Article:
Xing, J, Liu, R orcid.org/0000-0003-0627-3184, Zhang, Y et al. (3 more authors) (2023) Urban network-wide traffic volume estimation under sparse deployment of detectors. Transportmetrica A: Transport Science. ISSN 2324-9935

https://doi.org/10.1080/23249935.2023.2197511

Accepted for publication at TransportmetricA and accessible at:
https://doi.org/10.1080/23249935.2023.2197511

Urban network-wide traffic volume estimation under sparse deployment of detectors

Jiping Xing [a], *Ronghui Liu* [b], *Yuan Zhang* [a,c], *Charisma F. Choudhury* [b], *Shuyan Chen* [a,]

Qixiu Cheng [a,d*]

[a] *School of Transportation, Southeast University, Nanjing, China 211189*

[b] *Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, UK*

[c] *School of Cyber Science and Engineering, Southeast University, Nanjing, China 211189*

[d] *Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China.*

*Corresponding author. Email: qixiu.cheng@polyu.edu.hk.

Abstract

Sensing urban traffic information is fundamental for the sustainable development of urbanization and modernization. Accurate detection of network-wide traffic volumes is of significant importance in urban planning, traffic management, and vehicle emission control. However, owing to the limited budget and high device maintenance costs, the deployment of detectors is usually quite sparse, and thus it is neither effective nor practical for acquiring traffic volumes over the entire network merely using detectors. Considering the high coverage rate of taxi global position system (GPS) data, this study proposes a tailored transfer learning framework, called the transfer learning-based least square support vector regression model, to estimate the network-wide traffic volumes (especially on the undetected road segments) by fusing taxi GPS data and detector data. Furthermore, a similarity analysis based on the Jensen–Shannon divergence method was applied to identify similarities between road segments

installed with and without detectors. The numerical experiments were carried out on an experimental road network of Nanjing City, China, followed by a sensitivity analysis of the estimation. The results demonstrate that our proposed approach can be effectively used to estimate network-wide traffic volume under sparse deployment of detectors.

Keywords: Traffic volume estimation; Transfer learning; Data fusion; Similarity analysis; License plate recognition (LPR) data; Taxi GPS data.

1. Introduction

With the rapid development of urbanization and car ownership, modern cities suffer from many externalities from transportation systems, such as traffic congestion, traffic accidents, and carbon emissions (Majumdar et al., 2021). The accurate detection of traffic volume is fundamental to solving the problems in the urban transportation system, which can be an essential component in transport planning, traffic management, and control, and in turn, contributing to the improvement of urban sustainability.

In particular, network-wide traffic volume detection can be used to identify traffic congestion, optimize traffic signaling, and improve traffic routing plans (Li et al., 2020). Furthermore, the observed traffic volume data can serve as high-quality input data for driver path guidance (Lee et al., 2019), transit service scheduling (Huang et al., 2020), time-dependent origin-destination (OD) matrix inference (Zhou and Mahmassani, 2007), and vehicle emission estimation (Jia, 2021; Liu et al., 2019a).

Existing approaches for traffic volume detection are mainly based on fixed on-road or roadside detectors, such as ground loop detectors (Du et al., 2021), microwave sensors (Ma et al., 2015), and surveillance cameras (Cheng et al., 2019b; Zeroual et al., 2019). These detectors generate massive time-series traffic data at fine granularity and precisely reflect traffic conditions (Wilkie et al., 2013). Despite the growing adoption of traffic detectors, their coverage is still quite sparse because of the high installation and maintenance costs, particularly at suburban intersections or segments. Hence, the entire network-wide traffic volume information cannot be directly acquired by these fixed detectors. Furthermore, there are no historical traffic volume data on undetected segments. Estimating traffic volume on these segments without sensor data is a challenging problem, and solving such a challenging problem has a large market prospect that has been attempted by many emerging transportation companies, such as Gaode, TomTom, StreetLights, etc. It is important to highlight that the

network-wide traffic volume estimation is distinct from short-term traffic forecasting problems that merely apply historical detected data to predict future values.

Theoretically, there are two main groups of approaches for estimating traffic volumes, including statistical regression models and machine learning methods. The former includes the linear regression model (Shan et al., 2013), compressive sensing model (Liu et al., 2016), Kriging-based model (Yang et al., 2018), and particle filter model (Mihaylova et al., 2007). In contrast to statistical models based on prior assumptions of the input data, the principle of machine learning methods is to construct a nonlinear relationship without assumptions of prior knowledge. It includes tensor-based approaches (Tan et al., 2013), Bayesian network models (Zhan et al., 2017), Gaussian graphical models (Hara et al., 2018), and deep meta-learning (Pan et al., 2020). The aforementioned approaches were established mainly by using the spatiotemporal correlation of traffic volume data from adjacent segments or cascade segments with installed detectors. However, in general, the deployment of fixed detectors is sparse, and only partial segments with fixed detectors are installed. Thus, for network-wide traffic volume estimation, where most of the links do not install detectors, detector-based approaches are undermined. For example, the license plate recognition (LPR) detector is mainly used to monitor daily traffic operations and traffic accidents by the authorities, and these detectors are installed only at important road segments or key intersections, whereas they are much sparser in suburbs than in city centers.

On the other hand, with the advent of the big data era, many kinds of passive data are generated from floating vehicles, which can also be used to estimate traffic information with wider spatial coverage. These data include mobile phone records, private cars, and taxi GPS navigation trajectories. The network-wide traffic speed can be calculated from floating GPS data using a map-matching pre-process (Yu et al., 2020). Because traffic speed usually evolves from traffic volume, the similarities among road segments can be learned with respect to the

speed pattern (Cheng et al., 2021; Meng et al., 2017), and thereafter the extracted traffic features from floating GPS data can help to estimate the volume of road segments without fixed detectors.

This study presents a customized transfer learning approach, namely, the transfer learning-based least square support vector regression (TL-LSSVR) model, to cope with the dilemma that fixed detectors are usually too sparse to cover the network-wide traffic volume. To find segments similar to those installed with LPR detectors, the Jensen–Shannon divergence (JSD) method is proposed to identify the distribution of taxi speed based on the fusion of LPR data and taxi GPS data. The contributions of this study are twofold.

1. We propose a TL-LSSVR model to estimate the traffic volume on segments without LPR detectors. The accuracy of the proposed method was verified based on real-world data, and we compared it with other state-of-the-art estimation methods to demonstrate its superior performance.

2. To overcome the issue of insufficient sample size for training and variability by sparse LPR detector deployment, we propose an innovative method based on the JSD to determine similarities between segments with and without LPR detectors. The road segments installed with LPR detectors are divided into similar segment sets and auxiliary segment sets, where different penalty coefficients are set for these two types of segment sets.

The remainder of this paper is organized as follows. Section 2 provides an overview of relevant studies. Section 3 describes the research problem and introduces the proposed framework. Section 4 introduces the methodology, and Section 5 provides a case study to validate the performance of the proposed approach. Finally, conclusions are summarized in Section 6.

2. Literature Review

Studies on urban road network traffic volume/flow estimation are part of the problem of traffic dynamic information modeling in intelligent transportation systems, which has gained much popularity. Various data-driven approaches for predicting future traffic information have been developed. Statistical models (e.g., such as the autoregressive integrated moving average (ARIMA) and its variants (Ghosh et al., 2009)), machine learning methods (e.g., neural networks (Van Lint et al., 2005), support vector machines (SVMs) (Zhang et al., 2009), random forest (Zhang and Haghani, 2015)), and deep learning frameworks (Ke et al., 2017; Lv et al., 2015) can all be used for predicting short-term traffic flows with description and tracking of information. However, these studies attempt to extract the evolution patterns of traffic flow using historical observations, which is completely different from the traffic volume estimation problem.

Previous studies on network-wide traffic volume estimation can be divided into three categories based on different detector types: fixed detector-based methods, probe detector-based methods, and detector fusion-based methods (Aslam et al., 2012; Li et al., 2016; Liu et al., 2019b). In the fixed detector-based category, traffic volume can be aggregated from every vehicle detected by fixed detectors (i.e., loop detectors or surveillance cameras). Existing studies in this area have mainly focused on *missing volume imputation* from some missing spatial-temporal points/segments with a fixed detector installed, and *undetected volume estimation* from some segments without fixed detectors. Missing volume imputation was implemented using a multiple linear regression model in which the key information from correlated segments was extracted to interpolate traffic dynamics on undetected segments (Shan et al., 2013). Based on this study, other matrix-based approaches, such as the compressive sensing method (Liu et al., 2016) and principal component analysis (Qu et al., 2009), were used to retrieve the traffic flow features from the detected information for data

imputation. Similar to the factorization of principal components, tensor decomposition was further applied for missing volume imputation by the multidimensional inherent correlation of spatiotemporal traffic data (Tan et al., 2013). Chen et al. (2018) and Tang et al. (2018) extended this study and exploited the Bayesian probabilistic tensor factorization approach, which considers the effect of multivariate distribution of data.

However, these methods cannot estimate the traffic volume on road segments without fixed detectors. In other words, when the missing rate becomes high or the missing length becomes long, it is very challenging to estimate traffic volume data with satisfactory results (Laña et al., 2018). Inspired by kriging as a geostatistical interpolation model in the mining industry, several studies have attempted to apply this approach to estimate the traffic volume on the entire road segment. Yang et al. (2018) utilized the spatial location relationship of traffic-detected data to construct a variance function model and estimate the traffic volume on undetected segments. Bae et al. (2018) further developed co-kriging methods and extra utilized the characteristics of temporal dependencies that outperform only spatial dependencies. However, such parametric assumption models cannot explicitly model the inherent relationships between geo-graph attributes and various types of spatiotemporal correlations in a complex traffic environment, such as the influence of road grade level, surrounding infrastructure, and point of interest (POI).

With the development of machine learning in intelligent transportation systems, Shan et al. (2016) developed an online neural network to model insufficient detected traffic information. To avoid the weak correlation between detected and non-detected segments, Liu et al. (2019c) further exploited a dynamic artificial neural network framework with a graph-parallel processing approach, which could dynamically select similar road segments. To express multivariate normal distributions, Hara et al. (2018) proposed a mixture Gaussian graphical model to describe more complicated distributions and provide an implicit interpretation of

shifts in the traffic state. Inspired by the multi-view deep learning method, Li et al. (2020) illustrated a hybrid spatiotemporal imputation model that integrated time-series properties of the temporal part with the residual component of the spatial part for missing traffic imputation. Zhang et al. (2020a) applied a generative adversarial network (GAN) model to estimate the missing traffic information. Pan et al. (2019) proposed a deep meta graph attention network structure to address the challenges of complex spatiotemporal correlations using the encoder and decoder structures. However, traffic flow is constantly changing with time series, and these approaches are mainly modeled from stationary traffic environments. It is difficult to use data correlations from a single fixed detector to derive dynamic traffic information that fluctuates over time.

Because probe detector data can serve as low-cost and reliable data resources, a variety of studies have focused on traffic information estimation using probe detector data (Bwambale et al., 2019; Wang et al., 2018). Some classical probe detector-based traffic volume estimation methods have been derived from classical traffic flow theory. With the help of the relationship between the three parameters (i.e., traffic speed, density, and volume) in the traffic flow fundamental diagram (FD), traffic volume can be inferred by traffic speed or traffic density (Cheng et al, 2021) since 1930s, which can be easily detected by a probe detector, and taxi GPS data have been proven to be effective in this field. However, a large amount of traffic data is needed to calibrate FD models for each road segment (Kerner et al., 1998). Seo et al. (2015) proposed a new probe vehicle-based estimation method to obtain volume-related variables by assuming that a probe vehicle can measure the space distance to its leading vehicles. Affected by uncertain traffic demand and market penetration rate in probe vehicle operation, there is no guarantee that there will be enough vehicles to pass by at any moment. To mitigate the limitations of probe vehicle data, this study derived the minimum sample size of probe vehicles.

Based on the above discussion, the fixed detector devices can collect accurate traffic volume data, but they have low spatial coverage. By contrast, probe detector data have wide spatiotemporal coverage in urban road networks, but are marked by relatively low representability (or resolution) of all vehicles on the road. Thus, data fusion approaches that combine fixed detector data with probe detector data can be a valuable way to avoid the disadvantages and enhance the advantages of these two types of data. Recent studies have applied multi-source data fusion to improve the performance of traffic volume estimation and city sustainability (Zhou et al., 2021). Aslam et al. (2012) fused probe taxi data and fixed-ground loop data using a linear regression model for traffic volume estimation. Zhan et al. (2016) extended this study and proposed a hybrid volume re-interpretation framework to estimate city-wide traffic volume, which integrated Bayesian network structures with traffic flow theory. Similar studies were also conducted by Meng et al. (2017), and similar spatiotemporal features between non-detected segments and detected segments were extracted to build a graph-based semi-supervised learning method for city-wide traffic volume estimation. In addition, by applying a similar semi-supervised learning model, it is also possible to estimate the traffic volume on road segments that share a common path with detected segments by inferring the path of the probe detector data. Wang et al. (2021) extended this study and extracted taxi paths between LPR devices to identify hot paths and the optimal set of scaling factors. Zhang et al. (2020b) built a novel methodology to estimate network-wide traffic flow, which incorporates fixed detector records and crowdsourcing floating-car data into a geometric matrix completion model incorporated by the spatial affinity graph and temporal continuity constraint. With the fusion of mobile signaling data and loop detector data, Cui et al. (2017) proposed a Polaris compressive sensing model for highway traffic volume estimation. With similar data input, Liu et al. (2019b) exploited the urban network volume estimation model by

scanning the features of cell phone data and LPR data, which integrated a multi-grained ensemble learning model with a two-stage zero-shot learner approach.

These pioneering studies prove that integrating multiple data sources has significant advantages for accurate network-wide traffic volume estimation. However, only a few studies have considered the deployment status of fixed detectors in urban road networks, particularly when the detector deployment is sparse. Input training in the traffic estimation model is lacking when fixed detectors are deployed only in some road segments, which in turn affects the traffic estimation accuracy. This study aims to fill this gap and focus on network-wide traffic volume estimation, where fixed detectors are sparsely deployed.

3. Problem Statement

3.1 Preliminaries

We first illustrate the definitions and notations of the road network topology and traffic parameters, and then provide an explicit description of the problem formulation.

(1) Road network topology

As shown in Figure 1, the whole road segment set is denoted as $\mathcal{D}$, the road segment set installed with LPR detectors is denoted as $\mathcal{S} \subseteq \mathcal{D}$, and the road segment set that is not installed with LPR detector is represented as $\mathcal{S}^* \subseteq \mathcal{D}$.

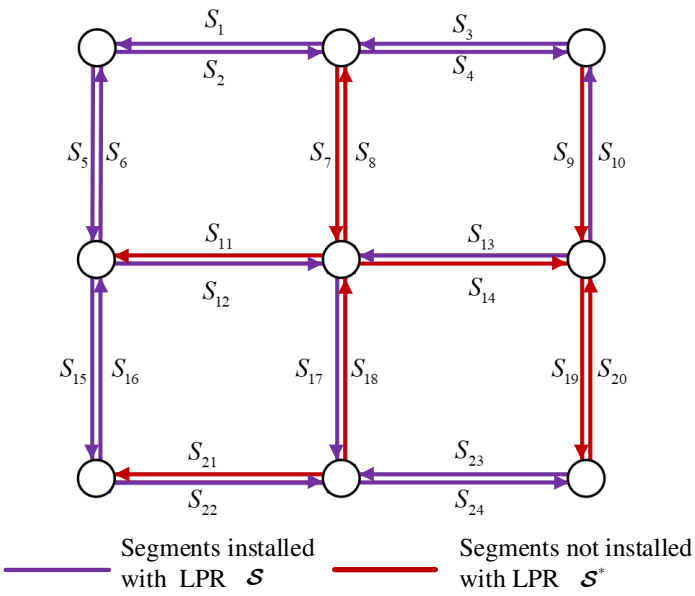


Figure 1. Topology of traffic networks

(2) Road attributes and traffic factors

Because the taxi GPS data is a type of crowdsourcing floating-vehicle trajectory data, we assume that this data can be viewed as full spatial and temporal coverage, and the taxi speed information for all road segments is available. The dataset on road segments installed with LPR detectors is denoted as $H^t = \left\{(\boldsymbol{x}_1^t, q_1^t), (\boldsymbol{x}_2^t, q_2^t), \cdots, (\boldsymbol{x}_m^t, q_m^t)\right\}$, where the dataset on road segments that are not installed with LPR detectors is denoted as $L^t = \left\{\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \cdots, \boldsymbol{x}_j^t\right\}$ ($m = 1, \cdots, M$ is the number of road segments installed with LPR detectors, and $j = 1, \cdots, J$ denotes the number of road segments without the LPR detector installed). $\boldsymbol{x}_i^t$ denotes the extracted taxi feature dataset at the t^{th} time slot (e.g., 15 min) of road segment S_i ($\forall S_i \subseteq \mathcal{D}$). It is a type of spatiotemporal taxi speed information matrix, which divides the taxi speed v_{itk} into several intervals, where $v_{itk} \in \mathbb{R}^{M \times T \times K}$ denotes the spatiotemporal taxi speed records for driving taxi $D_k (k = 1, ..., K)$ in road segment $S_i (i = 1, \cdots, M)$ and $t^{th} (t = 1, ..., T)$ time slots, respectively. q_i^t is the LPR-detected volume dataset at the t_j time slot of the road segment S_i, which is the total number of vehicles detected by the LPR detector on the road segment during

this time interval, which can be regarded as the actual traffic volume. Table 1 lists the sets, parameters, variables, and functions used in this study.

Table 1. List of notations.

Notation	Description
Sets	
$\mathcal{S}$	Road segment set with LPR detector installed
$\mathcal{S}^*$	Road segment set without LPR detector installed
H^t	Dataset on LPR detected segments
L^t	Dataset on LPR non-detected segments
H_p^t	Similar segment dataset
H_a^t	Auxiliary segment dataset
v_{itk}	Taxi average speed in a time slot t^{th} on a segment S_i from a taxi driver D_k
q_i^t	LPR-detected volume in t^{th} on segment S_i ($i=1,\cdots,M$, $t=1,\cdots,T$)
$\boldsymbol{x}_i^t$	Extracted taxi feature vector $x_i^t=\left[x_{i1}^t,x_{i2}^t,\cdots,x_{in}^t\right], n=1,\cdots,N$
Parameters	
PCC_i	Pearson correlation coefficient between taxi volume and LPR-detected volume on segment S_i
p_i	Probability distribution of taxi speed on LPR detected segment
p_j	Probability distribution of taxi speed on LPR non-detected segment
b	Constant value
γ_p	Regularization penalty coefficient for similar segment dataset
γ_a	Regularization penalty coefficient for auxiliary segment dataset
$\boldsymbol{a}$	Lagrange multiplier vector $\boldsymbol{a}=\left[a_1,a_2,\cdots,a_M\right]^T$
λ	RBF kernel coefficient
Variables	
$\boldsymbol{e}$	Error value
$\boldsymbol{\omega}$	Normal vector
Functions	
$\varphi(\boldsymbol{x})$	The feature-space transformation function

$K(\boldsymbol{x}_i, \boldsymbol{x}_j)$	Kernel function

3.2 Problem formulation

Based on model transfer in transfer learning theory (Pan and Yang, 2010), the goal of this study is to establish a traffic volume estimation model by using a dataset from road segments installed with an LPR detector and applying it to estimate the traffic volume in road segments without LPR detectors installed. The dataset on road segments installed with the LPR detector was set as $H^t = \left\{(\boldsymbol{x}_1^t, q_1^t), (\boldsymbol{x}_2^t, q_2^t), \cdots, (\boldsymbol{x}_m^t, q_m^t)\right\}$. The dataset on road segments without an LPR detector installed was set as $L^t = \left\{\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \cdots, \boldsymbol{x}_g^t\right\}$. The model aims to optimize the estimate function f_t using the knowledge from the estimate function f_s. Here, we want to find an optimization function $f_s\left(\boldsymbol{x}_i\right) = q \rightarrow f_t\left(\boldsymbol{x}_g\right) = q^*$, which is built on the dataset $\mathcal{D}_s$ to find the non-detected segment volume q^* for any new incoming detected sample $\boldsymbol{x}_g$ on the dataset L^t, where it can be represented as

$$\min\ J(h) = \sum_{i}^{M} L(f_t(\boldsymbol{x}_i), q_i) \tag{1}$$

However, in some cases, the deployment of LPR detectors is too sparse to find sufficient similar segments for modeling. To fill this gap, we divided the road segment set installed with the LPR detector into similar road segment set H_p and auxiliary road segment set H_a, which is classified based on the similarity to the target LPR non-detected road segments. Thereafter, we modified the objective function defined in equation (1) by adding the loss function of the auxiliary sets to the original objective function, which can be expressed as:

$$\min\ J'(h) = \gamma_p \sum_{i}^{H_p} L(f_t(x_i^p), y_i^p) + \gamma_a \sum_{H_p+i}^{H_a} L(f_t(x_i^a), y_i^a) \tag{2}$$

where H_p and H_a are the number of samples in the similar and auxiliary datasets, respectively. γ_p and γ_a are regularization penalty coefficients to balance losses.

3.3. Study framework

The study framework, which can be divided into the following five steps, is illustrated in Figure 2.

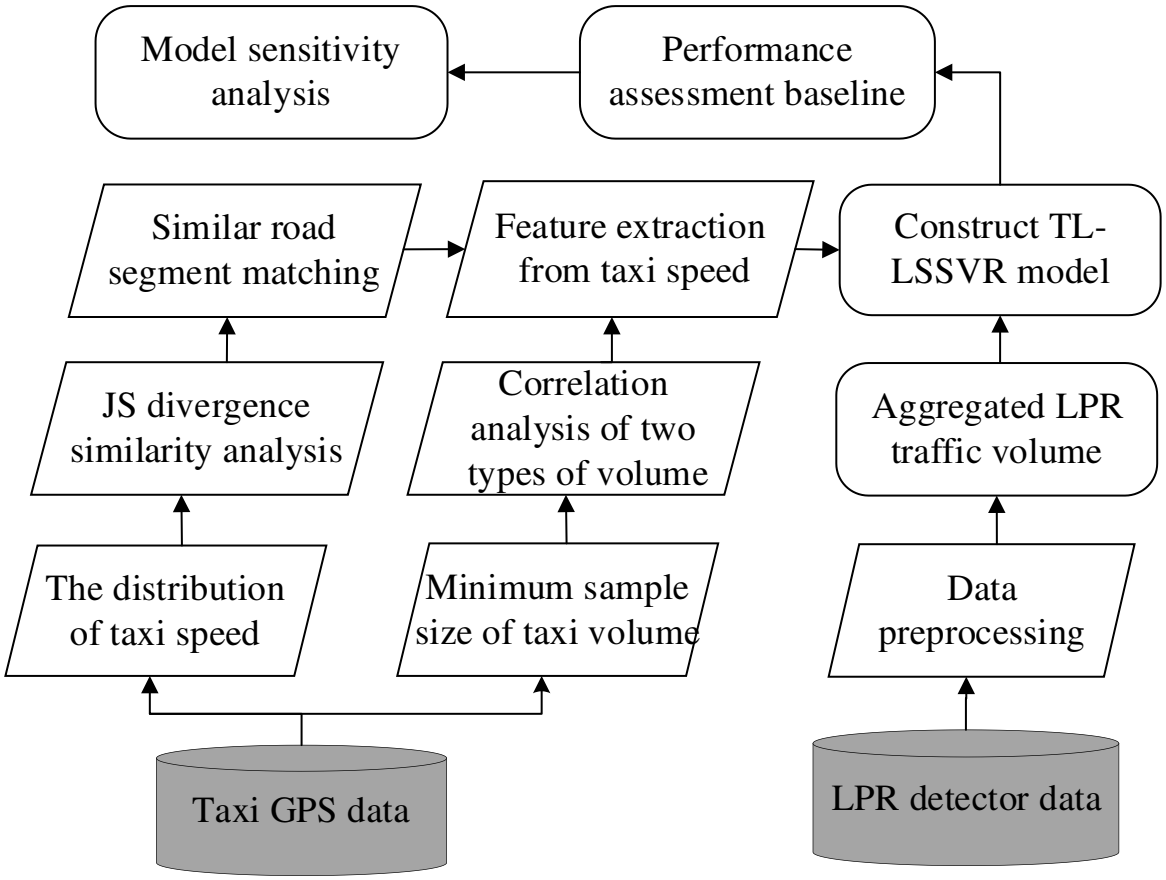


Figure 2. Flowchart of the study framework.

Step 1: Taxi GPS data and LPR data are pre-processed, and traffic speed and volume are extracted from the two types of data.

Step 2: The minimum sample size of taxis in each road segment and the correlation of traffic volume between taxi GPS data and LPR data in LPR-detected segments are calculated.

Step 3: The similarity between the unknown traffic volume segment and known traffic volume segment is analyzed, and these similar segments are divided into similar segment sets and auxiliary segment sets, which are all similar to the unknown traffic volume segment studied.

Step 4: The TL-LSSVR model is established to estimate the traffic volume on segments with no detectors installed in the urban network.

Step 5: Evaluation of the performance of traffic volume estimation using different scenarios and baselines.

4. Mathematical Models

In this section, we introduce the proposed network-wide traffic volume estimation architecture in the scenario of sparse deployment of LPR detectors and the TL-LSSVR model.

4.1 Data preparation

Taxi data are affected by uncertain travel demand (e.g., the points of interest, road grades) in the study area, which has spatial heterogeneity between the actual volume and taxi volume (Qian and Ukkusuri, 2015). To better utilize the taxi data and LPR data in traffic volume estimation, we first need to evaluate whether the data meets the criteria by calculating the minimum sample size of taxi volume, and then analyze the correlation between the two types of data.

When travel demand is low in a specific spatiotemporal range, a null value may appear that there is no taxi passing by. The taxi dynamic information is too low to reflect the actual traffic state at this moment, and cannot be regarded as input features for volume estimation modeling. Herein, we calculate the minimum sample size of the taxi to filter satisfied road segments based on whether the speed distribution of taxis corresponds to the overall speed distribution characteristics (Coburn, 2004). The taxi speed can be calculated using GPS positioning information, and the details can be found in Xing et al. (2019). The speed information of other vehicles can be recorded directly by the installed LPR detector. According to the t-test of a normal distribution, the distribution of all driving vehicles is approximately normal distribution $x \sim N(\mu, \sigma^2)$, and the values of μ and σ^2 are unknown (Coburn, 2004). At the a-significance level, the sample of mean speed μ meets hypothesis $H_0 : \mu = \mu_0$, and the form of the distribution of taxi speed can be formulated as follows:

$$P\left\{\left\{\left|\frac{\bar{x}-\mu_0}{S_n/\sqrt{n}}\right| \geq t_{a/2}(n-1)\right\}\right\} = a \tag{3}$$

where n denotes the minimum sample size of the taxi volume in each study segment. It can be queried by the statistical table of the t distribution within the confidence level $1-a$.

Furthermore, we also need to analyze the correlation between taxi volume and LPR-detected volume, which is affected by land use and the travel habits of drivers. Some road segments that do not have a significant correlation for modeling should be filtered out. We measure the correlation using the Pearson correlation coefficient (PCC) method, which is widely used to measure the correlation between two types of data and is defined in equation (4):

$$PCC_i = \frac{\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2}\sqrt{\sum_{t=1}^{T}(y_t-\bar{y})^2}} \tag{4}$$

where x_t and y_t represent the taxi volume and LPR-detected volume in t^{th} time slot, respectively, $\bar{x}$ and $\bar{y}$ denote the mean value of x_t and y_t, respectively. The PCC value ranges between +1 and -1, where a positive PCC value indicates a positive correlation, and vice versa. In practical applications, the PPC value indicates the correlation level (Holgado–Tello et al., 2010). For different ranges of PCC, it can be inferred that no correlation exists for the value of $|PCC| \leq 0.3$, a weak correlation for $0.3 < |PCC| \leq 0.5$, a significant correlation for $0.5 < |PCC| \leq 0.8$, or very strong correlation if $|PCC| > 0.8$.

4.2 Segment similarity analysis

Before modeling the traffic volume estimation problem on the segments without an LPR detector installed, we need to choose parts of LPR detected segments that are similar to LPR non-detected segments, and divide them into similar segment datasets and auxiliary segment datasets. Because actual traffic volume data do not have complete coverage in the study

network, by contrast, as a kind of probe data in the spatiotemporal range, taxi GPS data have unique coverage advantages. Furthermore, the taxi speed accurately reflects the actual traffic state. Herein, we select the taxi speed distribution as the criterion to measure the similarity between LPR detected segments and LPR non-detected segments. Similar segments and auxiliary segments are selected using the JSD method. This approach is based on measuring whether a probability distribution differs from another probability distribution. We set $P_i(i=1,...,M)$ and $P_j(j=1,...,J)$ as the probability distributions of taxi average speed from the LPR detected segment and LPR non-detected segment, respectively. The value of the JSD is defined as

$$Dist_{ij} = \frac{1}{2} d\left(P_i \middle\| \frac{P_i + P_j}{2} \right) + \frac{1}{2} d\left(P_j \middle\| \frac{P_i + P_j}{2} \right) \tag{5}$$

where $d(P_i \| P_j)$ is the Kullback–Leibler divergence (KLD) value, that is,

$$d(P_i \| P_j) = \int_0^{V\max} P_i(v) \log \frac{P_i(v)}{P_j(v)} dv \tag{6}$$

where $\sum_{i=1}^{L} P_i = \sum_{i=1}^{M} P_j = 1$. Compared with KLD, JSD considers the situation in which two types of distributions do not coincide. A lower JSD value represents a more similar taxi speed distribution. When it reaches zero, the two distributions are the same, and vice versa.

4.3 Traffic volume estimation based on the TL-LSSVR method

Transfer learning is a special machine learning technique that aims to solve the basic problem of unlabeled and insufficient data in the target domain by utilizing the available knowledge from the source domain (Weiss et al., 2016). This method could consider whether some part of the knowledge can be transferred across or not. Based on different leveraged knowledge on transfer learning, this technique can be categorized as instance transfer, feature transfer, or model transfer (Pan and Yang, 2010). Our proposed transfer learning method falls

into the category of model transfer and is inspired by a simplification of the tailor transfer learning model from Wang (2020), which was early used in handling missing data imputation in the field of gear fault diagnosis.

SVMs were proposed by Vapnik (1995) and are powerful tools for solving problems in nonlinear classification. It was originally illustrated in the context of the statistical learning theory and structural risk minimization. The SVM is to find a separating hyperplane that has the strongest generalization ability and the most robust classification results, which can be used for classification and regression. In this study, the basic model is implemented using the SVR approach. Furthermore, to overcome the problem of slow solution speed caused by the number of variable dimensions equal to the number of training samples in SVR, we apply an LSSVR model that simplifies the SVM solution problem to replace it (Suykens and Vandewalle, 1999). In this way, it is possible to translate the complex nonlinear quadratic programming for x and y into a simple linear relationship $\varphi(x)$ and y. Hence, the proposed traffic volume estimation model, which is based on the least squares SVM, is an easier solution method than the original SVR method. Let $\varphi(\boldsymbol{x})$ denote the mapped feature vector of $\boldsymbol{x}$, then the separating hyperplane model in the feature space can be expressed as

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^T \varphi(\boldsymbol{x}) + b + \boldsymbol{e} \tag{7}$$

where $\boldsymbol{\omega}$ is the model parameter, b is constant, and $\boldsymbol{e}$ is the estimated error. Specifically, $\boldsymbol{x}$ is the taxi traffic feature vector extracted from the taxi GPS data. According to the structural risk minimization principle, suppose we wish to learn a minimized risk bound that $f(\boldsymbol{x})$ and $\boldsymbol{y}$ are as close as possible to the establishment of the volume estimation model, which is equivalent to finding the parameters $\boldsymbol{\omega}$ and $\boldsymbol{e}$ that minimize $J(\boldsymbol{\omega}, \boldsymbol{e})$ (Chang and Lin, 2011), which is as follows:

$$\min_{\boldsymbol{\omega},\boldsymbol{e}} J(\boldsymbol{\omega},\boldsymbol{e}) = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + \frac{\gamma}{2}\sum_{i=1}^{T} e_i^2, \qquad \gamma > 0$$
$$s.t.\ \ y_i = \boldsymbol{\omega}^T\varphi(\boldsymbol{x}_i) + b + e_i, \qquad i = 1,2,...,T \tag{8}$$

where γ is the regularization penalty constant and e_i is the insensitive loss deviation. In the latter term of equation (8), if the estimated volume falls within the buffer region, $\left|f(\boldsymbol{x}) - y_i\right| \le e_i$ is considered as the correct result.

To increase the sample size in the model training process for a road network with the sparse deployment of the LPR detector and better utilize the different types of datasets, a transfer learning-based traffic volume estimation approach is proposed. In this approach, the auxiliary dataset is used to improve the estimation accuracy by adding the training dataset and simultaneously weakening the effect of redundant data, especially when similar road segments are insufficient to establish a traffic estimation model. We separate the regularization penalty terms into regularized penalty terms for similar road segment datasets and regularized penalty terms for nearly similar road segments based on the level of similarity between road segments. Thus, the optimization problem (8) can be transformed into the TL-LSSVR model (9) as follows:

$$\min_{\boldsymbol{\omega},\boldsymbol{e}} J(\boldsymbol{\omega},\boldsymbol{e}) = \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + \frac{\gamma_p}{2}\sum_{i=1}^{T_p} e_i^2 + \frac{\gamma_a}{2}\sum_{i=T_p+1}^{T_p+T_a} e_i^2$$
$$s.t.\ \begin{cases} y_i = \boldsymbol{\omega}^T\varphi(x_i) + b, i = 1,2,...,T_p \\ y_i = \boldsymbol{\omega}^T\varphi(x_i) + b_i, i = T_p+1,...,T_p+T_a \end{cases} \tag{9}$$

where γ_p and γ_a are the regularization penalty coefficients from the similar and auxiliary datasets, respectively. $\gamma = diag\left\{\underbrace{\gamma_p,...,\gamma_p}_{T_p}, \underbrace{\gamma_a,...,\gamma_a}_{T_a}\right\}$ represents the penalty weight of the insensitive error set when the sample is mapped to a certain feature space. In the setting of γ, the larger the value of γ, the greater the impact of the insensitive error in the training data.

When the value of γ tends to infinity, the estimation model strictly defines that samples with estimation errors are not allowed, and this would cause overfitting. By contrast, when the value of γ tends to 0, the estimation model will not obtain a meaningful solution, and the algorithm will not converge, which thus appears under-fitted. Because the auxiliary datasets and similar datasets do not have the same impact on the model estimator, the penalty factor terms γ_p and γ_a are set uniquely for the two types of datasets in the estimation model, respectively. Thereafter, the solution steps of optimization equation (9) are as follows:

Step 1: This constraint to minimize the function can be solved by applying Lagrange multipliers to (9). Specifically, introducing a Lagrange multiplier $a_i \geq 0$ to each constraint in equation (9) gives the Lagrange function:

$$L(\boldsymbol{\omega}, b, \boldsymbol{e}, \boldsymbol{a}) = J(\boldsymbol{\omega}, \boldsymbol{e}) - \sum_{i=1}^{T_p} a_i \left\{ \boldsymbol{\omega}^T \varphi(x_i) - b - y_i \right\} - \sum_{i=T_p+1}^{T_p+T_a} a_i \left\{ \boldsymbol{\omega}^T \varphi(\boldsymbol{x}_i) - b - y_i \right\} \tag{10}$$

where $\boldsymbol{a} = [a_1, a_2, \cdots, a_{T_p+T_a}]$, $i = 1, 2, \cdots, (T_p + T_a)$.

Step 2: To find the optimal solution, we exploit the well-known Karush–Kuhn–Tucker condition (Boyd and Vandenberghe, 2004), and set the partial derivatives of $L(\omega, b, e, a)$ with respect to $\boldsymbol{\omega}$, b, $\boldsymbol{e}$, and $\boldsymbol{a}$ to 0, the conditions for optimality

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{\omega}} = 0 \rightarrow \sum_{i=1}^{M} a_i \varphi(x_i) = \boldsymbol{\omega} \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{M} a_i = 0 \\ \dfrac{\partial L}{\partial \boldsymbol{e}} = 0 \rightarrow a_i = \begin{cases} \gamma_p e_i, & i = 1, 2, ..., T_p \\ \gamma_a e_i, & i = T_p + 1, T_p + 2, ..., T_p + T_a \end{cases} \\ \dfrac{\partial L}{\partial \boldsymbol{a}} = 0 \rightarrow \boldsymbol{\omega}^T \varphi(x_i) + b + e_i - y_i = 0 \end{cases} \tag{11}$$

Step 3: Substituting equation (11) into equation (10) eliminates the variables $\boldsymbol{\omega}$ and $\boldsymbol{e}$. Thereafter, the solution to equation (10) can be expressed in matrix form as follows:

$$\begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \Omega + I/\gamma \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{12}$$

where $Y = [y_1, y_2, \cdots, y_M]^T$, $1 = [1,1,\cdots,1]_{M\times 1}$; Ω is an $M \times M$ rank symmetric matrix, and $\Omega_{ij} = \varphi(\boldsymbol{x}_i)^T \varphi(\boldsymbol{x}_j) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1,2,\cdots,M$ ($M = T_p + T_a$), $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a kernel function, I is a unit matrix, γ is the regularization penalty factor matrix, and $a = [a_1, a_2, \cdots, a_M]^T$.

Step 4: Solve the kernel functions and regularization coefficients γ. In optimization problem solving, γ_p and γ_a are hyperparameters, which are pre-specified values. Because of the small dynamic variability between the traffic volume data in each segment and the efficiency of the solution, the values of γ_i in the regularization penalty term of γ_p or γ_a are set to the same in different mapping spaces in this study.

Among all types of kernel functions, the radial basis kernel function is used as the kernel function of the regression model in this study because of its advantages such as fewer parameters and better predictive performance, which is denoted as $K(x_i, x_j) = e^{-\lambda\|x_i - x_j\|}$. We set the radial basis function (RBF) as the kernel function, which is $K(x_i, x_j) = e^{-\lambda\|x_i - x_j\|}$, and λ is the RBF kernel coefficient, which defines the influence size of a single training sample and can be considered as the reciprocal of the influence radius when the training sample selected by the model is mapped to the surface space. That is, the smaller the value, the smaller the impact, and the larger the value, the larger the impact. In this study, the kernel parameter is also used as a hyperparameter, and the value of the kernel parameter λ plays an opposite role to that of γ, so the values of this parameter need to be solved together.

Step 5: Support $Q = \Omega + \gamma^{-1} I$ and $H = Q^{-1}$. By solving for equation (12), we obtain:

$$b = \frac{1^T H x}{1^T H 1} \tag{13}$$

$$a = H(x - 1 \times b) \tag{14}$$

Thus, the volume estimation model built using a similar segment dataset and auxiliary segment dataset is designed as follows:

$$f(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{i=1}^{T_p+T_a} a_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b\right) \tag{15}$$

5. Numerical Examples

In this section, the proposed method is quantitatively evaluated using real-world datasets. The validity of TL-LSSVR and a detailed discussion are presented in this section.

5.1 Studied area and data description

To demonstrate the effectiveness of the proposed model in the scenario of sparse fixed detector deployment, a real road network that includes an approximately 5 km radius around Nanjing South Railway Station was selected as our study area. As the border area between the city center and the suburbs in Nanjing, China, we can note that the southern area (suburbs) of Nanjing South Station, as a newly sprawl area, has a significantly lower density of LPR detector deployment than the northern area (city center). Herein, LPR detectors were mainly deployed on arterial and sub-arterial roads, and some LPR non-detected segments were not interconnected around other LPR detected segments. Furthermore, this particular scenario is common in some urban sprawling areas. Our area included 154 one-directional road segments (as shown in Figure 3), and 29 road segments were installed with LPR detectors (as shown in Figure 4). The overall coverage rate of the LPR detector in this study network was approximately 19%, and the number of detected segments was much smaller than that of undetected segments.

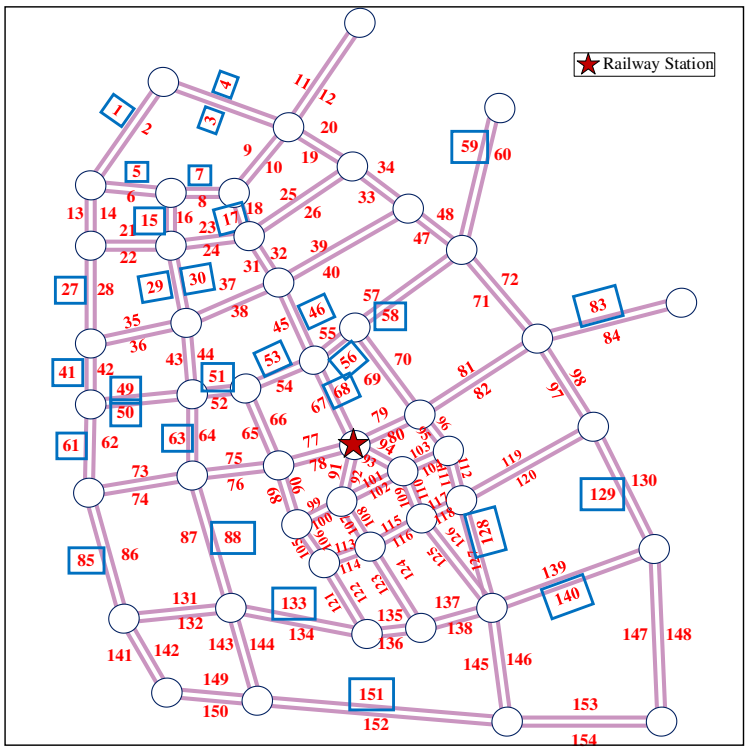


Figure 3. Road segments (numbered from 1 to 154) and the numbers of the LPR detectors (marked by the blue box) in the study area.

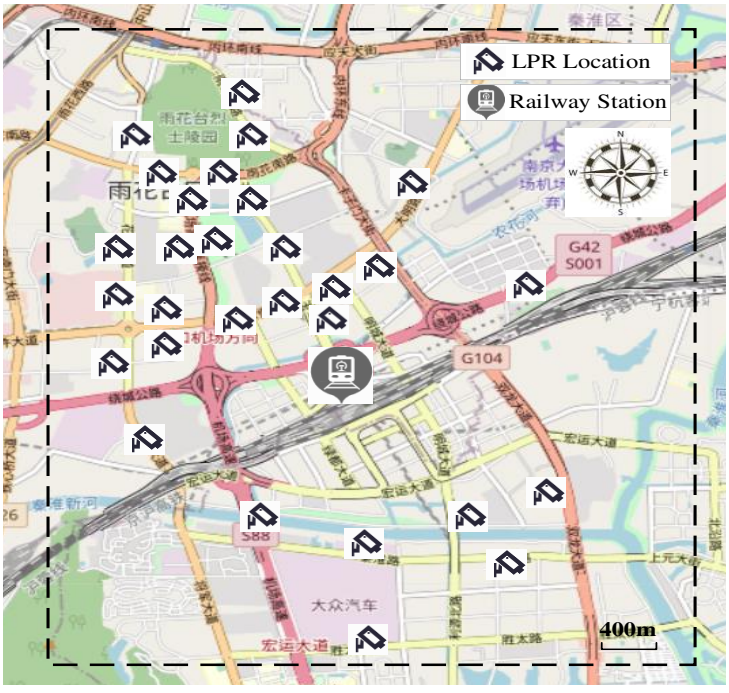


Figure 4. Study area with locations of LPR detectors marked.

Input data for this study include taxi GPS data and LPR detector data. Both sets of data were provided by the City Information Council of Nanjing for the same spatiotemporal

coverage. Taxi GPS data, as a type of floating-vehicle data, are produced by a global position system device that is attached to each driving taxi in the city. Geographic information can be generated at fixed intervals. In this study, the update interval of GPS transmission was 30 s, and the average travel distance for one taxi trip was almost 3.2 km. The taxi GPS data label includes the taxi ID number, pick-up time, drop-off time, longitude, latitude, and positioning record timestamp. Table 2 illustrates the details of the two data types used in this study. Due to space limitations, we do not provide details on the data pre-processing technique, and interested readers can refer to Xing et al.(2019).

Table 2. Data label for two types of detectors.

Data type	Data label	Data size	Time slots
Taxi GPS data	Taxi license ID, pick up time, drop-off time, longitude, latitude,	8 million data	26/09/2016~31/10/2016
LPR data	Encrypted vehicle plate ID, timestamp, vehicle type, LPR	0.24 million	01/10/2016~31/10/2016

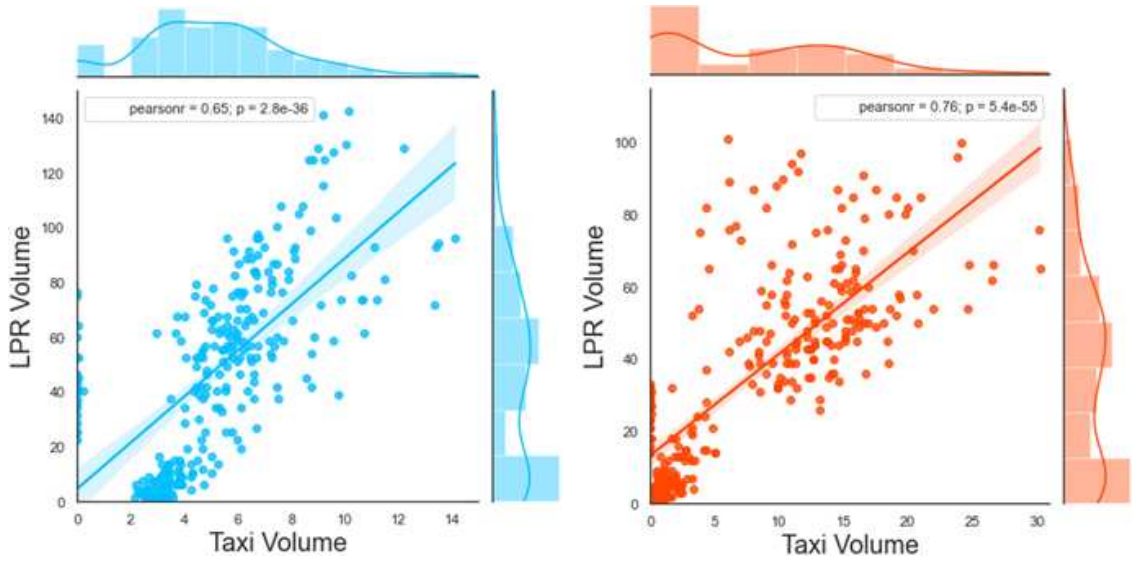
LPR detectors, as fixed detectors, are mainly located on key road segments or intersections. The location of the LPR detector in the study area is shown in Figure 4. This detector system, which uses cameras to capture vehicle data in real-time with high precision, is located along road segments and can recognize the license plates of passing vehicles using image processing techniques. In China, the recognition rate of LPR systems is as high as 95% during the daytime and no less than 90% at night. Furthermore, it can also identify vehicle types (i.e., large-size car, medium-size car, and small-size car) by license plate color and convert the aggregated volume into the form of passenger car units (PCUs). Hence, the LPR aggregated volume can be used as the actual traffic volume. The LPR data label mainly includes the record device ID, record time, license plate number, and instantaneous vehicle speed.

5.2 Experiments

5.2.1 Data correlation analysis

To analyze the spatial and temporal correlations, a day was divided into two time periods: peak hours (from 07:00 am to 10:00 am and from 4:00 pm to 8:00 pm) and non-peak hours (from 8:00 pm to 07:00 am and from 10:00 pm to 4:00 am). Additionally, considering the variability of travel demand between weekdays and weekends, we further divide a week into weekdays and weekends. Herein, four time slots are generated in this study.

In our study area, we set the minimum sample size of the taxi volume to 10 pcu/15 min. Furthermore, 25 segments from 29 LPR detected segments that met the requirements of the minimum sample size of taxi volume were viewed as the eligible segments for modeling. To illustrate the variation in taxi data and LPR data, the PCC values between taxi volume and LPR-detected volume in LPR detected segments were calculated during peak hours and non-peak hours in 15 min intervals. As shown in Figure 5, we find that two types of data show a significant correlation, and the correlation during non-peak hours is larger than that during peak hours.



(a) PCC during peak hours (b) PCC during non-peak hours

Figure 5. Correlation analysis between taxi volume and LPR-detected volume.

5.2.2 Similarity segment matching

As in the aforementioned method, we analyzed the similarity between each LPR detected segment using the JSD method. For the validation of our experiment (the following baseline models are included), the 25% road segments installed with LPR detectors were randomly regarded as undetected segments, and the traffic volume on the remaining road segments with LPR detectors were viewed as validation. The results of the JSD values are shown in Figure 6. From this matrix of JSD values, we obtained the corresponding similar segments and auxiliary segments by sorting the order of the JSD values.

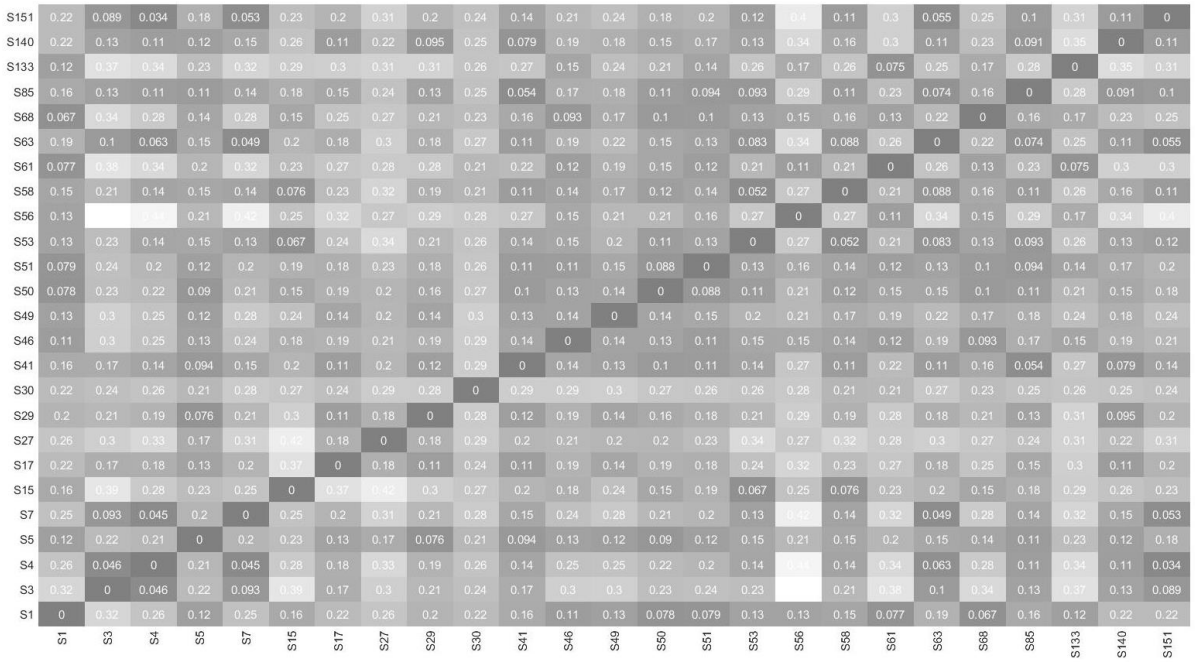
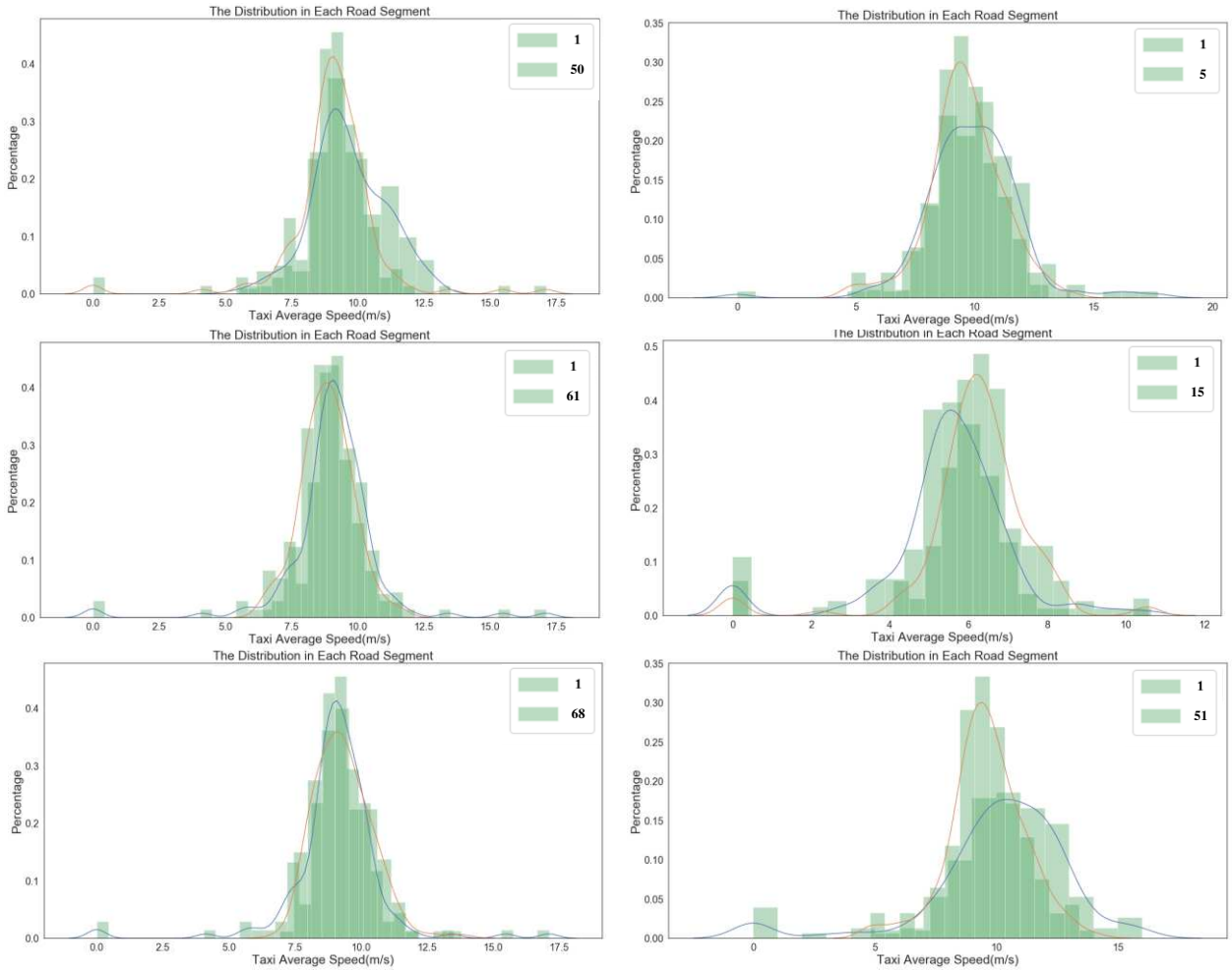


Figure 6. Value of JSD between each road segment.

The correlations of traffic flows between different segments are not negligible (Liu et al., 2016; Meng et al., 2017). In general, the traffic volumes of adjacent road segments are typically similar because vehicles frequently travel between them. The traffic flow from one road segment is related to many other segments. For example, the traffic volumes of two cascaded road segments are usually directly correlated (Liu and Smith, 2015; Liu et al., 2006). The traffic volume from one segment may not only relate to the traffic coming from its cascaded road segment, but also to the traffic redirected from intersecting road segments in an indirect way.



Similar segment (SS) Auxiliary segment (AS)

Figure 7. Comparison of taxi speed distribution in segment S_1.

Furthermore, limited by the variety of road network sizes, differentiation of internal structure, and coverage rates of fixed detectors, the criteria of similarity of road segments are different in the similarity analysis. To better evaluate the performance of our proposed model, four types of road segments were chosen as our object segments for the selection of similar segments and auxiliary segments. In this study, we chose four representative road segments as the object segment for comparison. They are as follows: (1) arterial road segment with LPR detected segment cascades (i.e., segment S_1); (2) arterial road segment without LPR section cascades (i.e., segment S_{151}); (3) branch road segment with LPR detected segment cascades (i.e., segment S_{51}); and (4) branch road segment without LPR detected segment cascades (i.e.,

segment S_{133}). In our experiment, the number of similar segments and auxiliary segments was set to three. By comparing the taxi speed distribution, the results of the selection of similar segment (SS) and auxiliary segment (AS) for segment S_1 are listed in Figure 7, and the results of the other section selections are listed in the appendix.

5.2.3 Experimental settings

To evaluate the effectiveness of the auxiliary dataset in our proposed method, we set up the following four scenarios with different applications of the auxiliary dataset.

Scenario 1: Set γ_a to 0 in objective function (9) and remove the second constraint. This translates to the ordinary LSSVR method.

Scenario 2: Set γ_p to 0 in equation (9) and remove the first constraint. The objective function of the operator can be formulated as

$$\begin{aligned} &\min_{\omega,e} J(\omega,e) = \frac{1}{2}\omega^T\omega + \frac{\gamma_a}{2}\sum_{i=1}^{T_a} e_i^2,\ \gamma_a > 0 \\ &s.t.\ y_i\left\{\omega^T\varphi(x_i)+b\right\} = 1-e_i,\ i = 1,2,\cdots,N \end{aligned} \tag{16}$$

Compared with Scenario 1, only the auxiliary dataset was used for training in the model construction.

Scenario 3: Set $\gamma_p = \gamma_a$ in equation (9) and keep the constraint. This model simultaneously inputs a similar dataset and auxiliary dataset for training and sets the same weights for both datasets.

Scenario 4: Complete the retention equation (9), and the coefficients γ_p and γ_a in the objective function are optimized by five cross-validations. This model is re-weighted by training a similar dataset and an auxiliary dataset.

In the setting of input features, taxi GPS data are full spatial and temporal coverage data, which can be used to extract features as input for our proposed model. Although the taxi volume cannot fully represent the complete traffic dynamics in the study road segment, taxi speed can

be used to evolve into an actual traffic state (Meng et al., 2017; Zhan et al., 2017). With the unique advantage of accurate taxi speed, we split the taxi volume into several intervals by taxi speed, and each interval was viewed as the model input. As shown in Table 3, we set the number of extracted taxi features n to 5, which is also equivalent to the number of model input layers. Limited by the maximum driving speed of 50 km/h for all traveling vehicles in Nanjing, China, five equivalent intervals are divided. The similar segment dataset and auxiliary segment dataset have the same input features. Furthermore, the LPR-detected volume is the actual traffic volume, which can be regarded as the output for modeling.

Table 3. Feature extraction.

Segment set	Input feature	Output label
Similar segment S_i^p	Taxi volume x_{i1}^p with the average speed in $(0-10]$ km/h	Actual volume q_i^p
	Taxi volume x_{i2}^p with the average speed in $(10-20]$ km/h	
	Taxi volume x_{i3}^p with the average speed in $(20-30]$ km/h	
	Taxi volume x_{i4}^p with the average speed in $(30-40]$ km/h	
	Taxi volume x_{i5}^p with the average speed in $(40-50]$ km/h	
Auxiliary segment S_i^a	Taxi volume x_{i1}^a with the average speed in $(0-10]$ km/h	Actual volume q_i^a
	Taxi volume x_{i2}^a with the average speed in $(10-20]$ km/h	
	Taxi volume x_{i3}^a with the average speed in $(20-30]$ km/h	
	Taxi volume x_{i4}^a with the average speed in $(30-40]$ km/h	
	Taxi volume x_{i5}^a with the average speed in $(40-50]$ km/h	

5.3 Assessment of results

5.3.1 Performance of the TL-LSSVR model

Because the traffic volume varies significantly on different road segments, the mean absolute error (MAE) and root-mean-squared error (RMSE) may not be reasonable measurements for this study. This is because the final errors are biased towards road segments with higher volume values, and many road segments with relatively low volumes are neglected (Meng et al., 2017). To eliminate the aforementioned bias, the performance of each method is judged through the mean absolute percentage error (MAPE) calculated by the mean of the

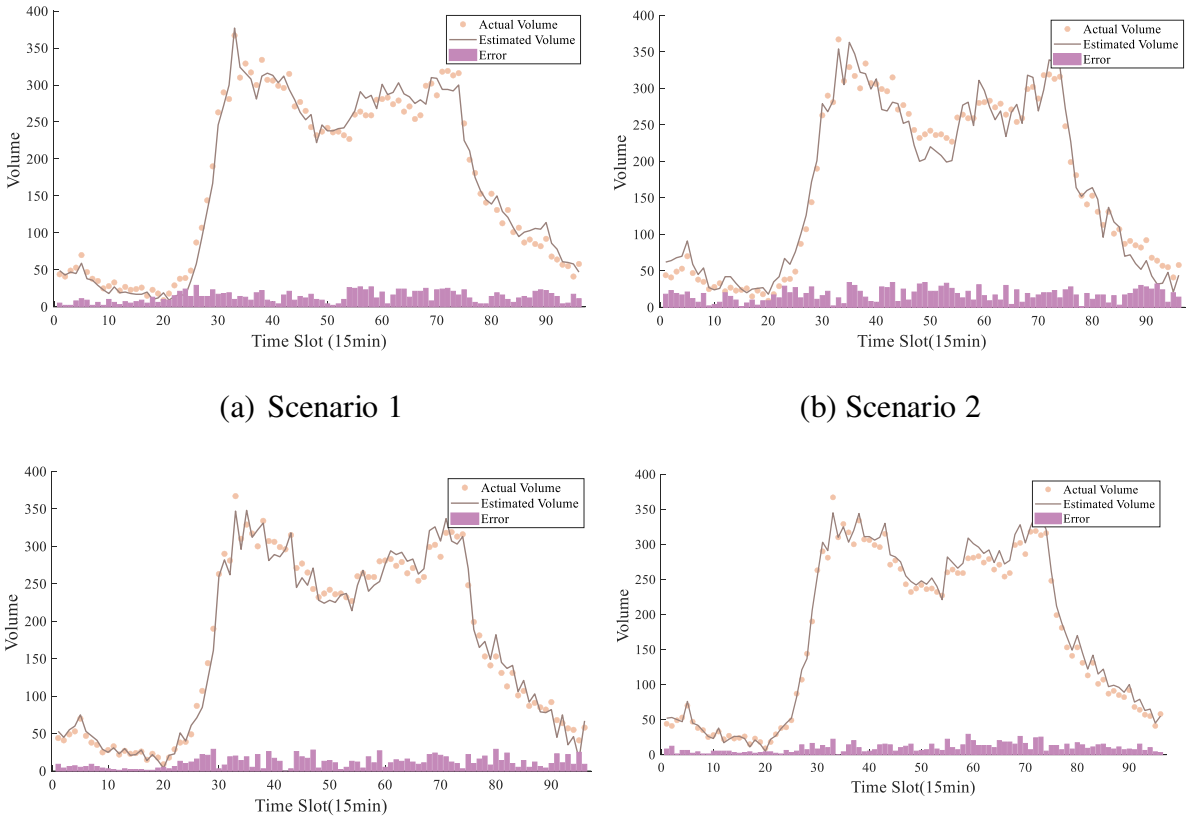
absolute differences between the estimated and detected volumes (Min and Wynter, 2011). The MAPE is defined as follows:

$$MAPE_j = \frac{1}{n}\sum_{i=1}^{N}\left[\frac{\left|q_{ij} - \widehat{q}_{ij}\right|}{q_{ij}}\right]\cdot 100 \tag{17}$$

where q_{ij} is the actual volume in the study segment, $\hat{q}_{ij}$ denotes the estimated volume in the study segment, $i(i = 1,2,..,n)$ represents the number of time slots, and $j(j = 1,2,..,L)$ is the number of study segments.

In this experiment, a personal computer with one CPU (Windows 10, 64-bit operating system, Intel (R) Core (TM) i7- i7-2600 @2.30 GHz, 4 cores), 16 GB RAM, and one GPU (NVIDIA GeForce 840 M, 4 GB memory) was configured.

In the process of setting the TL-LSSVR parameters, we selected the aforementioned four representative segments, in which the LPR detector was installed as our object segment. The time slot was 15 min (including 96-time slots). The training data were randomly divided at a ratio of 4:1. As all samples are addressed for solving in this SVR-based training process, this study uses the grid search method to solve for the hyperparameters γ_p, γ_a, and λ, which can avoid the large computational workload caused by the large training dataset for the hyperparameter solution of the model and improve the generalization performance of the model. We first roughly search for suitable parameters over a large range, and then tune the parameters through a grid search within the optimal scope. As a result, we set the kernel parameter $\lambda = 4$. In Scenario 1, we set the regularization coefficient $\gamma_p^1 = 0.05$. In Scenario 2, we set $\gamma_a^2 = 0.1$. In Scenario 3, γ_p^3 and γ_a^3 were all set as 0.05. $\gamma_p^4 = 0.05$ and $\gamma_a^4 = 0.1$ in Scenario 4.



(a) Scenario 1 (b) Scenario 2

(c) Scenario 3 (d) Scenario 4

Figure 8. Performance comparison between the actual and estimated volume with different scenarios.

As shown in Figure 8 (a)~(d), the results of the performance between the actual and estimated volumes in these four scenarios were compared. We can generally observe that Scenario 4 is the most effective, followed by Scenario 1. Compared with Scenario 2, Scenario 3 performs relatively well. For further intuitive contrast, we separated the study period into peak and non-peak hours on weekdays and weekends for analyses. The study time slots were also divided into 15 min time slots and 30 min time slots. As displayed in Figure 9, the MAPE in the 15 min time slots is generally higher than the value in the 30 min time slots. The performance during peak hours on weekdays and weekends was better than that during non-peak hours on weekdays and weekends. The model applied on weekdays has a better estimation result than that on weekends. It can be observed that a tiny fraction during the non-peak hours of weekends is underestimated more than the results during peak hours on weekdays, which

reflects that the correlation between the two types of traffic volumes during non-peak hours and weekend days are not perfectly modeled owing to the limited amount of taxi GPS data. This problem can potentially be solved by incorporating datasets that contain long periods.

From the above results, we can conclude that our approach outperforms other scenarios that do not distinguish similar segment datasets and auxiliary segment datasets, which presents significant improvements in the moment of large fluctuations in traffic dynamics.

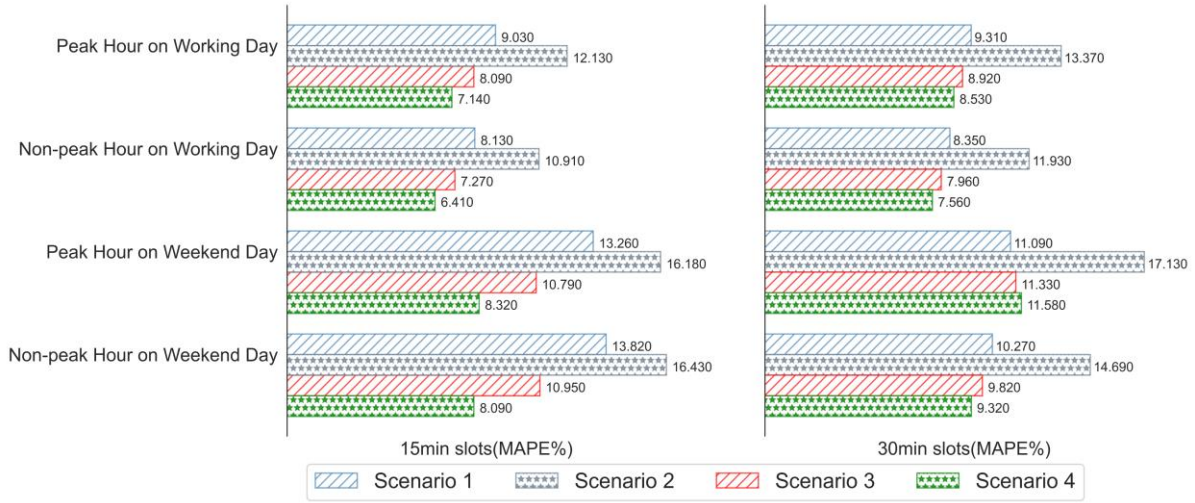


Figure 9. MAPE of scenarios with different auxiliary datasets in 15 min and 30 min slots.

5.3.2 Sensitivity analysis of the estimation model

The selected segments had different road grades and criteria for similarity analysis. The combination of different numbers of similar segments and auxiliary segments is viewed as an input for the sensitivity analysis of the proposed model. The comparison results are listed in Table 4. We find that with the increase in the two types of segment datasets, the accuracy of the model is improved. In addition, the effectiveness of increasing the number of similar road segment datasets outperforms that of the auxiliary road segment dataset.

Table 4. MAPE with the selection of input numbers of similar segments.

	15 min slots (%)		30 min slots (%)	
	Weekday	Weekend	Weekday	Weekend

Number of similar	Number of auxiliary	Peak hour	Non-peak	Peak hour	Non-peak	Peak hour	Non-peak	Peak hour	Non-peak
1	3	27.78	30.41	32.45	38.14	22.06	23.83	27.04	34.92
	2	27.98	31.29	33.82	39.14	24.24	24.82	29.14	37.43
	1	28.11	32.2	34.36	40.67	25.01	25.46	29.8	36.01
2	3	17.23	22.51	32.25	32.14	15.26	17.86	20.88	24.89
	2	17.35	23.44	33.78	32.77	16.19	19.03	22.66	26.73
	1	17.8	24.15	35.11	33.86	16.74	19.96	24.66	27.02
3	3	15.76	18.34	18.65	25.46	13.63	18.13	20.69	24.85
	2	16.29	19.07	20.19	28.38	14.42	18.93	22.54	26.19
	1	16.81	19.67	21.29	27.63	14.70	19.9	24.55	26.67

5.4 Performance of the baseline method

We selected the models of linear regression (LR), artificial neural network (ANN), and random forest (RF) as baseline models to compare them with our proposed method. The baseline models are classical statistical and machine learning methods.

LR: This model assumes that the input and output data have a linear relationship. Herein, we defined a linear regression model to approximate the relationship between taxi volume and LPR data, and a classical similarity study was conducted by Aslam et al. (2012). This model was applied to other non-detected segments to estimate the volume, which is given by

$$v_{i,t} = \gamma_0 + \gamma_1 v_{i,t-1} + \gamma_2 v_{i,t-2} + \cdots + \gamma_m v_{i,t-m} \tag{18}$$

where γ represents the weight coefficients, which is the correlation degree of adjacent road segments, and $v_{i,t}$ is the corresponding estimator at time slots t of the road segment i.

ANN: A multilayer perceptron of neural networks using backpropagation with a sigmoid activation function was trained (Park and Rilett, 1999). This baseline ANN model was built by adjacent segments with an LPR detector installed and applied to our target segment. This related study was developed by Liu et al. (2019c) and is given by:

$$y_j = \sum_{i=1}^{h} w_{ij} \exp(-\frac{1}{2\sigma^2} \left\| x_p - c_i \right\|^2) \quad (19)$$

where x_p is the p^{th} input value and $p = 1, \cdots, P$. We set the number of hidden layers h to 4. y_j denotes the estimated traffic volume in j^{th} time slots (set as 15 min slots in this test).

RF: The RF method was first proposed by Breiman (2001) and is mainly an ensemble learning method. Cheng et al. (2019a) proposed a similar study for traffic estimation. Herein, the same feature extraction method with our proposed model is applied to this baseline. In this model, the number of decision trees generated by the selection was set to 400, and the number of traffic estimators at each split node was set to 8.

The MAPE results from the four estimated methods are shown in Table 4, where the bold font denotes the best results and * marks the second-best results. + records the relative improvement of the best results compared with the second-best results. The following phenomena can be observed from the experimental results.

1. Comparing these four methods, we find that our approach outperforms the other baselines within the range of 2% to 5%. Among other baselines, RF is the second-best method, whereas it varies across different scenarios. LR methods are not available for a segment in which LPR detectors are sparely deployed.
2. Owing to the additional input of the extracted taxi feature, the performance of the RF model was better than that of the ANN model. It can be found that the incorporation of taxi feature information to increase the model input can substantially improve the model performance compared to models without the taxi feature.
3. Because the accuracy of this estimation model is mainly affected by the correlation of the two datasets, the correlation during peak hours is better than that during non-peak hours. The MAPE during peak hours outperformed that during non-peak hours. The MAPE on weekdays was higher than that on weekends because the volume on

weekdays had relatively higher volume values. The MAPE in the 15 min slots is not significantly different from that in the 30 min slots.

Table 4. The comparison of MAPE with baseline methods.

Scenario / MAPE	15 min slot				30 min slot			
	Weekday		Weekend		Weekday		Weekend	
	Peak hour	Non-peak hour	Peak hour	Non-peak hour	Peak hour	Non-peak hour	Peak hour	Non-peak hour
LR	27.2%	28.9%	30.5%	29.7%	27.3%	24.4%	25.6%	26.4%
ANN	24.6%*	26.8%	28.2%	27.4%*	22.3%	21.6%*	22.8%*	21.5%*
RF	25.7%	25.1%*	27.4%*	27.7%	21.6%*	22.7%	23.1%	22.3%
TL-LSSVR	**19.4%**	**22.9%**	**23.6%**	**25.1%**	**18.4%**	**19.2%**	**20.3%**	**19.6%**
+	4.2%	2.2%	3.8%	2.3%	3.2%	2.4%	2.5%	1.9%

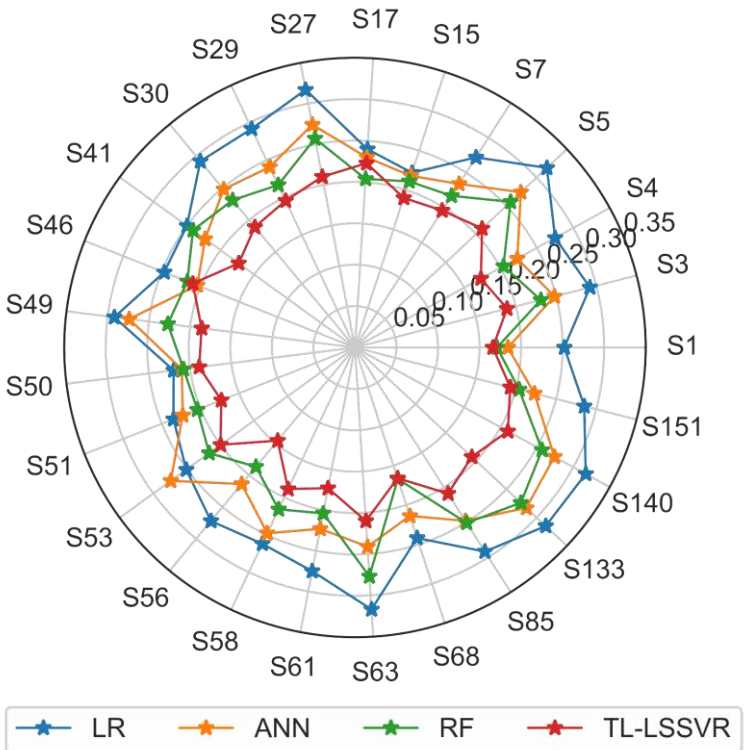
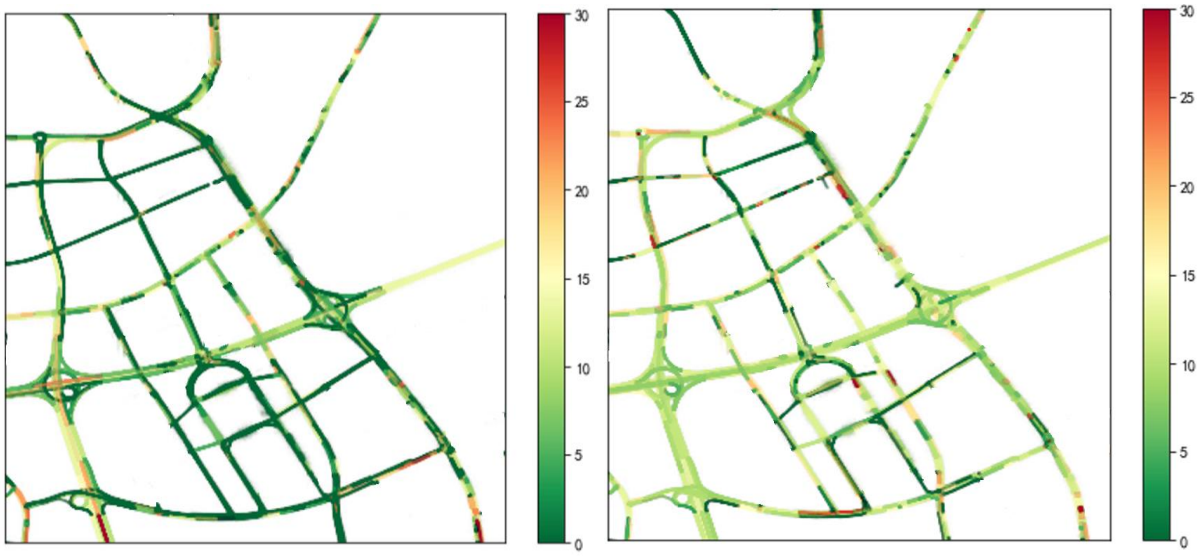


Figure 10. Comparison of segment-wise traffic volume estimation accuracy.

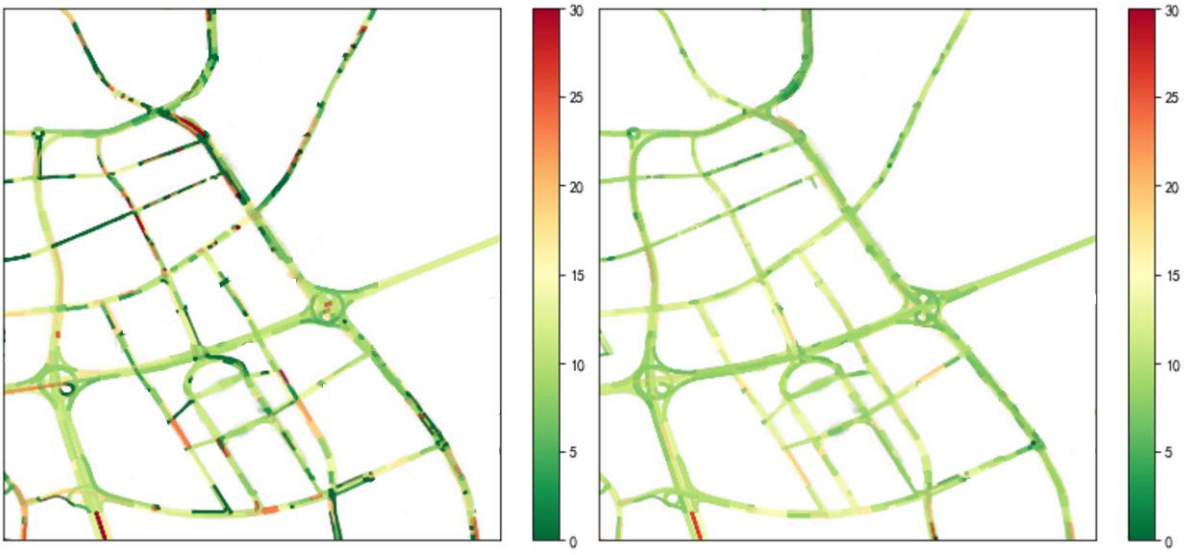
Figure 10 compares the estimated volume accuracy of the proposed model with the baseline models for each road segment in a network-wide area. The radius of the circle represents the MAPE, and the closer it is to the center, the better the estimation performance at this road segment. The LR and ANN models have more seriously biased estimations than the TS-LSSVR model on some road segments that do not have LPR detected segment cascades (e.g., road segment S_{27}, S_{63}, S_{133}, and S_{140}, respectively). For example, compared to other cascaded segments with LPR installed as input training dataset, segments that are cascaded with segment S_{133} do not have LPR installed. For the volume estimation on the segment S_{133}, LR and ANN make no distinction with their downstream segment S_{140} and parallel segment S_{150}, but the traffic dynamics on the road segment S_{133} are utterly dissimilar to both. In particular, when the fixed LPR detector is sparsely deployed, some road segments adjacent to the target segments with completely dissimilar road grades are still treated as the training dataset. By contrast, different degrees of similar road segments are searched network-wide, as well as the re-weighting process in our study, which results in rigorous results.

Based on our proposed approach, the estimated traffic volume for the remaining non-detected segments was conducted in our study area. The visualization of the estimation results is divided based on weekends and weekdays and peak and non-peak hours, as shown in Figure 11. It is evident that our model has its own advantages in estimating the traffic volume in the remaining network-wide area.



(a) Traffic volume visualization during peak hours on a weekday.

(b) Traffic volume visualization during non-peak hours on a weekday.



(c) Traffic volume visualization during peak hours on weekends.

(d) Traffic volume visualization during non-peak hours on weekends.

Figure 11. Visualization of estimated volume in partial study network-wide.

6. **Conclusion**

This paper proposes a framework to estimate the network-wide traffic volume using fixed LPR detector data and probe taxi GPS data. Based on this framework, a TL-LSSVR model is introduced, which solves the lack of training samples caused by the sparse layout of fixed

detectors. The input feature is extracted from the taxi GPS data, and the real traffic volume obtained from the LPR data was used as the output for model training. The model is generally established on LPR detected segments that are similar to object segments and applies it to other segments in which no LPR detector is installed. Considering the effectiveness of similar road segments and auxiliary road segments with different scenarios and periods, we find that combining two types of datasets with different regularization terms is valid. Furthermore, three baseline models were used for the verification, and the performance of the proposed model outperformed that of the baseline models, especially for locations with no LPR (sensor) data.

For future studies, improvements can be made in three areas: 1) diversity of detector data with multiple different forms of distributions under the effect of fixed detector coverage rate for volume estimation should be considered; 2) exploring the influence of different distribution forms of the two types of data and the characteristics of various data types; 3) in similarity analysis, other similarity analysis methods can be applied for analysis; and 4) the influences of different locations of urban areas and different deployments of detectors on the estimated results need to be evaluated.

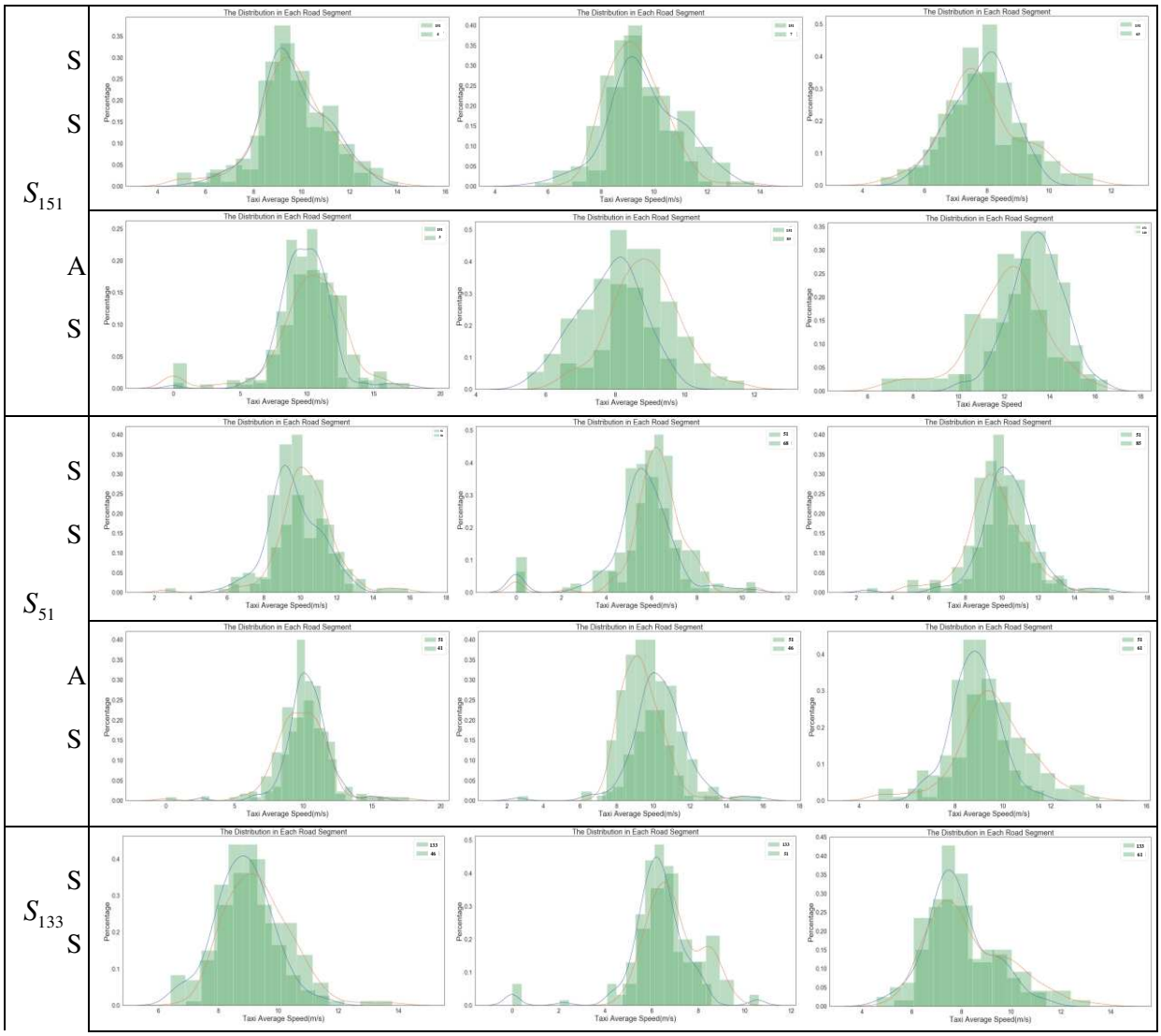
Authorship contribution statement

Jiping Xing: Conceptualization, Research Design, Visualization, and Writing-original draft; **Yuan Zhang**: Investigation, Validation; **Ronghui Liu**: Methodology, Supervision, and Writing-review & editing; **Charisma F. Choudhury**: Supervision, and Editing; **Shuyan Chen**: Supervision, and Resources; **Qixiu Cheng**: Data curation, Formal analysis, and Writing-review.

Acknowledge

This research was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 20YJAZH083), National Science Foundation of China (Project No. 71890972, 71890970), and the UK Rail Safety and Criteria Board (Project RSSB/494204565/aVSTP).

Appendix A. The taxi speed distribution plots of all selected segments



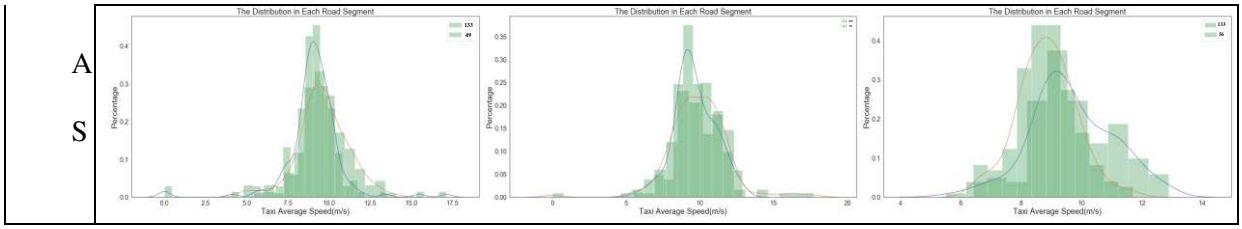


Figure A. The comparison of taxi speed distribution in segment S_{15}, S_{51}, and S_{133}.

References

Aslam, J., Lim, S., Pan, X., Rus, D., 2012. City-scale traffic estimation from a roving sensor network, *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*, pp. 141-154.

Boyd, S., Vandenberghe, L., 2004. *Convex Optimization*, Cambridge Univ. Press ed. Cambridge Univ. Press.

Breiman, 2001. Random forests. *Machine Learning* 45, 5-32.

Bwambale, A., Choudhury, C.F., Hess, S., 2019. Modelling trip generation using mobile phone data: A latent demographics approach. *Journal of Transport Geography* 76, 276-286.

Chang, C., Lin, C., 2011. LIBSVM: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology* 2,12-38.

Cheng, J., Li, G., Chen, X., 2019a. Developing a Travel Time Estimation Method of Freeway Based on Floating Car Using Random Forests. *Journal of Advanced Transportation, 2, 34-56.*.

Cheng, Q., Liu, Z., Lin, Y., Zhou, X., 2021. An s-shaped three-parameter (S3) traffic stream model with consistent car following. *Transportation Research Part B-Methodological* 153, 246-271.

Cheng, Q., Wang, S., Liu, Z., Yuan, Y., 2019b. Surrogate-based simulation optimization approach for day-to-day dynamics model calibration with real data. *Transportation Research Part C-Emerging Technologies* 105, 422-438.

Coburn, T.C., 2004. Statistical and Econometric Methods for Transportation Data Analysis. *Technometrics* 46, 492-493.

Cui, Y., Jin, B., Zhang, F., Han, B., Zhang, D., 2017. Mining Spatial-temporal Correlation of Sensory Data for Estimating Traffic Volumes on Highways, *The 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pp. 343-352.

Du, W., Zhang, Q., Chen, Y., Ye, Z., 2021. An urban short-term traffic flow prediction model based on wavelet neural network with improved whale optimization algorithm. *Sustainable Cities and Society* 69.

Ghosh, B., Basu, B., O'Mahony, M., 2009. Multivariate Short-Term Traffic Flow Forecasting Using Time-Series Analysis. *IEEE Transactions on Intelligent Transportation Systems* 10, 246-254.

Hara, Y., Suzuki, J., Kuwahara, M., 2018. Network-wide traffic state estimation using a mixture Gaussian graphical model and graphical lasso. *Transportation Research Part C-Emerging Technologies* 86, 622-638.

Holgado–Tello, F.P., Chacón–Moscoso, S., Barbero–García, I., Vila–Abad, E., 2010. Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity* 44, 153-166.

Huang, D., Gu, Y., Wang, S., Liu, Z., Zhang, W., 2020. A two-phase optimization model for the demand-responsive customized bus network design. *Transportation Research Part C Emerging Technologies* 111, 1-21.

Jia, S., 2021. Economic, environmental, social, and health benefits of urban traffic emission reduction management strategies: Case study of Beijing, China. *Sustainable Cities and Society* 67,23-39.

Ke, J.T., Zheng, H.Y., Yang, H., Chen, X.Q., 2017. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C-Emerging Technologies* 85, 591-608.

Kerner, B.S., 1998. Experimental features of self-organization in traffic flow. *Physical review letters* 81, 3797.

Laña, I., Olabarrieta, I.I., Vélez, M., Del Ser, J., 2018. On the imputation of missing data for road traffic forecasting: New insights and novel techniques. *Transportation research part C: emerging technologies* 90, 18-33.

Lee, H., Choi, S., Jung, H., Park, B.B., Son, S.H., 2019. A route guidance system considering travel time unreliability. *Journal of Intelligent Transportation Systems* 23, 282-299.

Li, L., Jiang, R., He, Z., Chen, X., Zhou, X., 2020. Trajectory data-based traffic flow studies: A revisit. *Transportation Research Part C-Emerging Technologies* 114, 225-240.

Li, Z., Liu, Z., Li, M., Xing, W., Lu, D., 2016. Mining road network correlation for traffic estimation via compressive sensing. *IEEE Transactions on Intelligent Transportation Systems* 17, 1880-1893.

Liu, J., Han, K., Chen, X., Ong, G.P., 2019a. Spatial-temporal inference of urban traffic emissions based on taxi trajectories and multi-source urban data. *Transportation Research Part C-Emerging Technologies* 106, 145-165.

Liu, R., Smith, M., 2015. Route choice and traffic signal control: A study of the stability and instability of a new dynamical model of route choice and traffic signal control. *Transportation Research Part B-Methodological* 77, 123-145.

Liu, R.H., Van Vliet, D., Watling, D., 2006. Microsimulation models incorporating both demand and supply dynamics. *Transportation Research Part A-Policy and Practice* 40, 125-150.

Liu, Z., Li, Z., Li, M., Xing, W., Lu, D., 2016. Mining road network correlation for traffic estimation via compressive sensing. *IEEE Transactions on Intelligent Transportation Systems* 17, 1880-1893.

Liu, Z., Liu, Y., Meng, Q., Cheng, Q., 2019b. A tailored machine learning approach for urban transport network flow estimation. *Transportation Research Part C-Emerging Technologies* 108, 130-150.

Liu, Z., Zhou, P., Li, Z., Li, M., 2019c. Think like a graph: real-time traffic estimation at city-scale. *IEEE Transactions on Mobile Computing* 18, 2446-2459.

Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.-Y., 2015. Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems* 16, 865-873.

Ma, X.L., Tao, Z.M., Wang, Y.H., Yu, H.Y., Wang, Y.P., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transp. Res. Pt. C-Emerg. Technol.* 54, 187-197.

Majumdar, S., Subhani, M.M., Roullier, B., Anjum, A., Zhu, R., 2021. Congestion prediction for smart sustainable cities using IoT and machine learning approaches. *Sustainable Cities and Society* 64.121-134.

Meng, C., Yi, X., Su, L., Gao, J., Zheng, Y., 2017. City-wide traffic volume inference with loop detector data and taxi trajectories, *25th Acm Sigspatial International Conference on Advances in Geographic Information Systems*, pp. 116-132.

Mihaylova, L., Boel, R., Hegyi, A., 2007. Freeway traffic estimation within particle filtering framework. *Automatica* 43, 290-300.

Min, W., Wynter, L., 2011. Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies* 19, 606-616.

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 1345-1359.

Pan, Z., Liang, Y., Wang, W., Yu, Y., Zheng, Y., Zhang, J., 2019. Urban traffic prediction from spatio-temporal data using deep meta learning, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1720-1730.

Pan, Z., Zhang, W., Liang, Y., Zhang, W., Zheng, Y., 2020. Spatio-Temporal Meta Learning for Urban Traffic Prediction. *IEEE Transactions on Knowledge and Data Engineering* PP, 1-1.

Park, D., Rilett, L.R., 1999. Forecasting freeway link travel times with a multilayer feedforward neural network. *Computer‐Aided Civil and Infrastructure Engineering* 14, 357-367.

Qian, X., Ukkusuri, S.V., 2015. Spatial variation of the urban taxi ridership using GPS data. *Applied Geography* 59, 31-42.

Qu, L., Li, L., Zhang, Y., Hu, J., 2009. PPCA-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Transactions on intelligent transportation systems* 10, 512-522.

Seo, T., Kusakabe, T., Asakura, Y., 2015. Estimation of flow and density using probe vehicles with spacing measurement equipment. *Transportation Research Part C-Emerging Technologies* 53, 134-150.

Shan, Z., Zhao, D., Xia, Y., 2013. Urban road traffic speed estimation for missing probe vehicle data based on multiple linear regression model, *16th International IEEE Conference on Intelligent Transportation Systems*, pp. 118-123.

Suykens, J., Vandewalle, J., 1999. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* 9, 293-300.

Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y.-J., Li, F., 2013. A tensor-based method for missing traffic data completion. *Transportation Research Part C-Emerging Technologies* 28, 15-27.

Tang, K., Chen, S., Liu, Z., Khattak, A.J., 2018. A tensor-based Bayesian probabilistic model for citywide personalized travel time estimation. *Transportation Research Part C-Emerging Technologies* 90, 260-280.

Van Lint, J.W.C., Hoogendoorn, S.P., van Zuylen, H.J., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C-Emerging Technologies* 13, 347-369.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. 34, 123-145.

Wang, P., Lai, J., Huang, Z., Tan, Q., Lin, T., 2021. Estimating Traffic Flow in Large Road Networks Based on Multi-Source Traffic Data. *IEEE Transactions on Intelligent Transportation Systems* 22, 5672-5683.

Wang, Z., He, S.Y., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society* 11, 141-155.

Weiss, K., Khoshgoftaar, T.M., Wang, D.D., 2016. A survey of transfer learning. *Journal of Big Data* 3, 1-40.

Wilkie, D., Sewall, J., Lin, M., 2013. Flow reconstruction for data-driven traffic animation. *ACM Transactions on Graphics (TOG)* 32, 89.

Xing, J., Liu, Z., Wu, C., Chen, S., 2019. Traffic volume estimation in multimodal urban networks using cellphone location data. *IEEE Intelligent Transportation Systems Magazine* 11, 93-104.

Yang, H., Yang, J., Han, L.D., Liu, X., Pu, L., Chin, S.-m., Hwang, H.-l., 2018. A Kriging based spatiotemporal approach for traffic volume data imputation. *PloS one* 13.120-134.

Yu, J., Stettler, M.E.J., Angeloudis, P., Hu, S., Chen, X., 2020. Urban network-wide traffic speed estimation with massive ride-sourcing GPS traces. *Transportation Research Part C-Emerging Technologies* 112, 136-152.

Zeroual, A., Harrou, F., Sun, Y., 2019. Road traffic density estimation and congestion detection with a hybrid observer-based strategy. *Sustainable Cities and Society* 46.94-112.

Zhan, X., Yu, Z., Yi, X., Ukkusuri, S.V., 2017. Citywide traffic volume estimation using trajectory data. *IEEE Transactions on Knowledge and Data Engineering* 29, 272-285.

Zhang, K., He, Z., Zheng, L., Zhao, L., Wu, L., 2020a. A generative adversarial network for travel times imputation using trajectory data. *Computer-Aided Civil and Infrastructure Engineering*, 3, 121-135.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. *Transportation Research Part C-Emerging Technologies* 58, 308-324.

Zhang, Y., Liu, Y., 2009. *Missing Traffic Flow Data Prediction using Least Squares Support Vector Machines in Urban Arterial Streets*. *IEEE Transactions on intelligent transportation systems* 3, 12-22.

Zhang, Z., Li, M., Lin, X., Wang, Y., 2020b. Network-wide traffic flow estimation with insufficient volume detection and crowdsourcing data. *Transportation Research Part C-Emerging Technologies* 121,23-50.

Zhou, X., Mahmassani, H.S., 2007. A structural state space model for real-time traffic origin-destination demand estimation and prediction in a day-to-day learning framework. *Transportation Research Part B-Methodological* 41, 823-840.

Zhou, Y., Yi, P., Li, W., Gong, C., 2021. Assessment of city sustainability from the perspective of multi-source data-driven. *Sustainable Cities and Society* 70, 95-120.