Matrix and Tensor Based Methods for Missing Data Estimation in Large Traffic Networks

Muhammad Tayyab Asif, *Student Member, IEEE*, Nikola Mitrovic, *Student Member, IEEE*, Justin Dauwels, *Senior Member, IEEE*, and Patrick Jaillet

***Abstract*—Intelligent transportation systems (ITSs) gather information about traffic conditions by collecting data from a wide range of on-ground sensors. The collected data usually suffer from irregular spatial and temporal resolution. Consequently, missing data is a common problem faced by ITSs. In this paper, we consider the problem of missing data in large and diverse road networks. We propose various matrix and tensor based methods to estimate these missing values by extracting common traffic patterns in large road networks. To obtain these traffic patterns in the presence of missing data, we apply fixed-point continuation with approximate singular value decomposition, canonical polyadic decomposition, least squares, and variational Bayesian principal component analysis. For analysis, we consider different road networks, each of which is composed of around 1500 road segments. We evaluate the performance of these methods in terms of estimation accuracy, variance of the data set, and the bias imparted by these methods.**

***Index Terms*—Missing data estimation, low-dimensional models.**

I. Introduction

With advancements in sensor technologies, intelligent transportation systems (ITS) can now collect traffic data from a wide range of stationary and mobile sensors [1]–[8]. Stationary sensors such as loop detectors and road side cameras tend to have limited spatial coverage, whereas mobile sensors such as GPS probes collect data with highly erratic spatial and temporal resolution. These issues make the problem of missing data unavoidable in traffic data sets. Furthermore, failures such as detector malfunction and lossy communication systems may also result in incomplete traffic information [4], [9]. This can result in situations, where a high percentage of data is missing. Consequently, missing data is a commonly reported problem in traffic data sets [9]–[14]. Different studies in this regard have reported that missing data percentages can be as high as 90% [13]. For traffic management systems, this is a critical issue [15], [16].

The methods proposed to solve the problem of missing data can be broadly divided into two categories: function estimation and matrix/tensor completion. In the first case, it is typically assumed that the problem of missing data is localized to certain links and time intervals. In this way, the historical data can be used to obtain the relationship function between the target road and its neighbors or past states of that road. For instance, Chen *et al.* [9] developed relationship models between neighboring loop detectors using historical data. This relationship function was then used to impute missing values for faulty detectors. Ming *et al.* [14] trained neural networks and used temporal features to estimate missing values. Yang *et al.* [17] also used a similar approach and applied least squares support vector machines to estimate missing values. The function estimation techniques require complete historical data to learn the relationship models. Hence, these methods will not work if historical data has missing values. In practical scenarios, uncorrupted historical data may not be available. On the other hand, matrix and tensor completion methods do not require training data to perform imputation [4]. Consequently, these methods have garnered considerable interest in the field of transportation studies [4], [13], [18]–[22].

Traffic states across neighboring roads tend to be strongly correlated [23], [24]. These relationships imply that road networks can be represented by low-dimensional models. Matrix and tensor completion methods utilize these patterns to estimate the missing values by obtaining a suitable low-rank approximation of the incomplete tensor/matrix. However, previous studies involving matrix/tensor completion methods for traffic data sets have mostly focused on data obtained from a few roads or intersections. For instance, Li Qu *et al.* [4], [18] used Bayesian principal component analysis (BPCA) to perform imputation for traffic flow data. They analyzed a small network consisting of around 50 road segments. Li Li *et al.* [19] used data from four detectors for analysis. Gang and Tongmin [20] compared the performance of various matrix completion methods on a test network of around 50 links. Huachun *et al.* [13] performed missing data imputation by tensor decomposition methods. For analysis, they considered four road segments and represented their data obtained from each road as a 3-way tensor.

Traffic conditions across city-scale networks also tend to have certain common global patterns [25]–[29]. Some studies

Manuscript received March 4, 2015; revised October 11, 2015 and November 25, 2015; accepted December 4, 2015. Date of publication January 18, 2016; date of current version June 24, 2016. This work was supported in part by National Research Foundation Singapore through the Centre for Future Urban Mobility, Singapore–Massachusetts Institute of Technology Alliance for Research and Technology (SMART). The Associate Editor for this paper was F.-Y. Wang.

M. T. Asif, N. Mitrovic, and J. Dauwels are with the School of Electrical and Electronic Engineering, College of Engineering, Nanyang Technological University, Singapore 639798 (e-mail: muhammad89@e.ntu.edu.sg; jdauwels@ntu.edu.sg).

P. Jaillet is with the Department of Electrical Engineering and Computer Science, School of Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA; with the Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02142 USA; and also with the Center for Future Urban Mobility, Singapore-Massachusetts Institute of Technology Alliance for Research and Technology, Singapore 138602 (e-mail: jaillet@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2015.2507259

[21], [22] have considered the problem of missing data in large networks, albeit in a limited manner. These studies did not analyze the performance of imputation algorithms for different road types (expressways, arterial roads, access roads) and during different days of the week. Furthermore, they did not analyze the bias and variance of the imputed traffic data.

In summary, function estimation methods for imputation have limited application for large networks due to their dependency on uncorrupted historical data [9], [14], [17], [30]. The previous studies which applied matrix and tensor completion methods mostly considered data from a single link or a few intersections [4], [13], [18]–[20]. These studies typically do not analyze the performance of imputation methods for different road types and during different days of the week [4], [18], [21]. Furthermore, analysis in terms of variance, bias, and the impact of the rank of the estimated low-dimensional model on the imputation performance also needs to be considered.

In this paper, we address the aforementioned limitations in previous studies by performing missing data imputation for large road networks comprising of expressways, arterial roads, access roads as well as slip roads. We propose various matrix and tensor based methods that can extract global traffic patterns from incomplete data. Our main contributions are as follows:

In transportation systems, the problem of missing data is typically handled by applying variants of the un-constrained weighted least square approach [20]. We propose the nuclear norm minimization based approach to solve the problem of missing data in large-scale transportation systems. We apply fixed point continuation with approximate singular value decomposition (FPCA) [31] to obtain a low-rank representation for large-scale road networks in presence of missing data. The results show that its performance is less sensitive to daily variations in traffic data. Furthermore, the method also provides better or similar performance (in terms of weighted relative error (WRE), root mean squared error (RMSE), variance and bias) in comparison to the other algorithms for different road categories.

Probabilistic PCA based methods have been previously used to estimate missing traffic information from incomplete data sets [4], [19]. However, these studies only considered small networks and did not evaluate the performance of probabilistic methods for different road categories and during different days of the week. Furthermore, analysis in terms of induced bias and variance in the recovered speed data also needs to be considered. Moreover, BPCA formulations used in these studies [4] do not scale well for large-scale systems. In this study, we consider a variant of variational Bayesian principal component analysis (VBPCA) [32], which is suitable for large-scale road networks.

We compare the performance of the above mentioned methods with baseline matrix and tensor decomposition methods such as weighted least squares and canonical polyadic (CP) decomposition. We analyze the performance of these methods for different road categories and for days of the week. We analyze the variance and bias induced by these methods in the imputed speed data. Furthermore, we discuss the impact of rank selection on the performance of different methods.

The rest of the paper is structured as follows. In Section II, we explain the data and different performance measures. In Section III, we review several matrix and tensor completion methods to estimate missing data in road networks. In Section IV, we analyze the performance of the proposed methods for different test networks. In Section V, we summarize our contributions and conclude the paper.

TABLE I
SIZE OF DIFFERENT TEST NETWORKS. EACH TEST NETWORK IS COMPOSED OF ROADS FROM A SPECIFIC CATEGORY. PRIMARY AND LOCAL ACCESS ROADS ARE REFERRED AS "OTHER ROADS" IN THE TABLE

CATA	CATB	CATC	Slip Roads	Other Roads
2175	2500	1428	1572	2221

II. TRAFFIC DATA SET AND PERFORMANCE MEASURES

A. Data Set

We represent the test road network of size p by a set E of road segments s_i, such that $E = \{s_i\}_{i=1}^{p}$. In this study, we consider average speed data. The average speed on a link s_i during the interval $(t_j - \Delta t, t_j)$ is represented by $z(s_i, t_j)$. The sampling interval Δt is 5 minutes. For each link s_i, we create a speed profile $\mathbf{a}_i \in \mathbb{R}^n$ such that $\mathbf{a}_i = [z(s_i, t_1), \ldots, z(s_i, t_n)]^T$. The speed profiles contain one day of speed data for each link. We stack these speed profiles to obtain the network profile matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ such that $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]$. Let $\mathbf{D} \in \mathbb{R}^{n \times p}$ be the corresponding incomplete observed data matrix. The set Ω contains the location of the entries in $\mathbf{D}$ for which speed data is available and the set $\Theta = \Omega^c$ represents the location of the missing speed values in $\mathbf{D}$. To generate the incomplete observed data matrix, we follow the procedure outlined in [13], [33].

For the tensor completion method, we create the network profile tensor $\underline{\mathbf{A}} \in \mathbb{R}^{n \times p \times q}$ by stacking together network profile matrices $\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_q\}$ from different days to form a 3-way tensor. To this end, we use $q = 7$ days of data. In this case, the incomplete tensor is represented by $\underline{\mathbf{D}} \in \mathbb{R}^{n \times p \times q}$.

For the analysis, we consider 5 test networks. The roads in each network belong to the city-state road network of Singapore for which sufficient data was available. The first test network consists of expressways (CATA). The second and third networks are composed of major and minor arterial roads respectively. We refer to major arterial roads as CATB and minor arterial roads as CATC. The fourth network contains slip roads, while the fifth network contains primary access and local access roads. Table I shows the size of each test network. The network consisting of primary/local access roads is referred as other roads in the table. The speed data was provided courtesy of Singapore's land transportation authority (LTA). In this study, we consider speed data from August 1, 2011 to August 7, 2011.

B. Performance Measures

In this section, we briefly describe different performance measures to assess the proposed methods. For matrices, we define the weighted relative error (WRE) between actual $\mathbf{A}$ and estimated network profile $\widehat{\mathbf{A}}$ as:

$$\text{WRE} = \frac{\left\| \mathbf{W} \circ (\mathbf{A} - \widehat{\mathbf{A}}) \right\|_F}{\|\mathbf{W} \circ \mathbf{A}\|_F} \tag{1}$$

where the symbol $\circ$ represents the element wise multiplication between the two matrices. The matrix $\mathbf{W} \in \mathbb{R}^{n\times p}$ is the weight matrix with values:

$$w_{ij} = \begin{cases} 0 & (i,j) \in \Omega \\ 1 & (i,j) \in \Theta. \end{cases} \tag{2}$$

The Frobenius Norm $\|\mathbf{A}\|_F$ of a matrix $\mathbf{A} \in \mathbb{R}^{n\times p}$ is defined as:

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} a_{ij}^2}. \tag{3}$$

Similarly, we define WRE for tensors as follows:

$$\text{WRE} = \frac{\left\|\underline{\mathbf{W}} \circ (\underline{\mathbf{A}} - \underline{\widehat{\mathbf{A}}})\right\|_F}{\|\underline{\mathbf{W}} \circ \underline{\mathbf{A}}\|_F} \tag{4}$$

where the symbol $\circ$ represents the element wise multiplication between the two tensors. The tensor $\underline{\mathbf{W}} \in \mathbb{R}^{n\times p\times q}$ is the weight tensor with values:

$$w_{ijk} = \begin{cases} 0 & (i,j,k) \in \Omega \\ 1 & (i,j,k) \in \Theta. \end{cases} \tag{5}$$

The Frobenius Norm $\|\underline{\mathbf{A}}\|_F$ of a tensor $\underline{\mathbf{A}} \in \mathbb{R}^{n\times p\times q}$ is defined as:

$$\|\underline{\mathbf{A}}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q} a_{ijk}^2}. \tag{6}$$

Weighted relative error is commonly used to evaluate the performance of matrix and tensor completion algorithms [31], [33]. We also compute root mean squared error (RMSE) of estimation algorithms as follows:

$$\text{RMSE}_{\text{mat}} = \sqrt{\frac{1}{|\Theta|}\sum_{(i,j)\in\Theta}(a_{ij} - \widehat{a}_{ij})^2} \tag{7}$$

$$\text{RMSE}_{\text{ten}} = \sqrt{\frac{1}{|\Theta|}\sum_{(i,j,k)\in\Theta}(a_{ijk} - \widehat{a}_{ijk})^2} \tag{8}$$

where $|\Theta|$ represents the size of the set Θ. We calculate the bias induced in the imputed speed data as follows:

$$\text{Bias}_{\text{mat}} = \frac{1}{|\Theta|}\sum_{(i,j)\in\Theta}(a_{ij} - \widehat{a}_{ij}) \tag{9}$$

$$\text{Bias}_{\text{ten}} = \frac{1}{|\Theta|}\sum_{(i,j,k)\in\Theta}(a_{ijk} - \widehat{a}_{ijk}). \tag{10}$$

Furthermore, we calculate the variance of the imputed values as follows:

$$\text{Variance}_{\text{mat}} = \frac{1}{|\Theta|}\sum_{(i,j)\in\Theta}(\widehat{a}_{ij} - \bar{a}_{\Theta})^2 \tag{11}$$

$$\text{Variance}_{\text{ten}} = \frac{1}{|\Theta|}\sum_{(i,j,k)\in\Theta}(\widehat{a}_{ijk} - \bar{a}_{\Theta})^2 \tag{12}$$

where $\bar{a}_{\Theta}$ represents the mean values of $\{\widehat{a}_{ij}\}_{(i,j)\in\Theta}$ and $\{\widehat{a}_{ijk}\}_{(i,j,k)\in\Theta}$ in (11) and (12) respectively.

III. Missing Data Estimation

In this section, we briefly discuss various matrix and tensor completion algorithms for estimation of missing data in matrices and tensors. We apply LS, FPCA and VBPCA to recover missing speed information in the incomplete matrices. For tensor completion, we apply canonical polyadic weighted optimization (CP-WOPT) to recover the missing traffic information.

A. Least Squares Method (LS)

Traffic parameters such as speed tend to behave similarly across an interconnected network [23], [27]. We aim to utilize these latent patterns to recover the missing speed information in the incomplete matrix $\mathbf{D}$. To this end, let us first consider the complete network profile matrix $\mathbf{A}$. By applying principal component analysis (PCA), we can obtain a low rank (with rank-r) approximation $\widehat{\mathbf{A}} = \mathbf{W}\mathbf{X} + \mathbf{M}$ of the network profile matrix $\mathbf{A}$, where $\mathbf{W} \in \mathbb{R}^{n\times r}$ and $\mathbf{X} \in \mathbb{R}^{r\times p}$ are two low-rank matrices and $\mathbf{M} \in \mathbb{R}^{n\times p}$ contains the row wise mean values of $\mathbf{A}$. This decomposition can be obtained by solving the following least squares optimization problem:

$$\begin{aligned} &\min_{\widehat{\mathbf{A}}} \sum_{i=1}^{n}\sum_{j=1}^{p}(a_{ij} - \widehat{a}_{ij})^2 \\ &\widehat{a}_{ij} = \mathbf{w}_i^T\mathbf{x}_j + m_{ij} \end{aligned} \tag{13}$$

with the constraint that the vectors $\{\mathbf{w}_i\}_{i=1}^{r}$ remain orthonormal [34]. In the case of incomplete matrix $\mathbf{D}$, we can reformulate the problem by minimizing the reconstruction error for observed speed data $\{d_{ij}\}_{(i,j)\in\Omega}$ only, where d_{ij} represents the speed value for road s_j at time t_i. Hence, the optimization problem will become [35]:

$$\begin{aligned} &\min_{\widehat{\mathbf{A}}} \sum_{(i,j)\in\Omega}(d_{ij} - \widehat{a}_{ij})^2 \\ &\widehat{a}_{ij} = \mathbf{w}_i^T\mathbf{x}_j + m_{ij}. \end{aligned} \tag{14}$$

In this study, we solve the optimization problem in (14) by the means of commonly applied gradient descent algorithm.

B. Variational Bayesian Principal Component Analysis (VBPCA)

In the previous section, we discussed the least squares method to obtain the low-rank approximation $\widehat{\mathbf{A}}$ of matrix $\mathbf{A}$ from the incomplete network profile matrix $\mathbf{D}$. However, the least squares approach is prone to over-fitting [32]. The problem of over-fitting can be avoided by using probabilistic methods to perform PCA on incomplete matrices.

We apply VBPCA to estimate missing speed data in the incomplete network profile matrix $\mathbf{D}$. VBPCA is more resilient to over-fitting in comparison with other probabilistic methods such as probabilistic principal component analysis (PPCA) and maximum a posteriori PCA (MAPPCA) [32]. In this study, we apply a variant of VBPCA proposed by Ilin and Raiko [32], which they termed as VBPCAd. This approach has faster

convergence rates as opposed to traditional VBPCA implementation [32].

VBPCA avoids the problem of overfitting by penalizing complex representation of data. Thus it has a built-in mechanism for rank regularization. However, this rank selection approach can sometimes lead to suboptimal solutions (local minima) [32]. Secondly, the network profile matrix $\mathbf{A}$ (or the incomplete profile matrix $\mathbf{D}$) is not a low-rank matrix in the strict sense. In Section IV-A, we will discuss the effect of the number of latent factors on the imputation performance of the algorithm.

C. Fixed Point Continuation With Approximate Singular Value Decomposition (FPCA)

In this section, we discuss an alternative way to estimate the missing traffic information. We aim to recover these missing speed values in the incomplete data matrix $\mathbf{D}$ by utilizing the common traffic behavior across different roads $\{s_i\}_{i=1}^{p}$. To this end, we need to obtain a suitable low-rank approximation $\widehat{\mathbf{A}}$ from the incomplete speed data $\{d_{ij}\}_{(i,j)\in\Omega}$. Furthermore, the estimated network profile $\widehat{\mathbf{A}}$ should also conserve the speed information already available in the incomplete data matrix $\mathbf{D}$ within a certain tolerance limit ε, such that $\{|\widehat{a}_{ij} - d_{ij}| < \varepsilon\}_{(i,j)\in\Omega}$. Hence, we can setup the optimization problem as follows:

$$\begin{aligned} &\min \text{rank}(\widehat{\mathbf{A}}) \\ &\text{s.t.} \quad |\widehat{a}_{ij} - d_{ij}| < \varepsilon, \quad \forall\,(i,j) \in \Omega. \end{aligned} \tag{15}$$

The above mentioned optimization problem tries to recover the missing speed data with the smallest number of latent components while preserving the speed information provided by the observed data $\{d_{ij}\}_{(i,j)\in\Omega}$. However, this is a non-convex and NP-hard problem [36], [37]. To make the problem tractable, we can replace $\text{rank}(\widehat{\mathbf{A}})$ by its convex envelope, which turns out to be the nuclear norm $\|\widehat{\mathbf{A}}\|_*$ of the estimated matrix $\widehat{\mathbf{A}}$ [36]. This way, the problem in (15) can be reformulated as:

$$\begin{aligned} &\min \|\widehat{\mathbf{A}}\|_* \\ &\text{s.t.} \quad |\widehat{a}_{ij} - d_{ij}| < \varepsilon, \quad \forall\ (i,j) \in \Omega \end{aligned} \tag{16}$$

where the nuclear norm of the matrix $\widehat{\mathbf{A}}$ of rank r is defined as $\|\widehat{\mathbf{A}}\|_* = \sum_{i=1}^{r} \sigma_i$, and σ_i is the ith singular value of the matrix $\widehat{\mathbf{A}}$. We consider fixed point continuation with approximate singular value decomposition (FPCA) to solve the optimization problem defined in (16) [31].

D. Tensor Decomposition

So far, we have discussed different matrix completion methods to extract the underlying traffic patterns in road networks. However, these methods cannot efficiently utilize multi-way dependencies in traffic data sets. For instance, consider the behavior of road traffic during different days of the week. Naturally, traffic parameters such as speed tend to follow similar daily patterns [38]. These temporal relationships can be extracted in a more efficient manner by creating a multi-way structure for traffic data. To this end, we represent the speed data in the form of a 3-way tensor $\underline{\mathbf{A}} \in \mathbb{R}^{n\times p\times q}$. This tensor profile is obtained by stacking together the network profile matrices $\{\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_q\}$ from different days. Canonical polyadic (CP) decomposition is commonly used to obtain low-rank approximations for tensors [39]. For the incomplete tensor profile $\underline{\mathbf{D}}$, we can obtain a suitable low-rank approximation $\widehat{\underline{\mathbf{A}}}$ by minimizing the reconstruction error for the observed speed data in the following manner:

$$\begin{aligned} &\min_{\widehat{\underline{\mathbf{A}}}} \frac{1}{2} \left\| \underline{\mathbf{W}} \circ (\underline{\mathbf{D}} - \widehat{\underline{\mathbf{A}}}) \right\|_F^2 \\ &\widehat{\underline{\mathbf{A}}} = \sum_{i=1}^{r} \mathbf{b}_i^{(1)} \otimes \mathbf{b}_i^{(2)} \otimes \mathbf{b}_i^{(3)} \end{aligned} \tag{17}$$

where $\mathbf{b}_i^{(m)}$ is the ith column vector of mode-m factor matrix $\mathbf{B}^{(m)}$ [33], [39]. In (17), the symbol $\otimes$ denotes the vector outer product, whereas the symbol $\circ$ represents element wise multiplication between two tensors [39]. The factor matrices $\mathbf{B}^{(1)}$, $\mathbf{B}^{(2)}$ and $\mathbf{B}^{(3)}$ contain the common traffic patterns across different modes of the tensor. These patterns include common traffic behavior across different days and between different roads. We apply CP weighted optimization (CP-WOPT) to obtain a suitable estimation $\widehat{\underline{\mathbf{A}}}$ from the incomplete network profile tensor $\underline{\mathbf{D}}$. We refer to this technique as CP (3D).

We also apply CP-WOPT on the unfolded tensor to study the impact of multi-way representation on the imputation performance. To this end, we create another network profile matrix $\mathbf{U} \in \mathbb{R}^{n\times pq} | \mathbf{U} = [\mathbf{A}_1, \ldots, \mathbf{A}_q]$ by combining speed data from multiple days. This network profile matrix $\mathbf{U}$ is essentially an unfolded representation of the network profile tensor $\underline{\mathbf{A}}$. In this case, the corresponding incomplete data matrix is represented by $\mathbf{D}_u$. Similar to CP (3D), the low-rank approximation $\widehat{\mathbf{U}}$ of the matrix $\mathbf{U}$ from the incomplete speed data $\mathbf{D}_u$ is obtained by minimizing the reconstruction error for the observed speed data:

$$\begin{aligned} &\min_{\widehat{\mathbf{U}}} \frac{1}{2} \left\| \mathbf{W} \circ (\mathbf{D}_u - \widehat{\mathbf{U}}) \right\|_F^2 \\ &\widehat{\mathbf{U}} = \sum_{i=1}^{r} \mathbf{b}_i^{(1)} \otimes \mathbf{b}_i^{(2)}. \end{aligned} \tag{18}$$

We apply CP-WOPT to obtain the estimated network profile matrix $\widehat{\mathbf{U}}$. This formulation will be referred to as CP (unfold).

IV. Results and Discussion

A. Latent Factors

In this section, we discuss the impact of the choice of rank (number of latent factors) on the imputation performance of the proposed algorithms. Fig. 1 shows the variations in reconstruction performance of different algorithms caused by the choice of rank for speed data obtained from expressways. Fig. 2 shows these variations for speed data obtained from major arterial roads. Let us first discuss the performance of LS, CP (3D) and CP (Unfold). These three methods try to extract common patterns in data by minimizing the mean squared error for the observed speed information. For large percentages of missing data, the reconstruction error of these algorithms can vary significantly, depending upon the choice of rank. Furthermore, the imputation performance of these algorithms fluctuates more in the case of arterial roads as compared to expressways

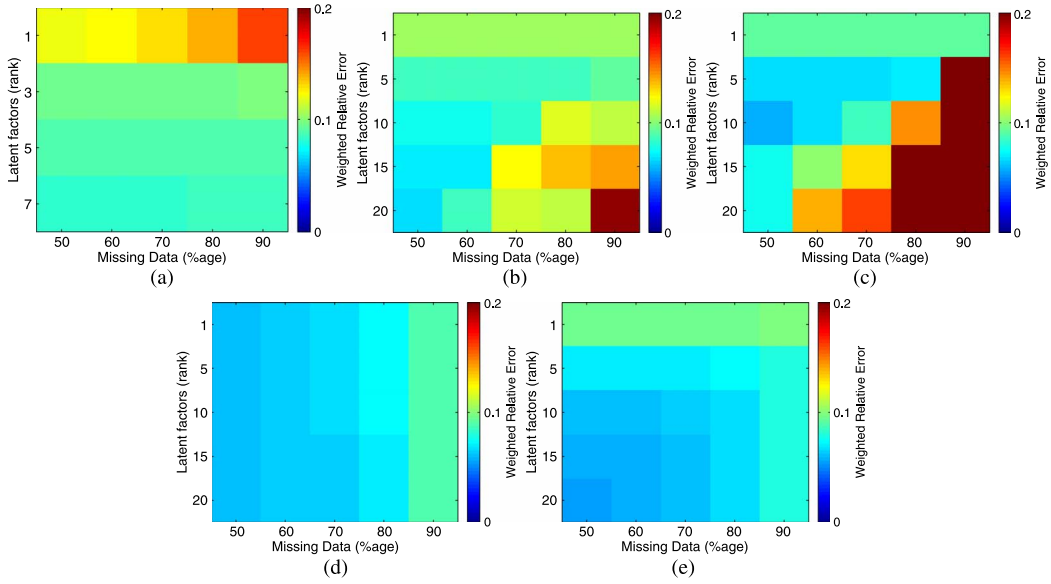

Fig. 1. Weighted relative error by considering different number of latent factors (rank) for different percentages of missing data. The test network is composed of expressways (CATA). (a) CP (3D). (b) CP (Unfold). (c) LS. (d) FPCA. (e) VBPCA.

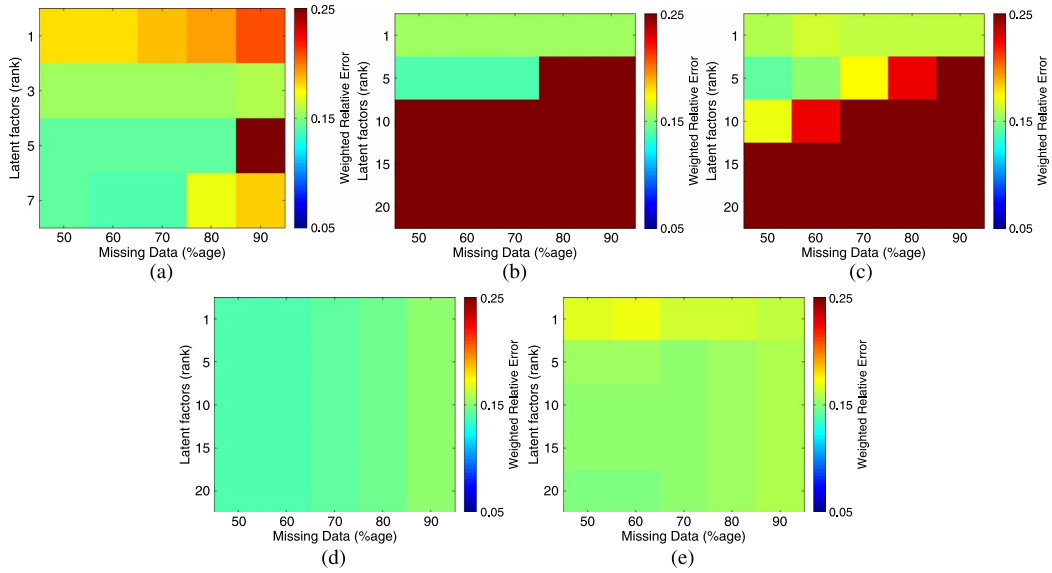

Fig. 2. Weighted relative error by considering different number of latent factors (rank) for different percentages of missing data. The test network is composed of major arterial roads (CATB). (a) CP (3D). (b) CP (Unfold). (c) LS. (d) FPCA. (e) VBPCA.

(see Figs. 1 and 2). On the other hand, the reconstruction error for FPCA and VBPCA does not vary significantly for different rank values. The rank values for VBPCA in Figs. 1(e) and 2(e) represent the limit on the maximum number of factors that can be used to reconstruct the estimated network profile matrix $\widehat{\mathbf{A}}$. VBPCA can automatically choose the optimal number of factors while estimating missing values in the incomplete speed data matrix $\mathbf{D}$. Hence, it might be tempting to assume that the information about the maximum number of factors (rank) is redundant. However, this assumption does not hold in all cases. Fig. 3 shows the impact of setting a limit on the maximum number of latent factors for VBPCA. The reconstruction error is shown for the scenario when 90% of speed data is missing. We can conclude that VBPCA is also prone to overfitting if a suitable cut-off value for the rank is not available.

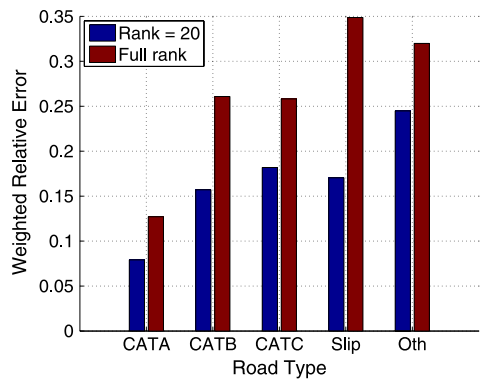

Fig. 3. Impact of restricting the maximum number of factors on the imputation accuracy of VBPCA. The WRE is calculated for the case when 90% of speed data is missing.

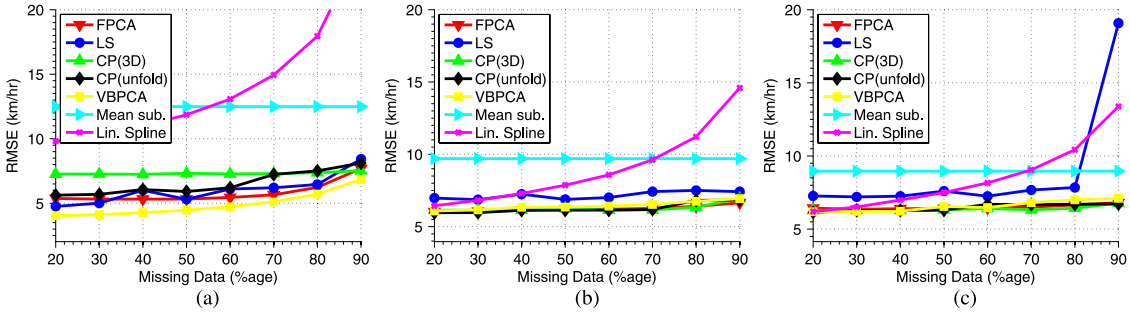

Fig. 4. RMSE of the proposed algorithms for different percentages of missing data and various road networks. (a) Expressway (CATA). (b) Major arterial roads. (CATB). (c) Minor arterial roads (CATC).

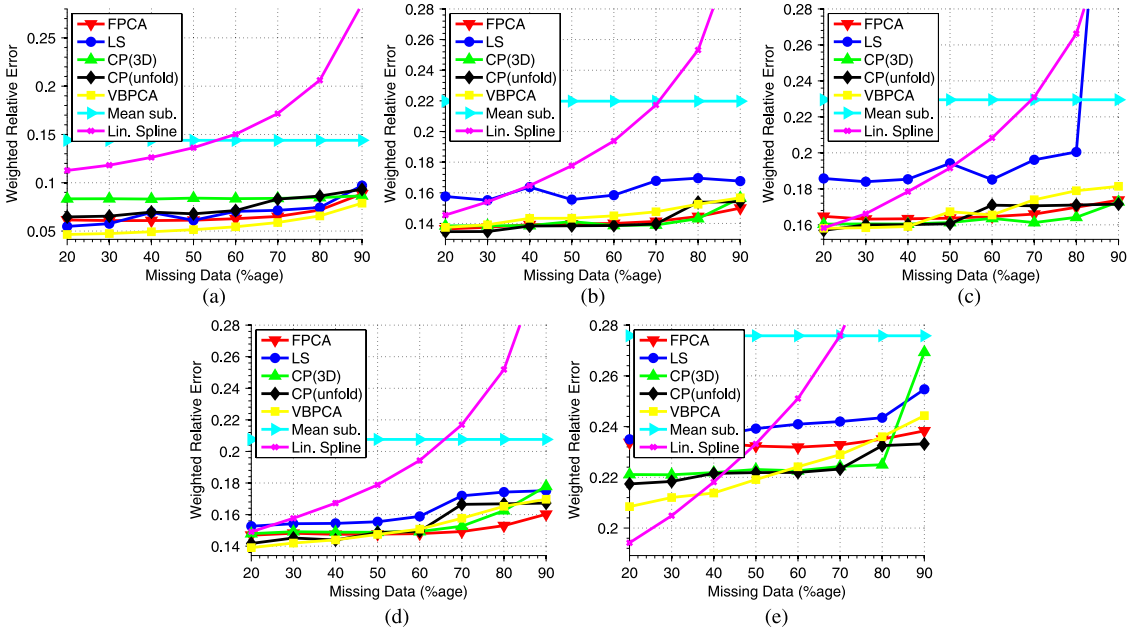

Fig. 5. Weighted relative error of the proposed algorithms for different percentages of missing data and various road networks. (a) Expressway (CATA). (b) Major arterial roads (CATB). (c) Minor arterial roads (CATC). (d) Slip roads. (e) Access roads.

B. Performance Analysis

In this section, we analysis the performance of various matrix and tensor completion methods for various road networks. As a baseline, we also estimate missing values using mean substitution and linear splines. These are commonly used baseline techniques for evaluating the performance of matrix/tensor completion methods [4].

Let us start by analyzing the estimation accuracy of the proposed algorithms in terms of weighted relative error (WRE) and RMSE. Figs. 4 and 5 show the imputation accuracy of these methods for different road types. For expressways, VBPCA achieves the lowest WRE followed by FPCA. The imputation error for expressways is lower for all the algorithms as compared to other road categories. For major and minor arterial roads, CP (3D) and FPCA provide slightly better performance as compared to other methods (see Fig. 5(b) and (c)). FPCA also achieves better performance for slip roads (see Fig. 5(d)). In the case of access roads, all the algorithms suffered from large imputation error.

CP (3D), CP (Unfold) and LS all try to impute missing values by finding those traffic patterns that can minimize the reconstructed squared error for the observed speed data $\{(d_{ij} - \widehat{a}_{ij})^2\}_{(i,j)\in\Omega}$. Out of these three least squared based methods, multi-way representation (tensor method) tends to achieve the best performance. Furthermore, for arterial roads and access roads, multi-way representation (tensor method) also achieves better imputation accuracy than other methods such as FPCA and VBPCA. However, in case of expressways, the advantage of considering multi-way representation is not that apparent (see Fig. 5(a)). It seems that tensor representation is more useful for smaller roads where traffic behaves more erratically. In such cases, multi-way representation of speed data is an efficient way to extract underlying traffic patterns.

Let us now analyze the performance of different imputation methods across the week. Fig. 6 shows the imputation error of FPCA, LS and VBPCA during different days of the week. The results are shown for different road categories with speed data obtained from August 1, 2011 to August 7, 2011. In this scenario, the missing data percentage was 70%. For expressways, VBPCA has lower WRE as compared to other methods during most of the days. This is expected as VBPCA has the lowest overall imputation error for speed data obtained

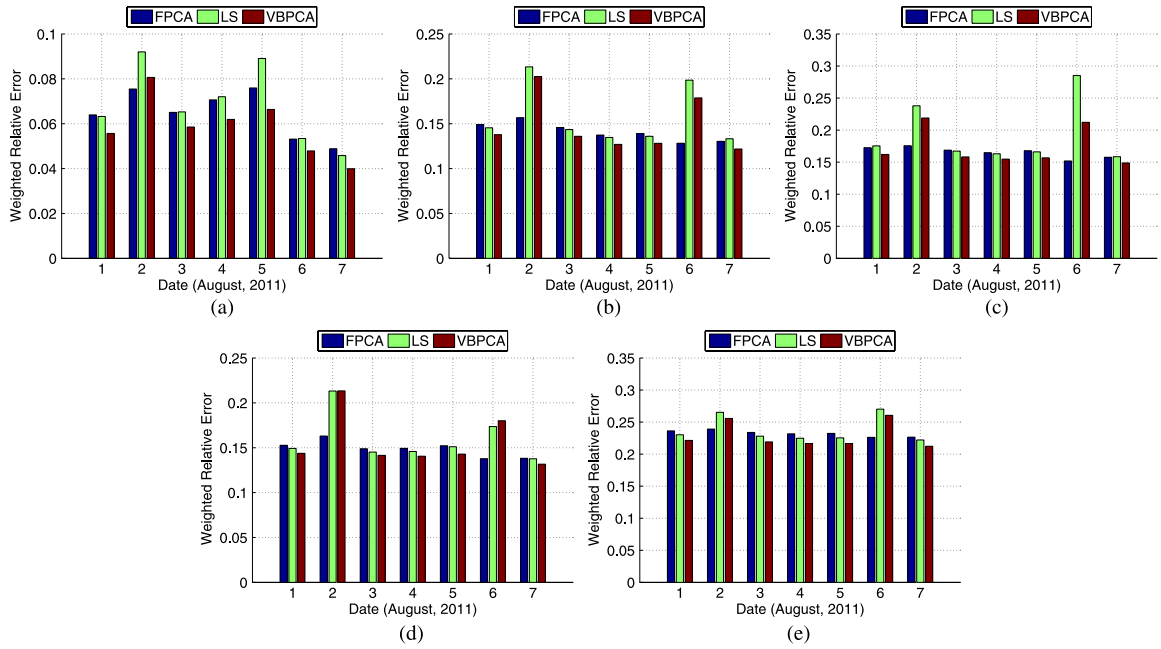

Fig. 6. Weighted relative error of the proposed algorithms during different days of the week for various road networks. The reconstruction error is for the case when 70% of data is missing. (a) Expressway (CATA). (b) Major arterial roads (CATB). (c) Minor arterial roads (CATC). (d) Slip roads. (e) Access roads.

TABLE II
VARIANCE OF THE IMPUTED SPEED DATA FOR DIFFERENT ROAD TYPES. THE UNITS FOR VARIANCE ARE km^2/hr^2. THE VALUES IN THE BRACKETS REPRESENT THE PERCENTAGE VARIANCE OF IMPUTED SPEED DATA w.r.t. THE ACTUAL SPEED DATA

Road Type	Missing Data	Variance (%age of actual variance)				
		FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
CATA Var = 153.73	20 %	118.30 (77)	128.47 (84)	94.39 (61)	117.95 (77)	128.75 (84)
	30 %	114.02 (74)	127.15 (83)	95.16 (62)	118.95 (77)	129.53 (84)
	40 %	112.64 (73)	109.53 (71)	95.37 (62)	114.43 (74)	129.13 (84)
	70 %	116.99 (76)	111.29 (72)	95.10 (62)	117.09 (76)	124.84 (81)
	80 %	112.10 (73)	114.81 (75)	95.52 (62)	96.37 (63)	121.89 (79)
Road Type	Missing Data	FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
CATB Var = 163.26	20 %	132.30 (81)	120.72 (74)	119.86 (73)	122.91 (75)	123.20 (75)
	30 %	128.66 (79)	121.22 (74)	119.60 (73)	122.83 (75)	122.98 (75)
	40 %	127.95 (78)	117.20 (72)	120.01 (74)	120.41 (74)	122.68 (75)
	70 %	128.76 (79)	106.49 (65)	119.95 (73)	120.61 (74)	118.31 (72)
	80 %	127.01 (78)	107.63 (66)	118.13 (72)	109.69 (67)	116.15 (71)
Road Type	Missing Data	FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
CATC Var = 112.15	20 %	79.28 (71)	74.91 (67)	67.87 (61)	71.10 (63)	73.78 (66)
	30 %	75.68 (67)	81.62 (73)	67.80 (60)	68.60 (61)	73.20 (65)
	40 %	80.67 (72)	67.78 (60)	67.70 (60)	68.68 (61)	72.90 (65)
	70 %	77.50 (69)	60.80 (54)	67.80 (60)	61.73 (55)	66.79 (60)
	80 %	75.11 (67)	62.10 (55)	66.38 (59)	61.81 (55)	63.95 (57)
Road Type	Missing Data	FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
Slip Var = 419.41	20 %	361.91 (86)	332.86 (79)	338.79 (81)	350.01 (83)	348.07 (83)
	30 %	366.98 (87)	333.50 (80)	337.95 (81)	347.75 (83)	347.32 (83)
	40 %	364.52 (87)	332.88 (79)	338.43 (81)	351.49 (84)	344.75 (82)
	70 %	355.76 (85)	315.75 (75)	342.67 (82)	312.50 (75)	336.64 (80)
	80 %	351.26 (84)	318.51 (76)	343.60 (82)	313.48 (75)	331.79 (79)
Road Type	Missing Data	FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
Oth Var = 205.56	20 %	132.67 (65)	115.86 (56)	109.47 (53)	118.33 (58)	123.75 (60)
	30 %	130.58 (64)	124.72 (61)	109.36 (53)	116.34 (57)	122.31 (59)
	40 %	130.42 (63)	109.35 (53)	109.29 (53)	110.28 (54)	121.10 (59)
	70 %	124.18 (60)	95.14 (46)	108.29 (53)	111.71 (54)	108.81 (53)
	80 %	119.51 (58)	96.87 (47)	107.93 (53)	99.67 (48)	105.20 (51)

from expressways (see Fig. 5(a)). For arterial roads, all three methods have similar performance during most of the days (see Fig. 6(b) and (c)). However, VBPCA and LS tend to suffer from large estimation error on certain days. On the other hand, the estimation performance of FPCA does not vary significantly from one day to another. We also observe similar trend in the performances of LS, FPCA and VBPCA for slip roads (see Fig. 6(d)). For primary and local access roads, all three

TABLE III
BIAS OF THE IMPUTED SPEED DATA FOR DIFFERENT ROAD TYPES. THE UNITS FOR BIAS ARE km/hr

Road Type	Missing Data	Bias				
		FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
CATA Avg. Speed = 86 km/hr	20 %	0.008	0.065	0.003	-0.009	0.065
	30 %	0.012	0.062	-0.006	-0.006	0.063
	40 %	0.008	0.069	-0.011	-0.004	0.068
	70 %	0.006	0.074	0.009	0.003	0.066
	80 %	0.004	0.062	-0.003	-0.010	0.064
Road Type	Missing Data	FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
CATB Avg. Speed = 43 km/hr	20 %	0.009	0.434	-0.002	0.004	0.409
	30 %	0.009	0.401	0.005	0.000	0.402
	40 %	0.004	0.465	0.014	0.008	0.425
	70 %	-0.004	0.377	0.003	0.007	0.360
	80 %	-0.006	0.377	0.004	0.018	0.344
Road Type	Missing Data	FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
CATC Avg. Speed = 38 km/hr	20 %	-0.007	0.478	0.015	0.013	0.400
	30 %	0.001	0.348	-0.006	-0.008	0.379
	40 %	0.000	0.416	-0.005	0.006	0.355
	70 %	-0.003	0.418	0.007	0.016	0.360
	80 %	-0.029	0.430	-0.006	0.001	0.334
Road Type	Missing Data	FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
Slip Avg. Speed = 55 km/hr	20 %	0.011	0.370	-0.012	-0.010	0.400
	30 %	0.011	0.385	-0.012	-0.019	0.406
	40 %	0.003	0.395	0.020	0.010	0.416
	70 %	-0.003	0.418	0.010	0.050	0.439
	80 %	-0.002	0.419	0.013	0.047	0.476
Road Type	Missing Data	FPCA	LS	CP (3D)	CP (Unfold)	VBPCA
Oth Avg. Speed = 41 km/hr	20 %	-0.023	0.446	-0.004	-0.005	0.391
	30 %	-0.025	0.423	0.021	0.016	0.419
	40 %	-0.024	0.411	-0.003	-0.002	0.386
	70 %	-0.022	0.361	0.005	0.003	0.367
	80 %	-0.031	0.359	-0.001	0.012	0.361

methods reported large imputation error during all seven days (see Fig. 6(e)). We can conclude that imputation performance of FPCA is more robust to daily variations in traffic conditions in comparison with other methods such as LS and VBPCA.

Table II shows the variance of the estimated data for different road types. It also shows the variance of the actual speed data. As expected, the imputation algorithms underestimate the variance of the imputed data. For instance, the actual variance of the speed data obtained from expressways was around 153 km^2/hr^2. However, the variance of the speed data obtained from different imputation methods was around 100–130 km^2/hr^2. Moreover, the difference between the variance of actual and imputed speed data becomes larger as the percentage of missing data increases. For expressways, VBPCA provided the best estimate of the variance in the speed data. For other road types such as arterial roads (CATA, CATB), access roads and slip roads, the variance of the imputed data obtained from FPCA was the closest to the variance of the actual speed data. Nonetheless, all the five methods had comparable performance in terms of conserving the variance of the speed data.

Table III shows the bias induced in the recovered speed data by various proposed methods. The results show that the proposed algorithms do not add significant bias in the imputed data as the bias-value remains less than 1 km/hr for all test cases. Still, the imputed speed data obtained from VBPCA and LS had slightly higher bias (≈ 0.5 km/hr) in comparison with other methods.

V. CONCLUSION

Missing data is a common problem faced by many transportation management systems. In this paper, we compared various methods to estimate missing traffic information in data sets obtained from large road networks. To this end, we extracted common global patterns from incomplete speed data by applying various tensor and matrix completion algorithms such as FPCA, VBPCA, LS and CP-WOPT. Matrix and tensor completion methods have been previously applied to solve the problem of missing data in transportation systems. However, these studies typically consider small test networks comprising of a few roads. Furthermore, the performance of these methods for different road categories as well as during different days of the week is usually not analyzed.

In this study, we considered five large test networks each comprising of around 1500 road segments. We analyzed the reconstruction accuracy of various matrix and tensor completion methods for different types of roads as well as during different days of the week. We also analyzed the impact of the choice of latent factors on the estimation accuracy of recovered speed data. Moreover, we also compared the variance and bias induced in the imputed speed data by the proposed methods.

The results show that the performance of least square based methods is highly sensitive to the choice of rank as compared to VBPCA and FPCA. FPCA is particularly useful for imputation of traffic data sets as its performance is least sensitive to daily

variations in traffic data. Furthermore, it also provides better or comparable performance to other algorithms for different road categories. In the future, we plan to develop ensemble methods that combine the outputs of the algorithms considered in the paper for various scenario-specific applications.

REFERENCES

[1] P.-A. Laharotte *et al.*, "Spatiotemporal analysis of Bluetooth data: Application to a large urban network," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1439–1448, Jun. 2015.

[2] J. Rodrigues, A. Aguiar, F. Vieira, J. Barros, and J. Cunha, "A mobile sensing architecture for massive urban scanning," in *Proc. 14th IEEE ITSC*, Oct. 2011, pp. 1132–1137.

[3] B. Płaczek, "Selective data collection in vehicular networks for traffic control applications," *Transp. Res. C, Emerging Technol.*, vol. 23, pp. 14–28, Aug. 2012.

[4] L. Qu, L. Li, Y. Zhang, and J. Hu, "PPCA-based missing data imputation for traffic flow volume: A systematical approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 512–522, Sep. 2009.

[5] T. Hunter, T. Das, M. Zaharia, P. Abbeel, and A. Bayen, "Large-scale estimation in cyberphysical systems using streaming data: A case study with arterial traffic estimation," *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 4, pp. 884–898, Oct. 2013.

[6] T. Cheng, G. Tanaksaranond, C. Brunsdon, and J. Haworth, "Exploratory visualisation of congestion evolutions on urban transport networks," *Transp. Res. C, Emerging Technol.*, vol. 36, pp. 296–306, Nov. 2013.

[7] Y. Wang, Y. Zhu, Z. He, Y. Yue, and Q. Li, "Challenges and opportunities in exploiting large-scale GPS probe data," HP Labs., Palo Alto, CA, USA, Tech. Rep. HPL-2011-109, vol. 21, 2011.

[8] T. Hunter *et al.*, "Scaling the mobile millennium system in the cloud," in *Proc. 2nd ACM SOCC*, Cascais, Portugal, 2011, pp. 28:1–28:8.

[9] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya, "Detecting errors and imputing missing data for single-loop surveillance systems," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1855, no. 1, pp. 160–167, 2003.

[10] B. L. Smith, W. T. Scherer, and J. H. Conklin, "Exploring imputation techniques for missing data in transportation management systems," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1836, no. 1, pp. 132–142, 2003.

[11] D. Ni, J. D. Leonard, A. Guin, and C. Feng, "Multiple imputation scheme for overcoming the missing values and variability issues in its data," *J. Transp. Eng.*, vol. 131, no. 12, pp. 931–938, Dec. 2005.

[12] J. Conklin and B. Smith, "The use of local lane distribution patterns for the estimation of missing data in transportation management systems," *Transp. Res. Rec.*, vol. 1811, pp. 50–56, 2002.

[13] H. Tan *et al.*, "A tensor-based method for missing traffic data completion," *Transp. Res. C, Emerging Technol.*, vol. 28, pp. 15–27, Mar. 2013.

[14] M. Zhong, P. Lingras, and S. Sharma, "Estimation of missing traffic counts using factor, genetic, neural, and regression techniques," *Transp. Res. C, Emerging Technol.*, vol. 12, no. 2, pp. 139–166, Apr. 2004.

[15] C. Chen, Y. Wang, L. Li, J. Hu, and Z. Zhang, "The retrieval of intra-day trend and its influence on traffic prediction," *Transp. Res. C, Emerging Technol.*, vol. 22, pp. 103–118, Jun. 2012.

[16] S. Sharma, P. Lingras, and M. Zhong, "Effect of missing values estimations on traffic parameters," *Transp. Planning Technol.*, vol. 27, no. 2, pp. 119–144, Apr. 2004.

[17] Y. Zhang and Y. Liu, "Missing traffic flow data prediction using least squares support vector machines in urban arterial streets," in *Proc. Symp. CIDM*, Mar. 2009, pp. 76–83.

[18] L. Qu, Y. Zhang, J. Hu, L. Jia, and L. Li, "A BPCA based missing value imputing method for traffic flow volume data," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2008, pp. 985–990.

[19] L. Li, Y. Li, and Z. Li, "Efficient missing data imputing for traffic flow by considering temporal and spatial dependence," *Transp. Res. C, Emerging Technol.*, vol. 34, pp. 108–120, Sep. 2013.

[20] G. Chang and T. Ge, "Comparison of missing data imputation methods for traffic flow," in *Proc. Int. Conf. TMEE*, Dec. 2011, pp. 639–642.

[21] Z. Li, Y. Zhu, H. Zhu, and M. Li, "Compressive sensing approach to urban traffic sensing," in *Proc. 31st ICDCS*, Jun. 2011, pp. 889–898.

[22] M. T. Asif, N. Mitrovic, L. Garg, J. Dauwels, and P. Jaillet, "Low-dimensional models for missing data imputation in road networks," in *Proc. IEEE ICASSP*, May 2013, pp. 3527–3531.

[23] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transp. Res. C, Emerging Technol.*, vol. 19, no. 4, pp. 606–616, Aug. 2011.

[24] J. Dauwels *et al.*, "Predicting traffic speed in urban transportation sub-networks for multiple horizons," in *Proc. 13th ICARCV*, Dec. 2014, pp. 547–552.

[25] M. T. Asif, K. Srinivasan, N. Mitrovic, J. Dauwels, and P. Jaillet, "Near-lossless compression for large traffic networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1817–1826, Aug. 2015.

[26] Y. Han and F. Moutarde, "Analysis of network-level traffic states using locality preservative non-negative matrix factorization," in *Proc. 14th IEEE ITSC*, 2011, pp. 501–506.

[27] A. Hofleitner *et al.*, "Large-scale estimation of arterial traffic and structural analysis of traffic patterns from probe vehicles," in *Proc. Transp. Res. Board 91st Annu. Meet.*, 2012, vol. 12-0598, pp. 1–24.

[28] C. Furtlehner *et al.*, "Spatial and temporal analysis of traffic states on large scale networks," in *Proc. 13th IEEE ITSC*, Sep. 2010, pp. 1215–1220.

[29] N. Mitrovic, M. T. Asif, J. Dauwels, and P. Jaillet, "Low-dimensional models for compressed sensing and prediction of large-scale traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2949–2954, Oct. 2015.

[30] M. Zhong, S. Sharma, and P. Lingras, "Genetically designed models for accurate imputation of missing traffic counts," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1879, no. 1, pp. 71–79, 2004.

[31] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and Bregman iterative methods for matrix rank minimization," *Math. Programm.*, vol. 128, no. 1/2, pp. 321–353, Jun. 2011.

[32] A. Ilin and T. Raiko, "Practical approaches to principal component analysis in the presence of missing values," *J. Mach. Learn. Res.*, vol. 11, pp. 1957–2000, 2010.

[33] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometr. Intell. Lab. Syst.*, vol. 106, no. 1, pp. 41–56, Mar. 2011.

[34] A. Cichocki and S.-i. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, vol. 1. Hoboken, NJ, USA: Wiley, 2002.

[35] T. Wiberg, "Computation of principal components when data are missing," in *Proc. 2nd Symp. Comput. Stat.*, 1976, pp. 229–236.

[36] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Aug. 2010.

[37] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.

[38] S. Yang, K. Kalpakis, and A. Biem, "Spatio-temporal coupled Bayesian robust principal component analysis for road traffic event detection," in *Proc. 16th IEEE ITSC*, Oct. 2013, pp. 392–398.

[39] T. Kolda and B. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

Muhammad Tayyab Asif (S'12) received the B.Sc. degree in electrical engineering from University of Engineering and Technology Lahore, Lahore, Pakistan. Previously, he was with Ericsson, working in the domain of mobile packet core networks. He is currently working toward the Ph.D. degree with the School of Electrical and Electronic Engineering, College of Engineering, Nanyang Technological University, Singapore. His research interests include sensor fusion, network optimization, and modeling of large-scale networks.

Nikola Mitrovic (S'14) received the bachelor's degree in traffic engineering from University of Belgrade, Belgrade, Serbia, in 2009 and the master's degree in civil engineering from Florida Atlantic University, Boca Raton, FL, USA, in 2010. He is currently working toward the Ph.D. degree with the School of Electrical and Electronic Engineering, College of Engineering, Nanyang Technological University, Singapore. His research topics are traffic modeling, intelligent transportation systems, and transportation planning.

Justin Dauwels (M'09–SM'12) received the Ph.D. degree in electrical engineering from Swiss Federal Institute of Technology in Zurich (ETH), Zurich, Switzerland, in 2005. In 2008–2010, he was a Research Scientist with Massachusetts Institute of Technology. He is currently an Assistant Professor with the School of Electrical and Electronic Engineering, College of Engineering, Nanyang Technological University (NTU), Singapore. He is the Deputy Director of the ST Engineering-NTU Corporate Laboratory on Autonomous Systems. His research on intelligent transportation systems (ITSs) has been featured by the BBC, Straits Times, and various other media outlets. His research on Alzheimer's disease was featured at a five-year exposition at the Science Center in Singapore. He has filed five U.S. patents related to data analytics. His research interests are in Bayesian statistics, iterative signal processing, and computational neuroscience. He was a Postdoctoral Fellow of the RIKEN Brain Science Institute (2006–2007). He has been a JSPS Postdoctoral Fellow (2007), a BAEF Fellow (2008), a Henri Benedictus Fellow of the King Baudouin Foundation (2008), and a JSPS Invited Fellow (2010 and 2011). His research team has won several best paper awards at international conferences.

Patrick Jaillet is the Dugald C. Jackson Professor with the Department of Electrical Engineering and Computer Science and a member of the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. He is also a Codirector of the MIT Operations Research Center. He is also with Centre for Future Urban Mobility, Singapore–MIT Alliance for Research and Technology, Singapore. His research interests include online optimization, online learning, and data-driven optimization with applications to transportation and to the Internet economy. He is a Fellow of INFORMS.