# Missing Value Imputation for Traffic-Related Time Series Data Based on a Multi-View Learning Method

Linchao Li, Jian Zhang, Yonggang Wang, and Bin Ran

***Abstract*— In reality, readings of sensors on highways are usually missing at various unexpected moments due to some sensor or communication errors. These missing values do not only influence the real-time traffic monitoring but also prevent further traffic data mining. In this paper, we propose a multi-view learning method to estimate the missing values for traffic-related time series data. The model combines data-driven algorithms (long-short term memory and support vector regression) and collaborative filtering techniques. It can consider the local and global variation in temporal and spatial views to capture more information from the existing data. The estimations of missing values from four views are aggregated to obtain a final value with a kernel function. Data from a highway network are used to evaluate the performance of the proposed model in terms of accuracy, precision, and agreement. The results indicate that our proposed model outperforms other baselines, especially for block missing pattern with a high missing ratio. Furthermore, the sensitivity of the parameters is analyzed. We can conclude that combining different views can improve the performance of the imputation.**

***Index Terms*— Collaborative filtering technique, highway, long-short term memory, support vector regression, temporal and spatial views.**

## I. Introduction

MANY sensors have been installed in the physical world to continuously and cooperatively monitor the traffic state, such as microwave sensors and loop detectors. These sensors generate massive time series data that aid operators and drivers to better understand traffic conditions on highways. However, due to many natural and man-made factors, such as incomplete observations, power outage malfunctioning devices and data transfer problems, some sensor readings are lost at unexpected moments [1]. Thus, missing data are very common in many traffic-related time series datasets in intelligent transportation systems. For example, in Texas Transportation Institute, the rate of missing data is between 16% and 93%. In Beijing, this rate is usually around 10% but sometimes it reaches 20% to 25% [2].

Missing data will affect real-time traffic conditions monitoring and compromise the performance of further data analysis like traffic forecasting. Many previous studies show that the performance of several forecasting models will dramatically reduce if missing data is present [3], [4]. Consequently, how to estimate the missing data in the database becomes a critical issue. Specifically, the missing value imputation problem of traffic time series data can be defined as follows: Given a data matrix $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_i}, \ldots, \boldsymbol{x_n}] \in \mathbb{R}^{m \times n}$ where $\boldsymbol{x_i} = [x_i^1, \ldots, x_i^m]^T \in \mathbb{R}^m$ denotes a time series of traffic parameters from the $i$th sensor during $m$ time intervals. Some elements in $X$, which are unobserved are called missing values. Therefore, missing value imputation tries to estimate the unobserved data through some meaningful models.

Increasing concerns have been undertaken about this issue during recent years. An extensive variety of models have been built to impute missing values. From the viewpoint of the underlying mechanisms and theoretical principles, we can roughly categorize the available approaches in two kinds: statistical learning-based and machine learning-based approaches [5].

The simplest method of statistical learning-based approaches is interpolation that fills the missing values based on the average values in the same time interval of the days with similar fluctuation patterns [6]. This method highly depends on the assumption that there are some similar days in the database. However, such an assumption sometimes fails in real the world [7]. Moreover, this method always ignores the stochastic variations and the spatial relationship of traffic flow data, thus missing important information and reducing the imputation performance [8]. Regression algorithm is another group of simple statistical learning-based approaches to impute missing values including linear regression model, local regression models, quadratic regression models, linear regression model solved by Bayesian theory, etc. [9], [10]. These regression approaches are easy to build and apply, but their accuracy was proved to be unreliable under different traffic conditions. Also, some advanced time series techniques were presented to cope with the missing value in the database. This type of methods can take full advantage of the temporal information in the historical data to estimate the missing values. They have been proved to have superior performance for traffic state prediction, however, ignoring

Manuscript received October 23, 2017; revised March 25, 2018; accepted September 7, 2018. Date of publication October 1, 2018; date of current version July 31, 2019. The Associate Editor for this paper was C. S. Ioakimidis. *(Corresponding author: Jian Zhang.)*

L Li and J. Zhang are with the School of Transportation, Southeast University, Nanjing, Jiangsu 210096, China (e-mail: lilinchao123@163.com; jianzhang@seu.edu.cn).

Y. Wang is with the School of Highway, Chang'an University, Xi'an 710064, China (e-mail: wangyg@chd.edu.cn).

B. Ran is with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: bran@seu.edu.cn).

Digital Object Identifier 10.1109/TITS.2018.2869768

1524-9050 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.

some differences between traffic prediction and imputation. For example, time series models fail to capture the patterns in the data after missing value [11]. Currently, a novel method called Kernel Probabilistic Principle Component Analysis (KPPCA) was proposed based on Probabilistic Principle Component Analysis (PPCA) and was proved to be effective [12]. It not only utilized the temporal information in the data but also utilized the information from other detecting points to improve imputing performance [2]. An imputation model based on traffic flow theory was also applied to impute missing data of detectors on the ramp [13].

Generally, most of the statistical learning-based approaches are able to take advantage of the statistic feature of the traffic flow. Assuming a special probability distribution of the observed data, the missing values can be the best fit. However, those approaches are principally based on some strong assumptions over the data which are practically irrelevant to the traffic flow in reality. Moreover, before the modeling, the form of the function is determined including parameters that we need to estimate [14]. Differently, machine learning-based approaches try to take full advantages of the data and the form of the function is usually determined by the data.

The most robust and widely used machine learning-based approach to impute missing values is $k$-Nearest Neighbors (KNN), which calculates the missing value from the $k$ closest data points among the whole dataset [15]. Besides, other machine leaning approaches, such as Support Vector Regression (SVR) [16], Multi-Layer Perception (MLP) [10], [17], regression tree [18], and random forest [19] have also been commonly applied. Recently, a novel machine learning based approach called tensor decomposition method was introduced and extended to estimate the missing values in the traffic flow by obtaining a suitable low-rank approximation of the incomplete matrix. These methods were good at dealing with the correlation across different highway segments, and thus, they are suited to impute missing values in the highway network level [4], [20], [21].

Although many models have been proposed to impute the missing values in the database, it is still challenging for the following two reasons. The first one is how to appropriately consider temporal and spatial dependence [22]. Statistical learning-based approaches such as time-series models are good at capturing temporal information. Machine learning-based approaches such as Artificial Neural Network (ANN) can deal with spatial information. However, no generalized model has been introduced considering global and local variations of both temporal and spatial information. The second challenge is the block missing, as shown in Fig. 1, in which the values are missed at some consecutive timestamps and multiple individual sites in a road section. In this circumstance, the difficulty of imputation is increased as we may not be able to find stable inputs for a model.

To tackle these challenges, in this paper a Multi-View Learning Method (MVLM) is proposed to impute the missing values for traffic-related time series data in the database. To evaluate the imputing performance, we compare MVLM with several commonly used models including Autoregressive Moving Average Model (ARMA), seasonal ARIMA (SARIMA), KNN, ANN, and Kriging. The results indicate that the proposed model is more powerful, since it does not only consider the temporal and spatial views but also global and local variations.

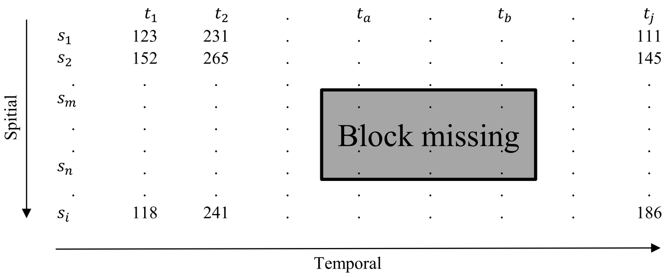


Fig. 1. Block missing pattern.

The rest of the paper is structured as follows. In Section II, the motivations and the proposed model are presented. Section III briefly introduces the benchmark models, dataset and performance measures. Section IV presents the imputing testing results and discusses the sensitivity of the parameters. Finally, we conclude the paper and summarize the contributions in Section V.

## II. Methodology

In this section, we firstly introduce the motivations of this study. Then, to capture the local and global variations of traffic-related time series data in temporal and spatial views, we propose the multi-view learning method.

### A. Motivations of MVLM

Due to the similarity of drivers' behaviors, environmental conditions, dynamic and uncertainty of demand and supply, traffic-related time series data is correlated in both space and time domains [23]. Fig. 2 illustrates the fluctuation of traffic volume along with timestamps and its frequency. From the red ellipse, it can be seen that there exists a change regularity over a long time period (global variation). Furthermore, from the blue ellipse, the traffic volume fluctuates tremendously from non-peak hours to peak hour (local variation) which is always ignored by the traditional imputation models. In spatial view, the first law of geography highlights: everything is related to everything else, but near things are more related than distant things [24]. Traffic state of a specific sensor is highly affected by the traffic state of adjacent sensors and proportional to the distance (local variation) [25]. However, in the block missing scenario, the data in adjacent locations is also missing, to consider the data from sensor at long distance (global variation) is also necessary. Motivated by the above global and local variations of both temporal and spatial views, we propose the MVLM to capture more information from the data to improve the accuracy.

### B. Global Variation in Temporal View (GVTV)

To model the global variation of traffic-related time series data in the temporal view, Long-Short Term Memory (LSTM)

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.

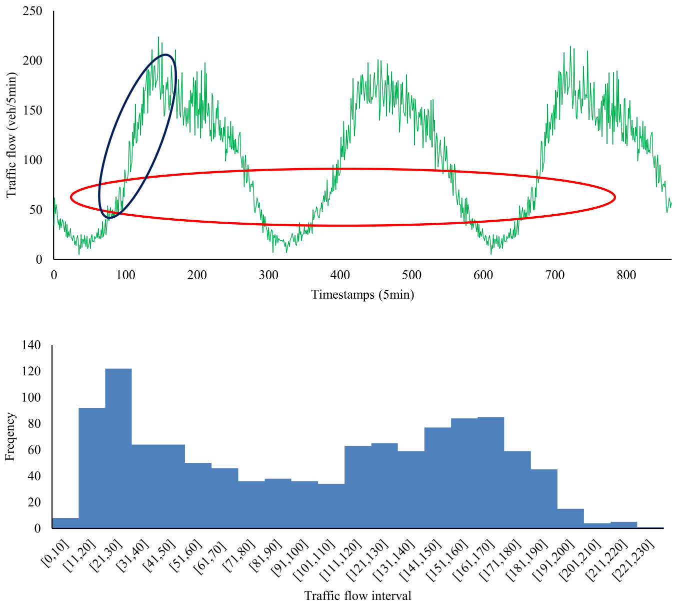


Fig. 2. Local variation and global variation in temporal view of traffic flow data.

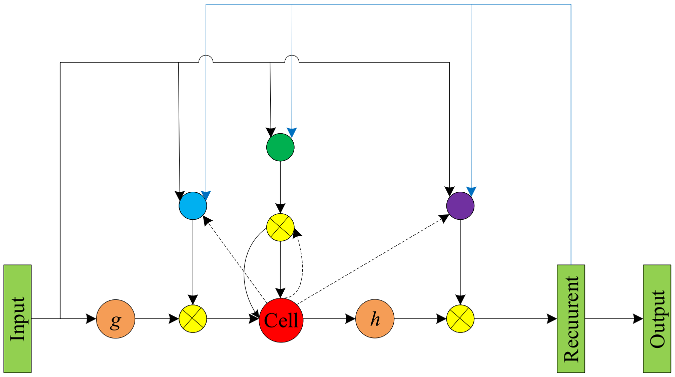


Fig. 3. Framework of LSTM.

is applied. It is a recurrent neural network architecture that is frequently applied in the time series domain and good at coping the time series data over a long-time period [26]. A LSTM contains one input layer, one recurrent hidden layer and one output layer. Different from the traditional neural network, it is composed of basic units called memory block in the recurrent hidden layer. The memory block contains two important parts. One is memory cell with self-connections to memory the temporal state. Another is three adaptive, multiplicative gating units that control the flow of information. The three gates are: input gate, output gate and forget gate, as depicted in Fig. 3. The forget gate can forget or reset the memory of the cell, while the input and output gate control the input and output activations into the block. Additionally, from the internal memory cell to its gate there exists peephole connections to learn precise timing of the outputs [27].

The input and output of the LSTM are defined as $x = (x_1, \ldots, x_T)$ and $y = (y_1, \ldots, y_T)$. In this context of missing value imputation, $x$ can be considered as historical complete traffic data, and $y$ is the missing values to be estimated. Then LSTM calculates the network unit activations using the following equations iteratively from $t = 1$ to $T$:

$$\begin{aligned} i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) && (1)\\ f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f) && (2)\\ c_t &= f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) && (3)\\ o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o) && (4)\\ m_t &= o_t \odot h(c_t) && (5)\\ y_t &= \phi(W_{ym}m_t + b_y) && (6) \end{aligned}$$

where $\odot$ is the scalar product of two vectors.

$\sigma(\cdot)$ is the standard logistics sigmoid function:

$$\sigma(x) = 1/(1 + e^{-x}). \tag{7}$$

$g(\cdot)$ donates a centered logistics sigmoid function with range $[-2, 2]$

$$g(x) = 4/(1 + e^{-x} - 2). \tag{8}$$

$h(\cdot)$ donates a centered logistics sigmoid function with range $[-1, 1]$

$$h(x) = (2/1 + e^{-x} - 1). \tag{9}$$

$i_t$, $o_t$, $f_t$ represents three gates: input gate, output gate, and forget gate, respectively. $W$ and $b$ are weight matrix and bias vector to connect the input layer, output layer and memory block. The activation vectors of each cell and memory block is represented by $c_t$ and $m_t$, respectively. The global variation in temporal view is captured by LSTM and its estimation is $v^1_{miss}$.

### C. Global Variation in Spatial View (GVSV)

In this view, support vector regression (SVR) is implemented to model the complex relationship among all the sites where to obtain traffic flow data. Given a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $y$ represents the traffic-related time series data in the target site and $x$ represents traffic-related time series data from all other sites, SVR approximates the unknown function as the following formula:

$$y(\mathbf{x}) = \boldsymbol{w}^{\mathbf{T}}\phi(\mathbf{x}) + b \tag{10}$$

where $\phi(\mathbf{x})$ represents a fixed feature-space transformation. To obtain coefficients $\boldsymbol{w}$ and $b$ it is necessary to solve a quadratic problem subject to a set of linear inequality constraints in which the convex object function for minimization is given by

$$C\sum\nolimits_{n=1}^{N} E_\epsilon(y(\mathbf{x}_n) - t_n) + \frac{1}{2}\|\boldsymbol{w}\|^2 \longrightarrow min \tag{11}$$

$$\text{subject to } \begin{cases} t_n \le y(\mathbf{x}_n) + \epsilon + \xi_n \\ t_n \ge y(\mathbf{x}_n) - \epsilon - \xi_n^* \\ \xi_n, \quad \xi_n^* \ge 0 \end{cases} \tag{12}$$

where $\xi_n$, $\xi_n^*$ are slack variables to allow data points to be classified incorrectly but with a penalty. $E_\epsilon$ is the penalty function called $\epsilon$-insensitive error function which equals zero if errors are smaller than $\epsilon$ where $\epsilon > 0$ and otherwise it has a linear cost associated with errors. Coefficient $C$ governs the relative importance of the penalty function compared with the

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.

regularization term which is the second part of the (11) used to control the over-fitting phenomenon.

The optimization problem can be solved by Lagrangian theory and its solution is given by

$$\boldsymbol{w}=\sum\nolimits_{n=1}^{N}\left(a_n-\hat{a}_n\right)\mathbf{x}_n \tag{13}$$

$$y=\sum\nolimits_{n=1}^{N}\left(a_n-\hat{a}_n\right)\boldsymbol{K}(\mathbf{x},\mathbf{x}_{\boldsymbol{n}})+b \tag{14}$$

$$\text{subject to }\begin{cases}0\le a_n\le C\\ 0\le \hat{a}_n\le C\end{cases} \tag{15}$$

where $a_n$ and $\hat{a}_n$ are Lagrange multipliers. $K(x,x_n)$ is kernel function to transform raw data into feature vector representation and compute inner products in the feature space to reduce computation.

Up until now, it is difficult to select the type of kernel functions for a specific data pattern. However, radial basis function (RBF) is easier to implement and can map the data into an infinite dimensional space and it was proved that is suitable to handle traffic flow prediction. Therefore, it was specified in this study and could be written as follows:

$$K\left(\mathbf{x}_i,\mathbf{x}_j\right)=\exp(-\gamma\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2) \tag{16}$$

where $i,j=1,2\ldots,N$. $\gamma$ is the width of the radial basis function. The grid search method is utilized to optimize the parameters in SVR [3]. The global variation in spatial view is captured by SVR and its estimation is $v^2_{miss}$.

### D. Local Variations in Temporal View (LVTV)

In this view, motivated by the collaborative filtering (CF) techniques that similar users make similar rating for similar items in recommender system, each traffic data point is regard as the item [28], [29]. The similarity between the two traffic data points can be calculated according to

$$c\left(t_l,t_p\right)=\left(\sum\nolimits_{i=1}^{m}\left(v_{i,l}-v_{i,p}\right)^2/m\right)^{-1/2} \tag{17}$$

where $m$ is the number of sensors that have readings at both the $l$th timestamp and the $p$th timestamp. $v_{i,l}$ is the reading of the $l$th timestamp from the $i$th sensor. The local variation in temporal view is captured by (18), shown below, and its estimation is $v^3_{miss}$.

$$v^3_{miss}=\sum\nolimits_{j=1}^{\omega}c\left(t_{miss},t_j\right)*v_j/\sum\nolimits_{j=1}^{\omega}c\left(t_{miss},t_j\right) \tag{18}$$

where $t_{miss}$ represents the timestamps at which the value is missing. $\omega$ is time windows.

### E. Local Variations in Spatial View (LVSV)

In this view, the sensors can also be regarded as the items. The similarity between the traffic time series data from two sensors can be calculated according to

$$c\left(s_{miss},s_j\right)=\left(\sum\nolimits_{k=1}^{\omega}\left(v_{miss,k}-v_{j,k}\right)/\omega\right)^{-1/2} \tag{19}$$

where $v_{j,k}$ represents the reading of the $k$th timestamp from the $j$th sensor. $\omega$ is the window size. The local variation in spatial view is captured by (20), shown below, and its estimation is $v^4_{miss}$.

$$v^4_{miss}=\sum\nolimits_{j=1}^{m}c\left(s_{miss},s_j\right)*v_j/\sum\nolimits_{j=1}^{m}c\left(s_{miss},s_j\right) \tag{20}$$

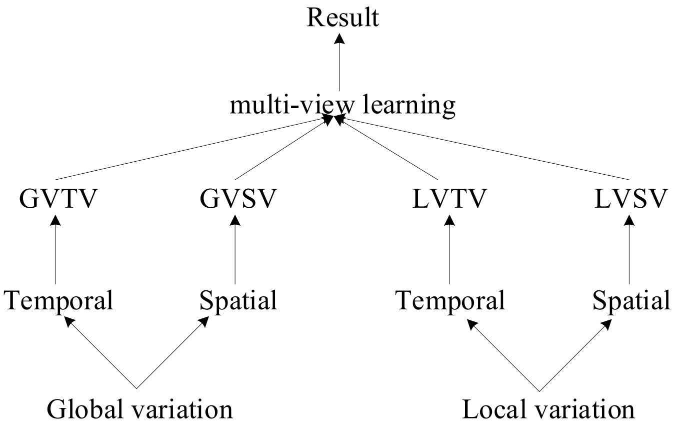


Fig. 4. Framework of the model.

TABLE I
PSEUDO-CODE OF THE PROPOSED MVLM

```
Algorithm: MVLM
Input: Origin data matrix $M, w$;
Output: Complete data matrix;
1.  $O$ ←Get the block missing values
2.  $M$ ←Initialization($M$,GVTV, GVSV);
3.  For each $v_{miss}$ in $O$
4.  $v^1_{miss}$ ←GVTV($M$)
5.  $v^3_{miss}$ ←GVSV($M$)
6.  $v^2_{miss}$ ←LVTV($M, w$)
7.  $v^4_{miss}$ ←LVSV($M$)
8.  $v_{mv}$ ←MVLM($v^1_{miss}, v^2_{miss}, v^3_{miss}, v^4_{miss}$)
9.  Impute $v_{mv}$ to $M$;
10. Return $M$;
```

### F. Multi-View Learning Method (MVLM)

Based on the above four different views, a multi-view learning method (see Fig. 4) is applied to integrate their results according to

$$v_{mv}=w_1*v^1_{miss}+w_2*v^2_{miss}+w_3*v^3_{miss}+w_4*v^4_{miss}+b_k \tag{21}$$

where $b_k$ is the residual and $w_1,w_2,w_3,w_4$ are weights assigned to the four views. The detail procedures of the MVLM are illustrated in Table I. In block missing pattern, LVSV and LVTV do not work well in the beginning. GVSV and GVTV are firstly implemented to generate some initial values to impute the missing values (line 2). Next, the missing value are estimated from four different views respectively (line 4-7). Finally, the four results from different views are combined used a linear kernel function based on multi-view leaning algorithm (line 8). The function (21) is solved by minimizing the least square error between the estimations and the ground truth [30].

## III. EXPERIMENTS

In this section, we first give a brief description of the data used in this study. Then, the method to generate the missing values is discussed. Moreover, we introduce the state-of-the-art imputation models and evaluation criteria for comparison.

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.

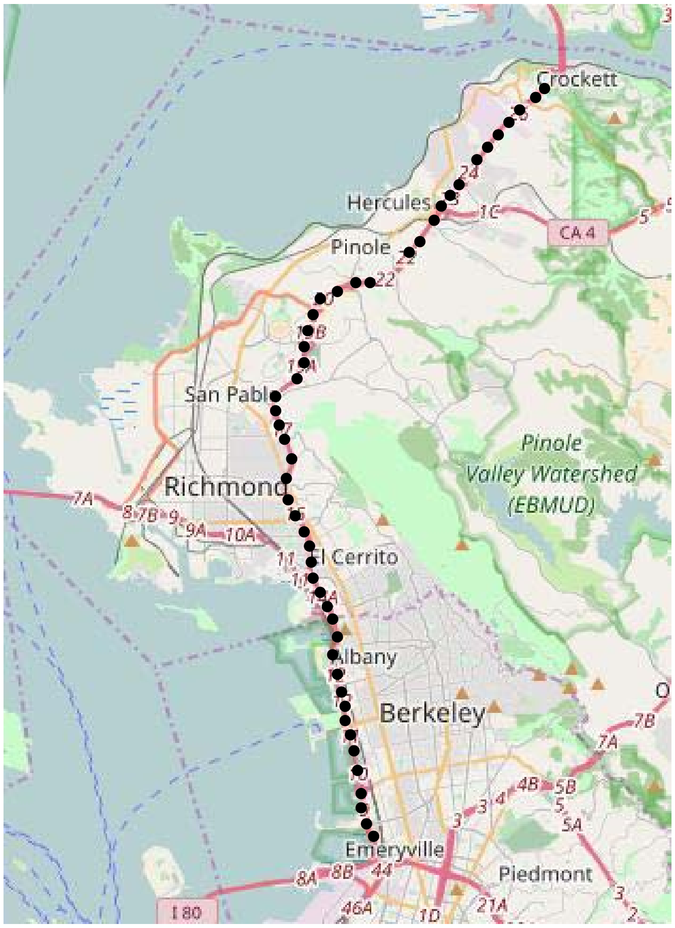


Fig. 5. Position of detectors.

### A. Data Description

In this study, the data is extracted from Performance Measurement System (PeMS) which provides an easy-to-access source of historical and real-time traffic data [35]. The dataset used in this paper was collected from Interstate Highway 80 in California. Traffic flow data are extracted during a period from July 4, 2012 to July 4, 2013. As shown in Fig. 5, the selected section has 50 locations that can obtain the traffic flow data measured by sensors. The data is updated every 5 minutes, and thus, each day contains 288-time intervals. The raw dataset is divided into two parts: a training dataset and a testing dataset. The data of the first 8 months (training dataset) is used to train the models and the rest of the data (testing dataset) is used to evaluate the models.

In previous studies, the missing patterns in intelligent transportation system can be generally classified into three classes: Missing Completely at Random (MCR), Missing at Random (MR), and Not MR (NMR) [32].

1) In MCR, the missing values may occur because of temporary power or communication failure. Thus, they are independent of each other completely. As shown in Fig. 6(a) Type 1, missing values are some isolated points randomly scattered.

2) In MR, the missing values may occur due to a physical damage or maintenance backlog. The missing values are related to their temporal or spatial neighboring readings. Thus, this missing pattern appear as some sequential points at the same time (Fig. 6(b) Type 2) or at the same sensor (Fig. 6(c) Type 3).

TABLE II

PSEUDO-CODE OF THE PROPOSED MVLM

| Method | Spatial | Temporal | Spatial+Temporal |
|---|---|---|---|
| **Global** | GVSV | GVTV | GVSV+ GVTV |
| **Local** | LVSV | ARMA,<br>LVTV | LVSV+ LVTV,<br>KNN, ANN |
| **Global+Local** | Kriging | SARIMA | MVLM |

3) In NMR, this missing pattern is often caused by a long-time malfunction of the sensors and missing values appears with certain patterns. As shown in Fig. 6 (d) Type 4, the values are missed like blocks.

In transportation information systems, Types 2, 3, and 4 are commonly observed [15], [32]. To test the robustness of the proposed imputation model, the four types missing patterns are generated in the dataset. On the other hand, to test the stability of the model across different missing ratios, the missing ratio in this study ranges from 5% to 50%.

### B. Baselines

The proposed model is compared with several widely used imputation models shown in Table II.

**ARMA**: This describes a stationary stochastic process parsimoniously in terms of two polynomials, one for the auto regression and another for moving average. The missing values is estimated based on the readings of a times ago [33].

**SARIMA**: SARIMA is an extension of ARMA to include more realistic dynamics, in particular, the seasonal behaviors. It not only captures information from the neighboring values but also the change regularity over long time period [33].

**ANN**: ANN is a computational approach inspired by the way the brain processes information. It solves specific problems through many highly interconnected processing elements (neurons) working in unison. Data from the neighboring sensors is imputed into ANN and it outputs the imputations [3].

**KNN**: KNN is a non-parametric method. The output is the average of the values of its $k$ nearest neighbors. In this study, the $k$ nearest spatial and temporal neighbors are used to calculate the missing values ($k = 6$).

**Kriging**: Kriging can estimate the value at a given sensor by computing a weighted mean of the known values at the geospatially adjacent sensors [31].

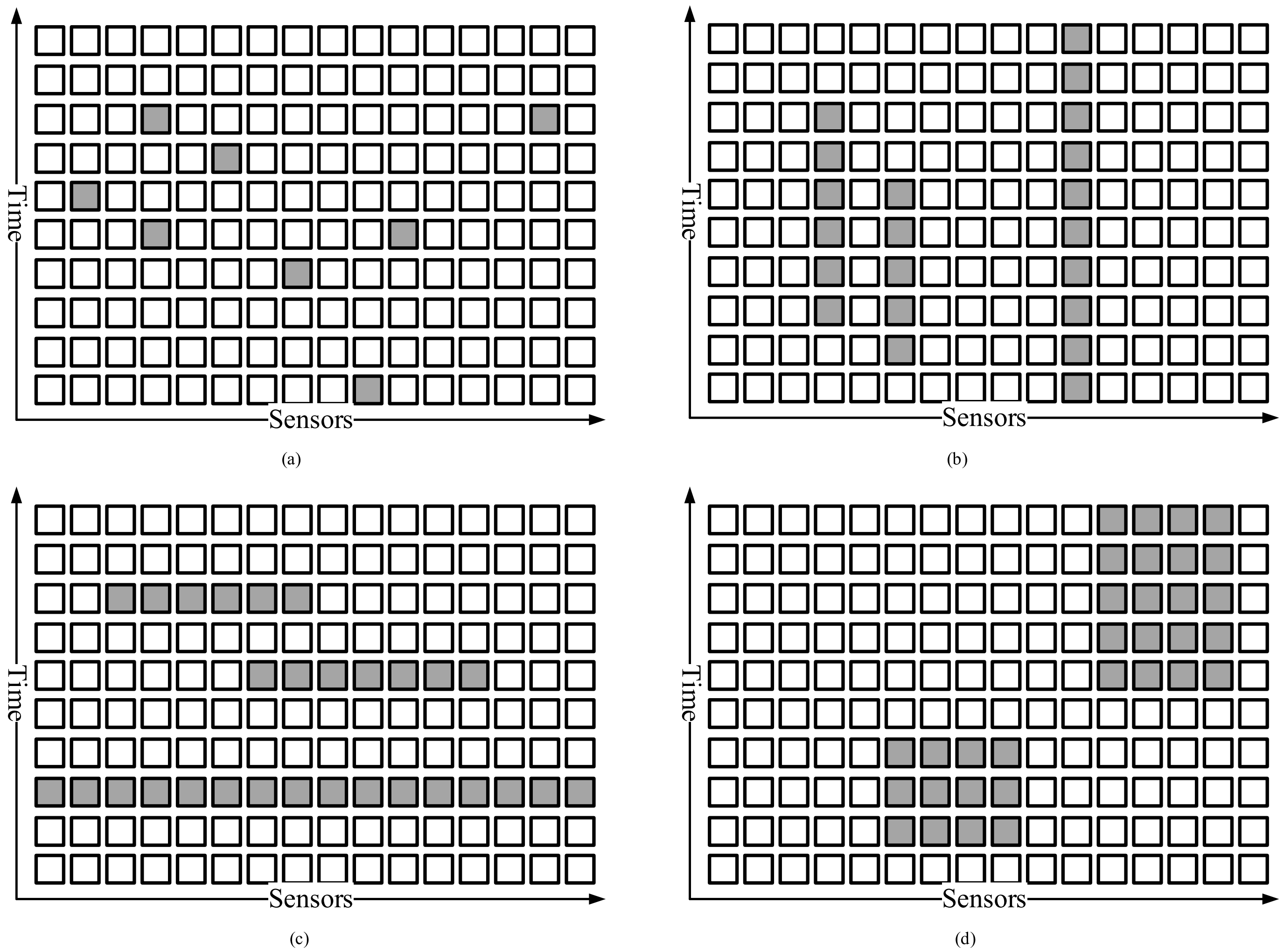
**GVSV + GVTV**: These fill missing values at a sensor by GVSV and GVTV respectively. The imputation is the mean of the two different models.

**LVSV + LVTV**: These fill missing values at a sensor by LVSV and LVTV respectively. The imputation is the mean of the two different models.

### C. Evaluation Criteria

The imputations were always evaluated in term of accuracy, precision, and agreement. The mean absolute percentage error

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c) (d)

Fig. 6. Patterns of missing value.

(MAPE; see (22) below) is applied to evaluate the mean error for each imputation procedure [34]–[36]. The precision of the models is assessed by the proportional variance (PV; see (23) below) comparing the sample variance of imputations with the variance of the ground truth [37]. Person's correlation coefficient ($r$; see (23) below), which is the most widely used indicator to evaluate the agreement when evaluating imputation procedures, is applied in this study.

$$\text{MAPE} = \sum\nolimits_{i=1}^{m} |(x_i - x_i')/x_i|/m \times 100\% \tag{22}$$

$$\text{PV} = abs(1 - var(x')/var(x)) \tag{23}$$

$$r = cov(x, x')/\sqrt{var(x') \times var(x)} \tag{24}$$

where $m$ is the total number of missing values. $i = 1, \ldots, m$. $x_i$ is the $i$th ground truth. $x_i'$ is the $i$th estimated value.

## IV. Results and Discussion

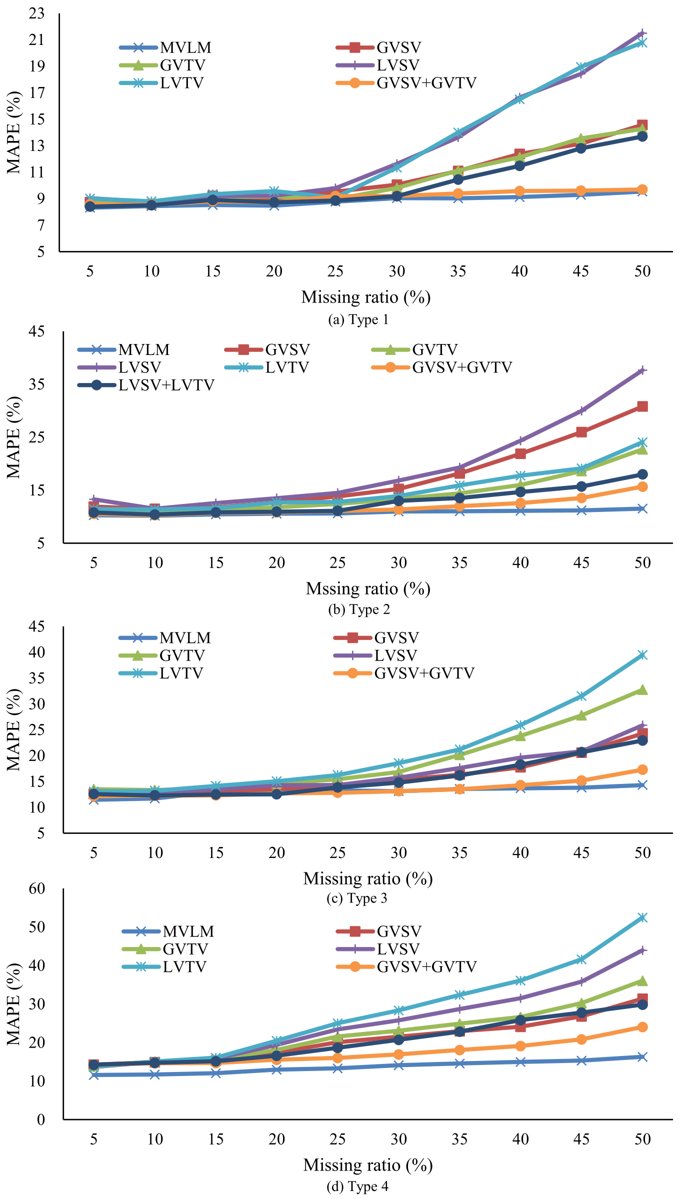
### A. Performance Analysis

In this section, we analyze the performance of the proposed ensemble model and baselines using real traffic flow data from a highway. The models are divided into two groups. Group 1 including MVLT, GVSV, GVTV, LVSV, LVTV, GVSV+GVTV, LVSV+LVTV is to evaluate the ensemble of multiple views. Group 2 including MVLT, ANN, ARMA, SARIMA, KNN, and Kriging is to compare the accuracy of the proposed MVLM with some traditional models.

Fig. 7 shows the MAPE of models in group 1 that evaluate the imputation accuracy for different missing types. For type 1 missing pattern, MVLM achieves the lowest MAPE followed by GVSV+GVTV. The MAPE of LVSV+LVTV is higher than the MAPE of GVSV+GVTV and similar to the MAPE of GVTV and GVSV. LVSV and LVTV have the largest imputation error. For type 2 missing pattern, our proposed model also has the lowest MAPE. When the missing ratio is higher than 20%, the MAPE in the figures indicate the five models can be ordered from low to high as GVSV+GVTV, LVSV+LVTV, GVTV, LVTV, GVSV and LVTV. For Type 3 missing pattern, besides the proposed model, the imputation error for GVSV+GVTV is also lower as compared to other models. MAPE of GVSV, LVSV+LVTV, and LVSV are similar and higher than MVLT and GVSV+GVTV but lower than GVTV and LVTV. For the last missing pattern, the MVLM shows its better performance even when the missing ratio is low.

From the above comparison for different missing patterns, we can get four important points: 1) For type 2 missing pattern, models from spatial view outperform these from temporal view. On the contrary, for type 3 missing pattern, models from

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.



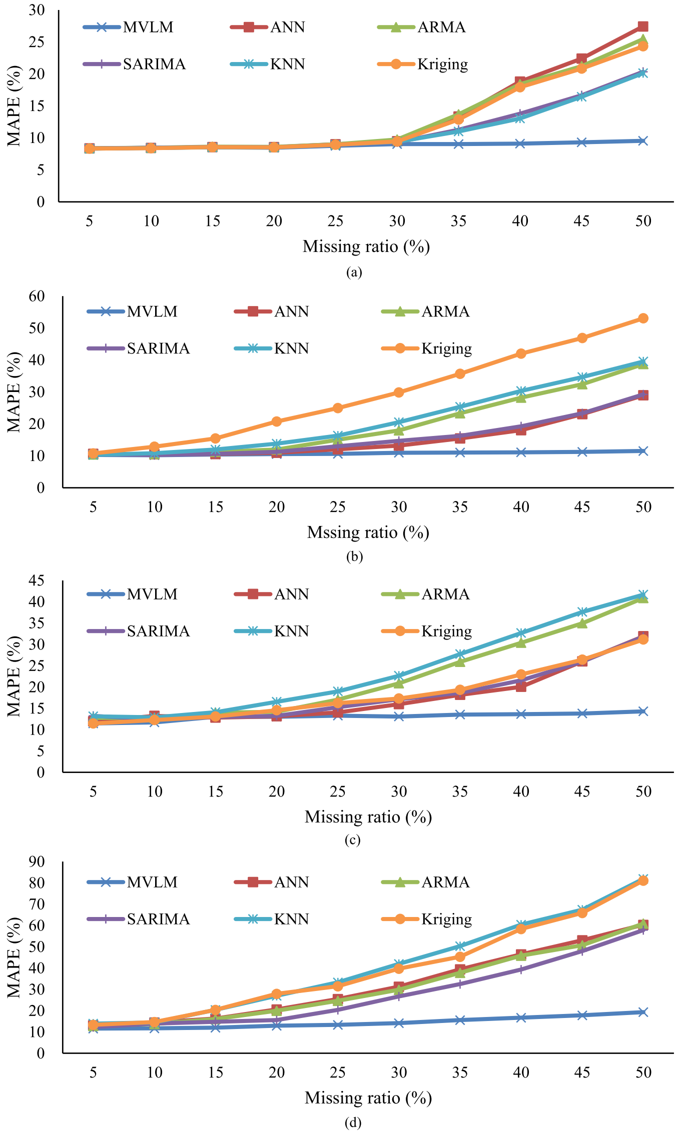
(a) Type 1

(b) Type 2

(c) Type 3

(d) Type 4

Fig. 7. MAPE of models in group 1 for different missing patterns.



(a)

(b)

(c)

(d)

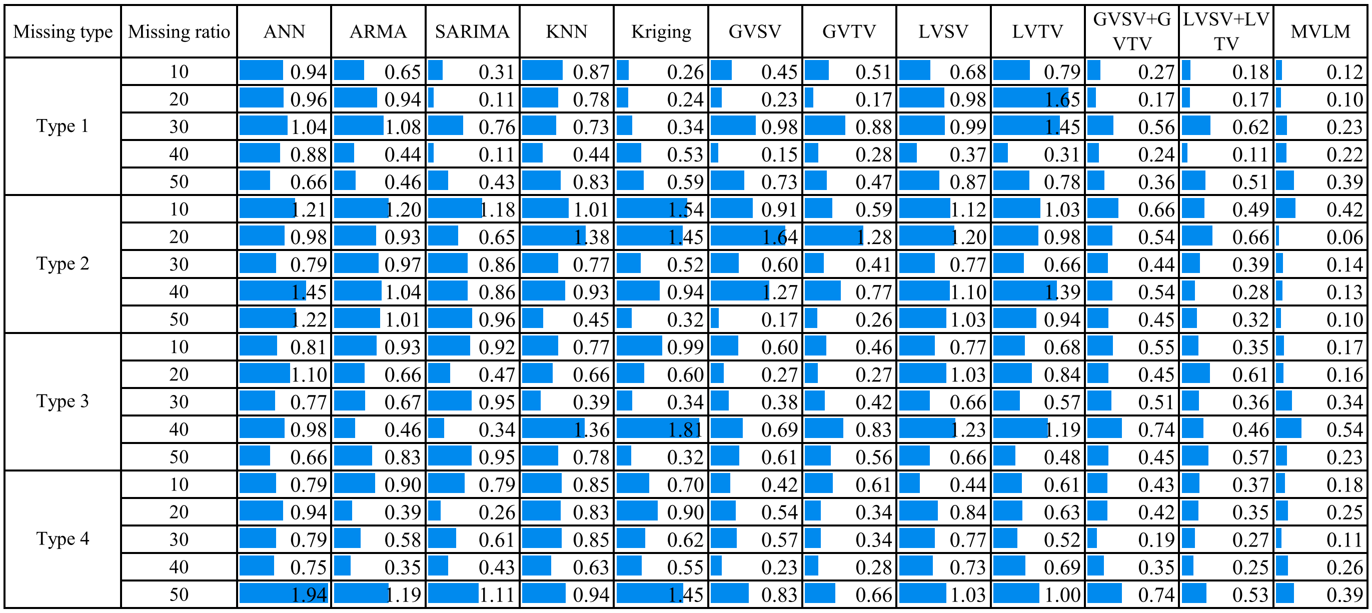
Fig. 8. MAPE of models in group 2 for different missing patterns.

temporal view outperform these from spatial view. (2) When the missing ratio is high (>25%), the models capture global variation outperform models capture local variation. (3) Overall, for non-block missing pattern (type 1, 2, 3), the imputation error of models in group 1 are similar when missing ratio is lower than 20%. However, as the missing ratio increases, imputation error of our proposed model remains steady while the imputation errors of other models are increasing. (4) For block missing pattern, the MVLM outperforms other models at all the missing ratio. Furthermore, although, other models achieve good performance for low missing ratio (<15%), their performance sharply degrades when the missing ratio is higher.

These results indicate the ensemble of the multiple views significantly improve the accuracy of the imputation, especially when the missing ratio is higher than 25%. This is mainly because the models from single view cannot use the full information from the data. The models from temporal view can only capture the correlation among timestamps but cannot use the correlation among sensors. The models from spatial view can only capture the correlation among sensors but cannot use the correlation among timestamps. The available information is reducing as the missing ratio is increasing. However, our proposed model can take so much information from different views that it can maintain stable performance.

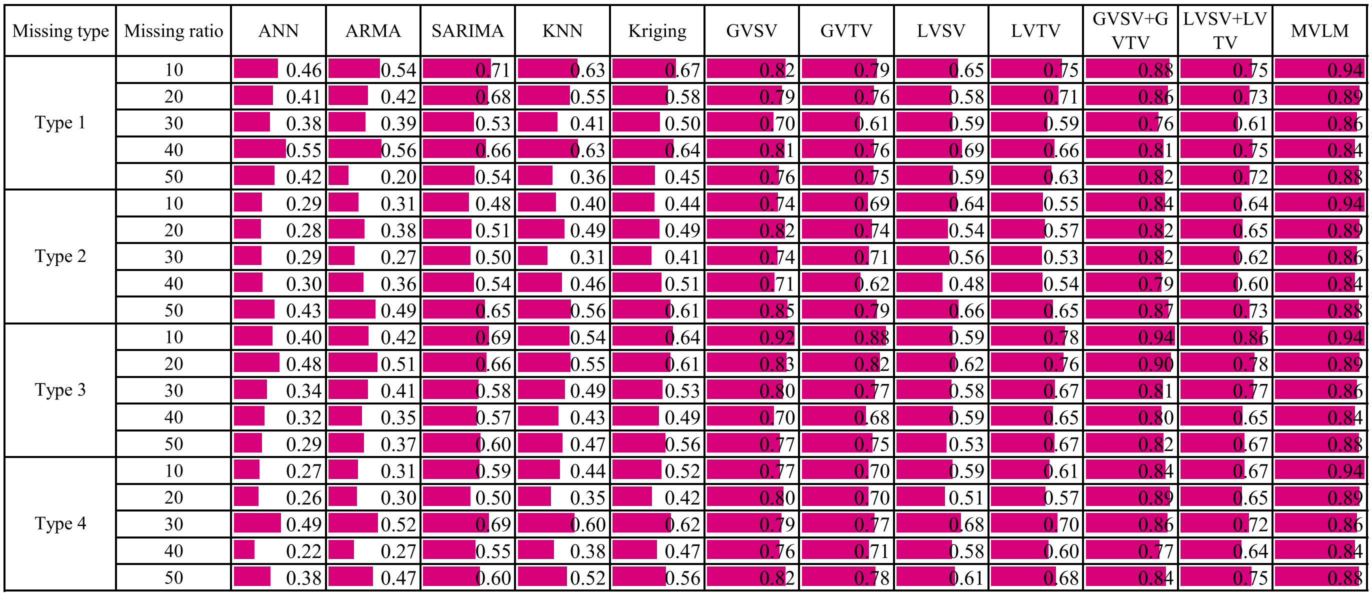
Fig. 8 illustrates the MAPE of models in group 2 for different missing types. For type 1 missing pattern, the MAPE of the models are similar when the missing ratio is lower than 30%. With the increase of missing values, the MAPE of the models except MVLM is increasing while MAPE of MVLM is still around 9%. For type 2 missing pattern, the MAPE of Kriging is the highest across all missing ratio. On the contrary, our proposed model possesses the lowest MAPE among all missing ratio. The MAPE of ARMA and KNN are similar and higher than SARIMA and ANN. For type 3 missing pattern, the MAPE of ARMA and KNN are higher than the MAPE

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.



| Missing type | Missing ratio | ANN | ARMA | SARIMA | KNN | Kriging | GVSV | GVTV | LVSV | LVTV | GVSV+GVTV | LVSV+LVTV | MVLM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Type 1 | 10 | 0.94 | 0.65 | 0.31 | 0.87 | 0.26 | 0.45 | 0.51 | 0.68 | 0.79 | 0.27 | 0.18 | 0.12 |
| | 20 | 0.96 | 0.94 | 0.11 | 0.78 | 0.24 | 0.23 | 0.17 | 0.98 | 1.65 | 0.17 | 0.17 | 0.10 |
| | 30 | 1.04 | 1.08 | 0.76 | 0.73 | 0.34 | 0.98 | 0.88 | 0.99 | 1.45 | 0.56 | 0.62 | 0.23 |
| | 40 | 0.88 | 0.44 | 0.11 | 0.44 | 0.53 | 0.15 | 0.28 | 0.37 | 0.31 | 0.24 | 0.11 | 0.22 |
| | 50 | 0.66 | 0.46 | 0.43 | 0.83 | 0.59 | 0.73 | 0.47 | 0.87 | 0.78 | 0.36 | 0.51 | 0.39 |
| Type 2 | 10 | 1.21 | 1.20 | 1.18 | 1.01 | 1.54 | 0.91 | 0.59 | 1.12 | 1.03 | 0.66 | 0.49 | 0.42 |
| | 20 | 0.98 | 0.93 | 0.65 | 1.38 | 1.45 | 1.64 | 1.28 | 1.20 | 0.98 | 0.54 | 0.66 | 0.06 |
| | 30 | 0.79 | 0.97 | 0.86 | 0.77 | 0.52 | 0.60 | 0.41 | 0.77 | 0.66 | 0.44 | 0.39 | 0.14 |
| | 40 | 1.45 | 1.04 | 0.86 | 0.93 | 0.94 | 1.27 | 0.77 | 1.10 | 1.39 | 0.54 | 0.28 | 0.13 |
| | 50 | 1.22 | 1.01 | 0.96 | 0.45 | 0.32 | 0.17 | 0.26 | 1.03 | 0.94 | 0.45 | 0.32 | 0.10 |
| Type 3 | 10 | 0.81 | 0.93 | 0.92 | 0.77 | 0.99 | 0.60 | 0.46 | 0.77 | 0.68 | 0.55 | 0.35 | 0.17 |
| | 20 | 1.10 | 0.66 | 0.47 | 0.66 | 0.60 | 0.27 | 0.27 | 1.03 | 0.84 | 0.45 | 0.61 | 0.16 |
| | 30 | 0.77 | 0.67 | 0.95 | 0.39 | 0.34 | 0.38 | 0.42 | 0.66 | 0.57 | 0.51 | 0.36 | 0.34 |
| | 40 | 0.98 | 0.46 | 0.34 | 1.36 | 1.81 | 0.69 | 0.83 | 1.23 | 1.19 | 0.74 | 0.46 | 0.54 |
| | 50 | 0.66 | 0.83 | 0.95 | 0.78 | 0.32 | 0.61 | 0.56 | 0.66 | 0.48 | 0.45 | 0.57 | 0.23 |
| Type 4 | 10 | 0.79 | 0.90 | 0.79 | 0.85 | 0.70 | 0.42 | 0.61 | 0.44 | 0.61 | 0.43 | 0.37 | 0.18 |
| | 20 | 0.94 | 0.39 | 0.26 | 0.83 | 0.90 | 0.54 | 0.34 | 0.84 | 0.63 | 0.42 | 0.35 | 0.25 |
| | 30 | 0.79 | 0.58 | 0.61 | 0.85 | 0.62 | 0.57 | 0.34 | 0.77 | 0.52 | 0.19 | 0.27 | 0.11 |
| | 40 | 0.75 | 0.35 | 0.43 | 0.63 | 0.55 | 0.23 | 0.28 | 0.73 | 0.69 | 0.35 | 0.25 | 0.26 |
| | 50 | 1.94 | 1.19 | 1.11 | 0.94 | 1.45 | 0.83 | 0.66 | 1.03 | 1.00 | 0.74 | 0.53 | 0.39 |

Fig. 9. Proportional variance of different models.



| Missing type | Missing ratio | ANN | ARMA | SARIMA | KNN | Kriging | GVSV | GVTV | LVSV | LVTV | GVSV+GVTV | LVSV+LVTV | MVLM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Type 1 | 10 | 0.46 | 0.54 | 0.71 | 0.63 | 0.67 | 0.82 | 0.79 | 0.65 | 0.75 | 0.88 | 0.75 | 0.94 |
| | 20 | 0.41 | 0.42 | 0.68 | 0.55 | 0.58 | 0.79 | 0.76 | 0.58 | 0.71 | 0.86 | 0.73 | 0.89 |
| | 30 | 0.38 | 0.39 | 0.53 | 0.41 | 0.50 | 0.70 | 0.61 | 0.59 | 0.59 | 0.76 | 0.61 | 0.86 |
| | 40 | 0.55 | 0.56 | 0.66 | 0.63 | 0.64 | 0.81 | 0.76 | 0.69 | 0.66 | 0.81 | 0.75 | 0.84 |
| | 50 | 0.42 | 0.20 | 0.54 | 0.36 | 0.45 | 0.76 | 0.75 | 0.59 | 0.63 | 0.82 | 0.72 | 0.88 |
| Type 2 | 10 | 0.29 | 0.31 | 0.48 | 0.40 | 0.44 | 0.74 | 0.69 | 0.64 | 0.55 | 0.84 | 0.64 | 0.94 |
| | 20 | 0.28 | 0.38 | 0.51 | 0.49 | 0.49 | 0.82 | 0.74 | 0.54 | 0.57 | 0.82 | 0.65 | 0.89 |
| | 30 | 0.29 | 0.27 | 0.50 | 0.31 | 0.41 | 0.74 | 0.71 | 0.56 | 0.53 | 0.82 | 0.62 | 0.86 |
| | 40 | 0.30 | 0.36 | 0.54 | 0.46 | 0.51 | 0.71 | 0.62 | 0.48 | 0.54 | 0.79 | 0.60 | 0.84 |
| | 50 | 0.43 | 0.49 | 0.65 | 0.56 | 0.61 | 0.85 | 0.79 | 0.66 | 0.65 | 0.87 | 0.73 | 0.88 |
| Type 3 | 10 | 0.40 | 0.42 | 0.69 | 0.54 | 0.64 | 0.92 | 0.88 | 0.59 | 0.78 | 0.94 | 0.86 | 0.94 |
| | 20 | 0.48 | 0.51 | 0.66 | 0.55 | 0.61 | 0.83 | 0.82 | 0.62 | 0.76 | 0.90 | 0.78 | 0.89 |
| | 30 | 0.34 | 0.41 | 0.58 | 0.49 | 0.53 | 0.80 | 0.77 | 0.58 | 0.67 | 0.81 | 0.77 | 0.86 |
| | 40 | 0.32 | 0.35 | 0.57 | 0.43 | 0.49 | 0.70 | 0.68 | 0.59 | 0.65 | 0.80 | 0.65 | 0.84 |
| | 50 | 0.29 | 0.37 | 0.60 | 0.47 | 0.56 | 0.77 | 0.75 | 0.53 | 0.67 | 0.82 | 0.67 | 0.88 |
| Type 4 | 10 | 0.27 | 0.31 | 0.59 | 0.44 | 0.52 | 0.77 | 0.70 | 0.59 | 0.61 | 0.84 | 0.67 | 0.94 |
| | 20 | 0.26 | 0.30 | 0.50 | 0.35 | 0.42 | 0.80 | 0.70 | 0.51 | 0.57 | 0.89 | 0.65 | 0.89 |
| | 30 | 0.49 | 0.52 | 0.69 | 0.60 | 0.62 | 0.79 | 0.77 | 0.68 | 0.70 | 0.86 | 0.72 | 0.86 |
| | 40 | 0.22 | 0.27 | 0.55 | 0.38 | 0.47 | 0.76 | 0.71 | 0.58 | 0.60 | 0.77 | 0.64 | 0.84 |
| | 50 | 0.38 | 0.47 | 0.60 | 0.52 | 0.56 | 0.82 | 0.78 | 0.61 | 0.68 | 0.84 | 0.75 | 0.88 |

Fig. 10. Person's correlation coefficient of different models.

of Kriging, SARIMA, and ANN, but all increase when the missing ratio over 20%. For type 4 missing pattern, the MAPE of models except our proposed model start to increase when the missing ratio is over 10%. The performance of models in this group indicates that our proposed model is more accuracy than other models. The dominance is even more apparent as missing ratio increases.

Fig. 9 demonstrates the precision of the imputation by the length of the blue bar. From the definition, it can be known that the PV value closes to zero indicates the imputation result has well captured the detail pattern in the historical data. As shown, the ANN, ARMA, KNN, LVSV and LVTV have higher PV because they can only extract information from the adjacent sensors or timestamps. The local effect makes the variance of the imputation increase. On the other hand, the GVSV+GVTV and LVSV+LVTV has smaller PV that is acceptable. It proves the ensemble of information from different views can improve the performance in term of conserving the variance of the imputation. The proposed model almost having the smallest PV among all missing types and all missing ratio also support this finding.

Fig. 10 presents the correlation $r$ between imputed values and actual values. A higher $r$ indicates a more consistent estimation. The attained $r$ of our proposed model is higher than other models that indicates imputations generated by MVLM are more consistent with the actual values.

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.

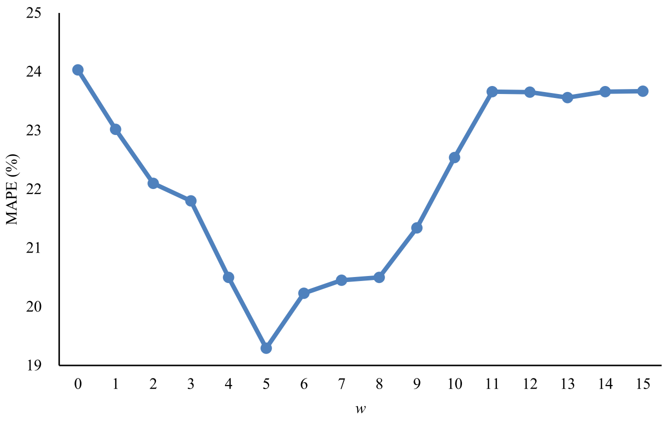
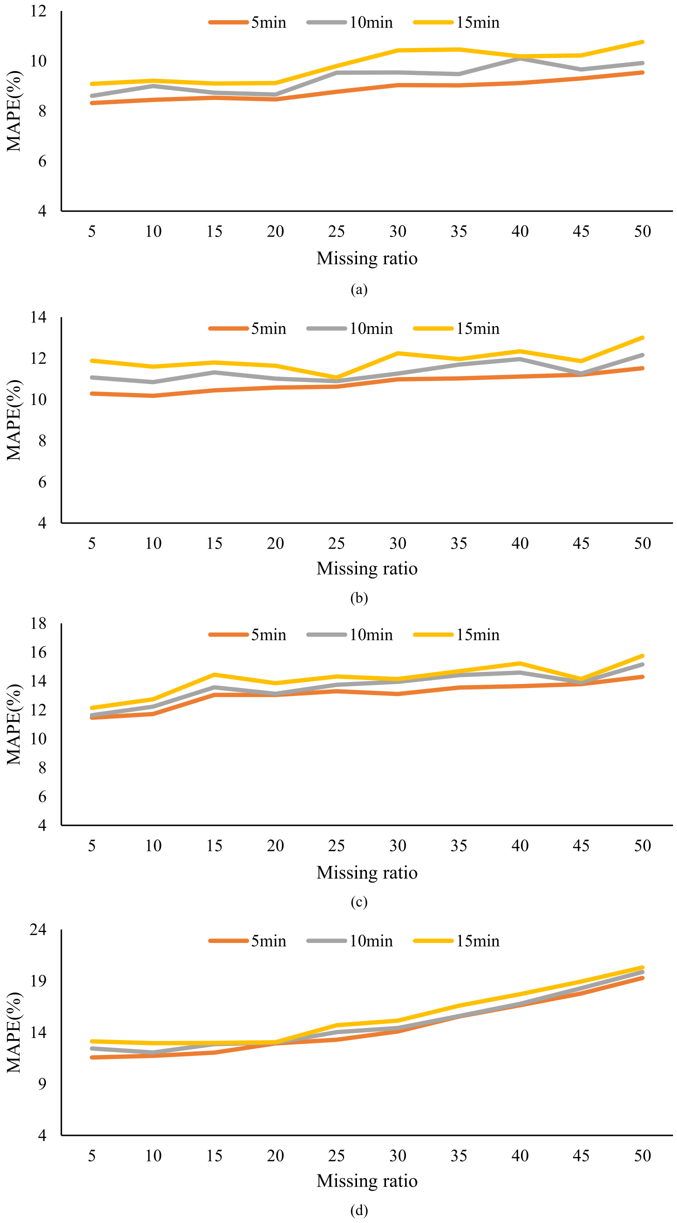


Fig. 11. Impact of window size.



(a)

(b)

(c)

(d)

Fig. 12. MAPE for different frequency.

### *B. Sensitivity of the Parameters*

Furthermore, the sensitivity of the parameters in the model are analyzed. Take block missing with 50% missing ratio for example, the effect of window size $\omega$ on the performance of the models are shown in Fig. 11. It can be seen during the initial stage, the accuracy decreases. After $\omega = 5$, the accuracy starts to increase. This varying pattern indicates that a large window size may fail to capture the local variation while a small window size may lose the correlation between sensors and timestamps.

### *C. Sensitivity of Frequency*

Moreover, to understand how sensitive the MVLM results are with respect to a different frequency of time series data, the origin data is aggregated into 10 min and 15 min and the missing values are generated and imputed the same as the above. After imputation, the MAPEs are shown in Fig.12. It can be found that MAPE is not decreasing or increasing significantly when the frequency is changing which indicates the proposed model is robust.

## V. Conclusion

Missing values in traffic time series data is a common problem of intelligent transportation system. In this paper, the concept of multi-view leaning is introduced into imputing the missing traffic data for the first time. The proposed method consists three widely used models: LSTM, SVR, and CF. The three models are applied to capture significant information in the data from different views to estimate the missing values, and then combine the results with a kernel function to improve the accuracy. The performance including accuracy, precision and agreement of the proposed model is compared to several widely used imputation methods (ANN, ARMA, SARIMA, KNN) as well as models from single view or double views (MVLT, GVSV, GVTV, LVSV, LVTV, GVSV+GVTV, LVSV+LVTV) in three different missing classes. The proposed MVLM outperforms in all types of missing patterns and provides robust results with different missing ratio. Especially in practice, sometimes the data is missed as block and with a high ratio. In this scenario, our method shows more advantages. Furthermore, the sensitivity of parameters in the models are analyzed.

In general, the contributions of this study are following four aspects: 1) MVLM simultaneously considers the temporal correlation between values at different timestamps and spatial correlations between sensors in highway network to estimate more accurate values, 2) MVLM integrates the global variation and local variation from temporal and spatial views based on data-driven algorithms to achieve better accuracy, 3) MVLM can handle different missing patterns even with high missing ratio, and 4) we evaluate our multi-learning framework using real data and compare it with several baselines.

Although the proposed method generates favorable imputation, it will be beneficial to consider some non-recurrent traffic conditions impacted by accidents, adverse weather conditions or some special events as input variables. Also, it will be interesting to incorporate the influence of the intersection into the method and apply it to roads in cities.

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.

## REFERENCES

[1] B. Ran, H. Tan, Y. Wu, and P. J. Jin, "Tensor based missing traffic data completion with spatial–temporal correlation," *Phys. A, Statist. Mech. Appl.*, vol. 446, pp. 54–63, Mar. 2016.

[2] 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.

[3] L. Li, S. He, J. Zhang, and B. Ran, "Short-term highway traffic flow prediction based on a hybrid strategy considering temporal–spatial information," *J. Adv. Transp.*, vol. 50, no. 8, pp. 2029–2040, Dec. 2016.

[4] H. Tan, Y. Wu, B. Shen, P. J. Jin, and B. Ran, "Short-term traffic prediction based on dynamic tensor completion," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2123–2133, Aug. 2016.

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

[6] J. Van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Accurate freeway travel time prediction with state-space neural networks under missing data," *Transp. Res. C, Emerg. Technol.*, vol. 13, nos. 5–6, pp. 347–369, Oct./Dec. 2005.

[7] C. Dong, C. Shao, S. H. Richards, and L. D. Han, "Flow rate and time mean speed predictions for the urban freeway network using state space models," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 20–32, Jun. 2014.

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

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

[10] M. Zhong, S. Sharma, and P. Lingras, "Genetically designed models for accurate imputation of missing traffic counts," *Transp. Res. Rec.*, vol. 1879, pp. 71–79, Jan. 2004.

[11] K. Henrickson, Y. Zou, and Y. Wang, "Flexible and robust method for missing loop detector data imputation," in *Proc. 94th Annu. Meeting Transp. Res. Board*, 2015, pp. 29–36.

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

[13] A. Muralidharan and R. Horowitz, "Imputation of ramp flow data for freeway traffic simulation," *Transp. Res. Rec.*, vol. 2099, pp. 58–64, Jan. 2009.

[14] C. Dong, S. H. Richards, Q. Yang, and C. Shao, "Combining the statistical model and heuristic model to predict flow rate," *J. Transp. Eng.*, vol. 140, no. 7, p. 04014023, 2014.

[15] S. Tak, S. Woo, and H. Yeo, "Data-driven imputation method for traffic data in sectional units of road links," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 6, pp. 1762–1771, Jun. 2016.

[16] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Inf. Sci.*, vol. 233, pp. 25–35, Jun. 2013.

[17] H. Chen, S. Grant-Muller, L. Mussone, and F. Montgomery, "A study of hybrid neural network approaches and the effects of missing data on traffic forecasting," *Neural Comput. Appl.*, vol. 10, no. 3, pp. 277–286, 2001.

[18] V. Ravi and M. Krishna, "A new online data imputation method based on general regression auto associative neural network," *Neurocomputing*, vol. 138, pp. 106–113, Aug. 2014.

[19] Y. Li, Z. Li, and L. Li, "Missing traffic data: Comparison of imputation methods," *IET Intell. Transp. Syst.*, vol. 8, no. 1, pp. 51–57, Feb. 2014.

[20] M. T. Asif, N. Mitrovic, J. Dauwels, and P. Jaillet, "Matrix and tensor based methods for missing data estimation in large traffic networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 7, pp. 1816–1825, Jul. 2016.

[21] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transp. Res. B, Methol.*, vol. 91, pp. 511–524, Sep. 2016.

[22] J. Wang, N. Zou, and G.-L. Chang, "Travel time prediction: Empirical analysis of missing data issues for advanced traveler information system applications," *Transp. Res. Rec.*, vol. 2049, pp. 81–91, 2008.

[23] P. Cai, Y. Wang, G. Lu, P. Chen, C. Ding, and J. Sun, "A spatiotemporal correlative $k$-nearest neighbor model for short-term traffic multistep forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 62, pp. 21–34, Jan. 2016.

[24] W. Tobler, "On the first law of geography: A reply," *Ann. Assoc. Amer. Geogr.*, vol. 94, no. 2, pp. 304–310, Jun. 2004.

[25] Y. Duan, Y. Lv, Y.-L. Liu, and F.-Y. Wang, "An efficient realization of deep learning for traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 72, pp. 168–181, Nov. 2016.

[26] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 187–197, May 2015.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Mar. 1997.

[28] X. Yi, Y. Zheng, J. Zhang, and T. Li, "ST-MVL: Filling missing values in geo-sensory time series data," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1–7.

[29] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, Jan. 2009, Art. no. 421425.

[30] C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1974.

[31] T. Choe, A. Skabardonis, and P. Varaiya, "Freeway performance measurement system: Operational analysis tool," *Transp. Res. Rec.*, vol. 1811, pp. 67–75, Jan. 2002.

[32] L. Li, J. Zhang, F. Yang, and B. Ran, "Robust and flexible strategy for missing data imputation in intelligent transportation system," *IET Intell. Transp. Syst.*, vol. 12, no. 2, pp. 151–157, Mar. 2018.

[33] J. D. Hamilton, *Time Series Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1994.

[34] Y. Wu, H. Tan, Y. Li, F. Li, and H. He, "Robust tensor decomposition based on Cauchy distribution and its applications," *Neurocomputing*, vol. 223, pp. 107–117, Feb. 2017.

[35] J. Tang, G. Zhang, Y. Wang, H. Wang, and F. Liu, "A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation," *Transp. Res. C, Emerg. Technol.*, vol. 51, pp. 29–40, Feb. 2015.

[36] S. He, J. Zhang, Y. Cheng, X. Wan, and B. Ran, "Freeway multisensor data fusion approach integrating data from cellphone probes and fixed sensors," *J. Sensors*, vol. 2016, Sep. 2016, Art. no. 7269382.

[37] W. Junger and A. P. de Leon, "Imputation of missing data in time series for air pollutants," *Atmos. Environ.*, vol. 102, pp. 96–104, Feb. 2015.

**Linchao Li** received the M.S. degree from Chang'an University, Xi'an, China, in 2013. He is currently pursuing the Ph.D. degree with the Research Center for Internet of Mobility, Southeast University, Nanjing, China.

His research interests are related to the use of machine learning in applications of transportation and predicting the traffic state.

**Jian Zhang** received the Ph.D. degree in 2011 from Southeast University, Nanjing, China, where he is currently the Vice Director of the Research Center for Internet of Mobility. His research interests include transportation application of mobile phone data, connected vehicles, and public transportation systems. He is also a member of the American Society of Civil Engineers.

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.

**Yonggang Wang** received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2009. He is currently an Associate Professor with the School of Highway, Chang'an University, Xi'an, China. He has authored or co-authored over 20 articles in international journals, including the *Transportation Letters*, *Traffic Injury Prevention*, and *Transport*. He serves as an Editorial Board Member of *Transport*, *Scientia Iranica*, *Promet Traffic & Transportation*, and the *International Journal of Advancements in Computing Technology*.

**Bin Ran** received the Ph.D. degree from the University of Illinois at Chicago, Chicago, IL, USA, in 1993.

He was the first Chairman of the Chinese Overseas Transportation Association, of which he is one of the co-founders. He is currently a Professor with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI, USA, and the Director of the Research Center for Internet of Mobility, Southeast University, Nanjing, China. He has authored or co-authored over 90 articles in international journals, including the *Transportation Science* journal, the *Transportation Research Part B* journal, and the *Transportation Research Part C* journal.

Authorized licensed use limited to: Central South University. Downloaded on December 12,2020 at 14:17:51 UTC from IEEE Xplore. Restrictions apply.