear correlation coefficient between the real-time speed data and their corresponding temporal data, and then the missing data are constructed by linear regression method. The imputing performances of these methods greatly depend on the surrounding data of missing points. That is, these methods utilize only a part of the correlations of traffic volume data. Thus, their performances are not satisfactory, especially when the ratio of missing data is high.

Different from the traditional methods, statistically principled methods depend on the statistical process and definition. BPCA and PPCA are classic methods for addressing the problem of missing traffic volume data (Qu et al., 2009). BPCA is a slight modification of PPCA. Indeed, both PPCA and BPCA are based on expectation maximization (EM) imputation methods (Dempster, Laird, & Rubin, 1977), and they make use of the correlations between observed data and latent variables for imputing the missing data. Generally, the correlations between observed data and latent variables are described as probabilistic model. In order to obtain the maximum probability of these parameters, the Bayesian model is introduced to estimate the missing values with respect to the estimated posterior distribution. The missing traffic volume data are gradually recovered along with the building of latent model. It feels like both of the methods balance the periodicity, spatial correlations, and other statistical correlations of traffic volume data. The statistically principled methods are based on correlation of traffic volume data; therefore, they outperform the traditional methods. However, there are other types of correlation missing from these models, such as the correlations between sample intervals for a single detector data, spatial correlations for multiple detectors data, and so on.
The methods just mentioned can only exploit partial correlations of traffic volume data due to the characteristic of vector or matrix pattern. In recent years, tensor completion methods have been proposed for imputing the missing data for multiway data and can exploit multicorrelation simultaneously. Liu et al. (2009, 2013) first proposed a tensor completion method based on trace norm minimization and applied it to image completion. A more general low-n-rank tensor completion method is proposed and evaluated on a magnetic resonance imaging (MRI) database (Gandy, Recht, & Yamada, 2011). Also, a first-order method is recently developed called CP-WOPT based on canonical polyadic (CP) decomposition of a tensor model and applied to imputing missing data (Acar, Dunlavy, Kolda, & Mørup, 2011). Traffic volume data have intrinsic multiway spatial–temporal correlations. For fully exploiting the spatial–temporal correlations and improving the performance of imputation methods, a multiway tensor model is utilized to construct the traffic volume data and a novel tensor completion method is proposed for imputing the missing data in this article.

Recently, Asif, Mitrovic, Garg, Dauwels, and Jaillet (2013) used low-dimensional models (including matrix and tensor model) to reconstruct data profiles for road segments, and impute missing values. However, they only focus on utilizing different methods for recovering the missing data. In this article, we make the following two contributions: (a) The multimode correlations such as day-to-day, hour-to-hour, and detector-to-detector are exploited and utilized in our proposed method, and (b) we propose a tensor-based method that is a multidimensional extension of the fixed point continuation (FPC) iterative algorithm (Ma, Goldfarb, & Chena, 2011) which is used as a common matrix-based method by Asif et al. (2013).

NOTATION

We denote the scalars in $\mathbb{R}$ with lowercase letters ($a$, $b$, $\ldots$) and the vectors with bold lowercase letters ($\mathbf{a}$, $\mathbf{b}$, $\ldots$). The matrices are written as uppercase italic letters, for example, $\mathbf{X}$, and the symbol for tensors are roman letters, for example, $\mathcal{X}$. The subscripts represent the following scalars: $(\mathcal{X})_{ijk} = x_{ijk}$, $(\mathcal{X})_{ij} = x_{ij}$. The superscripts indicate the size of the matrices or tensors. For example, there are traffic volume data that are recorded every 5 minutes for 16 days. Then, the data of 1 day preserves 288 data points (12 hours per day and 24 data points per hour). Therefore, the traffic data of 16 days can be constructed as a matrix model of size 16 $\times$ 288 or a tensor model of size 12 $\times$ 24 $\times$ 16. The Frobenius norm of matrix $\mathbf{X}$ is defined as $||\mathbf{X}||_F := (\sum_{i,j}|x_{ij}|^2)^{1/2}$. Let $\Omega$ be an index set; then $X_{\Omega}$ denotes the vector consisting of elements in the set $\Omega$ only. Define $X_{\Omega} := (\sum_{i,j\in \Omega} x_{ij}^2)^{1/2}$.

An n-way tensor can be rearranged as a matrix; this is called matricization, also known as unfolding or flattening a tensor. The "unfold" operation along the nth mode on a tensor $\mathcal{X}$ of size $I_1 \times I_2 \times \ldots \times I_N$ is defined as unfold($\mathcal{X}$, $n$) = $X_{(n)}$. The opposite operation, "fold," is defined as fold($X_{(n)}$) = $\mathcal{X}$. For example, the preceding tensor model $\mathcal{X}$ for traffic volume data that has size 12 $\times$ 24 $\times$ 16 can be unfolded along the third mode, to get a matrix $X_{(3)}$ of size 16 $\times$ 288. In addition, the mode-n rank of $\mathbf{X}$ is denoted as $\text{rank}_n(\mathcal{X})$, which is equal to the column rank of $X_{(n)}$.

SPATIAL AND TEMPORAL CORRELATION ANALYSIS

Traffic volume data are recorded as time series by fixed interval (such 5 minutes), and assembled with different detectors along the road. As commonly known and studied by traffic engineering researchers, spatial and temporal correlations are the most important factors impacting on performance for imputing missing traffic volume data (Qu et al., 2008; Zhang et al., 2004).

To illustrate the correlations intrinsic in traffic volume data, the data downloaded from PeMS open-access traffic flow database (PeMS, 2011) are analyzed. The traffic data of the PeMS system are collected in real time from more than 25,000 detectors that span the freeway system across all major metropolitan areas of the State of California. Specially, for our study, the particular data set used in this article was collected from the adjacent stations located at south bound freeway SR99, District 10, Stanislaus County, California (Figure 1). The freeway has three lanes under surveillance. The IDs of these detectors are 1017510, 1017610, 1017710, 1017810, 1017910, 1018110, 1018210, 1018310, 1018410, 1018510, and 1018610, respectively. The total link length is about 2.37 km. The average distance between these locations is about 300 m. There are no signalized intersections between these detectors. There exist one exit and one entrance between location 6 and location 7. The raw data, denoted as DA0, were obtained from March 14 to April 10, 2011, 28 days in total, and the selected field data are the total number of vehicles passing each loop detector during every 5 minutes.

In order to illustrate the temporal correlation of traffic volume data, the daily data during a week for a fixed detector are chosen randomly, denoted as DA1, and are plotted in Figure 2. Obviously, there are two peaks (morning and evening rush hours) and two valleys (a shallow one appears at high noon and a deep one at midnight), and the curves of traffic volume from Monday to Friday are very similar (correlated), but the curves of weekends are a little farther away than the curves of weekdays. Furthermore, quantitative correlation analysis of traffic volume data is conducted. Formally, Pearson’s correlation coefficient is applied to measure the data correlation, given as:

$$R = \frac{\text{cov}(X, Y)}{\sqrt{\text{cov}(X, X) \cdot \text{cov}(Y, Y)}}$$

where cov stands for the covariance of the two variables. Especially, the Pearson’s correlation for the data in a week is averaged as 0.92, and the value arises to 0.95 when we don’t consider the data in weekend.

Particularly, the traffic volume data for one day are selected randomly from DA, and modeled as a 12 $\times$ 24 matrix, which indicates that there are 12 data records in an hour and 24 hours for
1 day, as shown in Figure 3. The interval-to-interval correlation of traffic volume data during 24 hours is also obvious, where interval $i$ means the $i$th 5 minutes of traffic volume in an hour. The average Pearson’s correlation coefficient between interval and interval is calculated as 0.96.

In addition, for traffic volume data, the correlations include not only temporal correlations but also spatial correlations. The traffic data obtained from Loop 1 to Loop 7 detectors on March 14 are shown in Figure 4. Obviously, the spatial correlation is distinct, and the average Pearson’s correlation coefficient between two adjacent detectors is about 0.98.

The preceding analysis shows that the single-detector data have multiple correlations, which include not only day-to-day but also interval-to-interval correlation. Spatial correlation also is found in the data of multiple detectors, in addition to temporal correlation. There are many types of correlation that exist in traffic volume data sets.

---

**Figure 1** Eleven loop detectors in Wilson Way OC, California.

**Figure 2** The daily profile of the traffic volume from Monday to Sunday.

**Figure 3** The profile of the traffic volume in specified interval along hours.
LOW MULTILINEAR RANK TENSOR COMPLETION FOR MISSING TRAFFIC VOLUME ESTIMATION

Since traffic volume data are highly correlated, as mentioned earlier, they can be constructed as an approximate low rank tensor. Then, tensor completion can be applied for estimating missing traffic volume. According to tensor completion theory (Liu et al., 2009), the imputing target function is as follows:

\[
\min \text{rank}(X) \quad \text{s.t.} \ X_{(\Omega)} = T_{(\Omega)},
\]

(1)

where \(X\), \(T\) are the \(n\)-way tensors model, constructed based on the correlations of traffic volume data, and the entries (traffic volume) of \(T\) in set \(\Omega\) are given while the remaining entries are missing. Because there is no straightforward algorithm to determine the rank of a specific tensor, we unfold the tensor along each mode and make the mode-\(n\) rank of the tensor be low. Therefore, for a given \(n\)-way tensor \(T \in I_1 \times I_2 \times \cdots \times I_N\), the low mode-\(n\) rank tensor completion can be transformed as:

\[
\min \text{rank}(X_{(n)}) \text{ for } n = 1, 2, 3, \ldots, N \text{, respectively} \quad \text{s.t.} \ X_{(n)(\Omega)} = T_{(n)(\Omega)},
\]

(2)

where \(T_{(n)(\Omega)}\) is a matrix by matricizing \(T\) in the sample space \(\Omega\), and \(X_{(n)(\Omega)}\) is a matrix by matricizing the decision variable \(X\) in subset \(\Omega\). Similar to the definition of the trace norm of matrix, the trace norm of tensor can be given as:

\[
X_{\text{tr}} = X_{(n),r} = \sum_{\beta=1}^{I_n} \sigma^{(n)}_{\beta},
\]

(3)

where \(X \in I_1 \times I_2 \times \cdots \times I_N\), \(X_{(n)}\) is a matrix by mode-\(n\) matricization of the multiway traffic volume data. There are \(N\) trace norms of mode-\(n\) matricization for an \(n\)-way tensor. This is consistent with the fact that there are \(N\) mode-\(n\) ranks and \(N\) sets of mode-\(n\) singular values.

Thus, using the trace norms of the mode-\(n\) matricization for a tensor as an approximation to rank \((X_{(n)})\) in Eq. (2), we can get:

\[
\min X_{(n)} \text{ for } n = 1, 2, 3, \ldots, N, \text{ respectively} \quad \text{s.t.} \ X_{(n)(\Omega)} = T_{(n)(\Omega)},
\]

(4)

Generally the elements of the tensor \(T\) (i.e., traffic volume) may be contaminated by noise, and thus the constraint \(X_{(n)(\Omega)} = T_{(n)(\Omega)}\) must be relaxed, which then results in either the problem:

\[
\min X_{(n)} \text{ for } n = 1, 2, 3, \ldots, N, \text{ respectively} \quad \text{s.t.} \ X_{(n)} - T_{(n)(\Omega)}^2 \leq \theta,
\]

(5)

or its Lagrangian version:

\[
\min \mu X_{(n)} + \frac{1}{2} X_{(n)} - T_{(n)(\Omega)}^2 \text{ for } n = 1, 2, 3, \ldots, N, \text{ respectively},
\]

(6)

where \(\theta\) and \(\mu\) are parameters.
There are many available efficient algorithms to solve the preceding minimization problem in Eq. (6); the fixed-point continuation (FPC) iterative algorithm (Ma et al., 2009) is used in our implementation. In order to improve computational speed, we used the approximate SVD-based FPC algorithm (Ma et al., 2009). The pseudo-code of LMRTC algorithm is given in Algorithm 1:

Algorithm 1, LMRTC: Low Multilinear Rank Tensor Completion

1. Input: T
2. for n = 1 to N
   1. unfold T<sub>n</sub> along each mode, get T<sub>ikl</sub>
   2. Initialize: Given X<sub>t0</sub>, μ > 0. Select μ<sub>1</sub> > μ<sub>2</sub> > ··· > μ<sub>L</sub> = μ > 0.
   3. for μ = μ<sub>1</sub>, μ<sub>2</sub>, ···, μ<sub>L</sub>, do
      - while not converged, do
         —compute Y<sub>n</sub> = X<sub>n</sub> − T<sub>n</sub>(X<sub>n</sub> − T<sub>n</sub>), and
         SVD of Y<sub>n</sub>, Y<sub>n</sub> = U<sub>n</sub>Diag(σ) VT
         —compute X<sub>n</sub> = U<sub>n</sub>Diag(σ<sub>n</sub>) VT
      —end while
   4. X<sub>n</sub> = X<sub>n</sub>.
   5. fold X<sub>n</sub>.
3. Output: X<sub>0</sub> = \sum_{n=1}^{N} \text{weight} \times \text{fold}X_0(n),
   μ_{k+1} = \max(μ_kη_k, μ), k = 1, \ldots, L − 1

When the LMRTC procedure is finished, the values represented by the tensor model X<sub>0</sub> are regarded as the estimation of the missing traffic volumes.

**EXPERIMENTS**

To evaluate the flexibility and effectiveness of our LMRTC method for imputing missing traffic volume data, we have evaluated our framework in the application of highway traffic volume data from the PeMS open-access traffic flow database (PeMS, 2011). While the PeMS system often has missing data, it is difficult to compare our approach with other methods if we do not have ground-truth data. Thus, we chose a set of PeMS detector data that happened not have any missing data, as same as the data denoted in the fourth section. Then the missing data are simulated (as described in the following), and compared to the actual ground-truth data.

The data set DA0 with 11 detectors in the fourth section is used to evaluate the advantage of proposed LMRTC for multiple-mode spatial–temporal correlation data, and the only temporal correlation data set (denoted as DA2) is chosen as detector data from DA0. Three classical imputation methods, (a) mean-historical imputation (Kahaner, Moler, & Nash, 1989), (b) spline imputation (Boor, 1978), and (c) the BPCA-based imputation method (Qu et al., 2008), are selected as the benchmarks to compare with the proposed LMRTC algorithm.

For the mean-historical imputation method, the mean value of all the available data points belonging to the same detector at the same time interval in the last few days is calculated as the imputed value (Kahaner et al., 1989). For BPCA, the maximum number of iteration steps is set to 200 and the threshold of the approximate complexity is set to 10<sup>-6</sup> (Qu et al., 2008).

Our proposed method is an iterative algorithm that considers the mode-n ranks of a given tensor model. For LMRTC, the approximate SVD-based FPC algorithm is used to complete the matrices by matricization of the traffic tensor model along each mode. The parameters are set as follows: weight parameter as \( \bar{\mu} = 10^{-4} \), the relative error tolerance as xtol = 10<sup>-4</sup>, maximum number of iterations allowed for solving each subproblem in FPCA as \( I_m = 10 \), step length as \( \tau = 2 \), the stopping criterion for inner iterations as gtol = 10<sup>-4</sup>, initial estimator as \( X_0 = 0 \), and \( \eta_\alpha = 1/4 \). All the following results are averaged by 30 instances.

**Evaluation Indices**

To evaluate the performances of the proposed method, the following two indices are used in this article.

1. Mean absolute percentage error (MAPE): The index gives the evaluation of the average estimation error in terms of percentage:

   \[
   \text{MAPE} = \frac{1}{M} \sum_{m=1}^{M} \frac{|t_r^{(m)} - t_e^{(m)}|}{t_r^{(m)} \times 100}.
   \]

2. Root mean square error (RMSE): This index gives the evaluation of the variance in the estimation errors:

   \[
   \text{RMSE} = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (t_r^{(m)} - t_e^{(m)})^2},
   \]

where \( t_r^{(m)} \) and \( t_e^{(m)} \) are the nth elements and stand for the known real value and estimated value, respectively. \( M \) denotes the number of missing traffic points.

**Results on Temporal Correlation Data**

As analyzed in the fourth section, the traffic volume data for a detector have multimode temporal correlations, such as “day” mode, “hour” mode, and “interval” mode. Based on multiple correlations analysis of the traffic volume data, the chosen data set DA2 is formulated as a tensor model of size 12 × 24 × 28, which stands for 12 sample intervals (i.e., recorded by 5 minutes) per hour, 24 hours in a day, and 28 days. For the classical methods that are based on the matrix model, the data set is reshaped as a matrix of size 28 × 288 (which stands for 28 days and 288 records in a day) because the correlation of “day” mode is most remarkable and used generally for classical methods. The ratios of missing data are set from 5% to 80% and the
missing data are produced randomly because the case is most representative in real traffic.

RMSE curves and MAPE curves of those methods with randomly missing traffic volume data for temporal correlation data DA2 are shown in Figures 6 and 7. Obviously, RMSE and MAPE of proposed LMRTC are smaller than other approaches. RMSE of the mean historical method is higher than for other methods and flattening along the missing ratio. RMSE of BPCA algorithm increases sharply when the missing ratio exceeds 60%. LMRTC performs best even when the missing ratio reaches up to 80%. In addition, Figure 7 shows that the MAPE values of LMRTC are also smaller than other algorithms, which indicates that the estimated traffic volumes are more close to the actual traffic volumes on average. This is due to fewer temporal correlations being utilized in the previous method. For the mean historical method and spline-based method, traffic volume is modeled as a vector, which can only cover interval–interval correlation in a vector pattern, and only the one kind of correlation is employed for missing data estimation. BPCA models the traffic volume as a matrix, which can cover both day-to-day and interval-to-interval correlations. By exploring day and interval correlation, BPCA performs better than the mean historical method and spline-based method. For our proposed LMRTC, the tensor is used to model the traffic volume along day, hour, and interval modes, which can encapsulate more kinds of correlations simultaneously. With more temporal information taken into account simultaneously, LMRTC outperforms other methods.

For further illustrating the outperformance of LMRTC, examples of the estimated traffic volume data for LMRTC and BPCA where the missing ratio is 80% are shown in Figure 8. For simplicity, Figure 8 only shows the traffic volume data in a day. Figure 8a is the initial traffic volume data where 80% of the data are missing. Figures 8b and 8c demonstrate the imputed traffic volume and the ground-truth data. It can be found that LMRTC is better than BPCA. In fact, RMSE and MAPE are 28.12 and 12.31% for LMRTC, and 41.63 and 17.28% for BPCA. There are some abnormal values for traffic volume data recovered by BPCA, and the performance of BPCA gets worse sharply. This is because BPCA only utilizes two kinds of correlations by matrix pattern, and less information is employed when the missing ratio is high, thus resulting in
unsatisfactory recovery performances. However, for LMRTC, many more types of correlation are encoded in the tensor model, and much more temporal information can be explored than for BPCA even when the missing ratio is high.

In addition, the correlations of traffic volume data that are analyzed in the fourth section show that the characteristics of traffic volume data between weekend and weekday are different. For improving the performance of proposed LMRTC by using much more correlated traffic volume data, we choose only the weekdays’ data from the preceding data set and formulate this as a tensor model of size $12 \times 24 \times 20$, which stands for 12 intervals per hour, 24 hours in a day, and 20 weekdays. Using the proposed LMRTC approach, we get the following results for data missing ratios from 0.1 to 0.8, as shown in Table 1. From Table 1, RMSEs and MAPEs of the new tensor are all smaller than that of the original tensor of $12 \times 24 \times 28$. The results demonstrate that the performance of the LMRTC could be improved if the correlations of traffic volume data are strengthened.

### Results of Spatial-Temporal Correlation Data

The temporal–spatial correlations of traffic volume data are intrinsic characteristics in traffic theory. As analyzed in the fourth section, the data set DA0, collected from 11 detectors, has strong spatial–temporal correlations. According to the correlation analysis, the traffic volume data in DA0 can be constructed as a tensor model of size $12 \times 24 \times 28 \times 11$, which stands for 12 sample intervals per hour, 24 hours in a day, 28 days, and 11 loop detectors, respectively. In the newly constructed tensor, spatial and temporal information can be encoded simultaneously, and more spatial–temporal correlations can be employed by the proposed tensor completion algorithm to estimate missing traffic volume.

To evaluate the benefits of spatial–temporal correlations, the proposed method on the new constructed spatial–temporal tensor (the tensor model of size $12 \times 24 \times 28 \times 11$), which is denoted as LMRTC-ST, is tested. In principle, the performances of LMRTC-ST utilizing spatial–temporal information would be better than for those methods only using temporal correlation data due to more explored correlations. For validating this point, we also compare LMRTS-ST with LMRTC on the tensor only using temporal correlations ($12 \times 24 \times 28$ tensor model), which is called LMRTC-T, and with BPCA on a temporal matrix, termed BPCA-T, on 11 detectors separately. Experimental results show that LMRTC-TC outperforms both LMRTC-T and BPCA-T for all 11 detectors. The RMSE and MAPE of 11 detectors are averaged and listed in Table 2 for simplicity instead of all of the experimental results. From Table 2, it can be found that LMRTC produces the best results for all missing ratios, which can be used to prove that using more spatial–temporal information could improve the performance of imputing missing traffic volume. However, the recovery accuracy of LMRTC-TS degrades sharply when the data missing ratio is higher than 70%. The reason for this phenomenon may be that many rows of the decomposition matrix for the spatial mode are totally missed, and the decomposition matrix is flat when the missing ratio is very high, so the number of rows is farther greater than the number of columns, which can cause performance degrading sharply (Navasca & De Lathauwer, 2009).

### Results of Workdays/Weekends Tensor Pattern

Many studies figured out that the traffic flow patterns are significantly different on workdays and weekends due to the mobility pattern of people. Thus, many traffic modeling methods tend to separate the traffic data into two types before modeling.
To evaluate whether we can improve imputation performance of tensor-based methods by separating the workdays and weekends, the workday-tensor pattern \((12 \times 24 \times 20 \times 11)\) only containing data from workdays and the weekend-tensor pattern \((12 \times 24 \times 8 \times 11)\) are tested and compared to the performance of integrated tensor pattern. The results are given in Table 3.

Experimental results show that the separation of data from weekend/workdays can slightly promote the imputation accuracy. From Table 3, it can be found that the workday/weekend tensor pattern outperforms the integrated tensor on workdays/weekend under almost all missing ratios, which suggested that the separation of weekend and workdays while modeling the tensor pattern could improve the performance of imputing missing traffic volume. The reason may be that the integration of traffic data from both weekend and workday would decrease the temporal correlation of traffic volume. The results also indicate that the tensor modeling strategy plays an important role in tensor-based traffic data imputation methods.

### CONCLUSION

This article proposes a new traffic missing volume data imputation method. Based on traffic spatial–temporal correlation analysis, we introduce a tensor model to construct the traffic volume in a multiway matrix style, which allows for preserving the multiple correlations of traffic volume and exploring the intrinsic spatial–temporal information simultaneously along the tensor’s multimodes. Then a novel missing data estimation algorithm based on tensor completion, named low multilinear rank tensor completion, is presented to estimate the missing traffic entries by employing the multicorrelations in the tensor model. Two experiments are conducted on the PeMS database. The first experiment is to test the performances of only using temporal information; that is, the missing traffic volumes are imputed on a single detector. By segmenting traffic volume time series into a tensor, the multiple temporal correlations such as “day-to-day,” “hour-to-hour,” and “interval-to-interval” can be represented simultaneously. The proposed LMRTC outperforms previous methods in different missing ratios between 0.05 and 0.80, especially when the missing ratio is high. The second experiment tests the imputing methods on traffic volume data of 11 detectors. The results demonstrate that the proposed method using spatial–temporal information works better than those imputing methods only using temporal information. This also indicates that spatial correlation would benefit for imputing missing traffic volume. We also evaluate our methods on a weekend/workdays tensor pattern. Results show that the separation of weekend and workdays can promote the imputation accuracy. It suggests that tensor modeling strategy is important for tensor-based traffic data imputation methods.

In future work, we will study imputing other types of missing traffic data, such as speed and occupancy. The missing data issue for dynamic road network is interesting and still a challenge. Also, the proposed methods need improvement for real-time applications. Future work should also include some in-depth studies about the tensor modeling strategies for traffic data, such as day/night tensor pattern and peak hour tensor pattern.

### FUNDING

This research was supported by the NSFC (61271376, 51308115), the National Basic Research Program of China (973 Program: number 2012CB725405), the Beijing Natural Science Foundation (4122067), and the Basic Research Fund of the Beijing Institute of Technology (20120342026). We thank Guangdong Feng of Beijing Institute of Technology for the suggestive discussions.

### REFERENCES


