A Hybrid Data-Driven Framework for Spatiotemporal Traffic Flow Data Imputation

Peixiao Wang, Tao Hu, Fei Gao, Ruijie Wu, Wei Guo, and Xinyan Zhu

***Abstract*—An accurate estimation of missing data in traffic flow is crucial in urban planning, intelligent transportation, economic geography, and other fields. Thus, improving the data quality of traffic flow is a necessary step in data modeling. Most existing studies use data-driven models to determine spatiotemporal patterns in traffic flow data and fill in the missing information automatically. However, simple data-driven models have unsatisfactory results for describing complex patterns in traffic flow and filling in missing data. This study treated the complex patterns in traffic flow as integrating multiple simple patterns and proposed a hybrid missing data imputation framework called ST-PTD. We used a specific time-series analysis to mine periodic patterns and proposed a novel matrix decomposition method to describe the trend of the traffic flow data. Furthermore, we fused the periodic and trend characteristics of the missing data using a novel dendritic neural network. We applied the framework in actual traffic flow data sets collected in Wuhan, China. The results showed that the ST-PTD framework outperformed the eight existing baselines in terms of imputation accuracy.**

***Index Terms*—Dendritic net, matrix factorization, missing traffic flow data, spatiotemporal data imputation.**

I. Introduction

A. Motivation

AS SPATIOTEMPORAL data, traffic flow is widely used in many fields, such as urban planning, intelligent transportation, and economic geography [1]–[4]. However, due to limitations in data collection and privacy issues, incomplete traffic flow data are widespread, which restrict the performance of spatiotemporal data mining methods [5]–[8]. Spatiotemporal analysis of incomplete data sets may provide unreasonable inference and assumptions. However, directly omitting missing data is an inefficient data resource utilization [7], [9]. To solve this problem, the objective of this study is to propose a missing data imputation framework based on the patterns of traffic flow [10]–[12].

Manuscript received 5 November 2021; revised 6 January 2022; accepted 7 February 2022. Date of publication 14 February 2022; date of current version 24 August 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0502200; in part by the National Natural Science Foundation of China under Grant 41830645 and Grant 41701459; and in part by the National Science Foundation, USA, under Grant 1841403 and Grant 2027540. *(Corresponding author: Wei Guo.)*

Peixiao Wang, Fei Gao, Ruijie Wu, Wei Guo, and Xinyan Zhu are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: peixiaowang@whu.edu.cn; gaofei_gis@whu.edu.cn; jerrywu@whu.edu.cn; guowei-lmars@whu.edu.cn; xinyanzhu@whu.edu.cn).

Tao Hu is with the Department of Geography, Oklahoma State University, Stillwater, OK 74078 USA (e-mail: tao.hu@okstate.edu).

Digital Object Identifier 10.1109/JIOT.2022.3151238

B. Related Studies

Existing missing data imputation methods can be divided into two: 1) statistical learning and 2) data-driven methods. Classical statistical methods assume missing data obey certain mathematical rules in space and time dimensions and establish a specific parameter model to describe the pattern of missing data [13]. For example, the inverse distance weighting (IDW) model assumes the data distribution in the spatial dimension obeys the first law of geography and fills in the missing value by calculating the distance between the missing data and the observed data [14]. Kriging interpolation assumes the spatial distribution of the observation data satisfies the second-order stability and uses a covariance function to obtain the optimal linear unbiased estimation of the missing data [15]. The autoregressive (AR) integrated moving average (ARIMA) and simple exponential smoothing (SES) methods assume the observation data satisfy the time stationarity in the time dimension and infer the missing data based on the observation data of the preceding several moments [16], [17]. Some studies consider the characteristics of missing data in time and space and establish the corresponding statistical model, such as ST-IDW, ST Kriging, ST-ARIMA, and P-BSHADE [18]–[21]. Although classical statistical methods have been widely used in missing data imputation, achieving excellent results in traffic flow data set remains difficult. The classical statistical methods are based on strict assumptions, and the actual traffic environment deviates from this. In addition, the traffic flow data have complex spatiotemporal patterns, which are difficult to describe using specific mathematical formulas [22].

With the rapid development of artificial intelligence and high-performance computing, data-driven models have gradually become the mainstream in missing data imputation. Data-driven methods, such as machine learning, matrix factorization, neural networks, and deep learning, do not require data sets to obey specific mathematical rules but establish nonparametric models to automatically mine the spatiotemporal characteristics to impute the missing values [23]. For example, Chang and Ge [24] applied a variety of traditional matrix decomposition models for traffic flow data imputation and compared the performance of their models. Asif *et al.* [12] proposed a variant matrix factorization model based on traditional methods to extract global traffic patterns in large-scale road networks, estimating missing values in traffic flow data.

Yu *et al.* [25] integrated time dependence into the traditional matrix factorization model and proposed a new temporal regularized matrix factorization (TRMF) to estimate missing values in traffic flow data. Chen *et al.* [26]–[28] extended matrix factorization to tensor factorization, mining the missing patterns in traffic flow data from a higher dimensional perspective to impute the missing values in the data set. In addition, relevant studies apply deep learning algorithms to reconstruct missing data and achieve good results [29], [30]. For instance, Cheng *et al.* [31] used extreme learning machine (ELM) to integrate IDW and SES algorithms and proposed a lightweight missing data interpolation model. Li *et al.* [6] combined deep neural networks and P-BSHADE algorithm to propose a hybrid two-step estimation framework. Compared to classical statistical methods, data-driven methods do not require prior knowledge and explicit mathematical expressions and have reliable data imputation results. However, when the spatiotemporal pattern in the missing data set is more complex, simple data-driven models often cannot obtain satisfactory data filling results. The reason is that simple data-driven models cannot adequately describe the complex spatiotemporal patterns in missing data sets.

C. Strategy

Though missing data imputation gains interest, especially for traffic data quality enhancement, its reliability issues persist; i.e., spatiotemporal patterns in the traffic flow are not only difficult to describe using specific mathematical expressions but also to automatically capture by simple data-driven models [32]. To address this problem, inspired by multiview learning [33], [34], we treated the complex patterns in traffic flow as integration of multiple simple patterns, established a unique data-driven model for a single pattern, and finally, improved the performance of missing data imputation by fusing the data imputation results of different models.

The research shows that traffic flow data have typical trend and periodic characteristics [35]. To simultaneously characterize the trend and periodicity of the missing traffic flow, a hybrid traffic flow imputation framework called ST-PTD is proposed. First, a specific time-series model is used to mine the periodic patterns of the traffic flow. Then, a novel matrix factorization method considering bidirectional temporal dependence is proposed to describe the trend characteristic of traffic flow. Finally, we used a new dendritic neural network to obtain the final imputation results. This study provided the following significant contributions.

1) According to the trend and periodic characteristics of traffic flow, we defined two different processing strategies. Specifically, we used a specific time-series analysis to mine periodic patterns and proposed a novel matrix decomposition method to describe the trend of the traffic flow data. By manually extracting periodic and trending features, we alleviate the difficulty of the model to automatically mine complex traffic patterns.
2) We introduced a novel dendritic neural network (DD) [36] to fuse the results of multiple models. This improved the practicability of simple data-driven model in traffic flow data imputation.
3) We evaluated the performance of the ST-PTD framework using actual traffic flow data. The results demonstrated the advantages of our framework compared to eight baseline methods.

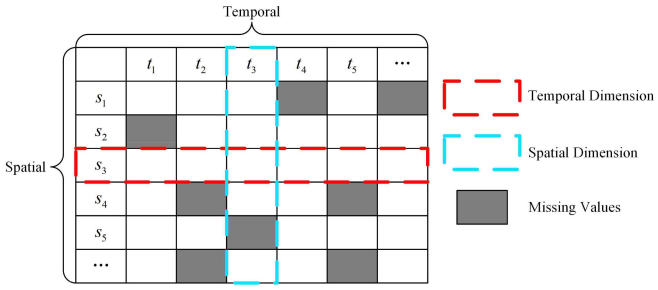


Fig. 1. Data structure of the traffic flow data.

II. Preliminaries and Problem Definitions

In this study, we regarded traffic flow data as a spatiotemporal matrix with missing data $X(S, T)$, where $S = \{s_i\}_{i=1}^{m}$ represents the spatial dimension index of data, $T = \{t_j\}_{js=1}^{n}$ represents the time dimension index of data, m represents the total number of spatial objects, and n is the total number of timestamps. As shown in Fig. 1, $X(s_2, t_1) = \phi$ indicates that the traffic flow data of the spatial object s_2 at the time stamp t_1 are missing.

This study aims to estimate the missing traffic flow data according to observed traffic flow data, and the process is shown in

$$\begin{cases} \hat{X}(S,T) = & \mathcal{M} \leftarrow X(S,T) \\ \forall (i,j) \in \Omega, & X\left(s_i, t_j\right) = \phi \\ \forall (i,j) \in \Omega, & \hat{X}\left(s_i, t_j\right) \neq \phi \end{cases} \tag{1}$$

where $X(S, T)$, $\mathcal{M}$, $\hat{X}(S, T)$, and Ω represent the spatiotemporal matrix with missing data, the missing data imputation framework proposed in this study, the spatiotemporal matrix obtained after imputation, and an indexed collection of missing data, respectively.

III. Methodology

In this section, we describe the proposed hybrid framework called ST-PTD for traffic flow data imputation, whose structure is presented in Fig. 2. The ST-PTD framework is mainly composed of three parts: 1) spatiotemporal periodic matrix modeling; 2) spatiotemporal trend matrix factorization; and 3) multiple fusion, which are discussed in Sections III-A–III-C, respectively. First, a specific time-series analysis was used to mine the periodic patterns of traffic flow data. Then, a novel bidirectional AR matrix decomposition (BiARMF) method was proposed to describe the trend information of traffic flow data. Finally, a novel DD network was used to fuse the periodic and trend characteristics of the missing data to obtain the final imputation result.

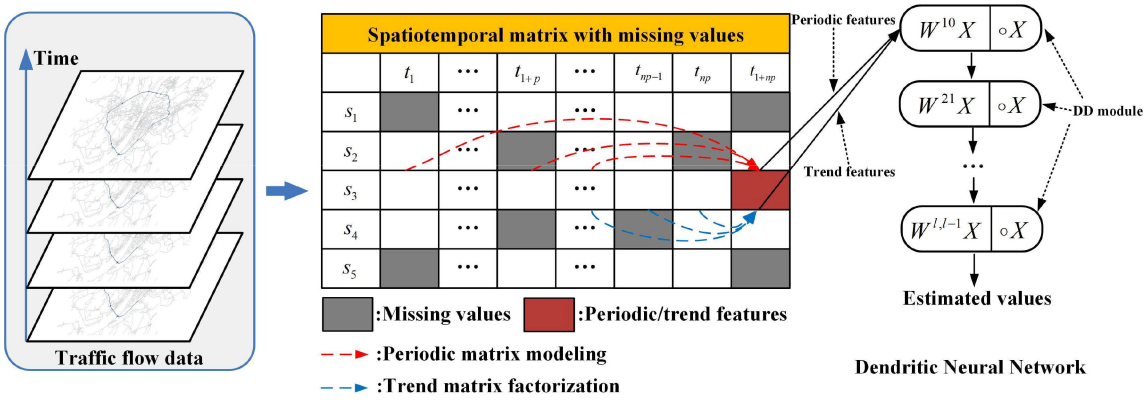


Fig. 2. Schematic of the overall ST-PTD framework process.

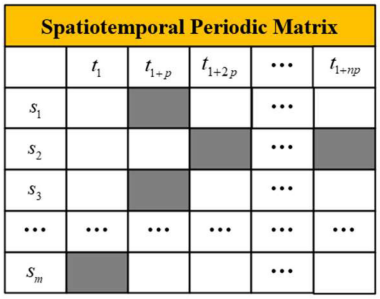


Fig. 3. Schematic of the spatiotemporal periodic matrix.

A. Modeling of the Spatiotemporal Periodic Matrix

Periodicity is a typical daily or weekly repetition pattern of traffic flow data. To use the traffic flow periodic pattern, we constructed a spatiotemporal periodic matrix of the traffic flow. Fig. 3 shows the data structure of the spatiotemporal periodic matrix, where p, m, and n represent the size of the period, the number of roads, and the number of time periods, respectively. The following are two advantages in exploring the missing pattern of traffic flow data from their periodicity: first, the data from each road in the spatiotemporal periodic matrix is stable in time; that is, the traffic flow data fluctuate around the mean value, making it conducive for mathematical modeling. Second, the adjacent data in the spatiotemporal periodic matrix have a time difference of period, which alleviates the impact on missing data imputation.

Based on the stable characteristics of the spatiotemporal periodic matrix, the filling value of traffic flow on different roads can be defined by

$$\begin{cases} \hat{v}_{1j}^{sp} = u_{1j} + b_{1j} \\ \hat{v}_{2j}^{sp} = u_{2j} + b_{2j} \\ \hat{v}_{3j}^{sp} = u_{3j} + b_{3j} \\ \quad \cdots \\ \hat{v}_{mj}^{sp} = u_{mj} + b_{mj} \end{cases} \tag{2}$$

where $\hat{v}_{mj}^{sp}$ represents the estimated value of road s_m at the time t_j with time step p as the period; u_{mj} represents the mean value of observable data of road s_m at the time $\{t_{j+ip}\}_{i=0}^{n}$ with time step p as the period, that is, the traffic flow of s_m at $\{t_{j+ip}\}_{i=0}^{n}$ fluctuates around u_{mj}; and b_{mj} represents the optimized parameter, that is, the bias of the traffic flow of road s_m at time $\{t_{j+ip}\}_{i=0}^{n}$ relative to the mean u_{mj}. In (2), $\hat{v}_{mj}^{sp}$, u_{mj}, and b_{mj} are scalars.

Taking road s_m as an example, the value of b_{mj} can be trained by minimizing the square loss between the estimated and true traffic volume. The loss function is defined by

$$\mathcal{L}(b_{mj}) = \frac{1}{2} \sum_{j \notin \Omega_m} \left(\hat{v}_{mj}^{sp} - v_{mj} \right)^2 \tag{3}$$

where Ω_m represents the index set of missing data in road s_m, that is, the only cumulative loss of the square error of the observable data; $\hat{v}_{mj}^{sp}$ represents the estimated traffic flow from the periodic view; and v_{mj} represents the expected output of the model.

B. Factorization of the Spatiotemporal Trend Matrix

Compared to the periodic characteristics of traffic flow, its trends are more complicated. Fig. 4 shows the data structure of the spatiotemporal trend matrix, and the adjacent elements of the spatiotemporal trend matrix in the time dimension differ by only one time window. That is, the traffic flow trend of a road at a specific time is affected by its surrounding roads and moments. In order to describe the correlation between time and space and to automatically extract the trend features in the traffic flow, a matrix factorization model is often used. However, the traditional matrix factorization model only determines the spatiotemporal trend matrix from spatial and temporal state matrices but does not explicitly define the trend characteristics in the traffic flow. Thus, BiARMF method was proposed to fill in the missing information in the traffic flow data set.

The BiARMF method is derived from the AR model, where the state at the current moment is equal to a linear combination of the state at previous moments [37]. Considering the time lag in the AR model, we introduced the trend characteristics of traffic flow from two directions (forward and backward). As shown in Fig. 4, the spatiotemporal trend matrix $\boldsymbol{STM} \in \mathbb{R}^{m \times n}$ is decomposed into the spatial state matrix $\boldsymbol{SSM} \in \mathbb{R}^{m \times k}$ and the temporal state matrix $\boldsymbol{TSM} \in \mathbb{R}^{k \times n}$. Additionally, the temporal state matrix $\boldsymbol{TSM}$ implies trend dependence. Taking time t_{n-2} as an example, the forward and backward trend

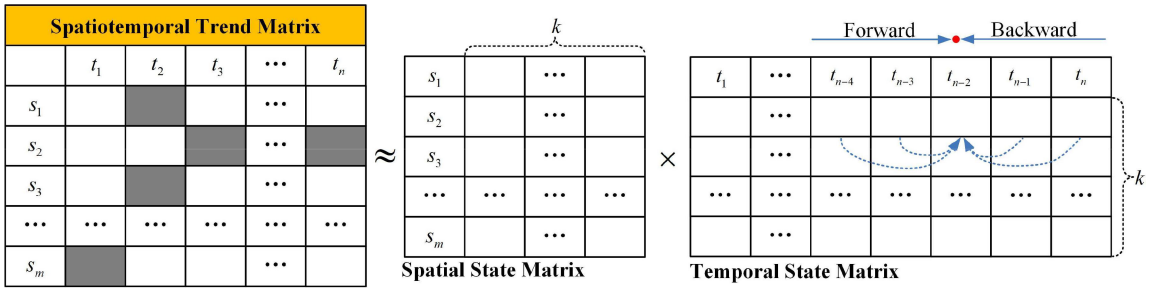


Fig. 4. Schematic of the spatiotemporal trend matrix factorization.

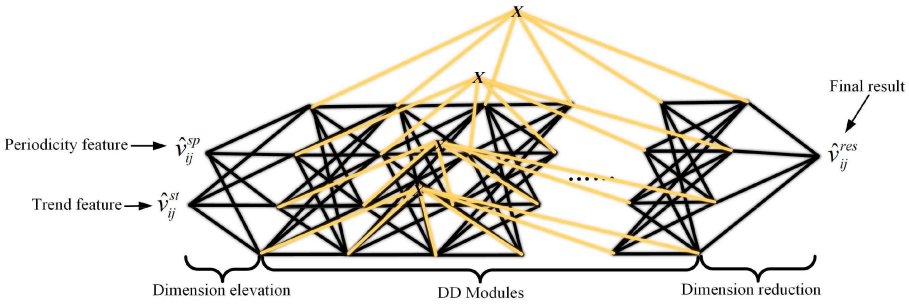


Fig. 5. Fusion of multiple results.

dependence of $\boldsymbol{tsm}_{n-2} \in \boldsymbol{TSM}$ is expressed

$$\boldsymbol{STM} = \boldsymbol{SMM} \odot \boldsymbol{TSM} \text{ s.t.} \begin{cases} \boldsymbol{tsm}_{n-2} \approx \sum_{i=n-2-l_f}^{n-3} \boldsymbol{\theta}_i^f \bigotimes \boldsymbol{tsm}_i & l_f \geq 1 \\ \boldsymbol{tsm}_{n-2} \approx \sum_{i=n-1}^{n-2+l_b} \boldsymbol{\theta}_i^b \bigotimes \boldsymbol{tsm}_i & l_b \geq 1 \end{cases} \tag{4}$$

where $\boldsymbol{tsm}_{n-2} \in \mathbb{R}^{k\times 1}$ represents the time state vector of traffic flow at time t_{n-2} and k is the hyperparameter in the traditional matrix decomposition; l_f and l_b are hyperparameters, which represent the time-dependent length of forward and backward directions, respectively; $\boldsymbol{\theta}_i^f \in \mathbb{R}^{k\times 1}$ and $\boldsymbol{\theta}_i^b \in \mathbb{R}^{k\times 1}$ are optimizable parameters, representing forward and backward time-dependent weights (vectors), respectively; $\bigotimes$ represents the multiplication of vectors; and $\odot$ means matrix multiplication.

The solution of matrix factorization considering trend constraints differs from that of traditional matrix factorization. The BiARMF model parameters were optimized by adding the bidirectional time dependency. To simplify the operation, we decomposed the bidirectional time dependence on two unidirectional time dependencies, and the loss function is expressed

$$\mathcal{L}(\boldsymbol{SSM}, \boldsymbol{TSM}, \boldsymbol{\Theta}_f, \boldsymbol{\Theta}_b) = \sum_{(i,j)\notin\boldsymbol{\Omega}} \frac{1}{2}\left(\hat{v}_{ij}^{st} - \boldsymbol{ssm}_i^T \bigotimes \boldsymbol{tsm}_j\right)^2 + R(\boldsymbol{TSM}|\boldsymbol{\Theta}_f, l_f) + R(\boldsymbol{TSM}|\boldsymbol{\Theta}_b, l_b) \tag{5}$$

$$\begin{cases} R(\boldsymbol{TSM}|\boldsymbol{\Theta}_f, l_f) = \sum_{t=l_f}^{n} \left(\boldsymbol{tsm}_t - \sum_{h=t-l_f}^{t-1} \boldsymbol{\theta}_h^f \bigotimes \boldsymbol{tsm}_h\right)^T \left(\boldsymbol{tsm}_t - \sum_{h=t-l_f}^{t-1} \boldsymbol{\theta}_h^f \bigotimes \boldsymbol{tsm}_h\right) \\ R(\boldsymbol{TSM}|\boldsymbol{\Theta}_b, l_b) = \sum_{t}^{n-l_b} \left(\boldsymbol{tsm}_t - \sum_{h=t+1}^{t+l_b} \boldsymbol{\theta}_h^b \bigotimes \boldsymbol{tsm}_h\right)^T \left(\boldsymbol{tsm}_t - \sum_{h=t+1}^{t+l_b} \boldsymbol{\theta}_h^b \bigotimes \boldsymbol{tsm}_h\right) \end{cases} \tag{6}$$

where $\boldsymbol{ssm}_i \in \boldsymbol{SSM}$ represents the spatial status vector of the road s_i; $\boldsymbol{tsm}_j \in \boldsymbol{TSM}$ is the time status vector at the moment t_j; l_f, l_b, $\boldsymbol{\theta}_h^b$, and $\boldsymbol{\theta}_h^f$ have the same definition in (4); $\boldsymbol{\Omega}$ represents the index set of missing data; and $\hat{v}_{ij}^{st}$ represents the estimated value (scalar) of the road s_i at time t_j considering the forward and backward trend.

C. Multiple Fusion

Based on the characteristics of traffic flow periodicity and trend, two different data-driven models have been established, but each model has a limited description of the missing pattern. Therefore, we attempted to fuse the results of different data-driven models to improve the accuracy of missing data filling. Considering the discernible logical relationship between data filling results from different perspectives (the results of data filling under multiple models are essentially different manifestations of the missing patterns of traffic flow), the DD network was used to fuse the different data imputation results.

The DD network is a new deep learning method based on gang neuron, which completes the function mapping from input to output by learning the logical relationship in the data [36]. Compared to the traditional cell body network, the DD network has low computational complexity and controllable precision. As shown in Fig. 5, the DD network adopted in this study is mainly composed of three parts. First, $\hat{v}_{ij}^{st}$ and $\hat{v}_{ij}^{sp}$ enter the DD modules through linear dimension elevation. Then, the final data fusion result $\hat{v}_{ij}^{res}$ is obtained through linear dimension reduction. The process of data fusion is expressed

$$\begin{cases} \boldsymbol{X} = \sigma\left(\boldsymbol{W}^e \odot \begin{pmatrix} \hat{v}_{ij}^{st} \\ \hat{v}_{ij}^{sp} \end{pmatrix}\right) \\ \boldsymbol{A}^2 = \boldsymbol{W}^{2,1}\boldsymbol{A}^1 \odot \boldsymbol{X} \\ \boldsymbol{A}^3 = \boldsymbol{W}^{3,2}\boldsymbol{A}^2 \odot \boldsymbol{X} \\ \quad \cdots\cdots \\ \boldsymbol{A}^L = \boldsymbol{W}^{L,L-1}\boldsymbol{A}^{L-1} \odot \boldsymbol{X} \\ \hat{v}_{ij}^{\text{res}} = \sigma\left(\boldsymbol{W}^r\boldsymbol{A}^L\right) \end{cases} \tag{7}$$

Algorithm 1 Training Process of ST-STD

Require: Training samples : $\boldsymbol{X} = \{v_{ij}\}$
Missing rate : c
Hyperparameters of BiARMF : l_f, l_b, k
Number of DD modules : num
Ensure: ST-STD model: $\mathcal{M}$
1: construct $\boldsymbol{X}_{\boldsymbol{\Omega}}$ and $\boldsymbol{\Omega}$ based samples $\boldsymbol{X}$ and miss rate c
//obtain imputation results from periodic and trend views
2: construct $\boldsymbol{SPM}$ and $\boldsymbol{STM}$ based $\boldsymbol{X}_{\boldsymbol{\Omega}}$
3: construct $\{\hat{v}_{ij}^{sp}\}$ by (2) and (3)
4: construct $\{\hat{v}_{ij}^{st}\}$ by (5) and (6) with l_f, l_b, k
//construct training instances
5: $\mathcal{D} \leftarrow \emptyset$
6: **for** each $(i, j) \notin \boldsymbol{\Omega}$ **do**
7: put a training instance $(\{\hat{v}_{ij}^{sp}, \hat{v}_{ij}^{st}\}, v_{ij})$ into $\mathcal{D}$
//train ST-STD framework
8: initialize the parameters $\boldsymbol{W}$
9: **repeat**
10: randomly select a batch of instances $\mathcal{D}_b$ from $\mathcal{D}$
11: find $\boldsymbol{W}$ by minimizing (8) with $\mathcal{D}_b$ and num
12: **until** stopping criteria is met
13: output the learned ST-STD model $\mathcal{M}$

where $\boldsymbol{W}^e$ represents the weight matrix of linear dimension elevation; $\boldsymbol{W}^r$ the weight matrix of linear dimension reduction; $\boldsymbol{W}^{L,L-1}$ the weight matrix of DD modules, where L is the number of DD modules; $\boldsymbol{A}^L$ the intermediate output of the DD module, where $\boldsymbol{A}^1 = \boldsymbol{X}$; $\odot$ represents the Hadamard product; and σ the sigmoid activation function. From (7), the largest difference between DD and cellular networks is that there is no activation function in the DD module.

The DD network can be trained by minimizing the square loss between the fused and true traffic volumes. The loss function is defined

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2} \sum_{(i,j) \notin \boldsymbol{\Omega}} \left(\hat{v}_{ij}^{\text{res}} - v_{ij} \right)^2 \tag{8}$$

where $\boldsymbol{W}$ represents all learnable parameters in the DD network, i.e., $\boldsymbol{W}^e$, $\boldsymbol{W}^r$, and $\boldsymbol{W}^{L,L-1}$; $\boldsymbol{\Omega}$ are the index set of missing data, $\hat{v}_{ij}^{\text{res}}$ the fusion result of DD network; and v_{mj} the expected output of the model.

D. Algorithms and Optimization

The basic principle of the ST-STD framework is to establish a supervised learning method, which takes the data filling results from periodic and trend perspectives as the input of the DD network to estimate the final missing traffic flow information. In order to train the ST-STD framework, the traffic flow data are divided into training and test samples. The training samples are used to train the parameters of the framework, while the test data are used to test the imputation performance. The training process of framework $\mathcal{M}$ is shown in Algorithm 1. Based on the training samples $\boldsymbol{X}$, we constructed a spatiotemporal matrix $\boldsymbol{X}_{\boldsymbol{\Omega}}$ with missing rate c, where $\boldsymbol{\Omega}$ records the index of missing data (line 1). Then, the interpolation results of the missing data are obtained from the periodic and trend perspectives (lines 2–4). Finally, the interpolation results from multiple perspectives and truth values are combined to form training instances (lines 6 and 7) to optimize the final ST-STD framework (lines 9–12).

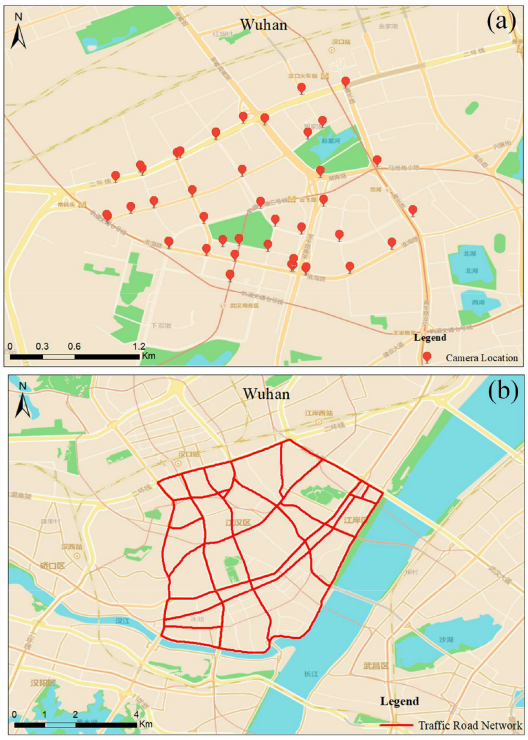


Fig. 6. Sketch map of the study area. (a) Spatial distribution of the experimental cameras. (b) Spatial extent of the road network.

TABLE I
FIELD INFORMATION OF AVI DATA

Field Name	Field Type	Description
Camera ID	Number	Camera unique identifier
Plate number	String	License plate number
Capture Time	Time	Time of photo shooting
Latitude	Number	Latitude of the camera
Longitude	Number	Longitude of the camera

IV. RESULTS AND DISCUSSION

A. Data Preparation

1) Data Sources: Two data sets were used to evaluate the performance of the ST-PTD framework: 1) the automatic vehicle identification (AVI) data and 2) vehicle trajectory data in an area in Wuhan, China.

The AVI data are based on the license plate recognition technology, and the geographic coordinates of the vehicle in the road space are automatically derived from the photos taken by the camera. The AVI data were gathered last March 1–28, 2021, covering 44 cameras. Fig. 6(a) shows the spatial distribution of the experimental cameras. Table I describes the field information of AVI data, in which the plate number is encrypted to obtain a unique identification of the vehicle.

The vehicle trajectory data are based on GPS technology, which automatically recognizes the position of a vehicle in the road space. The timespan of the GPS data is from August 1, 2018, to August 28, 2018. Owing to the large number of all

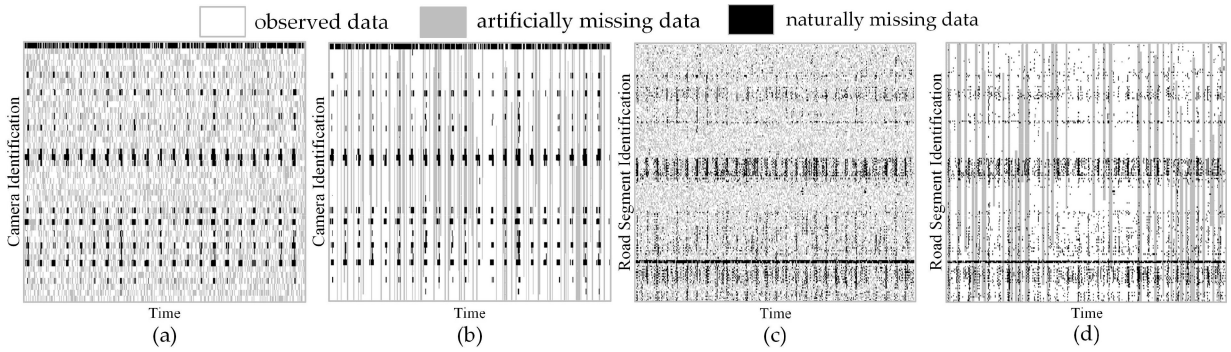


Fig. 7. Traffic status information after preprocessing: (a) traffic volume under random missing, (b) traffic volume under block missing, (c) traffic speed under random missing, and (d) traffic speed under block missing.

TABLE II
FIELD INFORMATION OF VEHICLE TRAJECTORY DATA

Field Name	Field Type	Description
Vehicle ID	Number	Vehicle unique identifier
Time	String	Data recording time
Latitude	Number	Latitude of the vehicle
Longitude	Number	Longitude of the vehicle

TABLE III
DESCRIPTION OF THE PROCESSED DATA SETS

Dataset	AVI Data	GPS Data
Indicator	Traffic volume	Traffic Speed
Time interval	5 min	15 min
Period step	288 (60/15*24)	96 (60/15*24)
Spatial objects	44	181
Temporal objects	8064 (28*288)	2688 (28*96)
Timespan	2021/3/1–2021/3/28	2018/8/1–2018/8/28

trajectory data points in Wuhan, we limited the study area of the trajectory data. Fig. 6(b) shows the limited spatial area of trajectory points in the experiment. Table II describes in detail the field information of the GPS data.

2) Data Preprocessing: Although the AVI and GPS data record vehicle trajectory, they do not directly count the traffic volume and traffic speed information on the road. Therefore, data preprocessing is conducted for the original data using the following preprocessing process.

1) For the AVI data, the traffic volume under different devices was counted at the time interval of 5 min. For the GPS data, we mapped the trajectory points to the traffic road segments and then calculated the average traffic speeds of different segments at the time interval of 15 min. Table III describes the numerical characteristics of the processed data set.
2) To train the ST-PTD model, based on the two missing types (i.e., random missing and block missing), partial traffic information was deleted at 20% and 40% missing rates, respectively. Among them, the random missing means that data missing is random, which is generally caused by poor equipment signal. The block missing indicates that data missing is continuity, which is generally caused by continuous equipment failure or power failure. Fig. 7 shows traffic volume and traffic speed information at 40% missing rate, where the natural missing data represents unobserved data caused by equipment failure.

B. Evaluation Metrics and Comparative Methods

1) Evaluation Metrics: In missing data imputation, a critical issue is how to evaluate the performance of the imputation model. In this study, the mean absolute error (MAE) and root mean square error (RMSE) between the imputation and true values were used as quantitative indicators to verify the imputing accuracy of the proposed model. The MAE and RMSE are calculated using

$$\mathrm{MAE} = \frac{1}{N}\sum_{(i,j)\in\mathbf{\Omega}} \left|v_{ij} - \widehat{v_{ij}}\right| \tag{9}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{(i,j)\in\mathbf{\Omega}} \left(v_{ij} - \widehat{v_{ij}}\right)^2} \tag{10}$$

where $\mathbf{\Omega}$ represents the index set of missing data; N represents the total number of missing data, i.e., $N = |\mathbf{\Omega}|$; v_{ij} represents the real traffic flow of road s_i at time t_j; and $\widehat{v_{ij}}$ represents the traffic flow estimated by the model on the road s_i at time t_j.

2) Comparative Methods: To comprehensively evaluate the performance of the ST-STD framework, we used eight baseline methods for comparison that are based on two missing types.

1) *ST-IDW [19]:* Spatiotemporal inverse distance interpolation (ST-IDW) is a statistical model that defines spatiotemporal distance and uses an IDW model to impute missing traffic volume in each road segment.
2) *ST-KNN [38]:* Spatiotemporal K-nearest neighbors (ST-KNNs) is a data-driven model, which fills the missing traffic condition by searching the k spatiotemporal nearest neighbors in the historical database.
3) *ST-2SMR [6]:* Spatiotemporal two-step missing data reconstruction (ST-2SMR) is a data-driven model that considers the missing patterns of the data set and uses a neural network to integrate coarse and fine interpolation results to improve the model's imputation performance.
4) *ST-ISE [31]:* Lightweight ensemble spatiotemporal interpolation (ST-ISE) is also a data-driven model that fills missing traffic volume in each road segment using ELM to integrate SES and IDW interpolation results.
5) *TRMF [25]:* TRMF is a data-driven model that incorporates temporal dependencies as a regularization term into commonly used matrix factorization to fill missing traffic volume in each road segment.

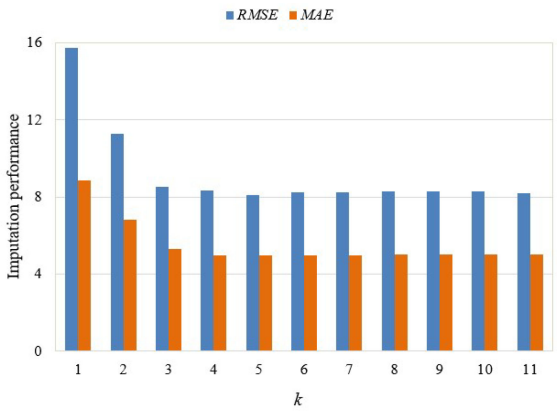


Fig. 8. Impact of parameter k on the imputation performance.

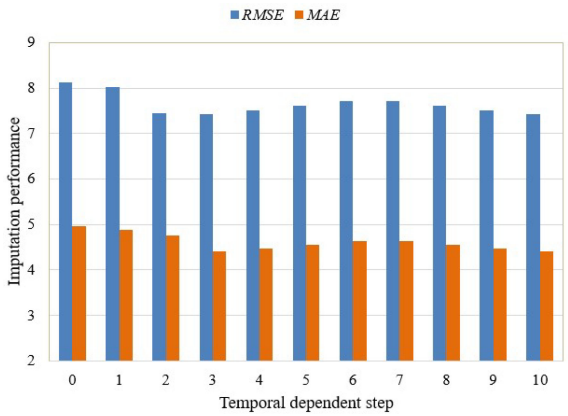


Fig. 9. Impact of parameters l_f and l_b on the imputation performance.

6) *BTMF [28]:* Bayesian temporal matrix factorization (BTMF) is a variation of the TRMF model that incorporates Bayesian theory into the solution of the TRMF model to fill missing traffic volume in each road segment.
7) *BGCP [26]:* Bayesian Gaussian CANDECOMP/PARAFAC (BGCP) is a data-driven model that extends the matrix factorization to the tensor factorization to improve the imputation performance.
8) *LRTC-TNN [27]:* Low-rank tensor completion with truncated nuclear norm (LRTC-TNN) is a data-driven method that fills the missing traffic condition by factorizing traffic tensor of location × day × time windows.

C. Variable Estimation

In this section, using the random missing rate of 20% as an example, the hyperparameter calibration process of the ST-PTD framework on the traffic volume data set is analyzed. The hyperparameters of the ST-PTD framework predominantly include the trend-forward dependent step size, l_f; trend-backward dependent step size, l_b; rank of matrix factorization, k; and the number of DD modules, *num*. To determine the optimal hyperparameter of the framework, the control variable method was used to obtain the combination of parameter values with the best imputation accuracy. In the parameter estimation stage, l_f, l_b, and k in the BiARMF algorithm were first determined. Then, the number of DD modules, *num*, was adjusted to test the imputation accuracy.

1) Calibrating the Parameters of BiARMF: In the matrix factorization model, the rank k of the matrix factorization plays an important role in the imputation process. During parameter calibration, we set the range of k to $[1, 2, \ldots, 11]$ and used cross-validation to obtain the best k and optimal combination of the parameters. Fig. 8 shows the effect of hyperparameter k on the imputation performance. The RMSE and MAE decreased initially and then stabilize with an increase of k. These results allow us to determine the optimal value for k for different data sets. When $k = 5$, the optimal imputation performance of the model is obtained on the traffic volume data.

After determining k, the time-dependent steps l_f and l_b are identified. To simplify the complexity of parameter adjustment, let l_f be identical to l_b. The optimal parameters of the time-dependent steps l_f and l_b are found in $[0, 1, 2, \ldots, 10]$. The effect of hyperparameters l_f and l_b on the imputation performance is shown in Fig. 9. These results allow us to determine the optimal l_f and l_b values for different data sets. The results show that MAE and RMSE decrease initially and then stabilize with the increasing step. When $l_f = l_b = 3$, the model has the best imputation performance.

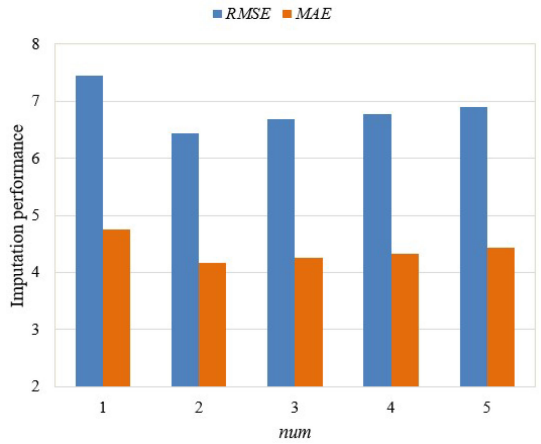


Fig. 10. Impact of the number of DD modules on imputation performance.

2) Calibrating the Number of DD Modules: The impact of the number of DD modules on the model performance was further verified. In the dendritic neural network, the optimal number of DD modules is found in [1, 2, 3, 4, 5]. The imputation results are shown in Fig. 10. As num increases, the imputation performance of the model initially decreases and then increases. However, when num is between [1, 2], the imputation performance of the model decreases. When num $>$ 2, the imputation performance of the model tends to increase. Therefore, when num $=$ 2, the optimal imputation performance of the model is obtained.

D. Comparison With Baselines

In this section, we analyze first the imputation performance of the ST-PTD and baselines under random missing and then the imputation performance of the ST-PTD and baselines under block missing.

Table IV shows the comparison results between ST-PTD and baselines under random missing in the traffic volume and

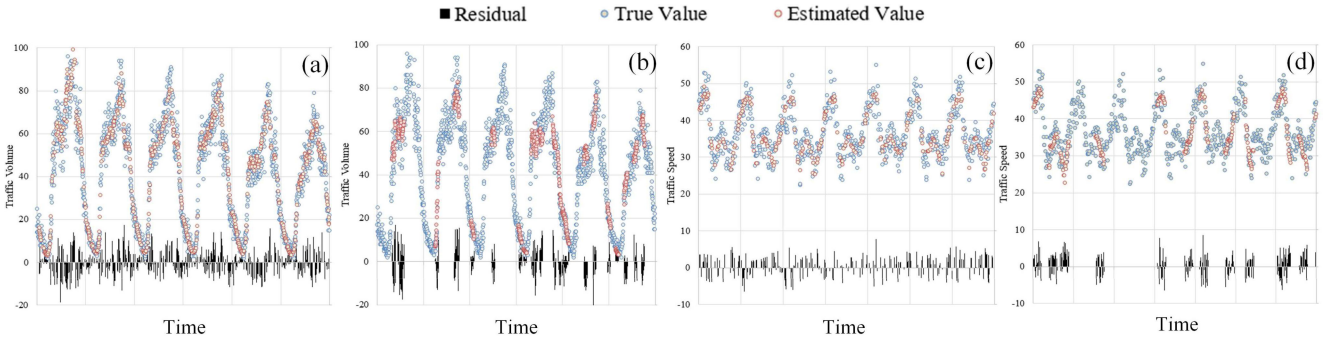


Fig. 11. Imputation values and corresponding actual values under 20% missing rate. (a) Random missing on traffic volume. (b) Block missing on traffic volume. (c) Random missing on traffic speed. (d) Block missing on traffic speed.

TABLE IV
COMPARISON RESULTS (IN MAE/RMSE) WITH BASELINES FOR RANDOM MISSING

Models	Traffic Volume		Traffic Speed	
	MR: 20%	MR: 40%	MR: 20%	MR: 40%
ST-IDW	7.74/14.06	9.11/16.56	7.38/9.61	8.90/11.24
ST-KNN	5.74/9.67	7.39/13.31	5.15/7.76	6.38/8.25
ST-ISE	4.91/7.30	5.31/7.87	3.26/4.34	3.38/4.71
ST-2SMR	4.95/7.15	5.48/8.95	3.44/4.71	3.82/5.19
TRMF	5.46/8.91	6.01/9.35	4.51/5.69	4.12/5.20
BTMF	5.28/7.73	6.05/9.69	3.12/3.78	3.13/3.80
LRTC-TNN	4.51/6.95	4.60/7.20	3.04/3.70	3.12/3.83
BGCP	4.67/7.26	4.75/7.45	3.05/3.70	3.06/3.72
ST-PTD	**4.17/6.43**	**4.15/6.43**	**2.91/3.42**	**2.93/3.43**

TABLE V
COMPARISON RESULTS (IN MAE/RMSE) WITH BASELINES FOR BLOCK MISSING

Models	Traffic Volume		Traffic Speed	
	MR: 20%	MR: 40%	MR: 20%	MR: 40%
ST-IDW	13.53/26.04	15.46/30.52	13.28/16.38	14.18/17.32
ST-KNN	10.96/19.39	11.66/23.04	8.38/11.40	8.94/13.26
ST-ISE	5.88/11.29	6.20/12.04	3.47/5.71	3.69/5.28
ST-2SMR	5.36/9.42	6.57/13.85	3.73/5.00	4.07/6.73
TRMF	7.65/14.31	9.39/18.17	5.01/6.88	5.25/6.93
BTMF	6.68/12.86	6.87/12.27	4.56/6.09	4.78/6.34
LRTC-TNN	5.56/8.80	5.80/9.13	3.16/3.88	3.24/4.03
BGCP	5.40/9.33	5.68/10.07	3.12/3.76	3.14/3.82
ST-PTD	**4.98/8.12**	**4.90/8.35**	**2.93/3.45**	**2.94/3.54**

traffic speed data sets. The results show that the imputation performance of all models decreases slightly with an increase of the missing rate under random missing. Among them, the imputation performance of the ST-IDW model is affected by the missing rate. Except for the ST-IDW model, other models were less affected by the missing rate. That is, the data-driven model is less affected by the missing rate than the statistical learning model. Based on the RMSE and MAE indicators, the ST-IDW model achieved worse imputation performance, while the models with better imputation performance are ST-2SMR, ST-ISE, TRMF, BTMF, BGCP, LRTC-TNN, and ST-PTD. That is, the imputation performance of the data-driven models is significantly better than that of the statistical learning models. In addition, in the data-driven model, ST-PTD obtains the best imputation performance.

Table V shows the comparison results between the ST-PTD model and the baselines under block missing in traffic volume data set and traffic speed data set. The results indicate that the imputation performance of all models show clear differences under block missing. Models ST-IDW, ST-KNN, BTMF, and TRMF achieved the worse imputation performance, while models ST-2SMR, ST-ISE, BGCP, LRTC-TNN, and ST-PTD achieved better imputation performance. That is, tensor factorization and deep learning models are more suitable for missing data imputation. In addition, in the data-driven model, compared to models ST-2SMR, ST-ISE, BGCP, and LRTC-TNN, ST-PTD also obtains the best imputation performance.

In general, the ST-PTD model shows good imputation performance under random missing, and block missing, thereby proving the superiority of the ST-PTD model.

E. Qualitative Analysis

In this section, the qualitative performance of ST-PTD is described by the scatter plots. Fig. 11 shows the imputation results of the ST-PTD with a single road segment with a missing rate of 40%. In the traffic volume data, the true value is closer to the estimated value under random missing and block missing. In addition, the residual under block missing is slightly larger than that under random missing. Compared to that of the traffic volume data set, the residual of the traffic speed data set is smaller. The results indicate that the imputation result of the ST-PTD has high imputation accuracy.

F. Effect of Different Components on Imputation Accuracy

The ST-PTD framework has three independent components: 1) modeling of spatiotemporal periodic matrix (MSPM); 2) decomposition of spatiotemporal trend matrix (BiARMF); and 3) fusion of multiple results (ST-PTD). Therefore, in this section, we further compare the impact of different components on imputation accuracy based on traffic volume data. To distinguish the performance of different components with different missing types and missing rates, we added specific suffixes to different components. For example, BiARMF/R/20% represents the filling result of the BiARMF components under random missing with a missing rate of 20%, and BiARMF/B/20% represents the filling result of the BiARMF components under block missing with a missing rate of 20%. The effects of different components on imputation accuracy are presented in Fig. 12. Overall, the imputation performance of the ST-PTD framework is better than that

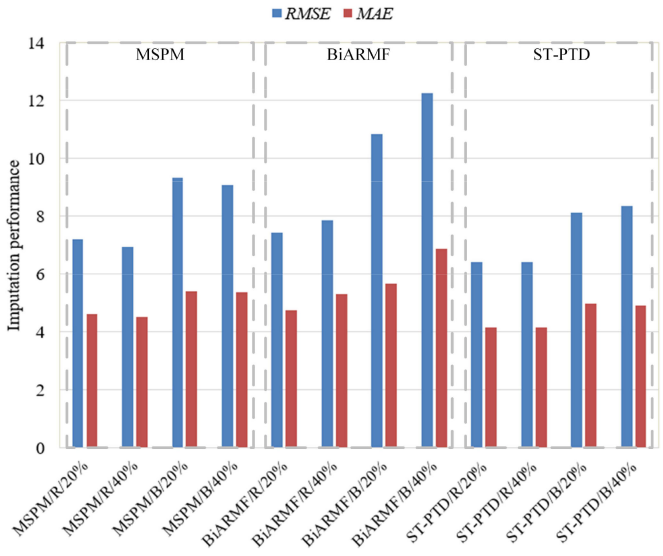


Fig. 12. Effect of different components on imputation accuracy.

of MSPM and BiARMF. BiARMF is greatly affected by the missing rate and missing type. Compared to that of the BiARMF component, the imputation results of MSPM are less affected by the missing rate and missing type, which is also why the ST-PTD has better performance.

V. Conclusion and Future Work

As an accurate estimation of missing data in traffic flow is crucial in urban planning, intelligent transportation, economic geography, and other fields, improving the data quality of traffic flow is often a necessary step in data modeling. In this study, we proposed a hybrid data-driven framework called ST-PTD to impute the missing information in traffic flow. First, we used a specific time-series analysis method to mine the periodic patterns of traffic flow data. Then, we proposed a novel matrix decomposition method called BiARMF to describe the trend information of traffic flow data. Finally, we applied a novel DD network that fused the periodic and trend characteristics of the missing data and obtained the final imputation result.

In the experimental study, we used two actual traffic flow data sets to verify the imputation performance of the ST-PTD framework. First, we applied the control variable method to obtain the optimal parameter combination of the ST-PTD framework. Second, we compared the ST-PTD to existing eight baseline methods, including ST-IDW, ST-KNN, ST-2SMR, ST-ISE, TRMF, BTMF, BGCP, and LRTC-TNN. Third, we used scatter plots to visually show the imputation results of ST-PTD. Finally, we tested the influence of different components on imputation accuracy, proving that the proposed method is suitable for traffic flow imputation.

For future work, the following limitations need further investigation: 1) verification of the proposed framework with a variety of data sources; 2) comprehensive comparison of this model with other imputation models; and 3) integration of more perspectives, such as seasonal trend and weekly periodicity, into the ST-PTD model to achieve a more robust model that further improves the accuracy of missing data imputation.

Acknowledgment

The authors are very grateful to the anonymous reviewers for their suggestions and the editor's careful revisions.

References

[1] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, "GeoMAN: Multi-level attention networks for geo-sensory time series prediction," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3428–3434.

[2] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Oct. 2014, doi: 10.1145/2629592.

[3] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1655–1661.

[4] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 468–478, Mar. 2020, doi: 10.1109/TKDE.2019.2891537.

[5] C. Staubach, V. Schmid, L. Knorr-Held, and M. Ziller, "A Bayesian model for spatial wildlife disease prevalence data," *Prev. Vet. Med.*, vol. 56, no. 1, pp. 75–87, Nov. 2002, doi: 10.1016/S0167-5877(02)00125-3.

[6] L. Jiang *et al.*, "A neural network method for the reconstruction of winter wheat yield series based on spatio-temporal heterogeneity," *Comput. Electron. Agr.*, vol. 154, pp. 46–53, Nov. 2018, doi: 10.1016/j.compag.2018.08.047.

[7] L. D. Cesare, D. E. Myers, and D. Posa, "Estimating and modeling space–time correlation structures," *Stat. Probab. Lett.*, vol. 51, no. 1, pp. 9–14, Jan. 2001, doi: 10.1016/S0167-7152(00)00131-0.

[8] P. Wang, T. Zhang, Y. Zheng, and T. Hu, "A multi-view bidirectional spatiotemporal graph network for urban traffic flow imputation," *Int. J. Geograph. Inf. Sci.*, Feb. 2022, doi: 10.1080/13658816.2022.2032081.

[9] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, Helsinki, Finland, 2008, pp. 880–887, doi: 10.1145/1390156.1390267.

[10] X. Chen, Z. He, and J. Wang, "Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 59–77, Jan. 2018, doi: 10.1016/j.trc.2017.10.023.

[11] M. Li, S. Gao, F. Lu, and H. Zhang, "Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data," *Comput. Environ. Urban Syst.*, vol. 77, Sep. 2019, Art. no. 101346, doi: 10.1016/j.compenvurbsys.2019.101346.

[12] 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, doi: 10.1109/TITS.2015.2507259.

[13] J. Y. Campbell and S. B. Thompson, "Predicting excess stock returns out of sample: Can anything beat the historical average?" *Rev. Financ. Stud.*, vol. 21, no. 4, pp. 1509–1531, Jul. 2008, doi: 10.1093/rfs/hhm055.

[14] P. M. Bartier and C. P. Keller, "Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW)," *Comput. Geosci.*, vol. 22, no. 7, pp. 795–799, Aug. 1996, doi: 10.1016/0098-3004(96)00021-0.

[15] L. Pesquer, A. Cortés, and X. Pons, "Parallel ordinary kriging interpolation incorporating automatic variogram fitting," *Comput. Geosci.*, vol. 37, no. 4, pp. 464–473, Apr. 2011, doi: 10.1016/j.cageo.2010.10.010.

[16] E. S. Gardner Jr., "Exponential smoothing: The state of the art—Part II," *Int. J. Forecast.*, vol. 22, no. 4, pp. 637–666, Oct./Dec. 2006, doi: 10.1016/j.ijforecast.2006.03.005.

[17] C. Yozgatligil, S. Aslan, C. Iyigun, and I. Batmaz, "Comparison of missing value imputation methods in time series: The case of Turkish meteorological data," *Theor. Appl. Climatol.*, vol. 112, no. 1, pp. 143–167, Apr. 2013, doi: 10.1007/s00704-012-0723-x.

[18] A. W. Aryaputera, D. Yang, L. Zhao, and W. M. Walsh, "Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging," *Solar Energy*, vol. 122, pp. 1266–1278, Dec. 2015, doi: 10.1016/j.solener.2015.10.023.

[19] L. Li, T. Losser, C. Yorke, and R. Piltner, "Fast inverse distance weighting-based spatiotemporal interpolation: A Web-based application of interpolating daily fine particulate matter $PM_{2.5}$ in the contiguous U.S. using parallel programming and k-d tree," *Int. J. Environ. Res. Public Health*, vol. 11, no. 9, pp. 9101–9141, Sep. 2014, doi: 10.3390/ijerph110909101.

[20] P. Duan, G. Mao, C. Zhang, and S. Wang, "STARIMA-based traffic prediction with time-varying lags," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 1610–1615, doi: 10.1109/ITSC.2016.7795773.

[21] C.-D. Xu, J.-F. Wang, M.-G. Hu, and Q.-X. Li, "Interpolation of missing temperature data at meteorological stations using P-BSHADE," *J. Climate*, vol. 26, no. 19, pp. 7452–7463, Oct. 2013, doi: 10.1175/JCLI-D-12-00633.1.

[22] S. Cheng, F. Lu, and P. Peng, "Short-term traffic forecasting by mining the non-stationarity of spatiotemporal patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 10, pp. 6365–6383, Oct. 2021, doi: 10.1109/TITS.2020.2991781.

[23] 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, 2014. [Online]. Available: https://doi.org/10.1049/iet-its.2013.0052

[24] G. Chang and T. Ge, "Comparison of missing data imputation methods for traffic flow," in *Proc. Int. Conf. Transp. Mech. Electr. Eng. (TMEE)*, Changchun, China, Dec. 2011, pp. 639–642, doi: 10.1109/TMEE.2011.6199284.

[25] H.-F. Yu, N. Rao, and I. S. Dhillon, "Temporal regularized matrix factorization for high-dimensional time series prediction," in *Proc. 30th Conf. Neural Inf. Process. Syst. (NIPS)*, Barcelona, Spain, 2016, p. 15.

[26] X. Chen, Z. He, and L. Sun, "A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 98, pp. 73–84, Jan. 2019, doi: 10.1016/j.trc.2018.11.003.

[27] X. Chen, J. Yang, and L. Sun, "A nonconvex low-rank tensor completion model for spatiotemporal traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 117, Aug. 2020, Art. no. 102673, doi: 10.1016/j.trc.2020.102673.

[28] X. Chen and L. Sun, "Bayesian temporal factorization for multidimensional time series prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 17, 2021, doi: 10.1109/TPAMI.2021.3066551.

[29] S. Cheng and F. Lu, "A two-step method for missing spatio-temporal data reconstruction," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 7, p. 187, Jul. 2017, doi: 10.3390/ijgi6070187.

[30] M. Deng, Z. Fan, Q. Liu, and J. Gong, "A hybrid method for interpolating missing data in heterogeneous spatio-temporal datasets," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 2, p. 13, Feb. 2016, doi: 10.3390/ijgi5020013.

[31] S. Cheng, P. Peng, and F. Lu, "A lightweight ensemble spatiotemporal interpolation model for geospatial data," *Int. J. Geograph. Inf. Sci.*, vol. 34, no. 9, pp. 1849–1872, Sep. 2020, doi: 10.1080/13658816.2020.1725016.

[32] Y. Gong, Z. Li, J. Zhang, W. Liu, Y. Yin, and Y. Zheng, "Missing value imputation for multi-view urban statistical data via spatial correlation learning," *IEEE Trans. Knowl. Data Eng.*, early access, Apr. 13, 2021, doi: 10.1109/TKDE.2021.3072642.

[33] S. Cheng, F. Lu, P. Peng, and S. Wu, "Multi-task and multi-view learning based on particle swarm optimization for short-term traffic forecasting," *Knowl. Based Syst.*, vol. 180, pp. 116–132, Sep. 2019, doi: 10.1016/j.knosys.2019.05.023.

[34] J. Sun, J. Zhang, Q. Li, X. Yi, Y. Liang, and Y. Zheng, "Predicting citywide crowd flows in irregular regions using multi-view graph convolutional networks," *IEEE Trans. Knowl. Data Eng.*, early access, Jul. 13, 2020, doi: 10.1109/TKDE.2020.3008774.

[35] S. Cheng, F. Lu, P. Peng, and S. Wu, "A spatiotemporal multi-view-based learning method for short-term traffic forecasting," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 6, p. 218, Jun. 2018, doi: 10.3390/ijgi7060218.

[36] G. Liu and J. Wang, "Dendrite net: A white-box module for classification, regression, and system identification," *IEEE Trans. Cybern.*, early access, Oct. 18, 2021, doi: 10.1109/TCYB.2021.3124328.

[37] S. R. Chandra and H. Al-Deek, "Predictions of freeway traffic speeds and volumes using vector autoregressive models," *J. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 53–72, May 2009, doi: 10.1080/15472450902858368.

[38] B. Yu, X. Song, F. Guan, Z. Yang, and B. Yao, "k-nearest neighbor model for multiple-time-step prediction of short-term traffic condition," *J. Transp. Eng.*, vol. 142, no. 6, Jun. 2016, Art. no. 4016018, doi: 10.1061/(ASCE)TE.1943-5436.0000816.

Peixiao Wang received the M.S. degree from the Academy of Digital China, Fuzhou University, Fuzhou, China, in 2020. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

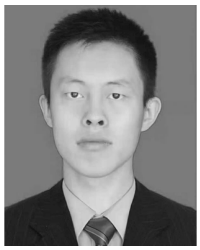
In the past years, he has published over ten refereed journal articles and conference papers as the first or corresponding author. His research focus on spatiotemporal data mining, social computing, and public health.

Tao Hu received the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2015.

He is an Assistant Professor with the Department of Geography, Oklahoma State University (OSU), Stillwater, OK, USA. Before joining OSU, he worked as a Postdoctoal Research Fellow with the Center for Geographic Analysis, Harvard University, Cambridge, MA, USA, and the Department of Geography, Kent State University, Kent, OH, USA. His research interests include geospatial big data analysis (i.e., social media), health geography, human mobility, and crime geography.

Fei Gao received the M.S. degree from Nanjing Normal University, Nanjing, China, in 2018. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

During the past years, he has published over five refereed journal articles. His primary research field is spatial optimization, simulation modeling, and spatiotemporal analysis.



Ruijie Wu received the B.S. degree from Chang'an University, Xi'an, China, in 2017, and the M.S. degree in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2020, where he is currently pursuing the doctorate degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing.

He focuses on the analysis and application of geographic information big data, and geographic information named entity recognition.

Wei Guo received the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010.

He is currently an Associate Professor with Wuhan University. During the past years, he has published over 40 refereed journal articles and conference papers. His research interests cover spatiotemporal data modeling and analysis, distributed spatial database, and public health.

Xinyan Zhu received the Ph.D. degree from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

He is currently a Professor with Wuhan University. During the past years, he has published over 100 refereed journal articles and conference papers. His research interests cover spatiotemporal data modeling and analysis, distributed spatial database, holographic position map and its application, indoor location and navigation, social geographic computing, and public health.