LOW-DIMENSIONAL MODELS FOR MISSING DATA IMPUTATION IN ROAD NETWORKS

Muhammad Tayyab Asif [1], *Nikola Mitrovic* [1], *Lalit Garg* [1,2], *Justin Dauwels* [1], *Patrick Jaillet* [3,4]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]Faculty of Information and Communication Technology, University of Malta, Malta
[3]Laboratory for Information and Decision Systems, MIT, Cambridge, MA
[4]Center for Future Urban Mobility, Singapore-MIT Alliance for Research and Technology, Singapore

ABSTRACT

Intelligent transport systems (ITS) require data with high spatial and temporal resolution for applications such as modeling, traffic management, prediction and route guidance. However, field data is usually quite sparse. This problem of missing data severely limits the effectiveness of ITS. Missing values are usually imputed by either using historical data of the road or current information from neighboring links. In most scenarios, information from some or all of neighboring links might not be available. Furthermore, historical data may also be incomplete. To overcome these issues, we propose methods which can construct low-dimensional representation of large and diverse networks, in presence of missing historical and neighboring data. We use these low-dimensional models to reconstruct data profiles for road segments, and impute missing values. To this end we use Fixed Point Continuation with Approximate SVD (FPCA) and Canonical Polyadic (CP) decomposition for incomplete tensors to solve the problem of missing data. We apply these methods to expressways and a large urban road network to assess their performance for different scenarios.

Index Terms— Missing data in large networks, low-dimensional models

1. INTRODUCTION

Data Driven Intelligent Transport Systems (D^{2}ITS) heavily rely on historical traffic data for applications such as traffic prediction, planning, management, and route guidance [1, 2]. These applications can improve the traffic conditions by avoiding potential congestions and traffic jams. The information about traffic parameters (speed, flow, travel time) is gathered by GPS probes and loop detectors. Loop detectors suffer from sparse coverage capability and high installation costs. GPS probes are cheaper. However, due to their dynamic nature, the collected data is usually sparse with highly irregular temporal resolution [3]. Consequently, the problem of missing data is prevalent in many transport management systems [4–7]. The methods employed to tackle this problem either use information from neighboring links [8, 9] or consider historical information of the road segment for imputation [5–7, 10]. These methods assume that the problem of missing data is localized to isolated links and time intervals. These assumptions are usually not valid when considering large interconnected road networks. The spatial and temporal distribution of missing data points is usually highly erratic [11]. Therefore, the methods which rely on complete historical or current information from neighbors for data imputation [5–10] may not work in such settings.

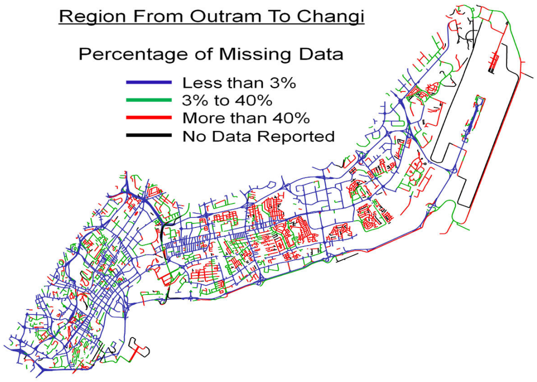


Fig. 1: Road network in Singapore (Outram to Changi).

Varying degrees of spatial-temporal correlations exist between links in urban networks [12, 13]. These relationships can be used to create low-dimensional models even for large networks [11, 14]. We propose to exploit these underlying structures for recovering missing data by reconstructing traffic profiles from low-dimensional representation of the network. Methods such as Singular Value Decomposition (SVD) and Canonical Polyadic (CP) decomposition are usually applied to find low-dimensional representation of multivariate systems. To perform the decomposition in presence of missing data, we use Fixed Point Continuation with Approximate SVD (FPCA) [15] and CP Weighted OPTimization (CP-WOPT) [16, 17]. For benchmarking, we compare their performance with Bayesian Principal Component Analysis (BPCA) [2] and historical averages. We compare the imputation accuracy of each algorithm for

The research described in this project was funded in part by the Singapore National Research Foundation (NRF) through the Singapore MIT Alliance for Research and Technology (SMART) Center for Future Mobility (FM).

heterogeneous large networks and for different percentages of missing data. We also provide comparison of computation times of the algorithms.

The paper is structured as follows. In section II, we propose different techniques for obtaining low-dimensional models for large networks in presence of missing data. In Section III, we explain the experimental setup. In section IV, we compare the accuracy and computational complexities of the models for different scenarios. In Section V, we summarize our contributions, in relation to prior work and suggest topics for future work.

2. LOW-DIMENSIONAL MODELS FOR MISSING DATA IMPUTATION

Definition 1: A road network is defined as a directed graph $\mathbf{G}$ = (N,E), where $E = \{s_i | i = 1,...,p\}$ represents the set of road segments/links.
Definition 2: Weight of edge/link s_i is represented by $z(s_i,t_j)$, which is a time varying function (average traffic speed for time interval $(t_j - t_0, t_j)$) representing the state of edge (link) at time t_j. For our study, sampling interval t_0 is 5 minutes.
In this section, based on the above definitions, we will develop data imputation methods for large road networks. Traditional formulations that use neighboring information for imputing data [8, 9] at the time t_e for the link s_i can be modeled as:

$$\hat{z}(s_i,t_e) = f_1(z(\theta_1,t_e),...z(\theta_k,t_e)) : \{\theta_j \in \Theta_{s_i}\}_{j=1}^{k}, \quad (1)$$

where $\Theta_{s_i} \subseteq E$ is the set of k neighboring links of s_i. Different methods are then applied to learn the function f_1 from historical relationships of the link s_i and its neighbors [8, 9]. In case, only the historical information of the link s_i is used [5–7, 10], we get:

$$\hat{z}(s_i,t_e) = f_2(z(s_i,\tau_1),...z(s_i,\tau_n)) : \{\tau_j \in T_{s_i}\}_{j=1}^{n}, \quad (2)$$

where T_{s_i} is the set of past similar temporal values, found in the speed profile of link s_i. To estimate $\hat{z}(s_i,t_e)$ in (1), it is assumed that $\{z(\theta_j,t_e)\}_{j=1}^{k}$ are available for imputation. Furthermore, $\{z(\theta_j,t < t_e)\}_{j=1}^{k}$ should also be reported so that f_1 can be estimated [8, 9]. Similarly in (2), enough historical data should be available to learn f_2 [5–7, 10]. In many practical scenarios, adequate historical and neighbor information is usually not available, for estimating parameters of relationship functions f_1 and f_2 [11]. For example, in this study, we consider a large subnetwork in Singapore (see Fig.1). The figure shows the percentages of missing speed data for different road segments for August, 2011, as provided by the Singapore Land Transportation Authority (LTA). It is quite evident from the figure that missing data problem is not restricted to isolated links.

To overcome such situations, we propose following methods, which can reconstruct data profiles for the whole network from low-dimensional models even in the presence of missing data.

2.1. Fixed Point Continuation with Approximate SVD

In this model, we create network profile $\mathbf{M}_G \in \mathbb{R}^{d\times p}$, for the road network G. We represent each link s_i by a speed profile $\mathbf{m}_i$, where $\{\mathbf{m}_i = [z(s_i,t_1)...z(s_i,t_d)]^T\}_{s_i \in E}$. The network profile $\mathbf{M}_G$ contains average speed values of all the p links in the network G from time t_1 to t_d, such that $\mathbf{M}_G$ = $[\mathbf{m}_1\mathbf{m}_2...\mathbf{m}_p]$. However, not all entries of $\mathbf{M}_G$ are known. Let $(i, j) \in \Phi$ be the set of entries in $\mathbf{M}_G$ for which data is available. Links in interconnected road networks exhibit strong temporal and spatial correlations [11, 13, 14]. Hence, we can model network G as a low-dimensional structure, without losing a great deal of information. Consequently, network profile $\mathbf{M}_G$ can be represented by another lower rank matrix $\mathbf{X}_G$ with minimal error. This can be easily achieved by SVD if all the entries of $\mathbf{M}_G$ are known. In the presence of set of missing values Φ, we can setup the problem as

$$\begin{aligned} &\min \ \text{rank}(\mathbf{X}_G) \\ s.t:\ &|x_{ij} - m_{ij}| \ \le \varepsilon \ \forall \ (i,j) \in \Phi. \end{aligned} \quad (3)$$

The parameter ε defines the error tolerance in case reconstructed value x_{ij} is different from the reported data value m_{ij}. However, matrix rank minimization is an NP-hard problem [11, 15]. It can be shown that convex envelope of rank($\mathbf{X}_G$) is the nuclear norm $\| \mathbf{X}_G \|_*$ of the matrix [15, 18]. So, we can redefine the problem in (3) in a more convenient manner as

$$\begin{aligned} &\min \ \| \mathbf{X}_G \|_* \\ s.t:\ &|x_{ij} - m_{ij}| \ \le \varepsilon \ \forall \ (i,j) \in \Phi, \end{aligned} \quad (4)$$

where $\| \mathbf{X}_G \|_*$ is defined as the sum of singular values $\{\sigma_i\}_{i=1}^{R}$ of $\mathbf{X}_G$ with rank R, where $R \ll p$.

We will solve the optimization problem defined in (4) using FPCA [15].

2.2. Missing Data Imputation using CP Decomposition

Traffic data often contains repetitive historical patterns. Traffic profiles for the weekdays/weekends usually show strong correlation with other weekdays/weekends. Also, there are distinct patterns during rush hours, and off peak hours [2]. Missing data imputation can also be achieved by exploiting the temporal correlations alongside spatial correlations for low-dimensional representation. Unlike other methods, which utilize temporal trends [2, 5, 6, 10], this method does not make any assumption on the distribution of missing data points, allowing us to apply it to more practical settings. In this approach, we add one more dimension to the network profile $\mathbf{M}_G$. This dimension contains data from similar day(s). For our study, we create a tensor $\mathscr{N}_G$ $\in \mathbb{R}^{d\times p\times w}$, where d is the number of speed data points in one day, and w is the number of similar days we use. The parameter p is the number of links the in network G. Similar to $\mathbf{M}_G$, $\mathscr{N}_G$ also contains instances of unreported data. Let

$(i,j,k) \in \Psi$ be the set of known tensor entries. To obtain low rank representation $\mathscr{X}_G$ of $\mathscr{N}_G$ in the presence of missing data, we use CP Weighted OPTimization(CP-WOPT) [16]. It has been shown to provide better performance as compared to competing algorithms such as INDAFAC [16]. CP-WOPT tries to minimize reconstruction error using the following formulation:

$$f_{\mathscr{N}}(\mathbf{A},\mathbf{B},\mathbf{C}) = \frac{1}{2}\sum\nolimits_{(i,j,k)\in\Psi}\Big(n_{ijk} - \sum_{r=1}^{R} a_{ir}b_{jr}c_{kr}\Big)^2, \qquad (5)$$

where n_{ijk} is the reported data value and a_{ir}, b_{jr}, c_{kr} represent the entries of factor matrices $\mathbf{A}_{d\times R}$, $\mathbf{B}_{p\times R}$ and $\mathbf{C}_{w\times R}$ respectively [16]. We also consider $\mathbf{M}_G \in \mathbb{R}^{d\times p}$ as a low-dimensional tensor, to observe the effect of additional information in $\mathscr{N}_G$ as opposed to $\mathbf{M}_G$. We use CP-WOPT to find low rank representation for $\mathbf{M}_G$ for data imputation. We refer to this as Low Dimensional CP Weighted OPTimization (LDCP-WOPT).

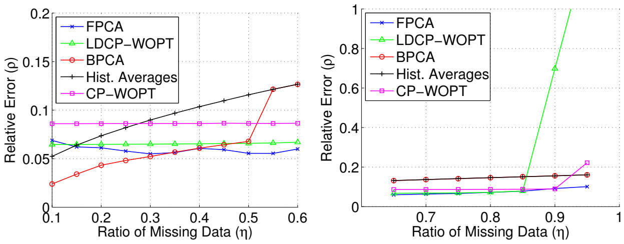
2.3. Missing Data Imputation using BPCA

BPCA has been previously applied to small road networks for traffic flow data imputation. It has been shown to provide superior imputation accuracy against competing methods [2]. However, those studies assumed that missing data locations only had temporal dependence, isolated to each link [2, 10]. In practice, such assumptions might not be valid [1, 11], particularly for speed data collected from taxi probes. Moreover, missing data percentage of only up to 50% was considered, where as field data may also contain higher proportion of missing data (see Fig.1). In this study, we assess the performance of BPCA by applying it to the network profile $\mathbf{M}_G$ of a large network G. Similar to above mentioned approaches, we use BPCA to find the low rank representation $\mathbf{X}_G$ of the network profile $\mathbf{M}_G$, in presence of missing data to impute missing values.

3. EXPERIMENTAL SETUP

In this section, we explain the data set we used, to compare the imputation accuracy and computational times of above mentioned algorithms. We use two scenarios to assess their performance.

In the first scenario, we consider a network G_1 comprising of three connected expressways (Pan Island Expressway, East Coast Parkway and Kallang Paya Lebar Expressway) in Singapore, spanning from Outram park to Changi. For analysis, we use data provided by LTA for the month of August, 2011. The data contains averaged space speed values per five minute interval for each individual road segment $\{s_i\}_{i=1}^{p}$. Fig.1 shows the percentage of missing data for each link. For analysis, we only consider those links for which missing data percentage was less than 3%. In this way, we can calculate imputation accuracy by using field data as ground truth. Based on this criteria, we obtained $p = 910$ road segments for the expressway network G_1. For FPCA, BPCA and LDCP-WOPT we use one day of data ($d = 288$) to construct $\mathbf{M}_{GH1}$. For $\mathscr{N}_{GH1}$, if we use data for a specific day of the week (e.g., Mondays), then we have $w = 4$, since we consider data for August 2011.



(a) Missing percentage: 10% to 60%.(b) Missing percentage: 65% to 95%.

Fig. 2: Relative Error: Expressway Network.

In the second scenario, we consider a large urban network G_2 (see Fig.1, colored blue) which has sufficient amount of data (more than 97%) for performance analysis. It contains road segments with different speed limits, capacities and lanes. It spans from Outram to Changi and also contains arterial roads carrying significant volumes of traffic, in the downtown region. For G_2, we obtain $p = 6024$ links. The complete network shown in Fig.1 contains a total of 15258 links, which gives a measure of severity of missing data problem in practical networks. Similar to $\mathbf{M}_{GH1}$ and $\mathscr{N}_{GH1}$, we construct $\mathbf{M}_{GH2}$ and $\mathscr{N}_{GH2}$ for G_2.

For performance analysis, we will consider matrices $\{\mathbf{M}_{GHi}\}_{i=1}^{2}$ and tensors $\{\mathscr{N}_{GHi}\}_{i=1}^{2}$ as ground truth. We will create $\mathbf{M}_{G1}$, $\mathscr{N}_{G1}$, $\mathbf{M}_{G2}$ and $\mathscr{N}_{G2}$ by randomly removing a proportion of data(η), from $\mathbf{M}_{GH1}$, $\mathscr{N}_{GH1}$, $\mathbf{M}_{GH2}$ and $\mathscr{N}_{GH2}$ respectively. As we will be reconstructing complete network profiles $\{\mathbf{X}_{GHi}\}_{i=1}^{2}$ and $\{\mathscr{X}_{GHi}\}_{i=1}^{2}$ from low-dimensional representations of $\{G_i\}_{i=1}^{2}$, hence a good error measure would be relative error [15]. Relative error measures for matrices $\rho_{\mathbf{M}}$ and tensors $\rho_{\mathscr{N}}$ are defined as

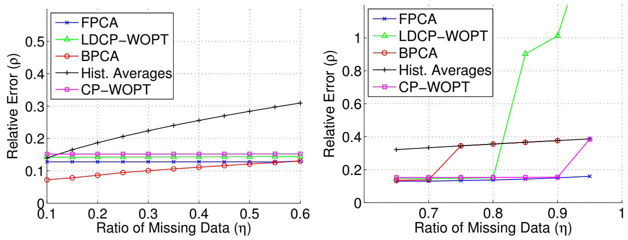
$$\rho_{\mathbf{M}} := \frac{\| \mathbf{X}_{GH} - \mathbf{M}_{GH} \|_{\mathbf{F}}}{\| \mathbf{M}_{GH} \|_{\mathbf{F}}} \qquad (6)$$

$$\rho_{\mathscr{N}} := \frac{\| \mathscr{X}_{GH} - \mathscr{N}_{GH} \|_{\mathbf{F}}}{\| \mathscr{N}_{GH} \|_{\mathbf{F}}}, \qquad (7)$$

where $\| \mathscr{Q} \|_{\mathbf{F}} := \Big(\sum\nolimits_{i_1,i_2\ldots i_n}(q_{i_1,i_2,\ldots i_n})^2\Big)^{1/2}$ is defined as Frobenius norm of tensor [19]. We also perform data imputation for networks G_1 and G_2 using historical averages, as benchmark for the proposed methods.

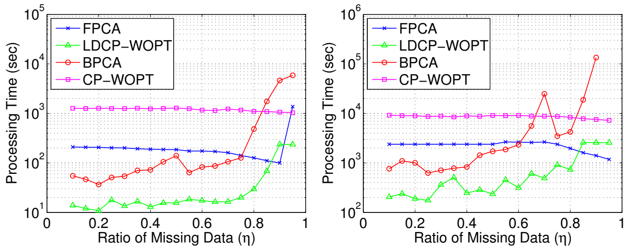
4. RESULTS AND DISCUSSION

In this section, we provide the imputation results for the two scenarios discussed above.



(a) Missing percentage: 10% to 60%.(b) Missing percentage: 65% to 95%.

Fig. 3: Relative Error: Large Urban Network.



(a) Expressway Network.
(b) Large Urban Network.

Fig. 4: Computation time for different methods.

Relative error values for the algorithms applied to expressways G_1 and large urban network G_2 are shown in Fig. 2 and Fig. 3 respectively. For low percentages of missing data, BPCA provides lower relative error in both scenarios. However, as the ratio of missing data increases, accuracy of BPCA starts to degrade. For higher percentages of missing data, its performance is only as good as historical averages (see Fig. 2b, 3b). LDCP-WOPT and CP-WOPT provide comparable performance for lower proportions of missing data (see Fig. 2a, 3a). For higher percentages of missing data, they perform better than BPCA. For sparse data sets, CP-WOPT outperforms LDCP-WOPT, although both methods have same underlying algorithm (see Fig. 2b, 3b). This seems to imply that additional temporal information (even incomplete) from similar days tends to provide more robust low-dimensional model for the network. This robustness, comes at additional computational price though (see Table 1, complexities reported for matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, and tensor $\mathscr{A} \in \mathbb{R}^{n\times n\times n}$), as CP-WOPT takes large time to converge to the solution (see Fig. 4). The difference in performance is more profound in case of large urban network. This can be attributed to the diverse nature of the network. FPCA provides comparable performance to other methods for small percentages of missing data. The method is also able to reconstruct network profiles with reasonable accuracy, even for sparse data sets (see Fig. 2b, 3b). For large percentages of missing data, FPCA provides best imputation accuracy as compared to other methods, for both scenarios.

It is also interesting to look at per iteration computational complexities [20] of these algorithms (see Table 1) alongside reported computational times (see Fig. 4). In Table 1, $c \leq n$ is the number of subspace components used for reconstruction [20, 21]. FPCA seems to report similar error measures (see Fig. 2,3) and convergence times (see Fig. 4) for a wide range of percentages of missing data. It may be so because a matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ with rank $r \ll n$ can be exactly reconstructed using (4), with only $O(nrlog(n))$ known entries [18]. Practical networks $\{\mathbf{M}_{Gi}\}_{i=1}^{2}$ might not be low ranked in the strict sense. However, as evident from Fig. 2and 3, we can construct low rank representations for $\{\mathbf{M}_{Gi}\}_{i=1}^{2}$ with low error using (4). Convergence time of BPCA seems to rely heavily on availability of data due to underlying EM algorithm (see Fig. 4 and [2, 20]). As expected, CP-WOPT has the highest convergence time, whereas LDCP-WOPT reports smaller convergence times (see Fig. 4) because it deals with $\mathbf{M}_G$ rather than $\mathscr{N}_G$ (see Table 1).

Table 1: Computation complexities of algorithms

CP-WOPT	LDCP-WOPT	FPCA	BPCA
$O(n^3)$ [16]	$O(n^2)$ [16]	$O(nc^2+c^3)$ [21]	$O(nc^3)$ [20]

From the results, we can conclude that FPCA and CP-WOPT can reconstruct traffic profiles with decent accuracy, even from very sparse data sets. The methods work well for expressway networks as well as large urban settings containing a diverse set of road segments.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed low-dimensional models for missing data imputation in large road networks. Missing data is a prevalent issue faced by intelligent transportation systems. Traditional methods require the availability of sufficient historical data and assume that missing data occurs at isolated instances [5–10]. However, due to highly erratic reporting patterns of sensors, these assumptions are usually not valid in practical road networks [1, 3]. As a consequence, these methods fail to deal with the problem of missing data in large interconnected urban settings. To overcome these limitations, we propose methods that can perform data imputation by constructing low-dimensional models of large and diverse networks in presence of missing data. We create these models by using FPCA and CP-WOPT for incomplete matrices and tensors respectively. To establish their imputation efficiency, we compared their performances with BPCA [2] and historical averages. We performed the comparison on expressways and a large generic urban network, for varying degrees of missing data. Performance evaluation showed that the low-dimensional models can perform data imputation with improved accuracy even in the presence of high percentage of missing data.

In the future, the imputation accuracy of the proposed methods can potentially be improved by developing kernel versions of the methods. Another application would be to assess the prediction accuracy of data driven traffic forecasting models, by performing traffic prediction, using the data obtained from the proposed imputation methods.

6. REFERENCES

[1] J. Zhang, F. Wang, K. Wang, W. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

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

[3] Y. Wang, Y. Zhu, Z. He, Y. Yue, and Q. Li, "Challenges and opportunities in exploiting large-scale gps probe data," *HP Laboratories, Technical Report HPL-2011-109*, vol. 21, 2011.

[4] B. Smith, W. Scherer, and J. Conklin, "Exploring imputation techniques for missing data in transportation management systems," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1836, pp. 132–142, 2003.

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

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

[7] M. Zhong, S. Sharma, and P. Lingras, "Genetically designed models for accurate imputation of missing traffic counts," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1879, pp. 71–79, 2004.

[8] C. Chen, J. Kwon, J. Rice, A. Skabardonis, and P. Varaiya, "Detecting errors and imputing missing data for single-loop surveillance systems," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1855, pp. 160–167, 2003.

[9] M. Treiber and D. Helbing, "Reconstructing the spatio-temporal traffic dynamics from stationary detector data," *Coop. Transp. Dyn.*, vol. 1, pp. 3.1–3.24, 2002.

[10] G. Chang and T. Ge, "Comparison of missing data imputation methods for traffic flow," in *IEEE International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, 2011, pp. 639–642.

[11] Z. Li, Y. Zhu, H. Zhu, and M. Li, "Compressive sensing approach to urban traffic sensing," in *31st IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2011, pp. 889–898.

[12] H. Zhang, Y. Zhang, Z. Li, and D. Hu, "Spatial-temporal traffic data analysis based on global data management using mas," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 4, pp. 267–275, 2004.

[13] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[14] F. Moutarde and Y. Han, "A new traffic-mining approach for unveiling typical global evolutions of large-scale road networks," in *18th World Congress on Intelligent Transport Systems*, 2011.

[15] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, no. 1, pp. 321–353, 2011.

[16] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41–56, March 2011.

[17] J. Dauwels, L. Garg, A. Earnest, and L. K. Pang, "Tensor factorization for missing data imputation in medical questionnaires," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2109–2112.

[18] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[19] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," in *IEEE 12th International Conference on Computer Vision*, 2009, pp. 2114–2121.

[20] T. Raiko, A. Ilin, and J. Karhunen, "Principal component analysis for sparse high-dimensional data," in *Neural Information Processing, Springer*, 2008, pp. 566–575.

[21] P. Drineas, R. Kannan, and M. Mahoney, "Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix," *SIAM Journal on Computing*, vol. 36, no. 1, pp. 158–183, 2006.