2021

Online Bus Speed Prediction With Spatiotemporal Interaction: A Laplace Approximation-Based Bayesian Approach

Haipeng Cui

Kun Xie
Old Dominion University, kxie@odu.edu

Bin Hu

Hangfei Lin

Original Publication Citation

Cui, H. P., Xie, K., Hu, B., & Lin, H. F. (2021). Online bus speed prediction with spatiotemporal interaction: A Laplace approximation-based Bayesian approach. *IEEE Access*, 9, 105205-105219. https://doi.org/10.1109/access.2021.3100261

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 14, 2021, accepted July 14, 2021, date of publication July 26, 2021, date of current version August 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100261

Online Bus Speed Prediction With Spatiotemporal Interaction: A Laplace Approximation-Based Bayesian Approach

HAIPENG CUI[1], KUN XIE[2], BIN HU[3], AND HANGFEI LIN[4]
[1]Department of Civil and Environmental Engineering, National University of Singapore, Singapore 117576
[2]Department of Civil and Environmental Engineering, Old Dominion University, Norfolk, VA 23529, USA
[3]Shenzhen Transport Information & Technology Company Ltd., Shenzhen 518040, China
[4]The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China

Corresponding author: Kun Xie (kxie@odu.edu)

This work was supported by the Chinese National Science Foundation under Grant 51238008.

ABSTRACT This study proposes a novel Bayesian hierarchical approach for online bus speed prediction by explicitly accounting for the spatiotemporal interaction (STI) of speed observations. The use of Laplace approximation can expedite the estimation of Bayesian models and enable the implementation of online prediction. Large numbers of trials are carried out to identify significant predictors and the optimal length of the look-back time window to achieve the highest prediction accuracy. The spatiotemporal interacting patterns are also explored, and results show that the Type IV model assuming the structured spatial effect interacts with the structured temporal effect can best accommodate the bus speed data. Besides, prediction errors of the Type IV model randomly distribute over time and space. The proposed model can achieve high prediction accuracy and computational efficiency without compromising the interpretability of the contributing factors and the unobserved spatiotemporal heterogeneity. The proposed model can be used to assist public transit operation and management, such as bus scheduling, congestion warning, and the development of proactive measures to mitigate bus delays.

INDEX TERMS Bus speed prediction, Bayesian hierarchical approach, spatiotemporal interaction, Laplace approximation, public transit operation.

I. INTRODUCTION

Travel speed is one of the effective and intuitive indicators for the operation and management of transportation systems. Large map service providers like Google Maps and Baidu Maps use the travel speed to represent congestion information. With travel speed, we can not only measure congestion but also derive the travel time and delay. The same holds true for the bus transit systems [1], [2]. Predicting bus speeds in a timely and accurate manner contributes to maximizing the benefits of an intelligent public transit system [3]. Prediction approaches can be divided into offline prediction and online prediction according to the utilization of streaming data. The offline prediction approaches train the model using static data and conduct prediction without real-time adjustment. An extensive range of reported works on traffic state prediction falls into this category [4]–[6]. On the other hand, the online prediction approaches update the model recurrently when a new data stream is acquired throughout the prediction horizon, making the prediction models adaptive to the latest information. Previous studies have recognized the adaptability advantage over the offline prediction approaches [7], [8]. To successfully implement the online prediction, the training process must be computationally efficient. However, there is limited research on the development of online approaches for bus speed prediction.

In light of the propagation of congestion and the disturbance caused by traffic incidents, bus speeds can be both spatially and temporally correlated. The investigation of spatial and temporal effects in bus travel speed/time is gaining increasing attention recently [6], [9]–[11]. However, most existing studies address the spatial and temporal effects separately without modeling the spatiotemporal interaction (STI) effects [12]. STI represents the pattern of the spatial correlations over time, or equivalently, the pattern of the temporal correlations over space. For example, spatial correlations of

The associate editor coordinating the review of this manuscript and approving it for publication was Michail Makridis.

bus speeds weaken in off-peak hours while strengthened in peak hours. Addressing STI properly can better accommodate the data and thus increase prediction performance.

A large number of related works in the literature utilize machine learning models to predict the traffic state, such as travel time, travel speed, and traffic volume [13]–[16]. Despite the superior predictive performance, machine learning models are criticized for the lack of interpretability [17]. Though spatial and temporal correlations can be captured to a certain extent by some machine learning models (e.g., deep neural network [6], [18], attention network [5], [19], convolutional neural network (CNN) [20], [21]), it is challenging to (i) explicitly interpret these correlations, and (ii) quantify their impacts on outcomes and to make relevant inferences. Few studies have developed models considering both predictive performance and interpretability in the traffic state prediction domain when STI is concerned [12], [17]. The Bayesian methods are widely used in the statistical domain for their great interpretability of the results [17]. The Markov Chain Monte Carlo (MCMC) simulation is commonly adopted to estimate the Bayesian models and takes a long time to converge [11], [22], especially for complex model specifications. Therefore, MCMC-based Bayesian models fail to meet the computation efficiency criterion of online prediction.

Motivated by the above challenges, this study aims to develop a novel online bus speed prediction model based on a Bayesian hierarchical framework that can fully capture the STI of bus speed data. Various STI patterns are explicitly characterized and compared. Furthermore, the Integrated Nested Laplace Approximation (INLA) technique is used to replace the MCMC method to greatly speed up the estimation of Bayesian models.

A. LITERATURE REVIEW AND RELATED WORK

1) BUS SPEED/TRAVEL TIME PREDICTION

Bus speed/arrival time prediction has been studied widely [9], [12], [17]. Some researchers focus on the travel time/speed of all links between two bus stops [13], [23], assuming the homogeneity between each link. This assumption can be easily violated due to different traffic conditions, road designs, etc. Furthermore, considering the segmentation of traffic flow by traffic signals, some researchers divide road segments based on signalized intersections [6], [9]. This segmentation method includes regular bus traveling, dwell at bus stops, and unexpected stoppages into the travel speed/time. We follow this segmentation method in this paper.

In the early stage, researchers utilized traffic flow theory for bus traffic state prediction [24]–[26]. For instance, Bie *et al.* [27] proposed an analytical model to predict bus arrival time considering the impact of the signalized intersection. The proposed model utilized the shock wave theory to characterize the impact of signals on travel time. The proposed model was tested on two signalized intersections, and it was found that the proposed model outperformed existing models in the literature Zhang *et al.* [28] proposed a prediction model for bus arrival time considering signalized control and surrounding traffic flow. The authors adopted the speed-flow-density relationship to derive the travel speed. The traffic flow theory-based models do not explicitly account for spatial, temporal, and STI effects. Later, researchers mainly utilize time series approaches or simple machine learning approaches for bus travel speed/time prediction [10], [23], [29]–[31]. Kumar and Vanajakshi [32] developed a pattern-specific model using time series data to predict bus arrival time and tested the model with data collected from a single bus route. Farooq *et al.* [33] developed a time series model for bus arrival time prediction based on GPS data and found that the prediction error decreases as buses operate further. However, these methods consider either the spatial effects or the temporal effects, which limits the ability to fit the data.

To improve prediction performance, researchers have focused on addressing spatial and temporal correlations in predicting traffic state recently. Some researchers have combined the spatiotemporal effects with conventional prediction models mentioned above [34], [35]. In these papers, spatiotemporal effects are often included as another feature along with other explanatory variables. However, this approach would lead to limited interpretability for the spatiotemporal terms [36], [37]. Lately, researchers have utilized state-of-the-art machine learning techniques to capture the spatial and temporal characteristics in bus travel speed/time prediction. For example, the recurrent neural network (RNN) models, the long short-term memory (LSTM) model [4], [6], [18], and the convolutional neural network (CNN) model [21], [38] Treethidtaphat *et al.* [39] focused on bus arrival time prediction using a deep neural network (DNN) model and compared the results with the ordinary least square (OLS) method Achar *et al.* [9] developed a spatial Kalman filter model for bus arrival time prediction. The spatial and temporal effects of the data were modeled in a linear state-space form. Liu *et al.* [6] addressed the spatial and temporal effects by combining LSTM and artificial neural network (ANN). Despite the variety of the abovementioned studies, predicting bus speed/travel time using Bayesian approaches is rarely reported in the literature.

2) BAYESIAN SPATIOTEMPORAL APPROACHES

Spatial analysis under the Bayesian framework originates from the work by Besag [40] Bernardinelli *et al.* [41] first incorporated the temporal effects into spatial analysis. Since then, this approach has been applied to many research domains, such as medicines [42], social science [43], and especially in the safety analysis domain [44]–[47]. The incorporated spatial and temporal effect terms in the Bayesian spatiotemporal models "borrow strength" from neighboring locations and contiguous time periods to better accommodate the data [48], [49]. Compared to the trained spatiotemporal features in the machine learning approaches, the Bayesian statistical approaches usually specify a structure for the

spatial and/or temporal effect terms, which would greatly increase the interpretability of the results [50]. The specified structures indicate how the data are spatially and/or temporally correlated and how to specify the corresponding probability distributions. As such, the Bayesian spatiotemporal models have superior performance in terms of both results interpretation and model estimation. Researchers in the transportation domain have applied this framework to various fields [11], [51], [52].

The aforementioned studies focus on modeling spatial and temporal correlations separately. However, these papers neglect the potential existence of STI among data. Namely, the spatial and temporal effects are considered independent among all locations and time periods in current literature. Besides, current Bayesian models are solved mainly by the MCMC method. However, this method is criticized for its high computation cost [17], which is unsuitable for online prediction. Therefore, how to address the STI effects when tackling the bus speed prediction problem with a fast solving process is still a challenge.

B. GAPS AND CONTRIBUTIONS

The gaps and contributions are summarized as follows:

(i) Most of the existing studies only focus on spatial and/or temporal effects of data, but the STI effects were rarely addressed in studies regarding bus traffic state prediction. This study adds to the literature by fully exploring the STI effects of bus data.

(ii) Most of learning-based models are expected to deliver satisfactory prediction accuracy but are often criticized due to lack of interpretability. By contrast, without compromising prediction accuracy, the proposed method allows us to assess the significance of explanatory variables, spatial, temporal, and STI effects and can infer to what extent each component of the model contributes to the variation of the response variable.

(iii) Bayesian statistical models are criticized for the large computation burden due to the Markov Chain Monte Carlo (MCMC) method used for estimation, which limits the potentials for online prediction. We propose an integrated nested Laplace approximation (INLA) technique to greatly expedite the estimation of the Bayesian models for online prediction.

In addition, this paper presents three major advancements comparing to our previous work in Hu *et al.* [53]: (i) We analyze all the four types of Bayesian spatiotemporal interaction (BSTI) patterns thoroughly by proposing new BSTI models, which are not studied in Hu *et al.* [53]; (ii) This study utilizes a novel computation method for Bayesian model estimation, which is the INLA method; (iii) This paper focuses on the potentials of the proposed BSTI model in online prediction by proposing an online prediction framework and conducting a large number of experiments regarding the major hyperparameters. All these aspects are not mentioned in Hu *et al.* [53].

The rest of the paper is organized as follows. In Section II, we present the Bayesian spatiotemporal modeling framework. Data and the studied region are described in Section III. In Section IV, results and discussions are presented. Conclusions are drawn in Section V.

II. METHODOLOGY

To investigate the characteristics of the STI and assess their effects on bus speed prediction, six models under a three-level Bayesian hierarchical framework are developed. These models account for different spatiotemporal effects, from no spatiotemporal effects to STI effects. A detailed introduction regarding each model is given in the following subsections, including the model specifications and the mathematical definition of the STI effects (Section A and B). Secondly, the framework of the Bayesian approach is presented, and the Markov Chain Monte Carlo (MCMC) method is then introduced to show how the Bayesian model is solved traditionally (Section C). Thirdly, the proposed INLA method is introduced to highlight the advantage of fast computation against the MCMC method while maintaining interpretability. Lastly, the prediction workflow is presented to show how the training and the testing process for bus speed prediction are accomplished (Section D).

A. BASIC MODELS

In the first level, for each road segment, bus speeds can be modeled using a Gaussian distribution as follows [11], [54]:

$$Y_i^t \sim Normal(\mu_i^t, \sigma_{it}^2) \tag{1}$$

where i denotes the road segment, $i = 1, 2, \ldots, n$, n is the total number of road segments, t denotes the time period, $t = 1, 2, \ldots, T$, T is the total number of time periods, μ_i^t is the mean bus speed for road segment i at time period t, and σ_{it} is the standard deviation of the bus speed for road segment i at time period t.

In the second level, the linear predictor μ_i^t is further formulated as a linear combination of observed covariates and spatial/temporal effects.

1) MODEL 1: PURE LINEAR MODEL (PL MODEL)

By breaking μ_i^t down into intercept and covariates (i.e., fixed effects), the benchmark model is formulated as follows:

$$\mu_i^t = \beta_0 + \sum_{p=1}^{P} \beta_p X_{pi}^t \tag{2}$$

where β_0 is the average bus speed when all covariates equal to 0, X_{pi}^t denotes the p^{th} observed covariate of road segment i in time period t, and P denotes the number of observed covariates. β_0 and β_p are the regression coefficients to be estimated. This model does not consider spatial or temporal correlations among bus speed data.

2) MODEL 2: SPATIOTEMPORAL MODEL WITHOUT INTERACTION (ST MODEL)

Spatial effects have been widely recognized as adjacent locations tend to present similar values at a given time period [48], [55]. Analogously, temporal effects are recognized as for a given location, previous time periods tend to present similar values [6], [56]. Following a classical framework by Knorr-Held [48], spatial/temporal effects are divided into two parts: the structured part and the unstructured part. The structured effect indicates that a predefined variance-covariance structure of spatial/temporal effects is used to restrict the estimation of hyperparameters, while the unstructured effect means that the variance and covariance of spatial/temporal effects are all freely estimated from the data. Random effect term u is included to capture the structured spatial effects. To further capture the unstructured spatial effects, a random effect term λ is included. Similarly, a random effect term γ is included to characterize the structured temporal effects, while a random effect term ζ is added to characterize the unstructured temporal effects. Intuitively, the spatial effects mentioned at the beginning of this paragraph represent the spatial structure. This spatial structure is captured by the structured spatial effect term u. On the contrary, the unstructured spatial effects can be considered as the noise of the structured spatial effect and are captured by λ. Similar understandings can be applied to the structured/unstructured temporal effects. Therefore, the ST model without interaction can be expressed as:

$$\mu_i^t = \beta_0 + \sum_{p=1}^{P} \beta_p X_{pi}^t + (u_i + \lambda_i) + (\gamma_t + \zeta_t) \tag{3}$$

where u_i denotes the structured spatial effect for road segment i and λ_i represents the unstructured spatial effect for road segment i. Similarly, γ_t and ζ_t are respectively the structured and unstructured temporal effect at time period t. These components account for spatial effects and temporal effects separately.

In the third level, prior distributions are assigned to each component in the linear predictor μ_i^t for each model separately. Prior distributions proposed by Besag *et al.* [55] are widely utilized in the literature due to better interpretability and simplicity [48], [57]. Following Besag *et al.* [55], the intrinsic conditional autoregressive (ICAR) model is selected for u_i. The full conditional distribution of u_i given u_{-i} follows a normal distribution and can be expressed as follows:

$$u_i | u_{-i} \sim Normal\left(\frac{1}{N_i}\sum_{j\in i\sim j, j\neq i} u_j, \frac{\sigma_u^2}{N_i}\right) \tag{4}$$

where u_{-i} is the set of u_j that $j \neq i, i \sim j$ denotes road segment j is a neighbor of road segment i. N_i is the number of the neighboring road segments of i and σ_u^2 is the global variance of u. In practice, road segments with numerous neighbors have more accurate estimates of u_i than those isolated. Therefore, for an arbitrary road segment i, the variance of u_i is set to the global variance σ_u^2 divided by the number of neighbors. The unstructured spatial effects λ_i are assumed independent and identically distributed (i.i.d.) among different locations.

For the structured temporal effects γ_t, we follow Clayton [58] by using a first-order random walk (RW) model to characterize the temporal dependence for better interpretability [48]. This specification assumes that the state for a time point depends on the previous one as follows:

$$\gamma_t | \gamma_{t-1} \sim Normal\left(\gamma_{t-1}, \sigma_\gamma^2\right) \tag{5}$$

where σ_γ^2 is the global variance of γ. The unstructured temporal effects ζ_t are assumed i.i.d. among different time periods.

B. BAYESIAN SPATIOTEMPORAL INTERACTION MODELS

Most of the works regarding spatiotemporal modeling in (transit) traffic state prediction propose their models with the idea similar to the ST model in Eq. (2) [9], [34], [35]. Namely, the spatial and temporal effects are included separately without considering their interactions. In this section, we propose four BSTI models. The BSTI model can further consider the interaction effects between the spatial and temporal effects.

Based on the independent spatial and temporal effects introduced in Eq. (3), STI can be formulated as a combination of one spatial effect and one temporal effect [48]. Consequently, given that two types of effects exist for either spatial or temporal effects (structured and unstructured), four types of interactions can be expected. STI characterizes the relations between the temporal effects in different spaces, or equivalently, the relations between the spatial effects at different time periods. Based on the ST model (Model 2), an STI random effect term δ_{it} is further introduced into the ST model as follows:

$$\mu_i^t = \beta_0 + \sum_{p=1}^{P} \beta_p X_{pi}^t + (u_i + \lambda_i) + (\gamma_t + \zeta_t) + \delta_{it} \tag{6}$$

where δ_{it} accounts for the interaction effect of road segment i in time period t. For computational convenience, u_i, λ_i, γ_t, ζ_t and δ_{it} are assumed to follow Gaussian distribution with separate precision matrix $\eta \boldsymbol{K}$, where η is an unknown scalar and $\boldsymbol{K}$ is a structure matrix.

Non-diagonal elements in $\boldsymbol{K}$ are equal to -1 for neighboring road segments/time periods, while diagonal elements are equal to the number of adjacent road segments/time periods. The other elements in $\boldsymbol{K}$ are filled with zero [48]. Note that for the structure matrix of the structured spatial effects u_i, the number of neighbors of a road segment is determined by the road network, while for the structure matrix of the structured temporal effects γ_t, given a first-order RW prior, each time period has at most two neighbors: (i) for the first or last time period, only one neighbor exists, while (ii) for the other time periods, two neighbors exist for each time period. Given the i.i.d. assumptions, the structure matrix of the unstructured spatial effects λ_i is an identity matrix $\boldsymbol{I}$, and the same applies to the unstructured temporal effects

ζ_t. For the STI effects δ_{it}, Clayton [58] suggested that the corresponding structure matrix can be obtained by producing the Kronecker product of the $\boldsymbol{K}$ of the interacting spatial effects and temporal effects. Further details can be found in Clayton [58] and Stone *et al.* [59].

1) MODEL 3: ST MODEL WITH TYPE I INTERACTION (TYPE I MODEL)

Type I model assumes the interactions exist between the two unstructured effects λ_i and ζ_t. Thus, their Kronecker product with Type I interaction can be given as follows:

$$\boldsymbol{K}_{\boldsymbol{\delta}} = \boldsymbol{K}_{\boldsymbol{\lambda}} \otimes \boldsymbol{K}_{\boldsymbol{\zeta}} = \boldsymbol{I} \otimes \boldsymbol{I} = \boldsymbol{I} \tag{7}$$

where $\boldsymbol{K}_{\boldsymbol{\delta}}$ denotes the structure matrix of the STI term δ, "$\otimes$" denotes the Kronecker product, $\boldsymbol{K}_{\boldsymbol{\lambda}}$ represents the structure matrix of the unstructured spatial term λ, and $\boldsymbol{K}_{\boldsymbol{\zeta}}$ denotes the structure matrix of the unstructured temporal effect term ζ. The prior distribution (abbreviate as "prior") of δ with Type I interaction is under the Gaussian family and can be given below:

$$p(\delta|\eta_\delta) \propto \exp\left(-\frac{\eta_\delta}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}(\delta_{it})^2\right) \tag{8}$$

Type I interactions account for the unobserved noise of each data in the spatial and temporal dimensions. No spatial or temporal structures exist in the Type I interactions. This distinguishes the Type I Model from the ST Model.

2) MODEL 4: ST MODEL WITH TYPE II INTERACTION (TYPE II MODEL)

Type II interaction assumes the interactions exist between the structured temporal effect γ_t and the unstructured spatial effect λ_i. Since γ_t follows an RW distribution, δ_i also follows this distribution which is independent of all other road segments. Their Kronecker product is:

$$\boldsymbol{K}_{\boldsymbol{\delta}} = \boldsymbol{K}_{\boldsymbol{\lambda}} \otimes \boldsymbol{K}_{\boldsymbol{\gamma}} = \boldsymbol{I} \otimes \boldsymbol{K}_{\boldsymbol{\gamma}} = \boldsymbol{K}_{\boldsymbol{\gamma}} \tag{9}$$

where $\boldsymbol{K}_{\boldsymbol{\gamma}}$ denotes the structure matrix of the structured temporal term γ. Therefore, the prior of δ with Type II interaction can be written as:

$$p(\delta|\eta_\delta) \propto \exp\left(-\frac{\eta_\delta}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}(\delta_{it}-\delta_{i,t-1})^2\right) \tag{10}$$

A Type II interaction denotes that temporal dependency exists in each road segment, and these dependencies are not spatially correlated.

3) MODEL 5: ST MODEL WITH TYPE III INTERACTION (TYPE III MODEL)

Type III interaction assumes the interactions exist between the structured spatial effect u_i and the unstructured temporal effect ζ_t. Following u_i, δ_i yields an intrinsic autoregression. Note that δ_i is independent of all other road segments. Their Kronecker product is given below:

$$\boldsymbol{K}_{\boldsymbol{\delta}} = \boldsymbol{K}_{\boldsymbol{u}} \otimes \boldsymbol{K}_{\boldsymbol{\zeta}} = \boldsymbol{K}_{\boldsymbol{u}} \otimes \boldsymbol{I} = \boldsymbol{K}_{\boldsymbol{u}} \tag{11}$$

where $\boldsymbol{K}_{\boldsymbol{u}}$ denotes the structure matrix of the structured spatial term u. Therefore, the prior of δ with Type III interaction can be written as:

$$p(\delta|\eta_\delta) \propto \exp\left(-\frac{\eta_\delta}{2}\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j\in i\sim j}(\delta_{it}-\delta_{jt})^2\right) \tag{12}$$

A Type III interaction denotes that spatial dependency exists among each road segment, and these spatial dependencies are not temporally correlated.

4) MODEL 6: ST MODEL WITH TYPE IV INTERACTION (TYPE IV MODEL)

Type IV interaction assumes that the interactions exist between the structured spatial effect u_i and the structured temporal effect γ_t. Since u_i yields an ICAR and γ_t yields an RW distribution, δ_{it} is utterly dependent on space and time. Therefore, the Kronecker product of δ is given below:

$$\boldsymbol{K}_{\boldsymbol{\delta}} = \boldsymbol{K}_{\boldsymbol{u}} \otimes \boldsymbol{K}_{\boldsymbol{\gamma}} \tag{13}$$

Therefore, the prior of δ with Type IV interaction can be written as:

$$p(\delta|\eta_\delta) \propto \exp\left(-\frac{\eta_\delta}{2}\sum_{t=2}^{T}\sum_{i=1}^{n}\sum_{j\in i\sim j}((\delta_{it}-\delta_{jt}) - (\delta_{i,t-1}-\delta_{j,t-1}))^2\right) \tag{14}$$

A Type IV interaction indicates that spatial dependency exists among each road segment for a time period, and these spatial dependencies from all time periods are temporally correlated. Equivalently, it can also indicate that temporal dependency exists among time periods for a certain road segment, and these temporal dependencies from all segments are spatially correlated. TABLE 1 summarizes these four types of STI.

With the cooperation of both predictors (covariates) and spatiotemporal effect terms, the model has been given the flexibility to interpret the contributing predictors that affect the bus speed and capture the STI effects simultaneously.

C. BAYESIAN MODELS USING INLA

1) ESTIMATIONS IN BAYESIAN FRAMEWORK

Two approaches can be considered when estimating statistical models, i.e., the frequentist method and the Bayesian method. Compared to the former, the latter aims to derive a probability distribution for each variable, where the estimates obtained in the former are the mean of the distribution [60]. With the distributions of the parameters derived, more comprehensive inferences on those parameters can be generated. Therefore, in this paper, we estimated the proposed models using the full Bayesian framework.

The Bayesian framework can be formulated as follows:

$$P(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}) \cdot L(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})} \tag{15}$$

TABLE 1. Summary of STI structures.

Model	STI prior	Interacting Effects	Interaction description
Model 3 (Type I)	$\delta_{it} \sim N(0, \sigma_{it}^2)$	λ and ζ	Interactions can be considered as unobserved noise in each space-time point and present no spatial or temporal correlations.
Model 4 (Type II)	$\delta_{it} \sim RW(\gamma_t, \sigma_t^2)$	λ and γ	Each road segment has a temporal correlation. Neighboring road segments have independent temporal correlations.
Model 5 (Type III)	$\delta_{it} \sim ICAR(u_i, \sigma_\delta^2)$	u and ζ	Road segments have spatial correlations at a particular time period. Contiguous time periods have independent spatial correlations.
Model 6 (Type IV)	$\delta_{it} \sim ICAR(u_i, \sigma_\delta^2) \otimes RW(\gamma_t, \sigma_t^2)$	u and γ	Road segments have spatial correlations at a particular time period. Contiguous time periods have correlated spatial correlations. Vice versa.

where $\mathbf{y}$ denotes the bus speeds, $\boldsymbol{\theta}$ denotes the model parameters, $P(\boldsymbol{\theta}|\mathbf{y})$ denotes the posterior probability distribution of $\boldsymbol{\theta}$ conditioned on $\mathbf{y}$, $p(\boldsymbol{\theta})$ denotes the prior distribution of $\boldsymbol{\theta}$, $L(\mathbf{y}|\boldsymbol{\theta})$ denotes the likelihood of $\mathbf{y}$ given $\boldsymbol{\theta}$, and $p(\mathbf{y})$ denotes the marginal probability distribution of the $\mathbf{y}$.

For better understanding, the likelihood of the STI models of road segment i at time period t is given as follow: (16), as shown at the bottom of the next page.

2) ALTERNATIVE TO MCMC METHOD: INLA METHOD

MCMC method [61]–[63] functions by performing a massive amount of simulation until the Markov chain is converged, and Bayesian inference can subsequently be produced, yet its computational cost is enormous [48]. As an alternative, Rue *et al.* [64] proposed the INLA method to estimate the Bayesian models. The INLA approach functions in a deterministic way and can significantly reduce the computation time [64].

To grasp the merit of the INLA method, here we demonstrate the deduction of the Laplace approximation. Given a random variable X and its density function $f(x)$, and suppose we are interested in the following integration:

$$\int f(x)dx = \int \exp(\log(f(x)))dx \tag{17}$$

By representing $\log(f(x))$ using a Taylor series expansion evaluated in $x = x^* = \text{argmax}_x \log(f(x))$, Eq. (17) becomes:

$$\int f(x)dx \approx \exp(\log(f(x^*))) \int \exp\left(\frac{(x-x^*)^2}{2} \times \frac{\partial \log(f(x))}{\partial x^2}\bigg|_{x=x^*} \right) dx, \tag{18}$$

By setting $\sigma^{2^*} = -1/\left.\frac{\partial \log(f(x))}{\partial x^2}\right|_{x=x^*}$, Eq. (18) becomes:

$$\int f(x)dx \approx \exp(\log(f(x^*))) \int \exp\left(-\frac{(x-x^*)^2}{2\sigma^{2^*}} \right) dx \tag{19}$$

where inside the integration is a Normal distribution $Normal(x^*, \sigma^{2^*})$ with mean equal to x^* and variance equal to σ^{2^*}. Finally, we have:

$$\int_n^m f(x)dx \approx f(x^*)\sqrt{2\pi\sigma^{2^*}}\,(F(m) - F(n)) \tag{20}$$

where $F(\cdot)$ denotes the cumulative density function of $Normal(x^*, \sigma^{2^*})$.

We can see that this method can approximate any density function using the normal distribution, which can reduce the computation time compared to the MCMC method [64].

D. WORKFLOW OF ONLINE PREDICTION

Offline prediction is usually performed using a static model trained with a massive amount of historical data. This leads to several disadvantages: (i) the computational cost for training a model is high. Thus (ii) it is not easy to update the model frequently when its performance drops. Besides, (iii) due to the extensive coverage in space and/or time dimension of the historical data, the ability to capture real-time data characteristics is limited. Compared to offline prediction, online prediction is usually performed by continuously updating the model with a smaller amount of real-time data. Online prediction ensures that the model is capable of utilizing the latest information at all times.

In this paper, the prediction procedure is divided into a training procedure and a testing procedure [34], [65]–[67]. Suppose there are n road segments and T time periods covered in the dataset. Time window (denoted as w), step ahead (denoted as a), starting period (denoted as p), the number of time periods extracted (denoted as c) and the sliding step (denoted as s) are the five critical hyperparameters defined in the prediction procedure. Concretely, time window (w) is defined as the number of consecutive time periods used to generate the training data. Step ahead (a) represents the prediction horizon and is defined as the number of time periods used for prediction testing. $a = 1$ indicates that the model predicts one time period into the future, i.e., the data from the first time period following the training set is used for testing. Starting period (p) is defined as the starting time period of w. For example, $w = 3$ and $p = 1$ indicate that the training dataset consists of all the data in the first three time periods. The number of time periods extracted (c) is defined as the total number of time periods needed to form training and test dataset, i.e., $c = w + a$. Sliding step (s) is a constant which equals 1 at all time, meaning that the time window

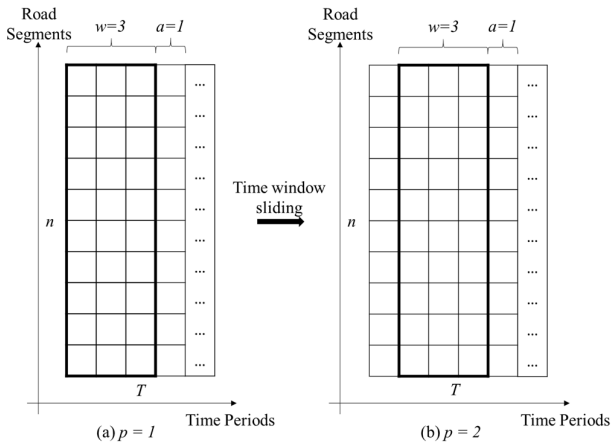


FIGURE 1. Time window (denoted as the bold box) sliding from $p = 1$ to $p = 2$ with $w = 3$ and $a = 1$. The data inside the time window are used as the training data, while the data in the next time period following the time window ($a = 1$) are used as test data.

will only slide one time period at a time to form new training dataset and test dataset.

Figure 1 illustrates how the time window slides to form the training dataset and test datasets. In Figure 1, the time window (denoted as the bold box) equals 3 ($w = 3$) and is sliding from $p = 1$ to $p = 2$. The step ahead equals 1 ($a = 1$). The data inside the bold box are the training dataset, and the data in the next time period following the bold box are the test dataset. To ensure a consistent spatial structure during the whole prediction process, data from all the road segments are chosen in each prediction procedure.

In order to achieve the best prediction performance, all the combinations of p, w and a for the six proposed models are tested. Each combination is considered a scenario. ALGORITHM 1 describes the procedures in detail.

Algorithm 1 Online Prediction Process

```
Input: Complete dataset
for w in min(w): max(w) do
    for a in min(a): max(a) do
        c = w + a;
        for p in 1: T − c + 1 do
            Extract column p to column p + w − 1 from the complete dataset as training dataset;
            Extract column p + w to column p + c − 1 from the complete dataset as test dataset;
            Perform training on each of the six proposed models;
            Compute prediction performance indexes for each model using test dataset and store the results;
Summarize all the output results for further analysis;
```

III. DATA AND STUDY AREA

Shennan Avenue in Shenzhen city is selected as the test road (Figure 2). This road is one of the major arterial roads that communicate between different districts. The test road reaches 17 km in mileage. Following Achar *et al.* [9] and Liu *et al.* [6], the test road is divided into 11 road segments separating at consecutive signalized intersections. The average length for road segments is 1,632 m. In practice, different travel directions of the same road segment may present distinct travel states, and thus should be treated separately. In this study, different directions of the same road segment are treated as two road segments. Therefore, 20 segments were generated in the dataset (data for two segments were missing due to quality issue).

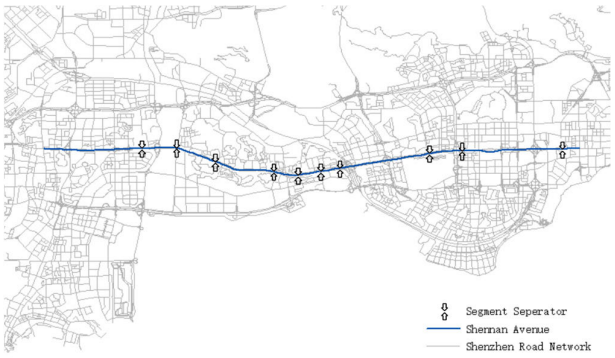


FIGURE 2. Location of the test road in Shenzhen city, with road segment displayed respectively.

The test road is covered with dedicated bus lanes (DBLs). DBLs are implemented on the roadside of all selected road segments and are only enabled during the morning peak hours 7:30-9:30 a.m. and the evening peak hours 05:30-07:30 p.m. on weekdays. DBLs are exclusively used by buses during the eligible periods.

A. GPS DATA PROCESSING

Bus GPS data was collected on 2015-06-03 from all the operating buses. The average bus speeds of selected road segments were calculated in 15-minute intervals. Thus, bus speeds were classified into 64 consecutive time periods covering from 6:00 to 22:00 and are denoted from 1 to 64. Varying

$$L_i^t(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_{it}} \exp\left(-\frac{(y_i^t - (\beta_0^t + \sum_{p=1}^{P_t} \beta_p^t X_{pi}^t + (u_i + \lambda_i) + (\gamma_t + \zeta_t) + \delta_{it}))^2}{2\sigma_{it}^2} \right) \tag{16}$$

generating intervals apply to different bus route categories (e.g., major, express, branch) with the range from 1 s to 59 s. Over 90% of the GPS records are generated with intervals under 30 s. In this study, GPS records with intervals under 30 s were used.

Map matching was conducted to match GPS data to road segments. To reduce computation workload, a grid-based searching algorithm was implemented to find the road segments within the grid of each GPS record. The grid size was 100 m $*$ 100 m. After locating the road segments, we identify the GPS records that were located within the range of 30 m of an intersection and tagged them for further verification using the subsequent matching records of the same vehicle. If the record was located outside the range of an intersection, it was then appended to the processed dataset with respect to plate number and timestamp. Another detection algorithm was then executed to deal with irregular circumstances like reverse driving by skipping to the next regular GPS record. The processed GPS records were eventually matched to the nearest road segments if they passed the anomaly detection, and match results were generated for speed calculation. The tagged records were verified using those match results. Compared to the ground truth, the matching accuracy reaches 98%, the 2% error is caused due to the lack of subsequent data for verification.

B. VARIABLE EXTRACTION

Road infrastructure data and bus service data are also included in this study, where road infrastructure data includes driveway density and the number of lanes. Driveway is defined as the set of property entrances and non-signalized intersections. Bus service data consists of the number of bus routes, bus stop density, bus stop configuration (harbor-shaped or not), bus stop location (near an intersection or not), and bus density. The bus volumes used for density computing were extracted using bus GPS data. The detailed information regarding these variables is summarized in TABLE 2. The abovementioned data were obtained from field surveys and the Baidu Street Map. For further details, please refer to Cui *et al.* [11].

IV. RESULTS AND DISCUSSIONS

Based on previous work, speed data with time intervals separated over 2 hours presents little temporal correlations [11]. Thus, the look-back time window (w) was set within the range of 1 to 8, and the step ahead (a) for prediction was also set within the range of 1 to 8. T was set to 64.

Over 4,000 scenarios were analyzed to characterize the model performance and to discover the best model and scenario. Using the INLA method, the average run time for the Bayesian spatiotemporal models is 5.3 seconds. Three evaluation metrics were used, including mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). Each metric quantifies the error from a distinctive perspective. Using all of them provides a comprehensive evaluation regarding the prediction results.

TABLE 2. Data description and descriptive statistics.

Variable	Description	Mean	SD*
Bus speed	Average bus speed in the road segment for 15 minutes interval (km/h)	30.11	8.32
Driveway density	Number of driveways per km along the road segment	3.60	0.98
Number of lanes	Number of lanes along the road segment	5.00	1.00
Number of bus routes	Number of bus routes operating on the road segment	49.75	10.29
Bus stop density	Number of stops per km along the road segment	3.33	1.78
Bus stop configuration	Whether bus stop is harbor-shaped or not -- 0 denotes no, 1 otherwise	0.90	0.30
Bus stop location	Whether bus stop locates near intersection or not -- 0 denotes no, 1 otherwise	0.40	0.50
Density	The ratio of bus volume and bus speed. (Veh/km)	9.73	8.66

* SD stands for standard deviation

The testing environment for the proposed models was macOS version 10.14.3 with 2.6 GHz processors and 16 GB random access memory.

A. PREDICTOR SELECTION AND INTERPRETATION

A sense-making statistic inference on predictors is the proof for properly accounted for the relations between the independents and the dependent. Two criteria were used to select predictors: (i) the coefficient must have the correct positive/negative sign, and (ii) the coefficient should be statistically significant. The 95% Bayesian Confidence Interval (BCI) was used in this study. Based on these rules, the variable selection was performed. Results show that only density met the criteria and was thus included into the model.

B. MODEL SELECTION

1) MODEL PERFORMANCE CHANGING PATTERN OVER STARTING PERIOD p

In order to discover both the best model and combination of p and w, step ahead should be controlled. In this case, a is set to 1, meaning that the testing procedure is only performed on the next time period following the training time periods. Density is included in the model as a predictor.

Figure 3 shows the 3D distribution of the one-step-ahead prediction MAPE versus start time and time window. Both the Type II model and the Type III model present severe prediction errors, where the maximum error of the Type III model reaches 1,000%, while its average error yields around 100%. The Type II model presents a completely different distribution pattern, and its maximum error reaches 100%. This indicates that the Type II model interaction and the Type III model interaction are not a good fit for the data. For the other models, the PL model and the Type I model have a similar maximum MAPE error of 50%. The ST model and the

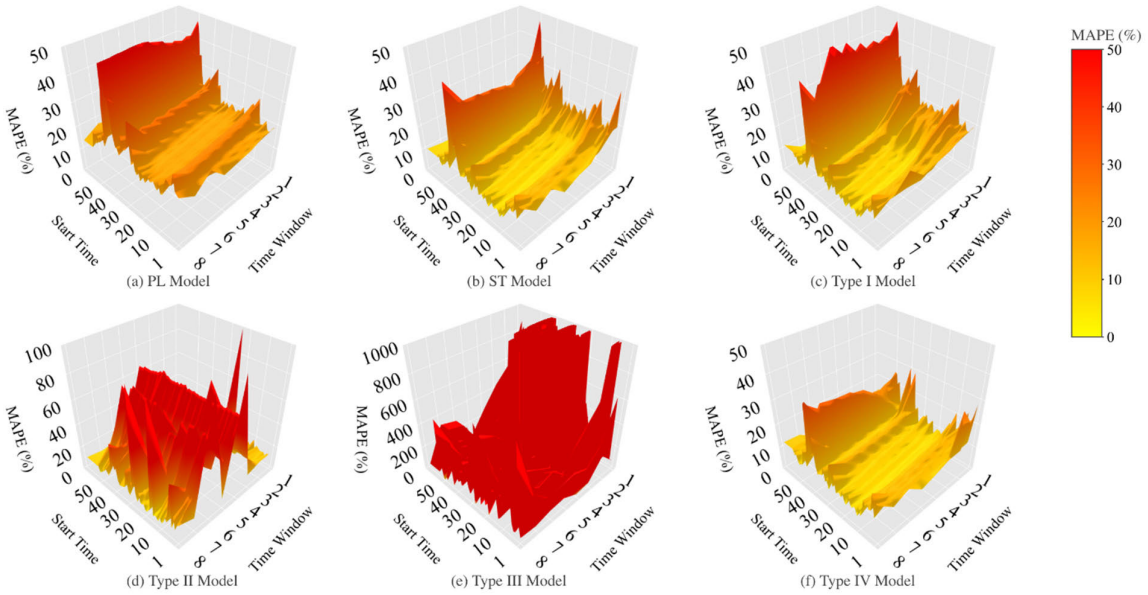


FIGURE 3. One-step ahead prediction MAPE vs. start time and time window in 3D for different models. MAPE over 50% shares the same displaying color as MAPE equals 50%.

Type IV model present smaller MAPE, where the maximum MAPE is 45% for the ST model and 30% for the Type IV model. In summary, the Type IV model presents the smallest MAPE.

2) MODEL PERFORMANCE CHANGING PATTERN OVER TIME WINDOW *w*

Figure 3 presents an intuitive way for model evaluation, and the Type IV model should be considered the best model based on this intuitive evaluation. This section presents more quantitative evaluation results for a drawn conclusion on model selection.

In the time series analysis framework, the number of time periods used for training a time series model is essential to the performance of the prediction. A window tracing back too much historical data can disrupt the actual pattern hidden in the time series data, misleading the model to learn from a false pattern. On the contrary, a window not connecting enough history data will lead to the incapability of prediction performance [68]. As such, the time window w plays a vital role in prediction performance.

TABLE 3 summarizes the MAPE, MAE, and RMSE results for the test dataset. The smallest MAPE is represented in bold. It is found that the PL model, the ST model, the Type I model, and the Type III model are inferior to the Type IV model and the Type II model due to more significant prediction errors for each evaluation measure. The abnormally high MAPE of the Type_III model indicates that the structure of the Type_III model cannot well accommodate the data. Comparing the Type II model with the Type IV model, we can see that the prediction performance of the Type II model is much less stable than the Type IV model. For example, for the Type II model, the MAPE values for cases with w equals 4 and 6 are much larger than the MAPEs for other w cases. This fluctuation is also observed in the MAE and RMSE for the Type II model. Instead, the performance of the Type IV model is rather stable. Therefore, the Type IV model is chosen as the best model.

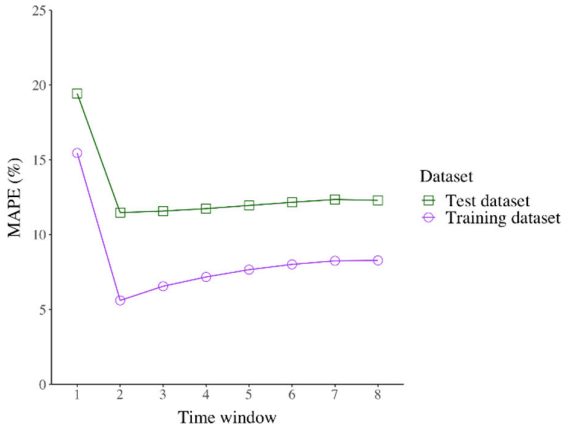


FIGURE 4. MAPE values of the Type IV Model versus time window from 1 to 8 using both training dataset and test dataset.

Subsequently, we determine the best choice of the time window (w). Figure 4 shows the MAPE values of the Type IV model versus the time window from 1 to 8 using the test dataset and training dataset. In both cases, the Type IV model achieved the lowest prediction MAPE when the time window equals 2 (the MAPE is 11.48% for the test dataset).

To conclude, the Type IV model is chosen as the best model, which denotes that spatial trends exist among different road segments, and these spatial trends are temporally correlated as well. Besides, time window 2 is proven to be the optimal time window for prediction. The above analysis proves that specifying a proper STI model on data will greatly improve the prediction performance. In addition, this finding will also provide new insights for transit agencies to form new policies regarding zonal management.

C. MODEL INTERPRETATION

The interpretability of a model can be defined as the ability to quantify to what extent each component (e.g., explanatory variables, spatial, temporal, and STI effects) of the model contributes to the variation of the response variable.

In this study, the response variable, i.e., the bus speed, is formulated using Eq. (6). Based on Section IV.A, density is selected as the explanatory variable. Therefore, Eq. (6) becomes:

$$\mu_i^t = \beta_0 + \beta_1 \cdot Density + (u_i + \lambda_i) + (\gamma_t + \zeta_t) + \delta_{it} \quad (21)$$

Based on the above definition of interpretability, the proposed BSTI model can quantitatively describe the contribution of each component on the right-hand side of Eq. (21).

For the explanatory variable, this interpretability can be achieved by analyzing the regression coefficients of the variables. The coefficient of density is -0.65 for morning peak hours, which indicates that one bus increment per kilometer would lead to a decrease of 0.65 km/h in bus speed. For non-peak hours from 14:00-16:00, the average coefficient for density is -1.15, while for evening peak hours, the average coefficient is -0.45.

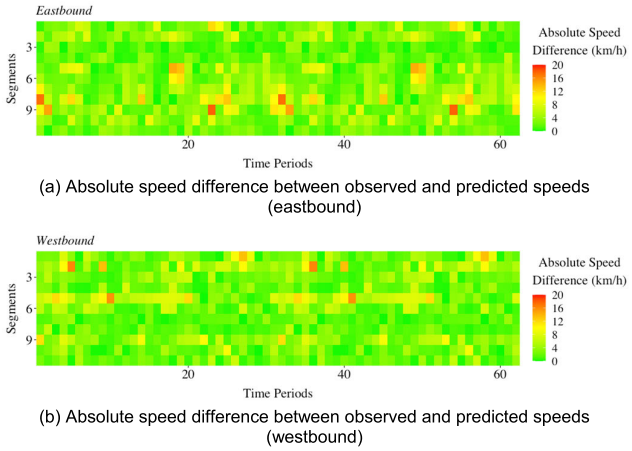
For the random effects, including spatial, temporal, and STI effects, this interpretability can be achieved by analyzing the variance of estimated values of each effect. TABLE 4 summarizes the mean and variance of the estimates of each random effect. From TABLE 4, only the STI effect is significant at the 85% confidence level, and other spatial/temporal effects are not statistically significant. The contribution of each component on bus speeds can be explained by comparing the variances. For example, let f_{STI} denote the ratio of the variance of the STI effect to the variance of all the random effects, and f_{STI} is calculated as follows:

$$f_{STI} = \frac{\sigma_\delta^2}{\sigma_\delta^2 + \sigma_u^2 + \sigma_\lambda^2 + \sigma_\gamma^2 + \sigma_\zeta^2} \quad (22)$$

Based on TABLE 4, f_{STI} equals 99.94%. This finding indicates that the STI effect dominates other random effects in addressing the speed variance.

D. BUS SPEED ONLINE PREDICTION

In this subsection, we explore the results of the Type IV model for online prediction. The results for the step ahead (a) equals 1 are discussed. In time series analysis, prediction performance usually decreases as the time period of interest for prediction extends further in time [68]. Thus, the impacts of the step ahead (a) on the prediction accuracy of Bayesian spatiotemporal models are also investigated.



(a) Absolute speed difference between observed and predicted speeds (eastbound)

(b) Absolute speed difference between observed and predicted speeds (westbound)

FIGURE 5. Distributions of the prediction errors of the bus speeds.

Figure 5 shows the distributions of the prediction errors of the bus speeds in both directions using a heat map. Figure 5(a) shows the absolute differences (residuals) between observed bus speeds and predicted bus speeds. Most of the prediction errors are limited to a small range of 0 to 4 km/h (denoted as bright green), which affirms the prediction performance of the proposed model. Another valuable finding is that the prediction residuals are randomly distributed, which indicates that the proposed Type IV model accounts for the spatiotemporal effects properly. Similar results are found for the westbound in Figure 5(b).

Compared to the conventional statistic approaches, the proposed model can interpret the significant dependent variables correctly and make precise predictions based on them (with a prediction MAPE of 11.48%). Furthermore, compared with classic ''black-box'' like machine learning models, the proposed model can adjust statistical inference and subsequently improve prediction performance based on the spatiotemporal effects. These edges again proved the superiority of the proposed model.

Figure 6 demonstrates the prediction MAPE using the Type IV model with $w = 2$ versus step ahead. In Figure 6(a), the Type II model and the Type III model present large MAPE values, where the former has a MAPE over 50% while the latter has a MAPE that nearly reaches 90%. Based on the aforementioned analysis, a possible reason is that the assumptions of the STI are incompatible with the data.

Figure 6(b) excludes the Type II model and the Type III model. From Figure 6(b), prediction MAPE increases when the step ahead increases. This finding is consistent with Das [68]. The Type IV model presents smaller MAPE at all possible values of the step ahead than any other models, which further proves that the Type IV model has superior compatibility with data. Another finding is that the ST model and the Type IV model present larger MAPE increments than

TABLE 3. One-step ahead prediction assessment results for test dataset.

Measures	Models	Time window (w)							
		1	2	3	4	5	6	7	8
MAPE (%)	PL	21.97	22.26	22.53	22.81	23.08	23.34	23.51	23.77
	ST	22.21	12.89	12.06	12.11	12.22	12.47	12.69	12.71
	Type_I	21.97	20.01	14.99	13.42	13.51	13.54	12.98	12.59
	Type_II	13.41	11.69	12.93	95.88	11.54	76.35	11.62	16.51
	Type_III	309.36	1,524.01	16,585.41	82.45	170.82	177.17	224.83	264.13
	Type_IV	19.44	**11.48**	11.58	11.74	11.95	12.17	12.35	12.30
MAE	PL	5.48	5.50	5.52	5.54	5.56	5.59	5.60	5.63
	ST	5.56	3.40	3.17	3.14	3.13	3.17	3.21	3.21
	Type_I	5.48	5.05	3.82	3.40	3.37	3.36	3.27	3.18
	Type_II	3.83	3.28	3.59	28.68	3.04	22.56	3.02	4.50
	Type_III	87.68	424.99	4,683.35	23.47	45.64	47.97	62.35	72.03
	Type_IV	5.79	3.15	3.10	3.09	3.10	3.13	3.16	3.15
RMSE	PL	6.78	6.79	6.79	6.81	6.84	6.89	6.91	6.96
	ST	6.88	4.39	4.12	4.11	4.14	4.19	4.24	4.24
	Type_I	6.78	6.26	4.83	4.37	4.37	4.40	4.30	4.23
	Type_II	4.92	4.34	4.59	29.64	4.01	23.53	4.02	5.55
	Type_III	88.21	425.61	4,683.85	24.64	46.06	48.44	62.67	72.36
	Type_IV	7.46	4.14	4.09	4.09	4.13	4.18	4.20	4.18

Note: bold denotes lowest MAPE.

TABLE 4. Estimate results of Type_IV model.

Effect terms	Effect types	Estimates	Significance
Spatial (structured)	Unobservable / random effect		
μ_u		-7.93E-08	-
σ_u^2		2.58E-03	-
Spatial (unstructured)	Unobservable / random effect		
μ_λ		-2.29E-08	-
σ_λ^2		1.16E-04	-
Temporal (structured)	Unobservable / random effect		
μ_γ		-1.34E-08	-
σ_γ^2		1.17E-04	-
Temporal (unstructured)	Unobservable / random effect		
μ_ζ		-2.21E-08	-
σ_ζ^2		1.17E-04	-
STI	Unobservable / random effect		
μ_δ		-1.73E-05	*
σ_δ^2		4.96E+00	*

* Significant under 85% confidence level
\- Not significant

the PL model and the Type I model as the step ahead increases from 1 to 8. This increment for the former two models is 8.3% on average, while that of the latter two models is 7.1%. A possible reason is that the ST model and the Type IV model account for structured temporal effects, which is no doubt sensitive to time changing; while the PL model and the Type I model do not account for structured temporal effects (See TABLE 1 for detail about four types of interaction).

Aside from the proposed model, three benchmark models are also tested, i.e., autoregressive integrated moving average (ARIMA) model [69], CNN model [70], and the back-propagation neural network (BPNN) model [69]. The hyperparameters for the benchmark models are listed in TABLE 5. CNN has two convolution layers, while BPNN has five regular layers. The input data of CNN is organized as a matrix, where each row represents a road segment, and each column denotes a time period, while the input of the BPNN is organized as a vector. The kernel size of the CNN model is set to 3×3, indicating that each convolution involves three connected road segments and three consecutive time periods. Note that BPNN does not perform convolution. Thus, ''Kernel size'', ''Pooling'', and ''Stride'' do not apply for BPNN. The dropout rate of the BPNN is larger than CNN due to more parameters are trained in the BPNN model. For ARIMA, the order of autoregression (p), difference (d), and moving average (q) are tuned to obtain the smallest Akaike information criterion (AIC), where AIC estimates the prediction error and is widely used for model selection [71]. After tuning, p, d, and q equal to 1, 0, and 1, respectively.

Based on the methodology of each model, CNN utilizes both spatial and temporal information via a 2-dimensional kernel filter. ARIMA utilizes temporal information and assumes linear relations between historical data and future data. BPNN addresses the non-linearity among data and cannot learn spatial or temporal features.

TABLE 6 summarizes the comparison results on MAPE, MAE, RMSE, and the computation time between the proposed Type IV model and three benchmark methods. Based

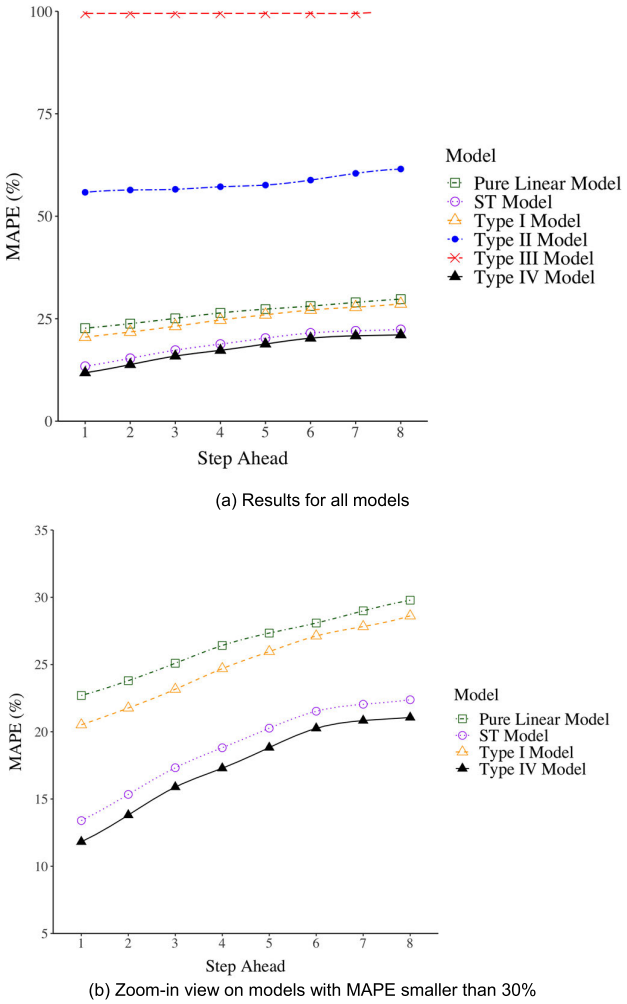


FIGURE 6. Prediction MAPE with $w = 2$ versus step ahead (a) categorized by different models: (a) Results for all models; (b) Zoom-in view on models with MAPE smaller than 30%.

TABLE 5. Specifics of benchmark models.

Model Specifics	CNN	BPNN	ARIMA
Number of hidden layers	2	5	-
Number of filters/neurons in each layer	[32, 64]	[12,12,9,6,3]	-
Kernel size	[3×3]	-	-
Pooling	Max Pooling	-	-
Stride	1	-	-
Dropout rate	0.10	0.15	-
Activation function	Relu	Relu	-
Optimizer	Adam	Adam	-
Learning rate	0.01	0.01	-
Order of autoregression (p)	-	-	1
Order of difference (d)	-	-	0
Order of moving average (q)	-	-	1

"-": Not applicable

on TABLE 6, our model outperforms the benchmark methods by presenting the lowest MAPE, MAE, and RMSE. Besides, our proposed model runs faster than the benchmark models.

TABLE 6. Bus speed prediction comparisons.

Algorithm	MAPE (%)	MAE (km/h)	RMSE (km/h)	Time (s)
ARIMA	13.75	3.92	4.98	5.4
CNN	12.43	3.37	4.76	873
BPNN	22.74	7.02	9.02	671
Proposed Type IV model	**11.48**	**3.15**	**4.14**	**5.3 (1044)**

Note: the time in parenthesis denotes the computation time using the MCMC method.

TABLE 7. Bus traffic state prediction accuracy in literature.

Papers	MAPE (%)
Sun, et al. [5]	14.8
Achar, et al. [9]	20.0
Pang, et al. [18]	16.32
Achar, et al. [72]	18.32
Chen [29]	23.17
Proposed model	11.48

The average computation time for our model is 5.3 seconds. Even a simple model like ARIMA costs more time than our model. In addition, the computation time for the MCMC method is 1044 seconds, which is much larger than that of our model. Furthermore, we conduct a two-sample t-test on the prediction results of the proposed model and ARIMA model. Based on the results, the p-value is $0.054 < 0.1$. Therefore, we can conclude that the proposed model performs significantly better than the ARIMA model at the 90% confidence level.

TABLE 7 presents the prediction MAPEs of the reviewed papers. Based on TABLE 7, the prediction MAPE observed in the literature is between 14.8% and 23.17%. Our proposed model reaches the MAPE of 11.48%.

V. CONCLUSION

This paper introduces a novel Bayesian hierarchical approach for bus speed prediction, which explicitly accounts for the STI of bus speed observations. Large-scale bus GPS data from Shenzhen, China are used to obtain bus speeds. All the developed Bayesian models are estimated using the integrated nested Laplace approximation technique, which expedites the estimation process. The optimal model specification is then identified. Predictor selection is conducted based on the sensibility of coefficient signs and their statistical significance. The most suitable STI pattern is identified based on prediction performance.

Key findings are summarized as follows: (1) The proposed Type IV model can best accommodate the bus speed data. The lowest MAPE for one-step-ahead prediction is 11.48% using data from two previous time intervals; (2) MAPE of the Type IV model grows as the steps ahead for prediction increases, and the Type IV model consistently outperforms the others; (3) Prediction errors of the Type IV model tend to be randomly distributed over time and space, indicating

that our proposed model is capable of addressing the spatial and temporal correlations among data; (4) The average run time for the Bayesian spatiotemporal models using Laplace approximation is 5.3 s, while the MCMC-based algorithm takes 17.4 mins; (5) Our model outperforms the benchmark statistical and machine learning models by a range from 0.95% to 11.26% in MAPE.

The proposed approach can achieve high prediction accuracy and computational efficiency without compromising its interpretability regarding the effects of contributing factors and the unobserved spatiotemporal heterogeneity. Given these advantages, the proposed model is expected to be able to handle arbitrary data with spatiotemporal correlations and STI effects and has the potential to be widely utilized for online bus speed prediction in intelligent public transit systems. In the future study, the proposed model will be further examined using bus speed data from a more extensive urban network. The transferability of the proposed model will also be tested using data from other cities

ACKNOWLEDGMENT

The authors would like to thank the Transport Commission of Shenzhen Municipality for providing the data.

REFERENCES

[1] C. E. Cortés, J. Gibson, A. Gschwender, M. Munizaga, and M. Zuñiga, "Commercial bus speed diagnosis based on GPS-monitored data," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 695–707, Aug. 2011.

[2] W. X. Hu and A. Shalaby, "Use of automated vehicle location data for route- and segment-level analyses of bus route reliability and speed," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2649, no. 1, pp. 9–19, Jan. 2017.

[3] M. M. Rahman, S. C. Wirasinghe, and L. Kattan, "Analysis of bus travel time distributions for varying horizons and real-time applications," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 453–466, Jan. 2018.

[4] J. Zhao, Y. Gao, Z. Bai, H. Wang, and S. Lu, "Traffic speed prediction under non-recurrent congestion: Based on LSTM method and BeiDou navigation satellite system data," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 2, pp. 70–81, Summer 2019.

[5] Y. Sun, G. Jiang, S.-K. Lam, S. Chen, and P. He, "Bus travel speed prediction using attention network of heterogeneous correlation features," in *Proc. SIAM Int. Conf. Data Mining*. Philadelphia, PA, USA: SIAM, 2019, pp. 73–81.

[6] H. Liu, H. Xu, Y. Yan, Z. Cai, T. Sun, and W. Li, "Bus arrival time prediction based on LSTM and spatial-temporal feature vector," *IEEE Access*, vol. 8, pp. 11917–11929, 2020.

[7] S. Taguchi, S. Koide, and T. Yoshimura, "Online map matching with route prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 338–347, Jan. 2019.

[8] W. Ding and S. Shen, "Online vehicle trajectory prediction using policy anticipation network and optimization-based context reasoning," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, A. Howard *et al.*, Eds., May 2019, pp. 9610–9616.

[9] A. Achar, D. Bharathi, B. A. Kumar, and L. Vanajakshi, "Bus arrival time prediction: A spatial Kalman filter approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1298–1307, Mar. 2020.

[10] B. Yu, H. Wang, W. Shan, and B. Yao, "Prediction of bus travel time using random forests based on near neighbors," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 4, pp. 333–350, 2018.

[11] H. Cui, K. Xie, B. Hu, H. Lin, and R. Zhang, "Analysis of bus speed using a multivariate conditional autoregressive model: Contributing factors and spatiotemporal correlation," *J. Transp. Eng. A, Syst.*, vol. 145, no. 4, Apr. 2019, Art. no. 04019009.

[12] L. Moreira-Matias, J. Mendes-Moreira, J. F. de Sousa, and J. Gama, "Improving mass transit operations by using AVL-based systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1636–1653, Aug. 2015.

[13] P. He, G. Jiang, S.-K. Lam, and Y. Sun, "Learning heterogeneous traffic patterns for travel time prediction of bus journeys," *Inf. Sci.*, vol. 512, pp. 1394–1406, Feb. 2020.

[14] W. Li, J. Wang, R. Fan, Y. Zhang, Q. Guo, C. Siddique, and X. Ban, "Short-term traffic state prediction from latent structures: Accuracy vs. efficiency," *Transp. Res. C, Emerg. Technol.*, vol. 111, pp. 72–90, Feb. 2020.

[15] Z. Zhang, M. Li, X. Lin, Y. Wang, and F. He, "Multistep speed prediction on traffic networks: A deep learning approach considering spatio-temporal dependencies," *Transp. Res. C, Emerg. Technol.*, vol. 105, pp. 297–322, Aug. 2019.

[16] Y. Yang, A. Heppenstall, A. Turner, and A. Comber, "Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems," *Comput., Environ. Urban Syst.*, vol. 83, Sep. 2020, Art. no. 101521.

[17] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 3, pp. 387–399, Jun. 2011.

[18] J. Pang, J. Huang, Y. Du, H. Yu, Q. Huang, and B. Yin, "Learning to predict bus arrival time from heterogeneous measurements via recurrent neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3283–3293, Sep. 2019.

[19] L. N. N. Do, H. L. Vu, B. Q. Vo, Z. Liu, and D. Phung, "An effective spatial-temporal attention based neural network for traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 108, pp. 12–28, Nov. 2019.

[20] B. Yu, Y. Lee, and K. Sohn, "Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (GCN)," *Transp. Res. C, Emerg. Technol.*, vol. 114, pp. 189–204, May 2020.

[21] S. Deng, S. Jia, and J. Chen, "Exploring spatial–temporal relations via deep convolutional neural networks for traffic flow prediction with incomplete data," *Appl. Soft Comput.*, vol. 78, pp. 712–721, May 2019.

[22] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transp. Res. A, Policy Pract.*, vol. 44, no. 5, pp. 291–305, Jun. 2010.

[23] Y. Bin, Y. Zhongzhen, and Y. Baozhen, "Bus arrival time prediction using support vector machines," *J. Intell. Transp. Syst., Technol., Planning, Oper.*, vol. 10, no. 4, pp. 151–158, 2006.

[24] C.-W. Tan, S. Park, K. Zhou, H. Liu, P. Lau, M. Li, and W.-B. Zhang, "Prediction of transit vehicle arrival times at signalised intersections for signal priority control," in *Proc. IEEE Intell. Transp. Syst. Conf.*, Sep. 2006, pp. 1477–1482.

[25] H. Chang, D. Park, S. Lee, H. Lee, and S. Baek, "Dynamic multi-interval bus travel time prediction using bus transit data," *Transportmetrica*, vol. 6, no. 1, pp. 19–38, Jan. 2010.

[26] X.-L. Sun and J. Lu, "Travel time prediction between bus stations," in *Proc. ICCTP*, 2011, pp. 2731–2739.

[27] Y. Bie, D. Wang, and H. Qi, "Prediction model of bus arrival time at signalized intersection using GPS data," *J. Transp. Eng.*, vol. 138, no. 1, pp. 12–20, Jan. 2012.

[28] H. Zhang, S. Liang, Y. Han, M. Ma, and R. Leng, "A prediction model for bus arrival time at bus stop considering signal control and surrounding traffic flow," *IEEE Access*, vol. 8, pp. 127672–127681, 2020.

[29] C.-H. Chen, "An arrival time prediction method for bus system," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 4231–4232, Oct. 2018.

[30] M. Chen, X. Liu, J. Xia, and S. I. Chien, "A dynamic bus-arrival time prediction model based on APC data," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 19, no. 5, pp. 364–376, 2004.

[31] S. Cheng, F. Lu, P. Peng, and S. Wu, "Short-term traffic forecasting: An adaptive ST-KNN model that considers spatial heterogeneity," *Comput., Environ. Urban Syst.*, vol. 71, pp. 186–198, Sep. 2018.

[32] S. V. Kumar and L. Vanajakshi, "Pattern identification based bus arrival time prediction," *Proc. Inst. Civil Eng., Transp.*, vol. 167, no. 3, pp. 194–203, Jun. 2014.

[33] M. U. Farooq, A. Shakoor, and A. B. Siddique, "GPS based public transport arrival time prediction," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2017, pp. 76–81.

[34] Q. Tang, M. Yang, and Y. Yang, "ST-LSTM: A deep learning approach combined spatio-temporal features for short-term forecast in rail transit," *J. Adv. Transp.*, vol. 2019, pp. 1–8, Feb. 2019.
[35] P. Liu, Y. Zhang, D. Kong, and B. Yin, "Improved spatio-temporal residual networks for bus traffic flow prediction," *Appl. Sci.*, vol. 9, no. 4, p. 12, Feb. 2019.
[36] F. Lin, Y. Xu, Y. Yang, and H. Ma, "A spatial-temporal hybrid model for short-term traffic prediction," *Math. Problems Eng.*, vol. 2019, Jan. 2019, Art. no. 4858546.
[37] A. Salamanis, P. Meladianos, D. Kehagias, and D. Tzovaras, "Evaluating the effect of time series segmentation on STARIMA-based traffic prediction model," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.* Los Alamitos, CA, USA: IEEE Computer Society, Sep. 2015, pp. 2225–2230.
[38] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, Apr. 2017, Art. no. 818.
[39] W. Treethidtaphat, W. Pattara-Atikom, and S. Khaimook, "Bus arrival time prediction at any distance of bus route using deep neural network model," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Piscataway, NJ, USA, Oct. 2017, p. 5.
[40] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Roy. Stat. Soc. B, Methodol.*, vol. 36, no. 2, pp. 192–236, 1974.
[41] L. Bernardinelli, D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini, "Bayesian analysis of space—Time variation in disease risk," *Statist. Med.*, vol. 14, nos. 21–22, pp. 2433–2443, Nov. 1995.
[42] S. M. Berry, B. P. Carlin, J. J. Lee, and P. Muller, *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL, USA: CRC Press, 2010.
[43] S. Jackman, *Bayesian Analysis for the Social Sciences*. Hoboken, NJ, USA: Wiley, 2009.
[44] W. Cheng, G. S. Gill, Y. Zhang, T. Vo, F. Wen, and Y. Li, "Exploring the modeling and site-ranking performance of Bayesian spatiotemporal crash frequency models with mixture components," *Accident Anal. Prevention*, vol. 135, Feb. 2020, Art. no. 105357.
[45] W. Cheng, G. S. Gill, J. L. Ensch, J. Kwong, and X. Jia, "Multimodal crash frequency modeling: Multivariate space-time models with alternate spatiotemporal interactions," *Accident Anal. Prevention*, vol. 113, pp. 159–170, Apr. 2018.
[46] L. T. Truong, L.-M. Kieu, and T. A. Vu, "Spatiotemporal and random parameter panel data models of traffic crash fatalities in Vietnam," *Accident Anal. Prevention*, vol. 94, pp. 153–161, Sep. 2016.
[47] D. Rohde, J. Corcoran, and P. Chhetri, "Spatial forecasting of residential urban fires: A Bayesian approach," *Comput., Environ. Urban Syst.*, vol. 34, no. 1, pp. 58–69, Jan. 2010.
[48] L. Knorr-Held, "Bayesian modelling of inseparable space-time variation in disease risk," *Statist. Med.*, vol. 19, nos. 17–18, pp. 2555–2567, 2000.
[49] M. B. Ulak, E. E. Ozguven, O. A. Vanli, M. A. Dulebenets, and L. Spainhour, "Multivariate random parameter Tobit modeling of crashes involving aging drivers, passengers, bicyclists, and pedestrians: Spatiotemporal variations," *Accident Anal. Prevention*, vol. 121, pp. 1–13, Dec. 2018.
[50] X. Ma, S. Chen, and F. Chen, "Multivariate space-time modeling of crash frequencies by injury severity levels," *Anal. Methods Accident Res.*, vol. 15, pp. 29–40, Sep. 2017.
[51] D. Geepthi and C. C. Columbus, "Network traffic detection for peer-to-peer traffic matrices on Bayesian network in WSN," *J. Ambient Intell. Hum. Comput.*, vol. 12, pp. 6975–6986, Jul. 2020.
[52] K. Xie, X. Wang, H. Huang, and X. Chen, "Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models," *Accident Anal. Prevention*, vol. 50, pp. 25–33, Jan. 2013.
[53] B. Hu, K. Xie, H. Cui, and H. Lin, "A Bayesian spatiotemporal approach for bus speed modeling," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 497–502.
[54] Z. Y. Yu, J. S. Wood, and V. V. Gayah, "Using survival models to estimate bus travel times and associated uncertainties," *Transp. Res. C, Emerg. Technol.*, vol. 74, pp. 366–382, Jan. 2017.
[55] J. Besag, J. York, and A. Mollié, "Bayesian image restoration, with two applications in spatial statistics," *Ann. Inst. Stat. Math.*, vol. 43, no. 1, pp. 1–20, 1991.
[56] B. Du, H. Peng, S. Wang, M. Z. A. Bhuiyan, L. Wang, Q. Gong, L. Liu, and J. Li, "Deep irregular convolutional residual LSTM for urban traffic passenger flows prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 972–985, Mar. 2020.
[57] S. Zhang, X. Liu, J. Tang, S. Cheng, Y. Qi, and Y. Wang, "Spatio-temporal modeling of destination choice behavior through the Bayesian hierarchical approach," *Phys. A, Stat. Mech. Appl.*, vol. 512, pp. 537–551, Dec. 2018.
[58] D. Clayton, "Generalised linear mixed models," in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson, and D. Spiegelhalter, Eds. New York, NY, USA: Chapman & Hall, 1996, pp. 275–301.
[59] C. J. Stone, M. H. Hansen, C. Kooperberg, and Y. K. Truong, "Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture," *Ann. Statist.*, vol. 25, no. 4, pp. 1371–1470, Aug. 1997.
[60] M. Blangiardo and M. Cameletti, *Spatial and Spatio-Temporal Bayesian Models With R-INLA*. Hoboken, NJ, USA: Wiley, 2015.
[61] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.
[62] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Boca Raton, FL, USA: CRC Press, 1995.
[63] K. Xie, K. Ozbay, and H. Yang, "A multivariate spatial approach to model crash counts by injury severity," *Accident Anal. Prevention*, vol. 122, pp. 189–198, Jan. 2019.
[64] H. Rue, S. Martino, and N. Chopin, "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 71, no. 2, pp. 319–392, Apr. 2009.
[65] Y. Ding and S. I. J. Chien, "The prediction of bus arrival times with link-based artificial neural networks," in *Proc. 5th Joint Conf. Inf. Sci. (JCIS)*, vol. 5. Atlantic City, NJ, USA: Duke Univ./Association for Intelligent Machinery, Mar. 2000, pp. 730–733.
[66] B. Yu, W. H. K. Lam, and M. L. Tam, "Bus arrival time prediction at bus stop with multiple routes," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 6, pp. 1157–1170, 2011.
[67] Z. He, H. Yu, Y. Du, and J. Wang, "SVM based multi-index evaluation for bus arrival time prediction," in *Proc. Int. Conf. ICT Converg. (ICTC)*, Oct. 2013, pp. 86–90.
[68] S. Das, *Time Series Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1994.
[69] J. Zheng and M. Huang, "Traffic flow forecast through time series analysis based on deep learning," *IEEE Access*, vol. 8, pp. 82562–82570, 2020.
[70] W. Li, X. Ban, J. Zheng, H. X. Liu, C. Gong, and Y. Li, "Real-time movement-based traffic volume prediction at signalized intersections," *J. Transp. Eng. A, Syst.*, vol. 146, no. 8, p. 15, Aug. 2020.
[71] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.
[72] A. Achar, R. Regikumar, and B. A. Kumar, "Dynamic bus arrival time prediction exploiting non-linear correlations," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.

HAIPENG CUI was born in Hainan, China, in 1990. He received the Ph.D. degree in transportation engineering from Tongji University in 2019. He is currently a Postdoctoral Research Fellow with the Department of Civil and Environmental Engineering, National University of Singapore (NUS). His research interests include intelligent transportation systems, traffic safety, traffic operation, and logistics. He has developed expertise in deep learning, big data analytics, and statistical modeling. He has served as a Reviewer for several scientific journals and international conferences, such as *Transportation Research Part E: Logistics and Transportation Review*, Transportation Research Board (TRB), and IEEE International Conference on Intelligent Transportation Systems (ITSC). He has served as a Core Participant for multiple research projects sponsored by national authorities/research agencies, such as the Ministry of Housing and Urban-Rural Development of China and the National Science Foundation of China.

KUN XIE received the Ph.D. degree in transportation planning and engineering from New York University. Prior to joining ODU, he worked as an Assistant Professor with the Department of Civil and Natural Resources Engineering, University of Canterbury, and as a Postdoctoral Associate with the Center for Urban Science and Progress (CUSP), Department of Urban and Civil Engineering, New York University. He is currently an Assistant Professor with the Department of Civil and Environmental Engineering, Old Dominion University (ODU). He has published over 90 refereed papers in scholarly journals and conference proceedings. His research interests concentrate on the use of data-driven approaches and emerging technologies to enhance the safety, efficiency, and resiliency of transportation systems. He has developed expertise in big data analytics, machine learning, and statistical modeling. He has worked as the Principal Investigator (PI)/Co-PI/Core Participant for multiple research projects sponsored by agencies/organizations, such as the National Academy of Sciences (NAS), the U.S. Department of Transportation (USDOT), the National Science Foundation (NSF), Virginia Department of Transportation (VDOT), and American International Group (AIG). In an ongoing NAS project that he leads, he harnesses computer vision techniques to automatically extract roadway marking information on a large scale from aerial images and roadway GIS data. He is recognized by research awards such as IEEE ITS Best Dissertation Award, CUTC Milton Pikarsky Memorial Award, and Transportation Research Board (TRB) Best Paper Award. He has served as a Guest/Area Editor, an Editorial Board Member, the Session Chair, and a Reviewer for over 30 scientific journals and international conferences. He also serves as an Expert Panel Member for TRB NCHRP and BTSCRP projects and as a referee for a few university transportation centers.

BIN HU was born in Zhejiang, China, in 1977. He received the Ph.D. degree in transportation engineering from Tongji University, in 2019. He Founded the Shenzhen Communications Investment Technology Company Ltd., in 2013. His research interests include building intelligent transportation, intelligent networks, and smart city business system with independent intellectual property rights. He has served as the Principal Investigator (PI) for multiple projects including six science and technology plan projects of the Ministry of Housing and Urban-Rural Development of China, two science and technology plan projects of the Shenzhen Municipal Government. He has presided over more than 50 intelligent transportation system projects in Shenzhen and participated in more than 60 software copyrights and patents.

HANGFEI LIN received the Ph.D. degree in transportation planning and management from Tongji University, in 2007. He is currently a Full Professor with the College of Transportation Engineering, Tongji University. His research interests include urban transportation planning, traffic network modeling, and intelligent transportation systems. He has served as the Principal Investigator (PI) for multiple national key projects sponsored by national authorities/research agencies, such as the Ministry of Housing and Urban-Rural Development of China and the National Science Foundation of China. He is recognized by multiple scientific/technological awards, including the Science and Technology Progress Award of the Ministry of Construction of China and multiple Science and Technology Progress Awards of Shanghai, China.