Physica A 446 (2016) 54–63

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Short communication

Tensor based missing traffic data completion with spatial–temporal correlation

Bin Ran [a], Huachun Tan [b,*], Yuankai Wu [b], Peter J. Jin [c]

[a] Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, 53706, USA
[b] Department of Transportation Engineering, Beijing Institute of Technology, Beijing 100081, PR China
[c] Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, 08854-8018, USA

HIGHLIGHTS

- We use tensor pattern to fully cover the spatial–temporal correlation of corridor freeway traffic volume.
- Various tensor patterns are tested on the traffic volume data collected from PeMS open database.
- We suggest that the tensor based method with spatial correlation achieves better performance than that without spatial information.
- The experimental results show that the proposed method can address the extreme case where the data of a long period of one or several weeks are completely missing.

ARTICLE INFO

Article history:
Received 28 February 2015
Received in revised form 12 May 2015
Available online 18 November 2015

Keywords:
Missing traffic data
Spatial correlation
Tensor completion

ABSTRACT

Missing and suspicious traffic data is a major problem for intelligent transportation system, which adversely affects a diverse variety of transportation applications. Several missing traffic data imputation methods had been proposed in the last decade. It is still an open problem of how to make full use of spatial information from upstream/downstream detectors to improve imputing performance. In this paper, a tensor based method considering the full spatial–temporal information of traffic flow, is proposed to fuse the traffic flow data from multiple detecting locations. The traffic flow data is reconstructed in a 4-way tensor pattern, and the low-n-rank tensor completion algorithm is applied to impute missing data. This novel approach not only fully utilizes the spatial information from neighboring locations, but also can impute missing data in different locations under a unified framework. Experiments demonstrate that the proposed method achieves a better imputation performance than the method without spatial information. The experimental results show that the proposed method can address the extreme case where the data of a long period of one or several weeks are completely missing.

1. Introduction

With the steady increase in travel demand, urban freeways worldwide have experienced increased congestion, but the problem can no longer be addressed by building new highways for economical and environmental reasons [1]. As a consequence, the optimization of existing traffic network [2] has increasingly become a more desirable alternative to the management of traffic congestion. Intelligent transportation systems (ITS) play a significant role in optimizing the existing

[*] Corresponding author. Tel.: +86 10 68914582; fax: +86 10 68914582.
E-mail addresses: bran@wisc.edu (B. Ran), tanhc@bit.edu.cn (H. Tan), 5433809@bit.edu.cn (Y. Wu), peter.j.jin@rutgers.edu (P.J. Jin).

http://dx.doi.org/10.1016/j.physa.2015.09.105
0378-4371/© 2015 Elsevier B.V. All rights reserved.

transportation networks. As a key input for ITS, real-time collected traffic data enable the optimization of transportation networks, ITS applications such as route planning and driver assistance systems [3]. With the development of data collection technology, traffic data that collected from multiple sources such as loop detectors, GPS and video sensors become more and more important in ITS [4–6].

Unfortunately, missing data problems are inevitable due to detector faults or transmission distortion [7,8], which seriously restricts the application and generalization of intelligent transportation system. For example, the traffic control system requires sufficient traffic flow data (i.e., traffic volumes, occupancy rates, and flow speeds) to generate appropriate traffic management strategies [9]. In traffic forecast areas, if there exists missing data, the predicting performance will reduce sharply [10,11]. Without proper imputation methods, traffic counts with missing values are usually either discarded or simply estimated, which may seriously affect the performance of ITS. Consequently, it is urgent to develop a method of better effect on estimating the missing data.

A considerable amount of literature has been published on missing traffic data imputation. These studies to date have tended to focus on missing traffic data imputation with temporal correlations rather than imputation with global information. For example, Smith et al. [12] studied several temporal imputation methods such as Historical Average and Average of Surrounding Time Periods. They concluded that missing traffic data could be imputed by sophisticated statistical techniques using global information. Zhong et al. [13,14] modeled traffic data of a single location as time series and then impute missing data based on the relationship identified from historical past-to-future data pairs. Ni and Leonard [15] use a Bayesian network to learn the temporal correlations encoded within traffic variables to solve the incomplete ITS data issue. Qu et al. [16] propose PPCA-based imputation method makes use of daily periodicity and interval variation for traffic flow volume data incompleteness. Tan et al. [17] developed a RPCA-based imputation method that not only utilizes the temporal correlation, but also considers the physical limitation of traffic data. While all the previously mentioned imputation approaches are powerful and useful methods in some special cases, a serious weakness of these methods, however, is that they do not cover the spatial information of traffic data. Another problem with these approaches is that they could not impute missing data from a different location in a unified framework.

The recent missing traffic data imputation studies have shown that spatial information could help reduce estimation errors. Zhang and Liu [18] demonstrated that making good use of the spatial and temporal information greatly helps the data imputation based on SVM. Yin et al. [19] analyzed and compared temporal and spatial Interpolation-based imputation methods and found that spatial method is robust under various conditions. Li et al. [20] have showed that spatial information could help reduce imputing errors significantly for PPCA and KPPCA methods. Recent evidences of traffic prediction area also suggest that spatial information is helpful for short-term traffic prediction [21,22]. However, most studies in the field of missing data imputation have only focused on utilizing spatial information for single point missing data imputation in transportation networks. Few studies have been able to draw on the development of unifying missing data imputation method for multiple detecting locations in transportation networks. Most studies have only been carried out by using information from a limited number of locations (upstream and downstream). Thus, the modeling and fusion of spatial correlation requires stronger mathematical tools.

Recently, tensor (multi-way array) based methods [23–27] have been introduced to traffic data processing. The tensor based methods construct traffic data into multi-way matrices to accurately capture the underlying multi-mode structure (day, week, time and space mode) of traffic data. For example, the global information of traffic data can be simultaneously taken into account by the $day \times week \times time \times space$ tensor pattern [23]. Then the missing data within the tensor can be solved by tensor completion algorithm, which takes advantage of the global property of the data. Compared with traditional imputation methods, the tensor based method can combine and utilize more multi-mode correlations. Consequently, the tensor completion method is more accurate and robust. Despite the success achieved by tensor based imputation methods, however, there has been limited in-depth discussion about the use of spatial information to upgrade imputation performance.

To date, various methods have been developed and introduced to the tensor completion. As tensor decomposition gives a concise representation of the underlying structure of the tensor [28], a variety of tensor decomposition based methods are applied to the tensor completion. For instance, Acar et al. [29] proposed a method based on CANDECOMP/PARAFAC decomposition. Since CANDECOMP/PARAFAC decomposition is a special case of Tucker decomposition, numerous Tucker decomposition based methods [23,30,31] have been proposed. For low-rank matrix completion, the nuclear norm was used as a convex envelop for the rank function [32]. Generalizing this program to tensor case, the tensor nuclear norm is defined as the weighted sum of unfolding modes of tensors and applied to the tensor completion [33,34]. The previous studies show that the nuclear-norm based methods outperform tensor decomposition based methods particular to high-rank problem and high missing ratio [34]. The goal of this paper is to apply the tensor completion for imputing missing data on multiple locations on a unified framework. The framework uses the spatial information to upgrade missing data imputation performance under the tensor completion. Considering the missing ratio is high in some special cases, and traffic data could not be approximated as a very low-rank tensor due to its intrinsic characteristics [17]. The nuclear-norm based method—HaLRTC [34] is selected in this paper.

This paper is organized as follows: The Related tensor completion backgrounds are provided in Section 2. Theoretic background of the method is presented in Section 3. Various tensor models for a freeway corridor is conducted and analyzed in Section 4. In Section 5, the experiment results are given. The conclusion and future works are discussed in Section 6.

2. Tensor completion backgrounds

The problem of missing data came up in many scientific areas. As a consequence, a great deal of efforts has been made to develop the tensor completion algorithms. The tensor completion methods can be roughly grouped into two groups: tensor decomposition based method and tensor completion based method. The tensor decomposition based method is to approximate the tensor decomposition with an estimated low rank or low-n-rank when only parts entries of a tensor is observed. In Ref. [29], the CP decomposition with missing data is formulated as a weighted least squares problem and applied to missing EEG data imputation and network traffic data modeling. Draw lessons from their work, Tan et al. [23] developed Tucker decomposition based methods to impute missing traffic data. Recently, a simultaneous Tucker decomposition and tensor completion (STDC) [40] method is proposed. Different from traditional decomposition based methods, the decomposition method is combined with low rank constraint on the factor matrices of Tucker decomposition. Several Bayesian scheme based decomposition methods are also investigated and developed [30,31]. Besides the success of decomposition methods, the decomposition scheme is prone to overfitting.

In another line of research, the tensor completion problem is based on the tensor trace norm-the convex relaxation of a tensor's rank. Liu et al. [34] firstly defined the tensor trace norm as the weighted sum of mode matrices. They translated tensor completion problem into a convex optimization problem and applied it to visual data inpainting. Followed by their works, a series modified and similar methods have been proposed [33]. In a nutshell, completion based method, which is an extend form of matrix completion [32], provide more reliable estimation of missing data than decomposition based methods. In this paper, we choose one of the most representative tensor trace norm based method—HaLRTC for our study.

3. Theory

This section provides the necessary theoretical background, and it highlights how these theories are applied to the imputation of missing traffic data.

Tensor, also called the multidimensional array, is the higher-order generalization of vector and matrix. Tensor representation is one of the most practical ways to estimate a multi-dimensional object whose entries are indexed by several variables. For example, a video is indexed by two spatial variables and one temporal variable [35]. Tensor is often used for extracting hidden structures and capturing underlying correlations between modes in the data with multi-mode structure. For example, tensor modeling of traffic data can capture the multi-mode correlations of traffic data with a day × week × time × space tensor representation [23]. In many applications, the missing data problem can be formulated by tensor completion problems. While tensors naturally have a high dimensional characteristic, the tensor of interest is often low-rank, or approximately so, and hence the low-rank approximation can be used for missing data estimation or the tensor completion.

Many researchers have studied how to recover tensors with small Tucker rank or n-rank [36]. The n-mode Tucker rank of a N-dimensional tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, denoted by r_n, is the rank of the mode-n unfolding matrix $A_{(n)}$.

$$r_n = rank_n\left(\mathcal{A}\right) = rank(A_{(n)}). \tag{1}$$

Here, the mode-n unfolding (also called matricization or flattening) of a tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ is defined as $A_{(n)}$. The tensor element $(i_1, i_2, \dots, i_N)$ is mapped to the matrix element (i_n, j), where

$$j = 1 + \sum_{\substack{k=1 \\ k \neq n}}^{N} (i_k - 1) J_k, \quad \text{with}\, J_k = \prod_{\substack{m=1 \\ m \neq n}}^{k-1} I_m. \tag{2}$$

Therefore, $A_{(n)} \in \mathbb{R}^{I_n \times J}$, where $J = \prod_{\substack{k=1 \\ k \neq n}}^{N} I_k$.

The n-mode (matrix) product of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ with a matrix $U \in \mathbb{R}^{J \times I_n}$ is denoted by $\mathcal{X} \times_n U$ and is of size $I_1 \times I_2 \times I_{n-1} \times J \times I_{n+1} \cdots \times I_N$. We have $\mathcal{Y} = \mathcal{X} \times_n U$ is equal to $Y_{(n)} = UX_{(n)}$.

The low-rank approximation for the tensor completion can be formulated as a minimization problem below:

$$\begin{aligned} &\min_{X \in \mathcal{A}} \sum_{n=1}^{N} rank(\mathcal{X}_{(n)}) \\ &\text{st. } \mathcal{X}_{\Omega} = \mathcal{A}_{\Omega} \end{aligned} \tag{3}$$

where the elements of $\mathcal{A}$ in the set Ω are given while the remaining elements are missing. Unfortunately, as matrix rank is difficult to minimize in general, this problem is a difficult non-convex problem. To relax this problem, one common approach is to use the nuclear norm $\| \, \|_*$ to approximate the rank of matrices [37]. The advantage of the nuclear norm is that $\| \, \|_*$ is the tightest convex envelop for the rank of matrices. The definition of the nuclear norm [32] for general tensor case is:

$$\|\mathcal{A}\|_* = \sum_{i=1}^{N} \alpha_i \left\| \mathcal{A}_{(i)} \right\|_* \tag{4}$$

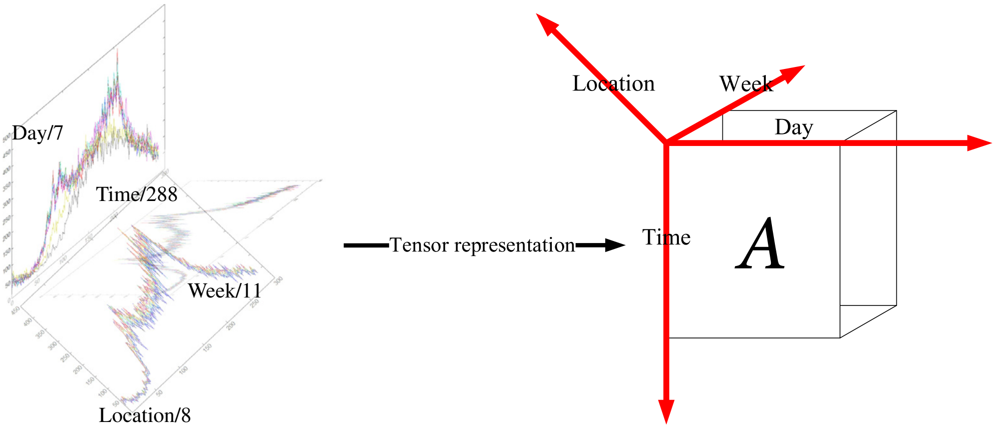

Fig. 1. The tensor representation of traffic flow data: Tensor representation can capture the hidden weekly, daily, spatial correlation of traffic flow data.

where α_i is the weighted parameter for i-mode. Thus the tensor completion problem can be relaxed as:

$$\min_{X \in \mathcal{A}} \ \|\mathcal{X}\|_* \qquad \text{s.t. } \mathcal{X}_\Omega = \mathcal{A}_\Omega. \tag{5}$$

As problem (5) is an equivalent to a convex problem, numerous nuclear-norm based tensor completion methods have been proposed [33,34]. In this paper, the traffic data is evaluated by algorithm proposed in Ref. [34], namely HaLRTC (High accuracy Low Rank Tensor Completion). In HaLRTC algorithm, accuracy of low-n-rank tensor completion is promoted by using ADMM framework.

According to the daily periodicity, weekly periodicity and spatial similarity, the traffic data can be naturally represented as a 4-way (*day* × *week* × *time* × *space*) tensor as shown in Fig. 1. By tensor representation, the missing data problem is formulated as a tensor completion problem. Then the tensor completion algorithms including HaLRTC can be used to estimate the missing traffic data.

4. Traffic data analysis

Obviously, the core problem of the missing traffic data imputation lies on how to build up the relationship between the known data and the unknown ones. In previous study, we have showed that using tensor completion to build multi-mode correlations of traffic data obtain better accuracy compared with matrix completion method [23]. The reason is that the tensor based methods can utilize the information of day, week and interval modes simultaneously, while matrix-based method only mines data correlations in two modes. Nevertheless, the former work has not give an in-depth discussion about the use of spatial correlation from adjacent locations for tensor based methods. To discuss and analyze the impact of spatial correlation on tensor completion, various traffic tensor patterns are constructed and analyzed in this paper.

The Performance Measurement System (PeMS) open-access traffic flow datasets [38] are used for this study. The objective of the PeMS project is to collect real-time freeway data from freeways in California and to measure traffic performance. The particular dataset used in this paper was collected from the adjacent stations located at south bound freeway SR99, District 10, Stanislaus County, California (Fig. 2). The index number of these detectors are 1017510, 1017610, 1017710, 1017810, 1017910, 1018110, 1018210, 1018310, 1018410, 1018510 and 1018610. The sampling period is from March 1, 2011 to May 29, 2011. Our analysis is based on the archived 5-min historical data, which means that we can get 77 (from March 1, 2011 to May 29, 2011) daily vectors of size 288 for each location.

Suppose that location F contains missing data, we can construct the traffic data into a 3-way tensor pattern of size 11(weeks) × 7(days) × 288(interval) by using only traffic flow data from local locations. With the traffic flows at neighboring points are tightly correlated, the information of flow at upstream/downstream locations can help estimate missing traffic data. For tensor based methods, it is very simple to take spatial information into account. The temporal 3-way tensor can be translated into a spatial–temporal 4-way tensor by adding a spatial location. Then the information from neighboring locations is covered. However, it is not easy to determine the number of neighboring location to use. In other words, the length of the spatial dimension of spatial tensor model is hard to choose. Previous work [39] showed that the added mode correlation of data had a great effect on the performance of the tensor completion (the spatial correlation for this paper). Obviously, quantitative analysis of the traffic tensor multi-mode correlations not only helps to construct tensor, but also helps to determine the parameters of tensor completion methods. Formally, the correlations of traffic data are measured by

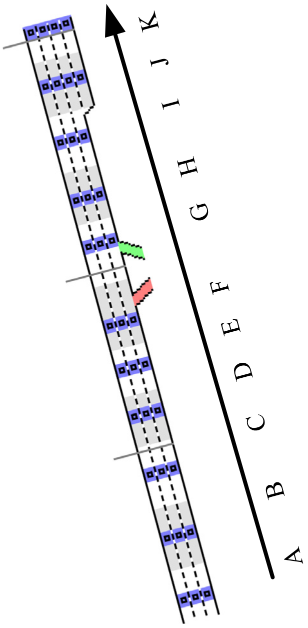

Fig. 2. The locations of this study: These detectors located on a freeway corridor.

Table 1
The week, day and space mode correlation of each tensor pattern.

Tensor model	Space mode coefficient	Week mode coefficient	Day mode coefficient
1	0.9928	0.9681	0.8978
2	**0.9975**	**0.9696**	0.8973
3	0.9945	0.9691	**0.8981**
4	0.9934	0.9685	0.8970
5	0.9915	0.9680	0.8973
6	0.9892	0.9673	0.8971
7	0.9881	0.9671	0.8970

similarity coefficient:

$$s_m = \frac{\sum\limits_{n_m > i > j > 1} R_m(i,j)}{n_m(n_m - 1)/2} \tag{6}$$

where n_m refers to the whole data points; $R_m(i,j)$ refers to the correlation coefficient matrix of the tensor m-mode unfolding.

To give a detailed analysis, we construct seven different tensor models:

1. 4-way tensor with information from downstream neighboring point E ($11 \times 7 \times 288 \times 2$)
2. 4-way tensor with information from upstream neighboring point G ($11 \times 7 \times 288 \times 2$)
3. 4-way tensor with information from both downstream/upstream neighboring points ($11 \times 7 \times 288 \times 3$)
4. 4-way tensor with information from 2 downstream/upstream neighboring points ($11 \times 7 \times 288 \times 5$)
5. 4-way tensor with information from 3 downstream/upstream neighboring points ($11 \times 7 \times 288 \times 7$)
6. 4-way tensor with information from 4 downstream/upstream neighboring points ($11 \times 7 \times 288 \times 9$)
7. 4-way tensor with information from 5 downstream/upstream neighboring points ($11 \times 7 \times 288 \times 11$).

The correlations of space, day, and week mode are given in Table 1. As can be seen from the table, all seven tensor patterns not only show a significant positive spatial correlation, but also keep a high week mode and day mode correlation. The results of this analysis indicate that the performance of tensor based methods can be promoted by fusing multiple locations information with a 4-way spatial–temporal tensor. Tensor model 2 obtains the highest spatial mode coefficient compared with other tensor models. It is reasonable because that the upstream location E is the nearest location from location F. Though downstream location G is also adjacent to F, due to the ramp entrance between F and G, tensor model 2 with information from G shows a lower spatial correlation. Interestingly, the week and the day mode correlation is also changed with different tensor models. The reason is that traffic flow data in a different location show slightly different daily and weekly periodicity. Overall, the finding suggests that the information of flow at upstream/downstream locations may benefit to estimate missing traffic data for tensor based methods.

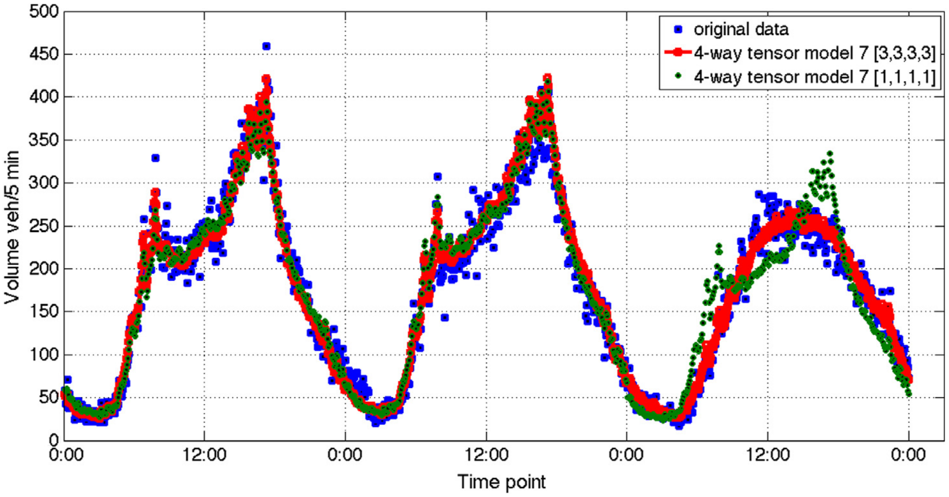

Fig. 3. The low-n-rank approximation for tensor model 7. Red: A low-n-rank with rank group [3,3,3,3] can roughly capture the principle variation of traffic flow data. Green: the rank group [1,1,1,1] approximation can only capture the variation of traffic flow on workdays. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The tensor completion based methods refer to the characteristics that the road traffic flow tensor with multi-mode correlations can be approximated as a low-n-rank tensor, which means that the multi-mode correlation can be captured by low-n-rank tensor completion. In order to illustrate how low-rank approximation works for tensor based missing traffic data imputation, the low-n-rank tensor approximation of some tensor models was performed by Tucker decomposition [36]. The Tucker decomposition is a form of higher-order principal component analysis. It decomposes a tensor into a core tensor multiplied (or transformed) by a matrix of each mode. A N-dimensional tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, we have:

$$\mathcal{A} \approx \mathcal{S} \times_1 U^1 \times_2 U^2 \cdots \times_N U^N. \tag{7}$$

Here, $U^n \in \mathbb{R}^{I_n \times R_n}$ are the factor matrices (which are usually orthogonal) and can be thought of as the principal components in each mode. If we perform an exact Tucker decomposition, R_n are the n-mode ranks where $R_n = rank_n(\mathcal{A})$. Usually the low-$n$-rank approximation can be achieved by a Tucker decomposition where $R_n < rank_n(\mathcal{A})$. In this paper, partial results of a low-n-rank approximation with rank-group [3,3,3,3] and rank-group [1,1,1,1] for tensor model 7 are given in Fig. 3.

From Fig. 3 we can see that a low-n-rank approximation with rank group [3,3,3,3] can basically keep the structure of a traffic tensor of size $11 \times 7 \times 288 \times 11$. The approximation can represent most of the characteristics of the road traffic flow. It indicates that the missing traffic data can be estimated by a good low-n-rank approximation. The low-n-rank approximation with rank group [1,1,1,1] can only keep the variation of traffic flow on workdays. This result may be explained by the fact that the traffic flow on workdays are significantly different from weekends. Only using one principle component on day mode is not sufficient to capture the whole variation on day mode.

5. Numerical experiments

In this section, we examine the imputing performance of different tensor models. The particular datasets mentioned above are used in following experiments. In the following tests, we simulate two common missing patterns in the test: (1) random missing cases: where the missing points are distributed randomly of each other. (2) extreme missing cases: several days' traffic data are missing. For the purpose of this paper, two parts' experiments are made. In the first case, we suppose that all 11 locations contain missing data, the 4-way tensor completion methods that impute all the missing data on a unified framework are compared with 2-way matrix-based method RPCA [25] and 3-way tensor based method that impute missing data for local location independently. In the second part, we suppose that only location F contains missing data and the 7 tensor patterns discussed above are tested. The tensor completion algorithm HaLRTC [34] was used in this paper. The weighted parameters for each mode nuclear norm of HalRTC are set to [1,1,1,1e−3] for 4-way tensor pattern and [1,1,1e−3] for 3-way tensor pattern. The parameters of RPCA are set according to Ref. [25].

The imputing performance is evaluated by the root mean squared error (RMSE) between the estimated missing points and the original data points. RMSE is a commonly used error criteria, which reflects the average performance for the missing data imputing:

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(t_{real}^{(m)} - t_{est}^{(m)})^2} \tag{8}$$

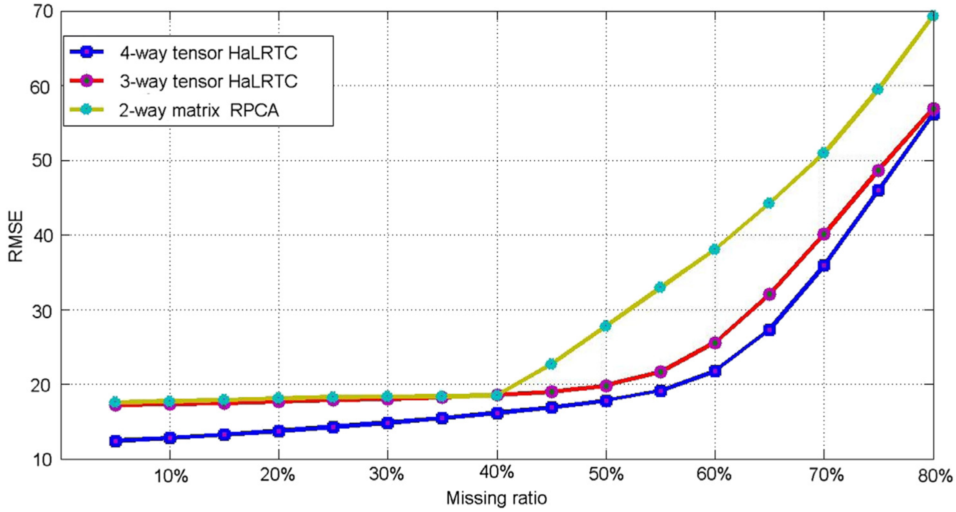

Fig. 4. Overall RMSE curves for 4-way tensor pattern, 3-way tensor pattern and 2-way matrix-based method RPCA.

Table 2
The RMSEs for 4-way tensor pattern for missing weeks on location F.

Missing weeks	1	2	3	4	5	6	7	8
RMSE	10.8793	11.8586	12.6595	13.6572	15.7206	19.0174	23.9593	33.6985

where ${t_{real}}^{(m)}$ and ${t_{est}}^{(m)}$ are the mth elements of the known real value and the estimated value, respectively. M is the number of missing data, which can be used to calculate the missing ratio. All the methods were performed using Matlab on a Windows Workstation with a Dual-Core Intel(R) Core(TM) 2.50 GHz CPU and 4 GB RAM.

5.1. Comparison of different methods

To compare with 3-way tensor based method and 2-way RPCA, the traffic flow data from all 11 locations is used to build a 11 × 7 × 288 × 11 4-way tensor. The first mode stands for week mode, the second mode stands for day mode, the third mode stands for time mode, and the fourth mode stands for space mode. We construct 11 11 × 7 × 288 3-way tensors and 11 77 × 288 matrices that estimate missing data just by local temporal information for comparison. Experiments are conducted on data of missing ratio from 5% to 80%. The results are given in Fig. 4.

As shown in Fig. 4, the 4-way tensor with spatial information reported significantly more accurate than the other two methods that only using temporal information. The reason is that tensor completion algorithm successfully captures the spatial correlation of traffic data through tensor modeling, while 2-way and 3-way methods only mine temporal correlation. Furthermore, as the same results from Ref. [23], although 2-way matrix-based method RPCA achieves good performance for low missing ratio data, its performances degrade sharply when the missing ratio is higher than 50%. A possible explanation for this might be that some day mode information is lost when missing ratio is higher than 50%. Thus using week mode and space mode information help a lot under high missing ratio. Overall, the accuracy of missing traffic data imputation can be improved by appropriate usage of spatial information.

The matrix-based method did not work when one or several days worth of traffic volume data are missing. The 3-way tensor based method could not impute the missing data when one or several weeks' traffic volume data on one location are missing. The underlying reason is related to matrix decomposition [32]. The 4-way tensor pattern extends the applicability of 3-way tensor pattern. It can also work in the extreme case when one or several weeks' traffic volume data on one or several locations are missing. The missing data on location F is set to 1–8 weeks (from the first week to the eighth week) to verify the ability of 4-way tensor pattern. Table 2 gives the result of 4-way tensor pattern, and Fig. 5 shows the partial imputed traffic flow when eight consecutive weeks' traffic data are missing.

From the graph and Table 2 we can see that the missing data within eight consecutive weeks can be successfully recovered by the proposed 11 × 7 × 288 × 11 4-way tensor. This result may be explained by the fact that the traffic flow data have the underlying multi-mode information. Even when the week or day mode correlations are missing under extreme case, the traffic data still can be successfully imputed by making a good use of the spatial mode correlation.

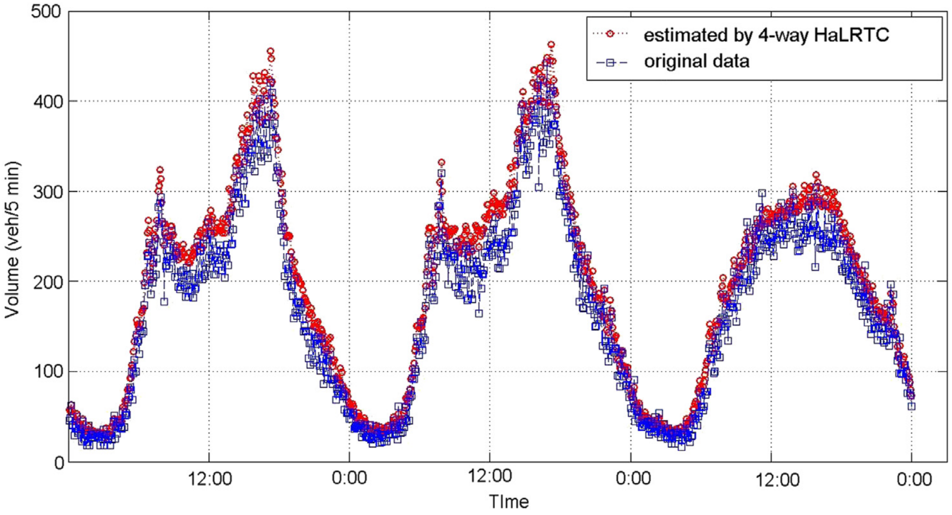

Fig. 5. Imputation results when eight consecutive weeks' traffic flow at location F is all missing.

Table 3
The RMSEs for various tensor pattern for missing weeks on location F.

Missing weeks	1	2	3	4	5	6	7	8
Model 1	Not work							
Model 2	Not work							
Model 3	16.3214	19.0516	21.5627	27.0748	38.5374	56.1463	75.8493	97.3731
Model 4	13.6430	15.4606	16.7764	19.3822	24.2868	32.8431	45.5602	66.0352
Model 5	12.3827	13.7798	14.7574	16.5374	19.7862	25.1407	33.2458	48.3672
Model 6	11.4813	12.6236	13.5189	14.8069	17.1705	21.1879	27.2528	39.0216
Model 7	10.8793	11.8586	12.6595	13.6572	15.7206	19.0174	23.9593	33.6985

5.2. Experiment with various tensor pattern

In the above discussion, we show that 4-way tensor pattern with spatial information outperforms the methods that only using temporal information under random and extreme cases. The experiments are conducted on data from locations on the freeway corridor. The performances of 4-way tensor pattern are inspiring. However, with locations on urban and transportation networks, caution for the tensor construction must be applied, as the performance might be correlated to the spatial correlation between different locations. The experiments on various tensor pattern constructed in Section 3 and the discussion about relationship between imputation performance and spatial correlation are made in this part.

In order to compare the performance of seven tensor models considering the spatial information from adjacent locations when missing data only occur in location F, experiments on both random missing cases and extreme cases are conducted. Similarly, the missing ratio for location F is set from 5% to 80% for random missing cases and 1–8 weeks for extreme cases. The results are given in Fig. 6 and Table 3.

The experimental results show tensor model 7 that uses maximum number of neighboring locations outperform the other models under both random and extreme cases. Tensor model 1 uses information from upstream neighboring point outperforms tensor model 2 with information from downstream point under random cases. The reason for this may be that the traffic flow on location F is more similar with upstream traffic flow than downstream traffic flow. Compare the results between all models under random missing cases, it can be found that, the more locations fused in the tensor models, the more the imputation accuracy it is. It seems possible that these results are due to the high similarities between neighboring locations on a freeway corridor.

Through the comparison of extreme cases provides three major observations. First, the tensor model 3 with three or more locations has the best performance. If several weeks worth of data on one location are missing, the best data sample size will be more than 2 locations from downstream/upstream. Only using one neighboring location cannot work. The reason is that a high ratio of successive missing data exits in such a tensor model. In this case, the observed entries cannot provide sufficient correlation for the imputation of missing data. Second, tensor model 6 and tensor model 7 can obtain a good imputation performance up to 6 consecutive days of missing data. Third, using more locations can greatly improve the imputation performance when missing weeks are above 4. The reason is that the extreme case could adversely affect the day mode correlation and the week mode correlation of the dataset when more than four weeks' traffic flow are missing. So, tensor completion algorithm needs sufficient space mode information to fill this gap.

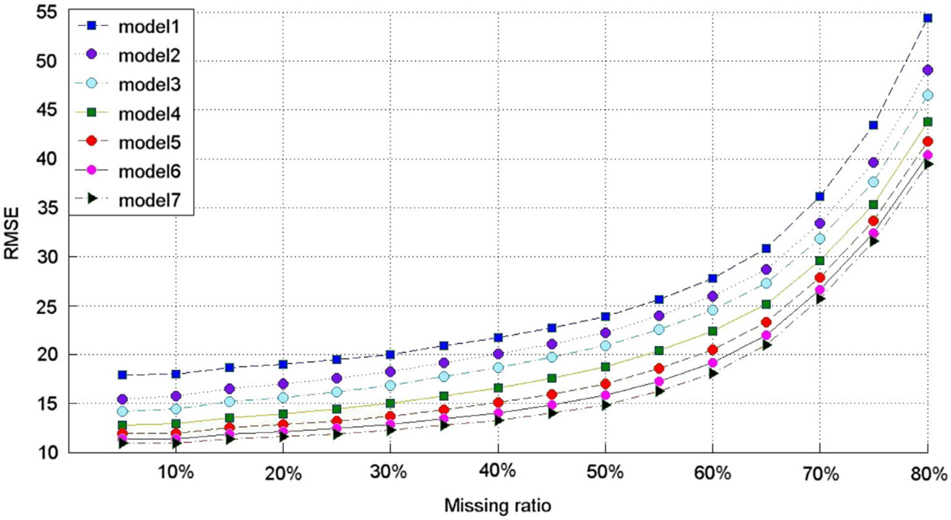

Fig. 6. RMSE curves for different tensor models.

6. Conclusion

In this paper, the use of spatial information for tensor based missing traffic data imputation methods are examined and discussed. Various tensor models are constructed to determine the appropriate usage of spatial information from neighboring detecting locations. The widely used low-n-rank tensor completion algorithm HaLRTC has been applied for missing traffic data. The proposed method formulates the traffic data into a 4-way dataset covers the spatial–temporal correlations. The analysis shows that low-n-rank approximation can successfully capture the underlying multi-mode structure of traffic flow data. Results show that using spatial information could help reduce imputing errors both under random and extreme cases. The tensor based method can synchronously impute the missing data from different locations under a unified framework with higher accuracy than methods that only using temporal correlations.

A number of caveats need to be noted regarding the present study. The experiments in this paper are conducted on traffic flow data on neighboring locations from a freeway corridor. The traffic flow data from these locations are very similar. The studies on tensor completion algorithm and tensor modeling strategies are still needed when fusing spatial information between locations that traffic flow data are significantly different. In the urban cities, the considerations about incident and traffic control light are also needed. Another more practical problem is how to find and use the most relevant data to impute missing data in the large-area transportation network. The traffic variables like speed and volume are highly correlated and often both speed, and volume data are missing at the same time on some locations. Future research should, therefore, concentrate on the development of unifying imputation algorithm for different traffic flow variables.

Acknowledgments

The research was supported by NSFC (61271376, 51308115), the National Basic Research Program of China (973 Program: No. 2012CB725405), and Beijing Natural Science Foundation (4122067).

References

[1] B.S. Kerner, Introduction to Modern Traffic Flow Theory and Control: The Long Road to Three-Phase Traffic Theory, Springer, 2009.
[2] Y. Liu, G.-L. Chang, J. Yu, An integrated control model for freeway corridor under nonrecurrent congestion, IEEE Trans. Veh. Technol. 60 (2011) 1404–1418.
[3] J. Wang, L. Zhang, D. Zhang, K. Li, An adaptive longitudinal driving assistance system based on driver characteristics, IEEE Trans. Intell. Transp. Syst. 14 (1) (2013) 1–12.
[4] B. Ran, P.J. Jin, D. Boyce, T.Z. Qiu, Y. Cheng, Perspectives on future transportation research: Impact of intelligent transportation system technologies on next-generation transportation modeling, J. Intell. Transp. Syst. 16 (2012) 226–242.
[5] J. Zhang, F.Y. Wang, K. Wang, W.H. Lin, X. Xu, C. Chen, Data-driven intelligent transportation systems: A survey, IEEE Trans. Intell. Transp. Syst. 12 (2011) 1624–1639.
[6] J. Wang, D. Ni, K. Li, RFID-based vehicle positioning and its applications in connected vehicles, Sensors 14 (3) (2014) 4225–4238.
[7] P.W. Lin, G.L. Chang, Modeling measurement errors and missing initial values in freeway dynamic origin–destination estimation systems, Transp. Res. C 14 (2006) 384–402.
[8] N.E.E. Faouzi, H. Leung, A. Kurian, Data fusion in intelligent transportation systems: Progress and challenges—a survey, Inf. Fusion 12 (2011) 4–10.
[9] R.C. Carlson, I. Papamichail, M. Papageorgiou, A. Messmer, Optimal mainstream traffic flow control of large-scale motorway networks, Transp. Res. C 18 (2010) 193–212.
[10] J.R. Xu, X.Y. Li, H.J. Shi, Short-term traffic flow forecasting model under missing data, J. Comput. Appl. 30 (2010) 1117–1120.
[11] J. Van Lint, S. Hoogendoorn, H.J. van Zuylen, Accurate freeway travel time prediction with state-space neural networks under missing data, Transp. Res. C 13 (2005) 347–369.

[12] B.L. Smith, W.T. Scherer, J.H. Conklin, Exploring imputation techniques for missing data in transportation management systems, Transp. Res. Rec.: J. Transp. Res. Board 1836 (1) (2003) 132–142.

[13] M. Zhong, P. Lingras, S. Sharma, Estimation of missing traffic counts using factor, genetic, neural, and regression techniques, Transp. Res. C 12 (2) (2004) 139–166.

[14] M. Zhong, S. Sharma, P. Lingras, Genetically designed models for accurate imputation of missing traffic counts, Transp. Res. Rec.: J. Transp. Res. Board 1879 (1) (2004) 71–79.

[15] D. Ni, J.D. Leonard II, Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data, Transp. Res. Rec.: J. Transp. Res. Board 1935 (1) (2005) 57–67.

[16] L. Qu, L. Li, Y. Zhang, J. Hu, PPCA-based missing data imputation for traffic flow volume: a systematical approach, IEEE Trans. Intell. Transp. Syst. 10 (3) (2009) 512–522.

[17] H. Tan, Y. Wu, B. Cheng, W. Wang, B. Ran, Robust missing traffic flow imputation considering nonnegativity and road capacity, Math. Probl. Eng. 2014 (2014).

[18] Y. Zhang, Y. Liu, Missing traffic flow data prediction using least squares support vector machines in urban arterial streets, in: IEEE Symposium on Computational Intelligence and Data Mining, 2009, CIDM'09, IEEE, 2009, pp. 76–83.

[19] W. Yin, P. Murray-Tuite, H. Rakha, Imputing erroneous data of single-station loop detectors for nonincident conditions: comparison between temporal and spatial methods, J. Intell. Transp. Syst. 16 (3) (2012) 159–176.

[20] L. Li, Y. Li, Z. Li, Efficient missing data imputing for traffic flow by considering temporal and spatial dependence, Transp. Res. C 34 (2013) 108–120.

[21] E.I. Vlahogianni, M.G. Karlaftis, J.C. Golias, Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks, Comput.-Aided Civ. Infrastruct. Eng. 22 (5) (2007) 317–325.

[22] W. Min, L. Wynter, Real-time road traffic prediction with spatio-temporal correlations, Transp. Res. C 19 (4) (2011) 606–616.

[23] H. Tan, G. Feng, J. Feng, W. Wang, Y.J. Zhang, F. Li, A tensor based method for missing traffic data completion, Transp. Res. C 28 (2013) 15–27.

[24] H. Tan, J. Feng, G. Feng, W. Wang, Y.J. Zhang, Traffic volume data outlier recovery via tensor model, Math. Probl. Eng. 2013 (2013).

[25] H. Tan, Y. Wu, G. Feng, W. Wang, B. Ran, A new traffic prediction method based on dynamic tensor completion, Procedia-Soc. Behav. Sci. 96 (2013) 2431–2442.

[26] H. Tan, J. Feng, Z. Chen, F. Yang, W. Wang, Low multilinear rank approximation of tensors and application in missing traffic data, Adv. Mech. Eng. 2014 (2014).

[27] H. Tan, L. Song, Y. Cheng, B. Cheng, B. Ran, A tensor completion-based traffic state estimation model, in: CICTP 2014@ s Safe, Smart, and Sustainable Multimodal Transportation Systems, ASCE, 2014, pp. 298–309.

[28] T.G. Kolda, B.W. Bader, Tensor decompositions and applications, SIAM Rev. 51 (3) (2009) 455–500.

[29] E. Acar, D.M. Dunlavy, T.G. Kolda, M. Mørup, Scalable tensor factorizations for incomplete data, Chemometr. Intell. Lab. Syst. 106 (1) (2011) 41–56.

[30] Z. Xu, F. Yan, Y. Qi, Bayesian nonparametric models for multiway data analysis, IEEE Trans. Pattern Anal. Mach. Intell. 99 (2013) 1.

[31] H. Shan, A. Banerjee, R. Natarajan, Probabilistic Tensor Factorization for Tensor Completion. Technical report, University of Minnesota, 2011.

[32] E.J. Candès, B. Recht, Exact matrix completion via convex optimization, Found. Comput. Math. 9 (6) (2009) 717–772.

[33] S. Gandy, B. Recht, I. Yamada, Tensor completion and low-n-rank tensor recovery via convex optimization, Inverse Problems 27 (2) (2011) 025010.

[34] J. Liu, P. Musialski, P. Wonka, J. Ye, Tensor completion for estimating missing values in visual data, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 208–220.

[35] H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, MPCA: Multilinear principal component analysis of tensor objects, IEEE Trans. Neural Netw. 19 (1) (2008) 18–39.

[36] L. De Lathauwer, B. De Moor, J. Vandewalle, A multilinear singular value decomposition, SIAM J. Matrix Anal. Appl. 21 (4) (2000) 1253–1278.

[37] J.F. Cai, E.J. Candès, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. 20 (4) (2010) 1956–1982.

[38] PeMS, California Performance Measurement System. http://pems.eecs.berkeley.edu.

[39] M. Signoretto, R. Van de Plas, B. De Moor, J.A. Suykens, Tensor versus matrix completion: a comparison with application to spectral data, IEEE Signal Process. Lett. 18 (7) (2011) 403–406.

[40] Y.L. Chen, C.T. Hsu, H.Y. Liao, Simultaneous tensor decomposition and completion using factor priors, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 577–591.