

Journal of Advanced Transportation

# Data-Driven Urban Mobility Modeling and Analysis

Lead Guest Editor: Xiaolei Ma

Guest Editors: Guohui Zhang and Xiaoyue Liu





---

# **Data-Driven Urban Mobility Modeling and Analysis**


Journal of Advanced Transportation

---

## **Data-Driven Urban Mobility Modeling and Analysis**

Lead Guest Editor: Xiaolei Ma

Guest Editors: Guohui Zhang and Xiaoyue Liu



---

Copyright © 2017 Hindawi. All rights reserved.

This is a special issue published in "Journal of Advanced Transportation." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



## Editorial Board

Francesco Bella, Italy  
K. Bogenberger, Germany  
Juan C. Cano, Spain  
Giulio E. Cantarella, Italy  
Anthony Chen, USA  
Steven I. Chien, USA  
Antonio Comi, Italy  
G. H. Correia, Netherlands  
Emanuele Crisostomi, Italy  
Luca D'Acierno, Italy  
Andrea D'Ariano, Italy  
Alexandre De Barros, Canada  
Luigi Dell'Olio, Spain  
Cédric Demonceaux, France

Sunder Lall Dhingra, India  
Yuchuan Du, China  
Juan-Antonio Escareno, France  
David F. Llorca, Spain  
J. Härri, France  
Serge Hoogendoorn, Netherlands  
Angel Ibeas, Spain  
Lina Kattan, Canada  
Victor L. Knoop, Netherlands  
Ludovic Leclercq, France  
Seungjae Lee, Republic of Korea  
Zhi-Chun Li, China  
Yue Liu, USA  
Jose R. Martinez-De-Dios, Spain

Andrea Monteriù, Italy  
Jose E. Naranjo, Spain  
Dongjoo Park, Republic of Korea  
Paola Pellegrini, France  
Luca Pugi, Italy  
P. N. Seneviratne, Philippines  
Wai Yuen Szeto, Hong Kong  
Richard S. Tay, Australia  
Martin Trépanier, Canada  
Pascal Vasseur, France  
Antonino Vitetta, Italy  
S. Travis Waller, Australia  
Chien-Hung Wei, Taiwan  
Jacek Zak, Poland

# Contents

---

## **Data-Driven Urban Mobility Modeling and Analysis**

Xiaolei Ma, Guohui Zhang, and Xiaoyue Liu  
Volume 2017, Article ID 8679827, 2 pages

## **Analysis and Prediction on Vehicle Ownership Based on an Improved Stochastic Gompertz Diffusion Process**

Huapu Lu, He Ma, Zhiyuan Sun, and Jing Wang  
Volume 2017, Article ID 4013875, 8 pages

## **Exploring the Influence of Attitudes to Walking and Cycling on Commute Mode Choice Using a Hybrid Choice Model**

Chuan Ding, Yu Chen, Jinxiao Duan, Yingrong Lu, and Jianxun Cui  
Volume 2017, Article ID 8749040, 8 pages

## **Clustering Vehicle Temporal and Spatial Travel Behavior Using License Plate Recognition Data**

Huiyu Chen, Chao Yang, and Xiangdong Xu  
Volume 2017, Article ID 1738085, 14 pages

## **Impact of Vehicular Countdown Signals on Driving Psychologies and Behaviors: Taking China as an Example**

Fuquan Pan, Lixia Zhang, Changxi Ma, Haiyuan Li, Jinshun Yang,  
Tao Liu, Fengyuan Wang, and Shushan Chai  
Volume 2017, Article ID 5838520, 11 pages

## **Large-Scale Demand Driven Design of a Customized Bus Network: A Methodological Framework and Beijing Case Study**

Jihui Ma, Yang Yang, Wei Guan, Fei Wang, Tao Liu, Wenyuan Tu, and Cuiying Song  
Volume 2017, Article ID 3865701, 14 pages

## **Compression Algorithm of Road Traffic Spatial Data Based on LZW Encoding**

Dong-wei Xu, Yong-dong Wang, Li-min Jia, Gui-jun Zhang, and Hai-feng Guo  
Volume 2017, Article ID 8182690, 13 pages

## **Dynamic Route Choice Prediction Model Based on Connected Vehicle Guidance Characteristics**

Jiangfeng Wang, Jiarun Lv, Chao Wang, and Zhiqi Zhang  
Volume 2017, Article ID 6905431, 8 pages

## **A Novel Trip Coverage Index for Transit Accessibility Assessment Using Mobile Phone Data**

Zhengyi Cai, Dianhai Wang, and Xiqun (Michael) Chen  
Volume 2017, Article ID 9754508, 14 pages

## Editorial

# Data-Driven Urban Mobility Modeling and Analysis

**Xiaolei Ma,<sup>1</sup> Guohui Zhang,<sup>2</sup> and Xiaoyue Liu<sup>3</sup>**

<sup>1</sup>*School of Transportation Science and Engineering, Beihang University, Beijing 100191, China*

<sup>2</sup>*Department of Civil & Environment Engineering, University of Hawaii, Manoa, HI, USA*

<sup>3</sup>*Department of Civil & Environment Engineering, University of Utah, Salt Lake City, UT, USA*

Correspondence should be addressed to Xiaolei Ma; [xiaolei@buaa.edu.cn](mailto:xiaolei@buaa.edu.cn)

Received 14 June 2017; Accepted 14 June 2017; Published 30 July 2017

Copyright © 2017 Xiaolei Ma et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With increasing economic and social activities, travel demand has increased significantly over the past several decades, overloading many already congested roadways. The widening gap between travel demand and infrastructure supply has worsened the levels of congestion worldwide, resulting in many urban mobility, safety, and environmental issues, such as severe congestion, lengthened travel time, increased risk of traffic accidents, excessive fuel consumption, increased air pollution, and significant public health issues. The concept of smart cities has been gaining popularity, which is to leverage big data analytics, sensing technologies, and Internet of Things (IoT) to move people and goods faster, cheaper, and more efficiently. As heterogeneous data and computational resources become available, the development of data-driven approaches has been advancing as well for modeling and analyzing urban mobility. This special issue serves as a major platform to facilitate the discussion and exchange of research ideas and technology development, encourage multidimensional knowledge sharing, and enhance research activities in data-driven urban mobility modeling and analysis. In total, seven papers are included in this special issue and are summarized as follows.

There are several articles focusing on developing data-driven approaches to improve public transit system efficiency and accessibility. J. Ma et al. proposed a methodological framework to address the issue of customized bus network design based on large-scale travel demand data. A route selection model considering operation cost and social welfare was built, followed by a branch-and-bound-based solution method for model solving. An empirical study in Beijing, China, validated the effectiveness of the proposed framework.

Z. Cai et al. proposed a Trip Coverage Index (TCI) based on mobile phone data to assess transit accessibility. TCI considered both the individual-level transit trip coverage and spatial distribution and was then applied to a transit network in Hangzhou, China.

From the perspective of car-based traffic operation and management, several articles talk about utilizing sensing technology and simulation data to analyze drivers' behaviors, route choices, or road traffic network structure. D. Xu et al. developed a compression method based on LZW encoding and principle component analysis. Six typical road segments in Beijing were tested using the proposed method and presented a high reconstruction accuracy. H. Chen et al. clustered drivers' travel characteristics based on license plate recognition data in Shenzhen, China. Each traveler's spatiotemporal variability and activity pattern are taken into account, resulting in six groups in weekdays and three groups in weekends. J. Wang et al. proposed a route choice prediction model in the context of connected vehicles. Five characteristics indexes including compliance rate, following rate, penetration rate, release delay time, and congestion level were built. A simulation scenario demonstrated the effectiveness of the proposed model with the average root mean square error as 3.19%. Y. Lu et al. introduced a novel improved stochastic Gompertz diffusion process to explain the relationship between vehicle ownership and GDP per-capita. Based on the data from US, UK, Japan, and Korea from 1960 to 2008, the proposed model performed well in the fitting process and predicted that China is still on the initial stage of motorization.

The remaining two articles investigated travelers' psychologies and behaviors using survey data. C. Ding et al. examined the impact of attitudes to walking and cycling on commute mode choice and used survey data to establish an integrated discrete choice model and structural equation model. A comparison confirmed that the proposed hybrid model outperforms other traditional models. F. Pan et al. explored the influence of vehicular countdown signals on driving psychologies and behaviors. An online survey with 1051 valid questionnaires was undertaken and analyzed. Results showed that most drivers prefer countdown signal controls and female drivers are more conservative before the green countdown ends.

*Xiaolei Ma*  
*Guohui Zhang*  
*Xiaoyue Liu*

## Research Article

# Analysis and Prediction on Vehicle Ownership Based on an Improved Stochastic Gompertz Diffusion Process

Huapu Lu,<sup>1</sup> He Ma,<sup>1</sup> Zhiyuan Sun,<sup>2</sup> and Jing Wang<sup>3</sup>

<sup>1</sup>*Institute of Transportation Engineering, Tsinghua University, Beijing 100084, China*

<sup>2</sup>*College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China*

<sup>3</sup>*School of Architecture and Urban Planning, Beijing University of Civil Engineering and Architecture, Beijing, China*

Correspondence should be addressed to He Ma; mah13@mails.tsinghua.edu.cn

Received 19 December 2016; Revised 19 April 2017; Accepted 27 April 2017; Published 11 July 2017

Academic Editor: Guohui Zhang

Copyright © 2017 Huapu Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper aims at introducing a new improved stochastic differential equation related to Gompertz curve for the projection of vehicle ownership growth. This diffusion model explains the relationship between vehicle ownership and GDP per capita, which has been studied as a Gompertz-like function before. The main innovations of the process lie in two parts: by modifying the deterministic part of the original Gompertz equation, the model can present the remaining slow increase when the S-shaped curve has reached its saturation level; by introducing the stochastic differential equation, the model can better fit the real data when there are fluctuations. Such comparisons are carried out based on data from US, UK, Japan, and Korea with a time span of 1960–2008. It turns out that the new process behaves better in fitting curves and predicting short term growth. Finally, a prediction of Chinese vehicle ownership up to 2025 is presented with the new model, as China is on the initial stage of motorization with much fluctuations in growth.

## 1. Introduction

The growth of vehicle ownership has witnessed a great change of transportation demand sector over the years and is an important part of urbanization. The study by Simonsen and Walnum [1] showed that transportation contributes nearly 30% of CO<sub>2</sub> emission in OECD countries and accounts for a critical cause of regional and local air pollutions. In addition, the prediction of future vehicle numbers is of great policy revelation. Therefore, the increasing pattern of vehicle ownership should be paid high attention to, especially for developing countries, for example, China, who are stepping into the fast growth stage [2].

Many factors have influence on vehicle ownership growth, such as economics factor, public transportation service level, policy restrictions, and urban layout, while the economic growth has been the dominant driven factor, which is GDP per capita in the present paper. The relationship of

growth of vehicle ownership and GDP per capita can be modeled in specific form, and an improvement has been carried out based on the most usually used Gompertz curve in order to obtain a better fitting projection. The introduced model solves two significant issues existing in the original Gompertz curve by introducing the stochastic diffusion process and a modification part. It reveals a new way of better prediction vehicle ownership based on limited data, with only the aggregate vehicle ownership condition and GDP per capita.

This paper is organized as follows. The next part reviews related studies of the specific topic. Section 3 proposes the model (improved stochastic Gompertz diffusion curve) with its structure, inference, and solution. In Section 4, applications to real data of the US, UK, Japan, and Korea have been carried out to evaluate comparisons, and projection of vehicle ownership growth of China is depicted based on the new model. Finally, we conclude our study and give suggestions on future study.

## 2. Literature Review

An external or internal vehicle ownership model is often used for various purposes, as mentioned by de Jong et al. [3]. For the aggregate level, vehicle ownership model can be used by car manufactory for market analysis, by national and local government in order to make policy incentives based on the forecasting results, or by energy industry who is concerned about the oil consumption related to vehicles. For the disaggregate level, the model is usually treated as an input to mode choice in transportation model systems and thus could have a more detailed output. Examples can be found as traditional disaggregate car choice model [4], panel models [5], and dynamic transaction models [6]. This paper focuses on the aggregate model, as we aim at the main pattern and future trend of vehicle ownership growth, rather than the detailed types or components of cars. In addition, the aggregate model has a much lighter requirement of data, while the disaggregate one relies heavily on the amount and types of data collected and sometimes even requires dataset with a long period of observation.

There are various aggregate models based on their different goals. Focusing on the market changing, K. U. Leuven, and Standard and Poor's DRI [7] studied how different structure of transportation modes and car prices influence the stock of personal vehicles. Focusing on vehicle ownership revolution, Van den Broecke [8] divided people into different categories by their age and predicted the growth pattern of vehicles by people becoming older, assuming that the behavior characteristics in each category would remain the same over the years. Based on product life cycle and diffusion theories, many studies use different models to depict the relationship between vehicle ownership and economics factors (e.g., per capita income or gross domestic product, GDP), which is more straightforward and is especially suitable for developing countries as they do not possess enough data related to vehicle ownership growth for other detailed types of studies. Early studies on this kind of models analyzed and described the relationship between vehicle ownership and time series, which is found as an S-shape curve [9]. Different variables have been shown to influence the development of vehicle ownership projection, however, given the difference in data sources, model of including too many variables could lead to the difficulty in comparison of different results from various countries and regions [2, 10, 11]. In addition, it is rather difficult to obtain and unify all the data of different variables in order to provide a complete dataset. Therefore, it is more appropriate to generate model based on simple dataset easily obtained, in order to present the projection of vehicle ownership development and give a unified prediction of different countries, especially of developing countries whose vehicle ownership remains rather small.

Hereafter, studies focus more on the economic factor, which is seen as the main driving force of vehicle ownership growth. Various kinds of models are developed to fit the S-shape curve, such as semi-log linear and log linear regression models by Dunkerley and Hoch [12], quasi-logistic function model by Button et al. (1993), elasticity analysis model by Stares and Liu [13], and Gompertz diffusion function model

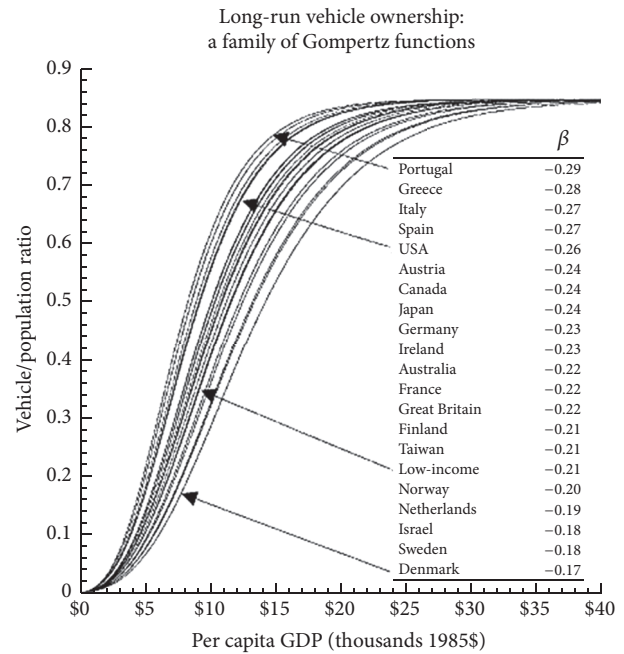


FIGURE 1: Projections of estimated vehicle Gompertz functions.

by Dargay and Gately [2]. The Gompertz model is found to be more flexible than logistic model and is suitable for analysis on both short term and long term prediction [14]. He has carried out a series of examples using a simple parametric method to choose between a Gompertz and a logistic equation and suggested that the Gompertz curve would be indeed appropriate for the stock of car series. In the present paper, we use the Gompertz function as the base of our model, with some improvements in order to make it fit better.

The Gompertz function was firstly used in biological field, with its good performance in predicting growth, mortality, and thus the lifespan (see example papers as those by Zwietering et al. [17] and Finch and Pike [18]). Acutt and Dodgson [19] used the Gompertz curve in his study to forecast the future car ownership. Dargay and Gately [2] applied the Gompertz function to countries of full range, with low-income countries and high-income ones. As the model is more flexible and suitable for developing countries, there are studies concerning the prediction of Chinese vehicle stock based on it. Wang [20] used the general Gompertz function to present the S-shape curve of vehicle ownership growth in China. Zhao [21] estimates the function with panel data of 21 countries and arears, 1963–2008, in order to predict the vehicle ownership in China up to year 2050. Although the Gompertz curve is widely used these years, it has some disadvantages in applications. As shown in Figure 1 [2], the shape of curve is smooth S-shape and remains the same after it quickly reaches the saturation level. This brings about two significant problems, firstly, the general Gompertz curve is unable to present the fluctuations existing in real data;



secondly, it cannot predict the remaining slow growth after the growth has reached its saturation level.

According to the slow remaining growth in the rear part of curve, a modification may be made to the model (see details in Section 3.2). In order to fit the fluctuations in real data, we use the stochastic differential equation (SDE) in the present paper. Gutierrez-Jaimez et al. [22] have tested the SDE on Gompertz equation and successfully proved that this new model performs well for random growth of rabbit weights. The new model proposed thus has advantages as follows:

- (1) The new model is able to present a better fitting result based on limited data of only vehicle ownership and GDP per capita.
- (2) It has overcome two main shortcomings of the original Gompertz curve as described before, by introducing a modification part as well as the stochastic diffusion process.

### 3. Methodology

As used in previous studies, the relationship of vehicle ownership to GDP per capita has been represented by Gompertz growth curve, modeled as follows:

$$X_{g_t} = a \cdot \exp(-b \cdot \exp(-c \cdot g_t)), \quad (1)$$

where  $X_{g_t}$  is the quantity of vehicle ownership per 1000 people in year  $t$ , and  $g_t$  is GDP per capita in year  $t$ , and  $a, b, c$  are parameters of the function to be calculated in regression.

Although there have been problems in applying this function to real data, as implied in the literature review; the S-shaped curve could successfully present the general growth pattern of the process and thus remains the main structure of the new model. In this part, we propose two kinds of improvement to the original model, as illustrated below.

*3.1. The Stochastic Differential Equation of Gompertz Growth Function.* In order to obtain a diffusion process related to

Gompertz curve (1), we should search for a process in which the solution of the Fokker-Planck equation without noise is such a curve, as proposed by Capocelli and Ricciardi [23], and is successfully conducted by Gutiérrez et al. [24] for a specific Gompertz-like curve used in biological phenomena. In this paper, we perform the procedure and define the stochastic Gompertz diffusion process (SGDP):

$$\frac{dX_{g_t}}{dt} = (\alpha_1 - \alpha_2 \cdot \ln(X_{g_t})) \cdot X_{g_t} + \alpha_3 \cdot X_{g_t} \cdot d\omega_t, \quad (2)$$

where  $X_{g_t}$  and  $g_t$  remain the same as before,  $\alpha_i, i \in (1, 2, 3)$ , are three parameters to be calculated in regression, and  $\omega_t$  is a one-dimensional Wiener standard process with zero mean and  $\text{var}(\omega_t - \omega_s) = (t - s)$ .

By applying the Fokker-Planck equation, this process has forward equation and infinitesimal moments as

$$\begin{aligned} \frac{\partial f}{\partial g} &= -\frac{\partial f}{\partial x} (ab \cdot \exp(-cg) \cdot x \cdot f) \\ &+ \sigma^2 \frac{\partial^2}{\partial x^2} (x^2 f), \end{aligned} \quad (3)$$

$$A_1(x, g) = ab \cdot \exp(-cg) \cdot x, \quad (4)$$

$$A_2(x, g) = \sigma^2 x^2,$$

where  $\sigma = 1$  for a standard Wiener process.

It is clear that when  $\sigma$  vanishes, the solution of (3) turns into the original equation (1). Thus the process we proposed fulfills the condition imposed.

After defining the SGDP function, we continue to its parameter estimation. There are three parameters in the function, with  $\alpha_1, \alpha_2$  being drifting parameters and  $\alpha_3$  being the noise coefficient. Ferrante et al. [25] have proposed Itô's stochastic differential equations from an observed continuous sample path. With the same method, the estimations are calculated as

$$\hat{\alpha}_1 = \frac{\left( \int_0^T (\log x_t)^2 dt \right) \left( \int_0^T (dx_t/x_t) \right) - \left( \int_0^T \log x_t dt \right) \left( \int_0^T (\log x_t/x_t) dx_t \right)}{T \int_0^T \log^2 x_t dt - \left( \int_0^T \log x_t dt \right)^2}, \quad (5)$$

$$\hat{\alpha}_2 = \frac{\left( \int_0^T \log x_t dt \right) \left( \int_0^T (dx_t/x_t) \right) - T \int_0^T (\log x_t/x_t) dx_t}{T \int_0^T \log^2 x_t dt - \left( \int_0^T \log x_t dt \right)^2},$$

where  $\{x_t; t \in [0, T]\}$  is the observed sample path.

In practice, as there is no continuous data for vehicle ownership, the estimation could only be based on discrete sample data  $(X_{g_1}, X_{g_2}, \dots, X_{g_T})$ . In the present study, we use Riemann integral instead of the continuous stochastic integral. The interval is divided with a small step (0.001 in

this paper), and each is applied with Itô formula, in order to approach the continuous function.

With the same procedure, we can obtain the noise coefficient with the following form:

$$\hat{\alpha}_3 = \frac{1}{T-1} \sum_{t=2}^T \frac{|x_t - x_{t-1}|}{\sqrt{t x_t x_{t-1}}}. \quad (6)$$

The conditional trend function of SGDP is presented as

$$\begin{aligned} m(g_t | g_s) &= E(X_{g_t} | X_{g_s} = x_s) \\ &= \exp\left(\frac{\alpha_1}{\alpha_2} - \exp(-\alpha_2(g_t - g_s))\right) \\ &\quad \cdot E \exp\left(\alpha_3 \int_s^t \exp(-\alpha_2(t - \eta)) d\omega_\eta\right). \end{aligned} \quad (7)$$

As  $\omega_t$  is a Wiener standard process with a variation of  $\alpha_3^2 \int_s^t \exp(-2\alpha_2(t - \eta)) d\eta$ , we can calculate its expectation as

$$\begin{aligned} E \exp\left(\alpha_3 \int_s^t \exp(-\alpha_2(t - \eta)) d\omega_\eta\right) \\ = \exp\left(\frac{\alpha_3^2}{2} \int_s^t \exp(-\alpha_2(t - \eta)) d\eta\right). \end{aligned} \quad (8)$$

Applying (8) to (7), and with the initial value, we can get the conditional trend function as (9), which should be used in the prediction of future values.

$$\begin{aligned} m(g_t) &= \exp\left(\frac{\alpha_1}{\alpha_2} - \exp(-\alpha_2(g_t - g_s))\right) \\ &\quad + \frac{\alpha_3^2}{4\alpha_2} (1 - \exp(-2\alpha_2 g_t)). \end{aligned} \quad (9)$$

**3.2. The Improved SGDP Model.** By applying the SGDP model to a set of data of vehicle ownership and GDP per capita in America, as presented in Figure 2, we can see that although the SGDP curve fits better than the original Gompertz curve when there is fluctuations in data, it still cannot present the slow increase when the curve begins to reach its saturation level in the rear. Therefore, an improvement in the deterministic part should be carried out for this problem.

In the present paper, an improvement is carried out in the deterministic part of the improved SGDP model, proposed as

$$X_{g_t} = a \cdot \exp(-b \cdot \exp(-c \cdot g_t)) + \beta_1 \frac{X_{g_t}}{X_{g_t} + \beta_2}, \quad (10)$$

where  $\beta_1$  and  $\beta_2$  are parameters to be estimated.

The improved SGDP model has a more complex function and thus is difficult for estimation by inference. In this paper, we use the SDE Toolbox of Matlab Package by Umberto Picchini (Umberto Picchini, SDE Toolbox: Simulation and Estimation of Stochastic Differential Equations with Matlab, <http://sdetoolbox.sourceforge.net>) to get the numerical results, illustrated in the next part.

## 4. Data and Applications

**4.1. The Estimation and Comparison in Sample Countries.** The growth pattern of vehicle ownership per 1000 people has been changing with the increasing of GDP per capita. In terms of elasticity (elasticity: the ratio of the average% growth in vehicle ownership to the average% growth in per

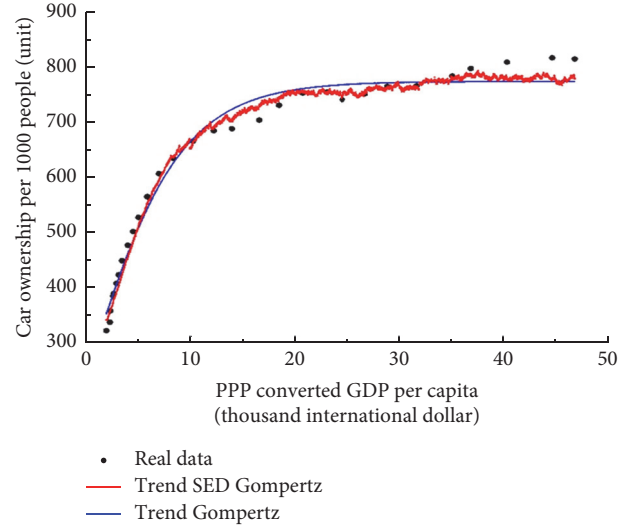


FIGURE 2: Real data versus SDE Gompertz versus Trend Gompertz in the United States.

capita income), the values are different in different stages of development, also in different countries and regions [13]. A country such as the United States has reached the saturation level with vehicle ownership per 1000 people of approximately 800 units, while some countries are in the stage of slow growth, like Japan, with the value of elasticity approximately 0.5, and other developing countries are in the stage of fast growth with a high elasticity of 1.7, taking China as an example [26]. In addition, some cities, for example, London and Singapore, have conducted policy restrictions on vehicle usage/purchase and thus have a different path from countries without any restrictions. Therefore, in the present paper, we choose four countries with quite different growth curve to illustrate the differences between a general Gompertz curve and the improved SDEG curve proposed.

The countries selected are the United States, the United Kingdom, Japan, and Korea. Due to data availability, the period of data estimation is 1960–2008 for the former three countries and is 1966–2008 for Korea. Three kinds of data are collected:

- (i) Population and economic index: we use the index of population, total and GDP, and PPP (current international dollar) from the database of World Bank (for population: <http://data.worldbank.org/indicator/SP.POP.TOTL>; for economics factor: <http://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD>).
- (ii) Vehicle ownership: we use the total number of vehicles in each country (unit). Different resources of data are given (US: Historical Highway Statistics (1945–1995): <http://www.fhwa.dot.gov/policy/ohpi/hss/hsspubsarc.cfm>; [https://www.census.gov/compendia/statab/cats/transportation/motor\\_vehicle\\_registrations\\_alternative\\_fueled\\_vehicles.html](https://www.census.gov/compendia/statab/cats/transportation/motor_vehicle_registrations_alternative_fueled_vehicles.html) (1990–2008); UK: <https://www.gov.uk/government/publications/tsgb-2011-vehicles>; Japan: 全国の自動車

TABLE 1: Estimated parameters of improved SGDP and general Gompertz in sample countries.

	Improved SGDP						General Gompertz		
	$a$	$b$	$c$	$\beta_1$	$\beta_2$	$\alpha_3$	$a$	$b$	$c$
US	589.623	1.453	0.406	432.828	40.879	1.13	774.996	1.167	0.205
UK	83.753	105.702	0.184	533.851	4.529	0.75	515.885	1.464	0.140
Japan	71.671	40.729	0.861	773.112	13.531	0.62	603.796	2.556	0.139
Korea	230.316	19.710	0.246	794.593	141.921	1.42cc	338.00	6.237	0.156

TABLE 2: Predict values by improved SGDP and general Gompertz of vehicle ownership per 1000 people in the United States.

	Real data	Gompertz	Improved SGDP	95% confidence interval	
2001	784.933	774.311	789.520	747.682	831.358
2002	808.761	774.517	794.774	749.836	839.712
2003	798.673	774.637	798.819	752.811	844.827
2004	797.935	774.760	804.472	757.099	851.845
2005	810.240	774.846	810.261	761.024	859.498
2006	815.994	774.899	815.489	765.278	865.700
2007	818.134	774.928	819.591	767.680	871.502
2008	820.714	774.934	820.557	766.282	874.832
<i>Standard error</i>		34.103	5.721	—	—

TABLE 3: Prediction standard errors of improved SGDP and general Gompertz in sample countries.

	General Gompertz	Improved SGDP
US	34.103	5.721
UK	41.671	5.702
Japan	23.753	4.729
Korea	4.267	4.956

車保有台数の推移. <http://www.city.osaka.lg.jp/kankyo/cmsfiles/contents/0000006/6885/101.1-1.pdf>;  
 Korea: [http://www.index.go.kr/egams/stts/jsp/potal/stts/PO\\_STTS\\_IdxMain.jsp?idx\\_cd=1257&bbs=INDEX\\_001](http://www.index.go.kr/egams/stts/jsp/potal/stts/PO_STTS_IdxMain.jsp?idx_cd=1257&bbs=INDEX_001)).

After the calculation of vehicle ownership per 1000 people and GDP per capita, we use the data of up to year 2000 as inputs variables and the parameters of proposed improved SGDP could be given by the SDE Toolbox of Matlab package, as presented in Table 1.

Then we predict the vehicle ownership per 1000 people based on the known GDP per capita and parameter estimated, from year 2001 to year 2008. Table 2 presents the real data, the results from the general Gompertz fitting curve, and the improved SGDP fitting curve as well as its 95% confidence interval. The improved SGDP curve has a much lower standard error and thus predicts better than the general Gompertz function. For simplicity, we only give the standard error of the other three countries, as shown in Table 3.

It should be highlighted that although improved SGDP function gives an obvious better performance for the data of the United States, the United Kingdom, and Japan, it behaves almost the same as general Gompertz for the data of Korea. To see more directly into this condition as well as

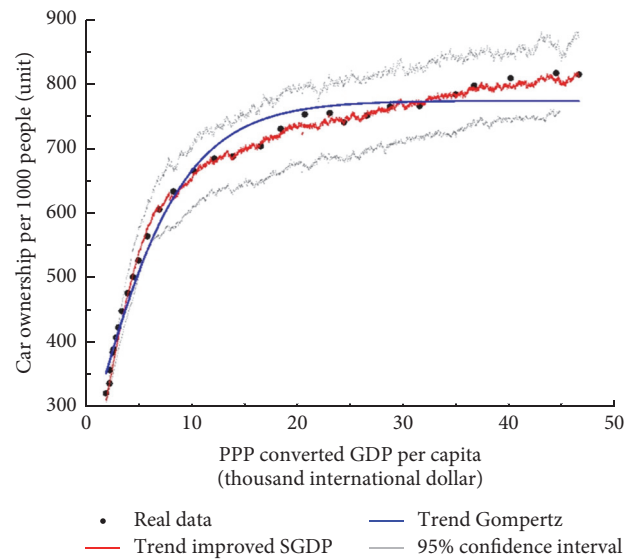


FIGURE 3: Real data versus Improved SDE Gompertz versus Trend Gompertz in the United States.

intuitive comparison of different fitting curves, each country's paths have been presented as Figures 3–6 (using data with a time span up to 2008 (when using data of full time span, the parameters regressed are a little different from those presented in Table 1; as the difference is little, for simplicity, we do not present the results specifically but pay more attention to the comparing figures)).

It is quite straightforward that the improved SGDP Gompertz curve proposed in this paper behaves better than the general one, which fits more close or nearly the same pattern as the real growth curve for all four sample countries, especially for the US, the UK, and Japan, whose real data either

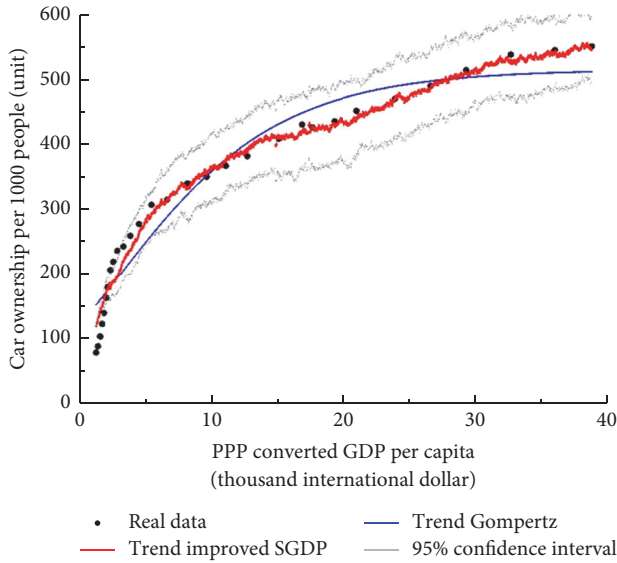


FIGURE 4: Real data versus Improved SDE Gompertz versus Trend Gompertz in the United Kingdom.

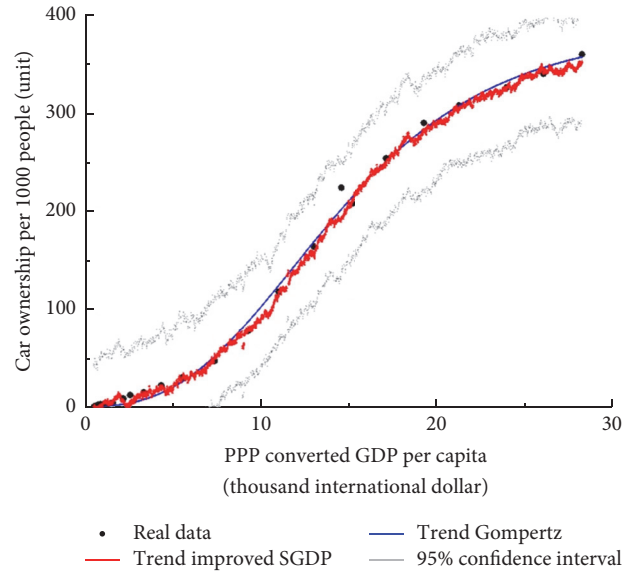


FIGURE 6: Real data versus Improved SDE Gompertz versus Trend Gompertz in Korea.

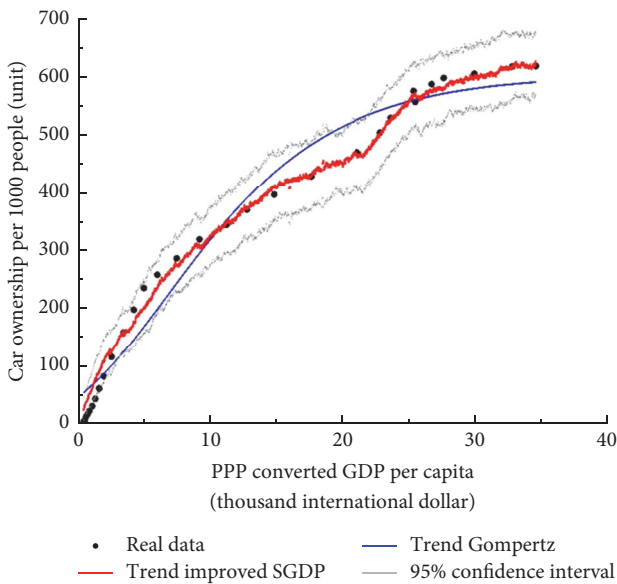


FIGURE 5: Real data versus Improved SDE Gompertz versus Trend Gompertz in Japan.

reaches the level of saturation or has quite large fluctuations in the path. As for Korea, as its path has a “perfect” feature of S-shaped curve, and almost no obvious fluctuations, the real data, general Gompertz curve, and the improved SGDP curve behave almost the same. Combining with the prediction results from Table 2, the general Gompertz behaves even slightly better than the improved SGDP curve.

Therefore, the improved SGDP model outstands when either there are features of the original curve: it has reached or almost reached the saturation level, or there is quite obvious fluctuations in the curve path. Otherwise, when the curve is a smooth S-shaped path, the general Gompertz performs

TABLE 4: Estimated parameters of improved SGDP in China.

$a$	$b$	$c$	$\beta_1$	$\beta_2$	$\alpha_3$ (noise)
30.790	8.308	0.028	1.190	0.028	1.26

approximately the same with the improved SGDP model, under which circumstance, we suggest using the general Gompertz as the estimation is much easier.

**4.2. The Projection of Vehicle Ownership to 2025 in China.** China is in the initial stage of mobility development and thus possesses a quite low quantity of vehicle ownership per 1000 people and a fast-growing trend. The vehicle ownership growth in the initial stage is quite unstable, with lots of possible fluctuations [13]. Therefore, we conduct the proposed improved SGDP model to the original data (civilian vehicles, with a time span of 1980 to 2014 (data resource: MarcoChina database (1980–2011) [http://www.macrochina.com.cn/macro\\_data/](http://www.macrochina.com.cn/macro_data/); National Bureau of Statistics (2010–2014) <http://www.stats.gov.cn/>)) in China. After regression, the prediction of vehicle ownership per 1000 people as well as the total vehicle quantity is made based on the parameters and predicted GDP per capita and population. As the improved SGDP model is suitable for short term predictions, thus we assume a predicted time span of up to year 2025. The estimated parameters and predicted values are presented in Tables 4 and 5, and the projection of vehicle ownership per 1000 people and total vehicle quantity in China up to year 2025 are depicted in Figures 7 and 8.

According to the prediction, China will maintain a fast-growing trend, and its vehicle ownership per 1000 people will reach 265 units in 2025, and a total vehicle quantity of over 350 million.

TABLE 5: Prediction of vehicle ownership per 1000 people and total ownership to 2025 in China.

Year	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
GDP <sup>1</sup> per capita	8451.81	9043.44	9631.26	10257.30	10924.02	11601.31	12320.59	13084.47	13895.70	14757.24
Population <sup>2</sup>	1381.53	1389.82	1398.16	1406.55	1414.99	1423.48	1427.75	1432.03	1434.89	1437.76
Vehicle per 1000 (unit)	133.60	148.56	164.01	181.07	199.90	219.73	241.53	249.02	256.74	264.70
Total vehicle (million units)	184.57	206.48	229.32	254.69	282.86	312.78	344.84	356.60	368.39	380.57

<sup>1</sup>GDP indicator (in international dollar) up to 2025 is obtained from the paper of Perkins and Rawski [15]; <sup>2</sup>population forecasting data (in million people) is obtained from the paper of Dadao [16].

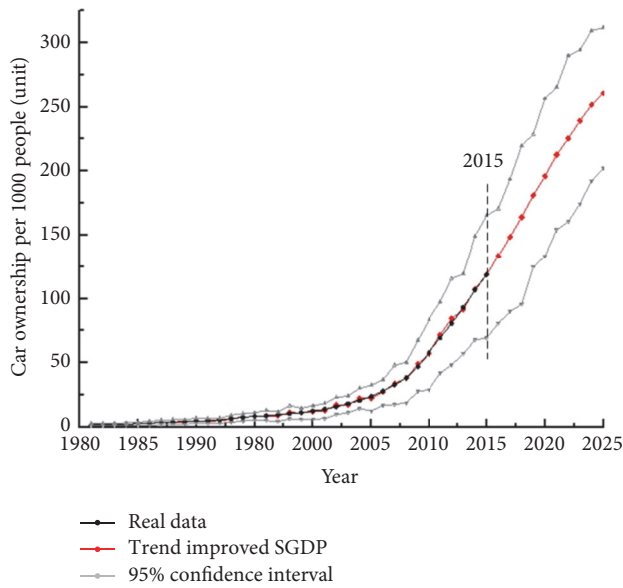


FIGURE 7: Projection of vehicle ownership per 1000 people to 2025 in China.

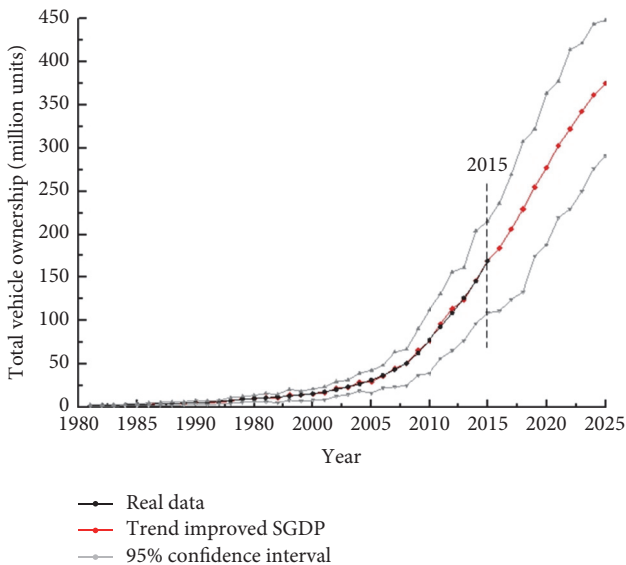


FIGURE 8: Projection of total vehicle ownership to 2025 in China.

### 5. Conclusion

This study introduces a new form of Gompertz function based on stochastic differential equation, aiming at depicting the growth pattern of vehicle ownership with economic driving. Different from previous models used for fitting the curve, the proposed improved SGDP model has an adjustment in the deterministic part and is transformed to the stochastic differential form. The general solution of SGDP model is presented in the present paper, and numerical studies are carried out based on a SDE Toolbox of Matlab Package. In comparison, the improved SGDP model has the following advantages and features:

- (i) Better fitting and predicting for countries with vehicle ownership reaching (or almost reaching) the saturation level: the improved SGDP model can reveal the slow growth in the rear and thus can obtain a better fitting curve and a more precise prediction in the short term.
- (ii) Better performance for vehicle ownership growth curves with fluctuations: when there is obvious fluctuations in the pattern, the stochastic nature and adjustment part can capture them in order to fit better and precisely capture the real projection; this is useful for the understanding of vehicle ownership growth.
- (iii) For growth patterns which is exactly the S-shaped curve, we suggest using the general Gompertz equation instead of the improved SGDP; they perform almost the same, and the computing work is much easier for the general Gompertz function.

Then we use the improved SGDP model to draw the prediction of Chinese vehicle ownership up to year 2025. The fitting curve performs quite well, and the prediction value of vehicle ownership per 1000 people and total vehicle quantity is over 250 units and over 350 million units in year 2025.

Future studies can focus on such aspects: firstly, the aggregate study usually follows on the relationship between vehicle ownership and economics factor; other influencing factors such as the public transportation service, vehicle policy restrictions, and car culture (proposed by research groups of IFMO (The Institute for Mobility Research by the BMW Group)) should be included to obtain a more comprehensive understanding of vehicle ownership revolution, a simple model with main influencing factors is highly recommended in the future study. Secondly, a mathematical



solution for the improved SGDP may be developed instead of the numerical outputs by Matlab Package. Last but not least, a more detailed model may be proposed according to different stages of growth, in order to better capture the features of vehicle ownership revolution.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This research was funded by the National Key Technology R&D Program (no. 2014BAG01B00) and The National Natural Science Funds (51408023).

## References

- [1] M. Simonsen and H. J. Walnum, "Energy chain analysis of passenger car transport," *Energies*, vol. 4, no. 2, pp. 324–351, 2011.
- [2] J. Dargay and D. Gately, "Income's effect on car and vehicle ownership, worldwide: 1960–2015," *Transportation Research Part A: Policy and Practice*, vol. 33, no. 2, pp. 101–138, 1999.
- [3] G. de Jong, J. Fox, A. Daly, M. Pieters, and R. Smit, "Comparison of car ownership models," *Transport Reviews*, vol. 24, no. 4, pp. 379–408, 2004.
- [4] D. Brownstone, D. S. Bunch, and K. Train, "Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles," *Transportation Research Part B: Methodological*, vol. 34, no. 5, pp. 315–338, 2000.
- [5] D. A. Hensher, P. O. Barnard, N. C. Smith, and F. W. Milthorpe, *Dimensions of Automobile Demand: A Longitudinal Study of Automobile Ownership and Use*, Elsevier, Amsterdam, Netherlands, 1992.
- [6] I. Hocherman, J. N. Prashker, and M. Ben-Akiva, "Estimation and use of dynamic transaction models of automobile ownership," *Transportation Research Record*, no. 944, 1983.
- [7] K. U. Leuven and Standard and Poor's DRI, *Auto-Oil II Cost-Effectiveness: Study Description of the Analytical Tools REMOVE 1.1*, Leuven, K. U., Standard and Poor's DRI, Leuven, Belgium, 1999.
- [8] Van den Broecke, "De mogelijke groei van het personenauto-bezit tot," Report for PbIVVS, BSR, Amsterdam, Netherlands, 2010.
- [9] J. C. Tanner, "A lagged model for car ownership forecasting," Tech. Rep. HS-036 436, 1983.
- [10] N. Meade and T. Islam, "Modelling and forecasting the diffusion of innovation—a 25-year review," *International Journal of Forecasting*, vol. 22, no. 3, pp. 519–545, 2006.
- [11] H. Liu, K. He, D. He et al., "Analysis of the impacts of fuel sulfur on vehicle emissions in China," *Fuel*, vol. 87, no. 13–14, pp. 3147–3154, 2008.
- [12] J. Dunkerley and I. Hoch, "The pricing of transport fuels," *Energy Policy*, vol. 14, no. 4, pp. 307–317, 1986.
- [13] S. Stares and Z. Liu, "China's urban transport development strategy: proceedings of a symposium in Beijing," Tech. Rep. 352, World Bank Publications, November 1995.
- [14] P. H. Franses, "A method to select between Gompertz and logistic trend curves," *Technological Forecasting and Social Change*, vol. 46, no. 1, pp. 45–49, 1994.
- [15] D. H. Perkins and T. G. Rawski, "Forecasting China's economic growth to 2025," *China's Great Economic Transformation*, pp. 829–886, 2008.
- [16] R. Q. H. Dadao, "Stochastic model for population forecast: based on leslie matrix and arma model," *Population Research*, no. 2, pp. 28–42, 2011.
- [17] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. Van't Riet, "Modeling of the bacterial growth curve," *Applied and Environmental Microbiology*, vol. 56, no. 6, pp. 1875–1881, 1990.
- [18] C. E. Finch and M. C. Pike, "Maximum life span predictions from the Gompertz mortality model," *Journals of Gerontology Series A Biological Sciences and Medical Sciences*, vol. 51, no. 3, pp. B183–B194, 1996.
- [19] M. Z. Acutt and J. S. Dodgson, "Transport and global warming: modelling the impacts of alternative policies," in *Transport Policy and the Environment*, pp. 20–37, 1998.
- [20] Y. N. Wang, "Car ownership forecast in china—an analysis based on gompertz equation," *Research On Financial and Economic Issues*, vol. 11, 2005.
- [21] H. M. Zhao, "Vehicle ownership per 1000 people prediction in mid-term and long-term in China based on Gompertz curve," *Industrial Technology and Economics*, vol. 7, pp. 7–23, 2012.
- [22] R. Gutierrez-Jaimez, P. Román, D. Romero, J. J. Serrano, and F. Torres, "A new Gompertz-type diffusion process with application to random growth," *Mathematical Biosciences*, vol. 208, no. 1, pp. 147–165, 2007.
- [23] R. M. Capocelli and L. M. Ricciardi, "Growth with regulation in random environment," *Kybernetik*, vol. 15, no. 3, pp. 147–157, 1974.
- [24] R. Gutiérrez, A. Nafidi, and R. G. Sánchez, "Forecasting total natural-gas consumption in Spain by using the stochastic Gompertz innovation diffusion model," *Applied Energy*, vol. 80, no. 2, pp. 115–124, 2005.
- [25] L. Ferrante, S. Bompadre, L. Possati, and L. Leone, "Parameter estimation in a Gompertzian stochastic model for tumor growth," *Biometrics*, vol. 56, no. 4, pp. 1076–1081, 2000.
- [26] H. Lu, H. Ma, and W. Kuang, "Analysis and prediction of mobility development patterns in China," in *Proceedings of the 15th COTA International Conference of Transportation Professionals*, July 2015.



## Research Article

# Exploring the Influence of Attitudes to Walking and Cycling on Commute Mode Choice Using a Hybrid Choice Model

Chuan Ding,<sup>1,2</sup> Yu Chen,<sup>1,2</sup> Jinxiao Duan,<sup>1,2</sup> Yingrong Lu,<sup>1,2</sup> and Jianxun Cui<sup>3</sup>

<sup>1</sup>*School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure System and Safety Control, Beihang University, Beijing 100191, China*

<sup>2</sup>*Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Si Pai Lou No. 2, Nanjing 210096, China*

<sup>3</sup>*School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150090, China*

Correspondence should be addressed to Jianxun Cui; [cuijianxun@hit.edu.cn](mailto:cuijianxun@hit.edu.cn)

Received 2 February 2017; Accepted 19 April 2017; Published 21 May 2017

Academic Editor: Guohui Zhang

Copyright © 2017 Chuan Ding et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Transport-related problems, such as automobile dependence, traffic congestion, and greenhouse emissions, lead to a great burden on the environment. In developing countries like China, in order to improve the air quality, promoting sustainable travel modes to reduce the automobile usage is gradually recognized as an emerging national concern. Though there are many studies related to the physically active modes (e.g., walking and cycling), the research on the influence of attitudes to active modes on travel behavior is limited, especially in China. To fill up this gap, this paper focuses on examining the impact of attitudes to walking and cycling on commute mode choice. Using the survey data collected in China cities, an integrated discrete choice model and the structural equation model are proposed. By applying the hybrid choice model, not only the role of the latent attitude played in travel mode choice, but also the indirect effects of social factors on travel mode choice are obtained. The comparison indicates that the hybrid choice model outperforms the traditional model. This study is expected to provide a better understanding for urban planners on the influential factors of green travel modes.

## 1. Introduction

In recent years, the increasing automobile ownership and usage cause serious traffic problems and lead to a large amount of greenhouse gas emissions. The commuting trip, as one of the most important travel demands, often occurs at fixed times and contributes regular pressures to the traffic system. Unlike automobile modes, the nonmotorized travel modes (e.g., walking and cycling) are widely considered as sustainable patterns with low emissions [1]. Besides, using the travel modes of walking and cycling to work provides opportunities for people to get physical exercise. A reasonable amount of physical activities related to travel can help people keep healthy, which appears to be important for people in modern society. Bassett et al. [2] found that higher levels of active travel are usually correlated with a lower obesity rate. Briefly, the modes of walking and cycling may be

the attractive alternatives for the short-distance commuting travels, especially in the compact China cities.

A literature review has identified a wide range of factors influencing an individual walking and cycling mode choice, such as sociodemographic factors, road infrastructure features, and environmental variables [3–10]. There is a growing body of research that focuses on the influence of psychological factors on the individual travel mode choice decision [11]. For example, the attitudes to the nonmotorized travel mode are perhaps one of the most important factors influencing the decision to walk or cycle. In developing countries like China, in order to improve the air quality, promoting sustainable travel modes to reduce the automobile usage is gradually recognized as an emerging national concern. Though there are many studies related to the physically active modes (e.g., walking and cycling), the research on the influence of attitudes to active modes on travel behavior is limited,

especially in China. To fill up this gap, this paper focuses on examining the impact of attitudes to walking and cycling on commute mode choice and capturing the mediating role of attitudes played in travel mode choice decision.

The remainder of this paper is organized as follows. The following section presents a literature review related to our study. The third section describes the modeling approach used in this study. Data sources and description are provided in the fourth section. In the following section, empirical model results are analyzed. In the end, the conclusions and limitations of this study are provided.

## 2. Literature Review

With the environmental and healthy advantages in mind, the nonmotorized travel modes gain a growing attention. The influences of sociodemographic factors and built environment on the modes of walking and cycling have been widely investigated. An and Chen [8] found that the nonmotorized mode choice was strongly influenced by the factors of employment density, household income, and average sidewalk length. Using the logit model, Plaut [12] found that higher household income and housing price were correlated with lower propensity to walk or bicycle. As to the other sociodemographic factors (e.g., gender, age, and car ownership), similar empirical studies were conducted [6, 8, 12–15]. With regard to the influence from built environment, it is widely confirmed that the factors of density, diversity, and mixed land use have significant influences on the travel modes of walking and cycling [8, 9, 16, 17]. Furthermore, there is an undeniable fact that travel mode choice behavior is affected, not only by the attributes of the modes themselves, but also by the unobserved factors. The cognitive psychology theory claims that preferences and behavior are correlated with perceptions and attitudes [18]. Generally, the attitudes determine the behavioral intentions, which are associated with the individual heterogeneity. The individual heterogeneity is a reflection of individual tastes, needs, values and goals, which is affected by experience, education, and so forth [19–22].

In recent years, many exiting studies have adopted travel choice model by involving psychological variables (e.g., attitude and perception). Using a discrete choice model with latent variables, Johansson et al. [21] examined the commuter travel mode choice behavior. It is found that the environmental preferences increase the possibility of selecting a train mode. Using the data collected from a stated preference survey, Maldonado-Hinarejos et al. [23] incorporated the latent variables of attitudes, perceptions, and security concerns on bicycle use into the travel choice model. These study results indicate the necessary role of latent variables played in the multinomial logit choice model, and probike attitudes have positively significant effect on the cycling mode choice. Similar findings relate to the study conducted by Dill and Voros [24]. Moreover, cyclists usually take cycling as a healthy and environmental travel mode [25]. Using the survey data collected in the San Francisco Bay Area, Choo and Mokhtarian [26] found that travel attitudinal factors and

personality characteristics significantly influenced individual vehicle type selections. Nurul Habib [14] examined the effects of willingness to walk, walking trip propensity, and walking distance on walking trip and indicated that young people are less possible to walk and females are more possible to walk. In particular, using advanced stated preferences survey data, Kamargianni and Polydoropoulou [27] examined the influence of teenagers' attitudes towards walking and cycling on mode choice behavior. It is found that willingness to walk and to cycle has a positive effect on the choice of those alternatives and a negative effect on the choice of a car.

With respect to the modeling methods, incorporating the latent variable into the discrete choice model is widely used. Though there is a growing literature on the influence of latent variable on travel mode choice, limited efforts have been made to capture the mediating role of the attitudes to travel mode. Different to the traditional discrete choice modeling approach, structural equation model (SEM) can not only obtain the direct effect, but also gain the indirect effect and total effect [28, 29]. As shown in Figure 1,  $X$ ,  $Y$ , and  $M$  are exogenous variable, endogenous variable, and mediating variable, respectively. Assume that the direct effect of exogenous variable  $X$  on the endogenous variable  $Y$  is  $\gamma$ , and the indirect effect of exogenous variable  $X$  on the endogenous variable  $Y$  is  $\alpha \times \beta$ . Hence, the total effect of exogenous variable  $X$  on the endogenous variable  $Y$  is the sum of direct effect and indirect effect (i.e.,  $\gamma + \alpha \times \beta$ ). In the previous studies, though the latent factors are considered in the travel choice model, only the direct effects are taken into account without the indirect effect path. In this context, it would lead to inaccurate results, especially when the signs of the direct effect and indirect effect of the factor are opposite.

According to the literature, it is found that most existing studies only used the objective variables to investigate travel mode choice. A growing body of research is conducted to account for the objective variables (e.g., sociodemographic, modal attributes) and subjective variables (e.g., attitudes, perceptions) simultaneously. However, limited studies have been made on the influences of attitudes to walking and cycling on commute mode choice. Moreover, most studies are empirically examined in the western countries, while for the eastern countries like China, where the urban development level, traffic conditions, and living habits are quite different from western countries, more attention should be obtained. To fill up this gap, this paper focuses on examining the impact of attitudes to walking and cycling on commute mode choice. Using the survey data collected in China cities, an integrated discrete choice model and the structural equation model are proposed. By applying the hybrid choice model, not only the role of the latent attitude played in travel mode choice, but also the indirect effects of social factors on travel mode choice are obtained. The comparison indicates that the hybrid choice model outperforms the traditional model.

## 3. Modeling Approach

As a new kind of discrete choice models, hybrid choice model (HCM) combines the discrete choice model and the

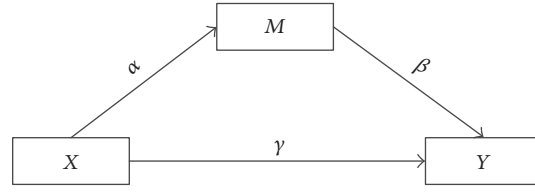


FIGURE 1: Modeling indirect effect diagram.

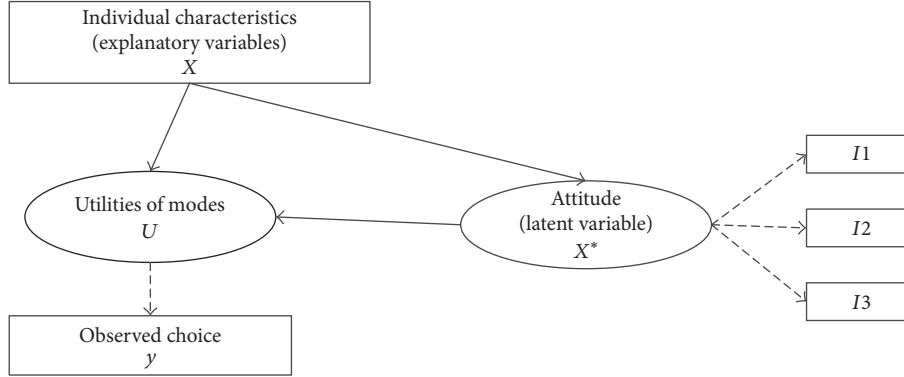


FIGURE 2: Modeling framework.

latent variable model in one framework. The most general framework has been proposed by Ben-Akiva et al. [30, 31] and it consists of two components. The measurement model component describes the relationship between the indicators and its corresponding latent variable, while the structural model component describes the complex relationships among the exogenous variable and endogenous variable. In this paper, we aim to examine the influences of objective variables and psychological variables on choosing the modes of walking or cycling to work. For this purpose, we construct the HCM framework with a latent variable “attitude towards walking or cycling.” To measure the latent variable, three attitudinal indicators are used. The modeling framework is presented in Figure 2.

For the HCM framework, the structural model and the measurement model are described as follows.

Structural equations part:

$$\begin{aligned} X_n^* &= \gamma X_n + \sigma_n, \quad \sigma_n \sim N(0, 1), \\ U_n &= \beta_n X_n + \Gamma X_n^* + \varepsilon_n, \end{aligned} \quad (1)$$

where  $X^*$  is the latent variable,  $X_n$  is the exogenous variable, and  $\sigma_n$  is the random error term.  $U_n$  is the utility of the nonmotorized travel mode, and  $\varepsilon_n$  is the independently, identically distributed (i.i.d.) extreme value.  $\beta_n$ ,  $\gamma$ , and  $\Gamma$  are the estimated parameters.

Measurement equations part:

$$\begin{aligned} I_n &= \alpha_n + \lambda_n X_n + v_n, \\ y_n &= \begin{cases} 1, & \text{if } U_i = \max(U_j) \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \quad (2)$$

where  $I_n$  is the indicator of the latent variable ( $X^*$ ),  $v_n$  is the random error term, and  $y$  is a choice indicator, taking the value one if the nonmotorized travel mode is chosen, and 0 otherwise. In this study, the maximum likelihood techniques are used to estimate the model parameters. For the HCM method, the likelihood function for a given observation is the joint probability of observing the travel mode choice and the attitudinal indicators that can be obtained as follows:

$$\begin{aligned} f(y_n, I_n | X_n; \alpha, \beta, \lambda) &= \int_{X^*} P(y_n | X_n, X^*; \beta) \\ &\cdot f(I | X_n, X^*; \alpha) f(X^* | X_n; \lambda) dX^*. \end{aligned} \quad (3)$$

In order to examine the role of latent attitude variable played in travel mode choice behavior, a comparison was conducted between the traditional model and the proposed model. Based on the hybrid choice model, the indirect effects and total effects of social factors on travel mode choice through the mediating latent variable are calculated. Therefore, the intermediary nature of the attitudes to walking and cycling on commute mode choice would be confirmed.

## 4. Data Sources and Description

**4.1. Questionnaire Design.** The survey questionnaire is mainly composed of three parts, as shown in Table 1. The first two parts are to collect the individual and household characteristics (e.g., age, gender, education, occupation, income, bus card ownership, driver license, household children, bicycle ownership, and car ownership). The third part aims to collect the respondents' attitudes towards walking and cycling. Considering the immeasurability of subjective factors, three indicators were designed to measure

TABLE 1: Objective variables used in models and description.

Variable name	Variable description
<i>Individual characteristics</i>	
Age	1 = below 35 years old; 2 = 35 to 55 years old; 3 = over 55 years old
Gender	1 = male; 2 = female
Education	1 = low (junior school); 2 = medium (junior college); 3 = high (bachelor, master, or Ph.D.)
Occupation	1 = government-related job; 2 = others
Income	1 = less than 2000 ¥; 2 = 2000–8000 ¥; 3 = more than 8000 ¥
Bus card	1 = individual with a bus card; 2 = without a bus card
Driver license	1 = individual with a driver license; 2 = without a driver license
<i>Household characteristics</i>	
Household children	1 = household with one or more children; 2 = without children
Car ownership	1 = household with one or more cars available; 2 = without
Bicycle ownership	1 = household with one or more bicycles available; 2 = without
<i>Mode choice</i>	
Mode choice	1 = walking and cycling; 2 = others

the respondents' attitudes towards walking and cycling in this study. The indicators are described as follows: nonmotorized travel mode can help to improve environmental pollution ( $I_1$ ), walking or cycling to work can get physical exercise and keep fit ( $I_2$ ), and nonmotorized travel mode can satisfy daily travel ( $I_3$ ). Agreements or disagreements of all those descriptions were measured by five-point Likert Scale: disagree strongly, disagree a little, neither agree nor disagree, agree a little, and agree strongly which were coded as 1, 2, 3, 4, and 5, respectively.

The survey was conducted in Zhenjiang city during the year 2015 in China. Zhenjiang city is located in the southwest of Jiangsu Province, the lower reaches of the Yangtze River. With regard to the political status, economic development, and city scale, Zhenjiang belongs to China's third-tier cities. The survey data used in this paper was obtained from the OpenITS website by Jiangsu University. At last, 2941 respondents completed the surveys, and 2660 valid samples were selected. It is worth mentioning that nearly half (41.7%) of the respondents live within five kilometers from workplace. Generally speaking, there is a possibility for the respondents to choose walking or cycling to commute. As is well known, for the long-distance travel, it is more likely to choose motorized travel modes to commute [6, 32]. For the small cities like Zhenjiang, it is meaningful and valuable to investigate the walking or cycling travel behavior. In this study, considering the general travel distance of walking and cycling, the respondents with more than five kilometers' commuting distance were removed. Finally, 1110 respondents were selected for further study.

**4.2. Descriptive Statistics.** In the final sample, 59.7% of the respondents are males, and 40.3% are females. 74.3% of the respondents are below 35 years old, 22.0% are between 35 and 55 years old, and 3.7% are above 55 years old. With regard to the education level, 11.9% of the respondents are below junior college level, 39.8% have undergraduate degrees, and 48.3%

have graduate degrees. In terms of the occupation, 8.0% of the respondents are government staffs. As to the income, 40.8% of the respondents are low income people, 55.1% are the middle-income group, and 4.1% are high-income group. As for the household children, 49.3% of the respondents' household have children and 50.7% of the respondents have no kids. According to the survey results, 51.5% of the respondents choose walking or cycling to commute. Figure 3 presents the respondents' responses to the attitude towards nonmotorized travel. Descriptive statistics of three indicators of the latent variable are shown in Table 2.

## 5. Result Analysis

**5.1. Model Fit.** The model parameters were estimated using the maximum likelihood method based on the *M*-plus software with 1,000 bootstrap draws. The model fit information is listed in Table 3, indicating that the hybrid choice model obtains a good fit [33]. With regard to the internal consistency of the indicators, as shown in Table 4, Cronbach's alpha value is larger than the threshold value of 0.60, indicating that the selected indicators are reliable to measure the underlying latent variable [34]. Meanwhile, for each indicator variable, the factor loading coefficient is significant at the 99% level. In other words, all the observed indicator variables contribute to capture of the unobservable latent variable.

**5.2. Modeling Results.** Using the survey data collected in China cities, a comparison between the traditional model and the hybrid choice model is conducted. The estimation results of both models are shown in Table 5. Comparing the traditional model and the hybrid choice model, it is obviously seen that the integrated model indeed provides greater explanatory power with respect to the travel mode choice, indicating that incorporating the latent variable into the discrete choice model improved the overall fitness of

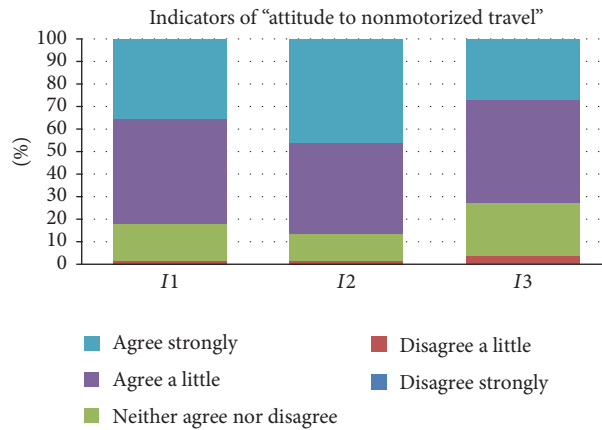


FIGURE 3: Indicators of latent variable.

TABLE 2: Latent variable and indicators.

Indicators	Indicator description	Mean	St. Dev.
$I_1$	Nonmotorized travel mode can help to improve environmental pollution	4.16	0.747
$I_2$	Walking or cycling to work can get physical exercise and keep fit	4.31	0.738
$I_3$	I would like to choose nonmotorized travel mode to satisfy daily travel	3.95	0.812

Note. 1 = disagree strongly, 2 = disagree a little, 3 = neither agree nor disagree, 4 = agree a little, 5 = agree strongly.

TABLE 3: Goodness-of-fit of model.

Indicators	Description	Values	Cut-off value
$\chi^2$	Measuring the differences between the observed covariance matrices and model-based covariance matrices.	22.88	Smaller $\chi^2$ shows better model fit.
CFI	Measuring noncenter parameter improvement.	0.97	>0.90
TLI	Measuring the discrepancy between the observed sample matrix and the theory matrix.	0.99	>0.90
RMSEA	Measuring the difference of each degree of freedom.	0.01	<0.05
SRMR	Measuring the approximation error of each degree of freedom.	0.01	<0.05

Note. CFI is Comparative Fit index; TLI is Tucker Lewis index; SRMR is Standardized Root Mean Square Residual; RMSEA is Root Mean Square Error of Approximation.

TABLE 4: Factor loading coefficients for the indicators of the latent variable.

Indicators	Cronbach's alpha	Attitude towards nonmotorized travel	
		Parameter	t-stat
$I_1$	0.620	0.590	18.727
$I_2$		0.622	19.736
$I_3$		0.645	20.309

the model. Specifically, the likelihood ratio index improves from 0.313 to 0.364. Besides, the Akaike information criterion (AIC) and adjusted Bayesian information criterion (BIC) of the integrated model are lower [35]. And more importantly, the influence of the latent attitude variable is positively significant at the 90% level, indicating that the attitude to walking and cycling plays a critical role in the nonmotorized

travel mode choice. From this point, this finding might be very helpful to encourage the green modes [36].

As to the individual characteristics, there is no difference between the male and female for the nonmotorized travel mode choice. However, it is significantly related to the factors of age, education, occupation, income, and household children. Specifically, younger and older travelers are more



TABLE 5: Estimation results for the traditional and the hybrid choice model.

Variables	Traditional model		Hybrid choice model	
	Parameter	<i>t</i> -stat	Parameter	<i>t</i> -stat
Constant	0.547	4.433**	0.547	4.433**
<i>Household characteristics</i>				
Household children	-0.058	-1.959*	-0.058	-1.969**
Bicycle ownership	0.243	8.444**	0.245	8.518**
Car ownership	-0.046	-1.471	-0.045	-1.431
Bus card	-0.042	-1.445	-0.043	-1.485
Driver license	-0.064	-2.105**	-0.066	-2.148**
<i>Individual characteristics</i>				
Gender	0.016	0.573	0.013	0.489
<i>Age</i>				
Age-1	0.115	3.368**	0.113	3.310**
Age-3	0.059	2.086**	0.059	2.056**
<i>Education</i>				
Education-1	0.005	1.624	0.053	1.711*
Education-3	0.089	2.780**	0.087	2.713**
Government	-0.054	-1.969**	-0.056	-2.036**
<i>Income</i>				
Income-1	0.171	5.494**	0.165	5.259**
Income-3	-0.011	-0.395	-0.013	-0.453
<i>Latent variables</i>				
Attitude	—	—	0.056	1.673*
Observations		1110		1110
LRI		0.313		0.364
AIC		8600.808		8267.361
Adjust BIC		8670.571		8311.986

Note. LRI is likelihood ratio index,  $LRI = 1 - (LL\kappa/LL\kappa_0)$ , and  $LL\kappa_0$  is the log-likelihood value when all the parameters are set equal to zero; AIC is Akaike information criterion; BIC is Bayesian information criterion; \* indicates significant values at the 90% level; \*\* indicates significant values at the 95% level.

likely to choose the walking or cycling mode to commute than middle-aged travelers. With respect to the educational attainment, it is found that higher level and lower level of them were both associated with a larger likelihood of choosing the walking or cycling mode. In addition, as expected, the people with low income tend to use the active travel mode choices. Government-related people are less likely to use the active travel mode choices. In terms of the household characteristics, it is found that the people from the household with children show less intention to choose the walking or cycling mode to commute. This may be due to the fact that it would be more convenient for parents to pick up their children on the way from/to work. As expected, the people owning a bicycle are more likely to choose active travel mode to commute, while the factor of driver license has a significantly negative effect. Therefore, in this context, in order to promote the walking or cycling mode, it is a feasible strategy to provide vast bicycles and improve the bike-sharing service in China cities (e.g., Ofo, Mobike).

*5.3. Indirect Effects and Total Effects.* With the advantage of hybrid choice model in mind, as displayed in Table 6, the indirect effect and the total effect are also calculated. For the

model results, though the indirect effects of all the observable variables are insignificant, they do have negligible influences on the total final effects of the variables on walking or cycling mode choice. The total effects are the outcome of the direct effects and the indirect effects. Due to the mediating effect of the latent variable, the effect of variables on the mode choice may be strengthened or weakened [37]. As the model results shown, the indirect effects of education and bicycle ownership are negative, while the direct effects and the total effects are both positive. It means that the positive effect of low level education and bicycle ownership are both weakened for the intermediary role of the attitude to walking or cycling. Therefore, the indirect effect of the factors on travel mode choice cannot be ignored. Similar examples relate to the factors of job type, children, bus card, and driver license.

## 6. Conclusions

Transport-related problems, such as automobile dependence, traffic congestion, and greenhouse emissions, lead to a great burden on the environment. In developing countries like China, in order to improve the air quality, promoting sustainable travel modes to reduce the automobile usage



TABLE 6: Estimation results for the indirect and total effects of observed variables.

Variables	Indirect effect		Walking and cycling		Total effect	
	Parameter	<i>t</i> -stat	Parameter	<i>t</i> -stat	Parameter	<i>t</i> -stat
Gender	0.002	0.907	0.016	0.573		
Age						
Age-1	0.002	0.686	0.115	3.368**		
Age-3	0.001	0.394	0.059	2.086**		
Education						
Education-1	-0.003	-0.931	0.050	1.624		
Education-3	0.002	0.763	0.089	2.780**		
Government	0.002	0.766	-0.054	-1.969**		
Income						
Income-1	0.006	1.396	0.171	5.494**		
Income-3	0.002	0.676	-0.011	-0.395		
Household children	0.000	0.104	-0.058	-1.959**		
Bicycle ownership	-0.002	-0.803	0.243	8.444**		
Car ownership	-0.001	-0.500	-0.046	-1.471		
Bus card	0.001	0.477	-0.042	-1.445**		
Driver license	0.001	0.507	-0.064	-2.105**		

Note. \* indicates significant values at the 90% level. \*\* indicates significant values at the 95% level.

is gradually recognized as an emerging national concern. Though there are many studies related to the physically active modes (e.g., walking and cycling), the research on the influence of attitudes to active modes on travel behavior is limited, especially in China. Hence, this paper focuses on examining the impact of attitudes to walking and cycling on commute mode choice.

Using the survey data collected in China cities, an integrated discrete choice model and the structural equation model are proposed. By applying the hybrid choice model, not only the role of the latent attitude played in travel mode choice, but also the indirect effects of social factors on travel mode choice are obtained. The comparison indicates that the hybrid choice model outperforms the traditional model. This study is expected to provide a better understanding for urban planners on the influential factors of green travel modes. For this study, it should be noted that the built environment has an important effect on travel mode choice [38]. However, due to the availability of land use data, the built environment measurements are not included in the model. For future studies, it is necessary to incorporate the built environment factors and attitudes into the travel behavior model.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (71503018, U1564212, and U1664262).

## References

- [1] N. Cavill, H. Rutter, and A. Hill, "Action on cycling in primary care trusts: results of a survey of Directors of Public Health," *Public Health*, vol. 121, no. 2, pp. 100–105, 2007.
- [2] D. R. Bassett, J. Pucher Jr., R. Buehler, D. L. Thompson, and S. E. Crouter, "Walking, cycling, and obesity rates in Europe, North America and Australia," *Journal of Physical Activity and Health*, vol. 5, no. 6, pp. 795–814, 2008.
- [3] A. Goodman, "Walking, cycling and driving to work in the english and welsh 2011 census: trends, socio-economic patterning and relevance to travel behaviour in general," *PLoS ONE*, vol. 8, no. 8, Article ID e71790, 2013.
- [4] R. Cervero, "Built environments and mode choice: toward a normative framework," *Transportation Research Part D: Transport and Environment*, vol. 7, no. 4, pp. 265–284, 2002.
- [5] Y. O. Susilo and K. Maat, "The influence of built environment to the trends in commuting journeys in the Netherlands," *Transportation*, vol. 34, no. 5, pp. 589–609, 2007.
- [6] G. Vandenbulcke, C. Dujardin, I. Thomas et al., "Cycle commuting in Belgium: Spatial determinants and 're-cycling' strategies," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 2, pp. 118–137, 2011.
- [7] C. L. Antonakos, "Environmental and travel preferences of cyclists," *Transportation Research Record*, vol. 1438, pp. 25–33, 1994.
- [8] M. An and M. Chen, "Estimating Nonmotorized Travel Demand," *Transportation Research Record*, vol. 2002, no. 1, pp. 18–25, 2007.
- [9] D. A. Rodríguez and J. Joo, "The relationship between non-motorized mode choice and the local physical environment," *Transportation Research Part D: Transport and Environment*, vol. 9, no. 2, pp. 151–173, 2004.

- [10] E. Heinen, K. Maat, and B. van Wee, "The effect of work-related factors on the bicycle commute mode choice in the Netherlands," *Transportation*, vol. 40, no. 1, pp. 23–43, 2013.
- [11] J. Anable, "Complacent Car Addicts; or 'Aspiring Environmentalists'? Identifying travel behaviour segments using attitude theory," *Transport Policy*, vol. 12, no. 1, pp. 65–78, 2005.
- [12] P. O. Plaut, "Non-motorized commuting in the US," *Transportation Research Part D: Transport and Environment*, vol. 10, no. 5, pp. 347–356, 2005.
- [13] T. Ryley, "Estimating cycling demand for the journey to work or study in West Edinburgh, Scotland," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1982, pp. 187–193, 2006.
- [14] K. Nurul Habib, X. Han, and W. H. Lin, "Joint modelling of propensity and distance for walking-trip generation," *Transportmetrica A: Transport Science*, vol. 10, no. 5, pp. 420–436, 2014.
- [15] J. Parkin, M. Wardman, and M. Page, "Estimation of the determinants of bicycle mode share for the journey to work using census data," *Transportation*, vol. 35, no. 1, pp. 93–109, 2008.
- [16] M. Khan, K. M. Kockelman, and X. Xiong, "Models for anticipating non-motorized travel choices, and the role of the built environment," *Transport Policy*, vol. 35, pp. 117–126, 2014.
- [17] J. D. Hunt and J. E. Abraham, "Influences on bicycle use," *Transportation*, vol. 34, no. 4, pp. 453–470, 2007.
- [18] I. Ajzen, *Attitudes, Personality, and Behavior*, McGraw-Hill Education, UK, 2005.
- [19] R. A. Daziano and D. Bolduc, "Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model," *Transportmetrica A: Transport Science*, vol. 9, no. 1, pp. 74–106, 2013.
- [20] D. Bolduc, N. Boucher, and R. Alvarez-Daziano, "Hybrid choice modeling of new technologies for car choice in Canada," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2082, pp. 63–71, 2008.
- [21] M. V. Johansson, T. Heldt, and P. Johansson, "The effects of attitudes and personality traits on mode choice," *Transportation Research Part A: Policy and Practice*, vol. 40, no. 6, pp. 507–525, 2006.
- [22] D. Mcfadden, M. Ben-Akiva, and T. Morikawa, "Discrete choice models incorporating revealed preferences and psychometric data," *Advances in Econometrics*, vol. 16, no. 1, pp. 29–55, 2002.
- [23] R. Maldonado-Hinarejos, A. Sivakumar, and J. W. Polak, "Exploring the role of individual attitudes and perceptions in predicting the demand for cycling: a hybrid choice modelling approach," *Transportation*, vol. 41, no. 6, pp. 1287–1304, 2014.
- [24] J. Dill and K. Voros, "Factors affecting bicycling demand: Initial survey findings from the Portland, Oregon, region," *Transportation Research Record*, vol. 2031, pp. 9–17, 2007.
- [25] B. Gatersleben and K. M. Appleton, "Contemplating cycling to work: attitudes and perceptions in different stages of change," *Transportation Research Part A: Policy and Practice*, vol. 41, no. 4, pp. 302–312, 2007.
- [26] S. Choo and P. L. Mokhtarian, "What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice," *Transportation Research Part A: Policy and Practice*, vol. 38, no. 3, pp. 201–222, 2004.
- [27] M. Kamargianni and A. Polydoropoulou, "Hybrid choice model to investigate effects of teenagers' attitudes toward walking and cycling on mode choice behavior," *Transportation Research Record*, no. 2382, pp. 151–161, 2013.
- [28] K. G. Jöreskog and A. S. Goldberger, "Estimation of a model with multiple indicators and multiple causes of a single latent variable," *Journal of the American Statistical Association*, vol. 70, no. 351, part 1, pp. 631–639, 1975.
- [29] Y. Zhu, Y. Wang, and C. Ding, "Investigating the influential factors in the metro choice behavior: evidences from Beijing, China," *KSCE Journal of Civil Engineering*, vol. 20, no. 7, pp. 2947–2954, 2016.
- [30] M. Ben-Akiva, D. Mcfadden, K. Train, J. Walker, C. Bhat, M. Bierlaire et al., "Hybrid choice models: progress and challenges," *Marketing Letters*, vol. 13, no. 3, pp. 163–175, 2002.
- [31] M. Ben-Akiva, J. Walker, A. T. Bernardino, D. A. Gopinath, T. Morikawa, and A. Polydoropoulou, "Integration of choice and latent variable models," *Perpetual Motion: Travel Behaviour Research Opportunities and Application Challenges*, pp. 431–470, 2002.
- [32] N. C. McDonald, "Children's mode choice for the school trip: The role of distance and school location in walking to school," *Transportation*, vol. 35, no. 1, pp. 23–35, 2008.
- [33] R. C. MacCallum, M. W. Browne, and H. M. Sugawara, "Power analysis and determination of sample size for covariance structure modeling," *Psychological Methods*, vol. 1, no. 2, pp. 130–149, 1996.
- [34] R. P. Bagozzi and Y. Yi, "On the evaluation of structural equation models," *Journal of the Academy of Marketing Science*, vol. 16, no. 1, pp. 74–94, 1988.
- [35] C. Ding, S. Mishra, G. Lu, J. Yang, and C. Liu, "Influences of built environment characteristics and individual factors on commuting distance: a multilevel mixture hazard modeling approach," *Transportation Research Part D: Transport and Environment*, vol. 51, pp. 314–325, 2017.
- [36] X. Ma, Y. J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart car data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.
- [37] C. Ding, D. Wang, C. Liu, Y. Zhang, and J. Yang, "Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance," *Transportation Research Part A: Policy and Practice*, vol. 100, pp. 65–80, 2017.
- [38] C. Ding, Y. Lin, and C. Liu, "Exploring the influence of built environment on tour-based commuter mode choice: a cross-classified multilevel modeling approach," *Transportation Research Part D: Transport and Environment*, vol. 32, pp. 230–238, 2014.

## Research Article

# Clustering Vehicle Temporal and Spatial Travel Behavior Using License Plate Recognition Data

Huiyu Chen, Chao Yang, and Xiangdong Xu

*Key Laboratory of Road and Traffic Engineering, Tongji University, 4800 Cao'an Road, Shanghai 201804, China*

Correspondence should be addressed to Chao Yang; [tongjiyc@tongji.edu.cn](mailto:tongjiyc@tongji.edu.cn)

Received 27 December 2016; Revised 13 March 2017; Accepted 2 April 2017; Published 24 April 2017

Academic Editor: Guohui Zhang

Copyright © 2017 Huiyu Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Understanding travel patterns of vehicle can support the planning and design of better services. In addition, vehicle clustering can improve management efficiency through more targeted access to groups of interest and facilitate planning by more specific survey design. This paper clustered 854,712 vehicles in a week using *K*-means clustering algorithm based on license plate recognition (LPR) data obtained in Shenzhen, China. Firstly, several travel characteristics related to temporal and spatial variability and activity patterns are used to identify homogeneous clusters. Then, Davies-Bouldin index (DBI) and Silhouette Coefficient (SC) are applied to capture the optimal number of groups and, consequently, six groups are classified in weekdays and three groups are sorted in weekends, including commuting vehicles and some other occasional leisure travel vehicles. Moreover, a detailed analysis of the characteristics of each group in terms of spatial travel patterns and temporal changes are presented. This study highlights the possibility of applying LPR data for discovering the underlying factor in vehicle travel patterns and examining the characteristic of some groups specifically.

## 1. Introduction

The trip starting and ending time, travel distance, travel frequency, activity duration, and some analogous features are the typical form of vehicle travel behaviors. All these aspects have a significant effect on the traffic condition in a direct or indirect way [1, 2]. For example, the distribution of the trip starting and ending time of all vehicles will decide the peak-hour time. Better understanding of these characteristics will be helpful to analyze the travel pattern and travel mode of vehicles. Identifying homogeneous travel behavior groups has been the research subject in several prior studies and the travel behavior analysis has always attracted great interest of transport authorities, since vehicle travel behavior has a vital impact on strategic and operational decisions [3–5].

Clustering is one of the most important methods to count and mine meaningful information in large amount of data since understanding the main differences between groups can contribute to a better understanding of their travel behaviors, which can provide valuable information for transportation planning [6]. Meanwhile, clustering vehicles based on their travel characteristics is one of the vital methods for studying

the representativeness of specific groups among the whole vehicle population and the travel profile of each group provides an aggregated characterization for the vehicles of a group as a whole [7]. It can also provide transportation planners with richer travel demand information for improving the system performance or better assessing network investments.

In the field of transportation, clustering has been widely accepted in dealing with big data and traffic problems [8, 9]. Reference [10] investigated the determination of historical traffic patterns by means of Ward's hierarchical clustering procedure. It classifies the traffic patterns in highways with the data collected by automatic vehicle identification (AVI) system into four groups and the resultant weekday traffic patterns can be used as input for macroscopic traffic models and as a basis for traffic management. Moreover, when predicting traffic flows based on historical data, a preclassification (e.g., holidays, Mondays, core weekdays, and Fridays) can be made to guide the authorities, and these patterns can be used to detect and replace erroneous data and to impute missing data.

Besides, [11, 12] utilized the density-based clustering algorithms to classify trajectories using GPS data. The study

of trajectory data can reveal individual trajectory patterns, understand the characteristics of human dynamics, and thus support trajectory prediction, urban planning, traffic monitoring, and so forth. The characteristics will be similar inside each group and significantly different outside the groups. According to the similarity, the individual similar trajectory recognition can be achieved; and by clustering, the abnormal trajectory mode detection can also be conducted. Similarly, literature [13, 14] combined DBSCAN and SVM (support vector machines) cluster algorithm to sort the GPS trajectories to identify the activity stop locations, which has significance in analyzing human urban mobility.

In the analysis of the time series characteristics of traffic flow data [15], clustering method is popular too. According to the similarity of traffic flow characteristics, the traffic sections are divided into different groups and in the literature [16]; performance of the proposed approach and the stability of the clustering technique are evaluated using the extensive simulation for different traffic densities.

Numerous researches concerning traffic and travel have been conducted by previous studies but there are some drawbacks at the same time. It is difficult to obtain the large amount of data. The acquisition is mostly based on artificial method but at the expense of consuming lots of manpower and resources. Worse still, there are much error and abnormalities in the information usually; thus, the research results always show lower reliability and higher deviation.

Recently, a number of well-established technologies for collecting vehicle related data have emerged, including loop detectors, GPS data, and probe car data [17, 18]. Loop detectors have the merit that once they are installed, there will be continuous record when every vehicle is passing the monitored road section. However, the share of segments in the network equipped with these sensors is typically low and cannot represent the urban network as a whole, which will leave the traffic conditions in most of the network unknown. Dedicated probe vehicles, meanwhile, are used to collect the travel time and other data for designated routes in the network. Nevertheless, due to cost considerations, the number of traffic studies with probe vehicles is typically small and the number of vehicles involved is very few. Hence, they can only cover a limited number of routes for a limited duration of time.

A number of limitations mean that new sophisticated methods are needed to process the data and generate useful information, compared to traditional sensors [19]. Most recently, with the emerging technologies and advanced devices, image recognition technology has been greatly improved. License plate recognition (LPR) system provides the opportunity to study in detail vehicle travel patterns. Compared to manual data collection techniques, LPR provides lower marginal costs, more detailed and disaggregated information, large sample size, and real-time data availability [20, 21]. LPR data is mainly applied in LPR data is mainly applied in solving three kinds of problems in the field of transportation, that is, (1) road network state discrimination, (2) vehicle microscopic characteristics mining, and (3) vehicle travel time/path estimation [22, 23]. Zhan et al.

[24] proposed a lane-based real-time queue length estimation model applying the LPR data. By using ground truth information of the maximum queue length from the city of Langfang in China, the model is validated. In addition, a novel trip route estimation method was given by researchers to estimate the vehicle travel path [25]. Similarly, based on LPR data, an approach for forecasting urban short-term OD matrix which can be used to obtain the original OD information was came up with, and then the OD amount between the detection points can be inferred and finally the OD information between fast track ramps is obtained [26, 27]. All of that mentioned above has proved that the massive amount of LPR data has been created and provides us with rich information and thus can be an effective analytical data source.

Methods for clustering are usually divided into two categories, supervised and unsupervised. Supervised methods use the past data as training samples or previously known outputs to create and learn a clustering rule that allows the clustering of future or new observations [28]. Because the form of the data is not fitting for this study, unsupervised methods are more applicable. Unsupervised cluster algorithms include the hierarchical algorithms and the partition algorithms. Hierarchical clustering algorithms have high computational complexity and cost, limiting their application to large-scale data sets and the shortcomings and advantages of these algorithms will be explained in the following paragraph.

*K*-means clustering algorithm, which belongs to the distance-based clustering algorithms, is not only the most classic, but also the most widely used. It has the property of rapid computing speed, easily explained principle, and high efficiency. *K*-means clustering algorithm is tested using load profiles of 100 residential smart meters collected over the interval extending from July 20th until August 9th, 2009. The method has shown high accuracy in dealing with traffic problems, which proved its great applicability [29].

In this paper, data from LPR system in Shenzhen, China, from November 4th to 10th, 2013, during seven days (a week) in total are analyzed. Variables chosen for clustering include the proportion of different starting/ending points, maximum/minimum/average travel distance for one trip, days of travel within a week, the number of trips per day, the average start time of the first trip, the average end time of the last trip, and activity duration [30]. Firstly, data cleaning is conducted to remove the wrong and repeated data. Then, deviation standardization is utilized to normalize each value for eliminating the error caused by dimension and considerable differences of magnitude. After preliminary treatment, data is divided into two groups, namely, the weekdays and weekends. Finally, to measure the optimal number of clusters, Davies-Bouldin index (DBI) [31] and Silhouette Coefficient (SC) [32] are employed.

In general, the purpose of this study is to classify vehicles into several categories based on some variables and determine travel behavior consistency over time and space by analyzing the vehicle temporal and spatial variability. It can support the study of representing specific groups among the



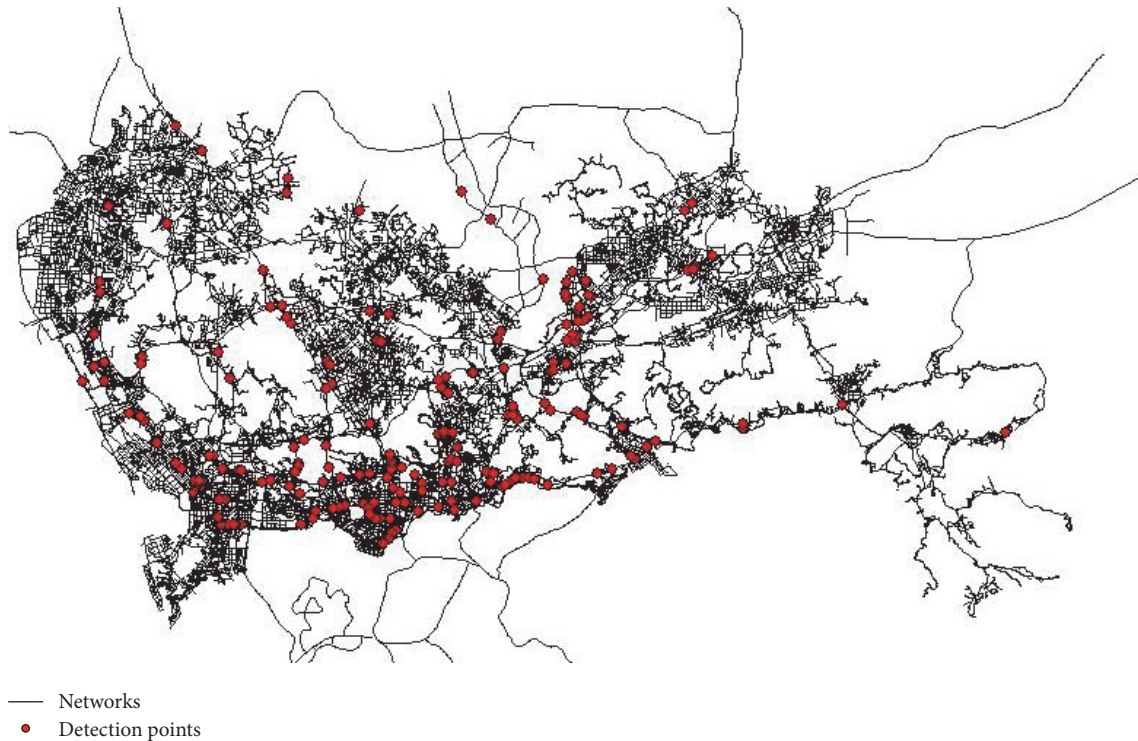


FIGURE 1: Road network and distribution of LPR system in Shenzhen.

TABLE 1: Raw data sample.

Vehicle ID	Detector ID	Lane number	Date and time
2a0adafb3bedf3ad09730c47e0b195b5	20400801	3	2013/11/04 23:00
686ded4bef182aeb03cf361eb6ac6f65	101A0753	2	2013/11/05 00:13
88765784c6cd4464ded7f2336c24eaf2	20602701	1	2013/11/06 14:26
14c397a9bc79e52854194e6c49a69b3e	30606705	2	2013/11/07 15:03
6c43c9bb61e730194de304f0ce9e99ae	206A0480	1	2013/11/08 09:21
fb567b73e8e3db9fa03a4d265b6bddbe	20605302	2	2013/11/09 04:18
45e8dca9594b91c5fdb40706b6f0abc7	10100210	3	2013/11/10 19:14

total population and help establish the predictive level of vehicle trips.

The rest of the paper is organized in the following way. Section 2 offers a brief description of data source. The methodology is introduced in Section 3. Section 4 displays the variables chosen for clustering. Section 5 shows the results of the clustered data and Section 6 is the conclusion and findings.

## 2. Data Description

**2.1. Data Overview.** The potential of LPR system has been explored for planning, managing, and assessing the performance of traffic systems. Further, data collected by these systems allows more comprehensive view of vehicle travel patterns and travel behaviors.

**2.1.1. Data Source.** The LPR system in Shenzhen, China, covers majority of parking lots and expressways for this city. Over

0.9 million vehicles are detected in a week and according to Shenzhen Statistical Yearbook in 2013, the total number of vehicles in Shenzhen is about 2.1 million, implying 42.86% of vehicles are detected by the LPR system. After data cleaning, there are still almost 128,000 recorded vehicles each day. Figure 1 is the sketched network of Shenzhen, where the red points represent the detectors installed on roads and the black lines show the roads.

LPR detectors are mainly installed in the expressways of the city unevenly, most of which are on the intersection or the pedestrian bridge nearby. They are denser in the city center area, while more are dispersed in the rest of the region. The sample of raw data is given in Table 1.

It is worth noting that the detector ID has two types, 10100610 and 101A0753. If “A” is contained in the ID, the detector is a parking lot. Otherwise, it represents a detector on road. Table 2 shows the amount of detectors for each day from November 4th to November 10th, 2013, for which more than 83% detectors are parking lots.

TABLE 2: Amount of detection ID for each day.

Date	Nov. 4	Nov. 5	Nov. 6	Nov. 7	Nov. 8	Nov. 9	Nov. 10
All detectors	918	942	936	934	933	910	873
Parking lots	759	783	777	775	774	751	717
On roads	159	159	159	159	159	159	156

There are three main types of parking lots, (1) residential parking lots (including residential and office buildings, commercial places, and shared parking lots), (2) temporary parking lots, and (3) public parking lots. The parking lots with detection data account for about 20% of all the parking lots in Shenzhen.

*2.1.2. Data Cleaning.* The data cleaning is conducted before vehicle clustering and there are two main steps.

(1) *Extract the Data by Day.* The whole dataset is for seven days (a week), which has been separated into seven files by date thus each file contains the data of the same day.

(2) *Verify the Original LPR Data*

- (1) Delete erroneous LPR data: there are two kinds of erroneous data in our study: (a) the detected time of the record is beyond the range of [0:00–24:00] and (b) the latitude and longitude of the detection site of the record are beyond the scope of Shenzhen.
- (2) Remove duplicated LPR data records: if there are two identical records, only one needs to be kept.
- (3) Extract the trip chain in accordance with the definition of one trip: that is, the data has been processed into the following form.

*Vehicle a-time(1)- location(1), vehicle a-time(2)- location(2),... ,vehicle a-time(n)- location (n)*  
*Vehicle b-time(1)- location (1), vehicle b-time(2)- location(2),... ,vehicle b-time(n)- location (n)*

Based on the trip chain of the vehicles, all of the mentioned variables can be calculated, such as the trip starting time, ending time, the whole activity duration, and the travel distance.

*2.2. Identification of Taxi.* The purpose of this paper is to cluster all the vehicles in the dataset according to some temporal and spatial variables. Each group will have some characteristics different from the other groups, so as to explore vehicle travel patterns we may not know before. Traffic researchers have always paid much attention to taxi, due to its special travel mode. It has the following characteristics [33]:

- (1) There are no fixed route and running time.
- (2) Operation is for 24 hours and can be located in any place of the city.
- (3) The origins and destinations of taxi are completely determined by passengers.
- (4) The operating routes are up to the driver, such as his experience and hobbies.

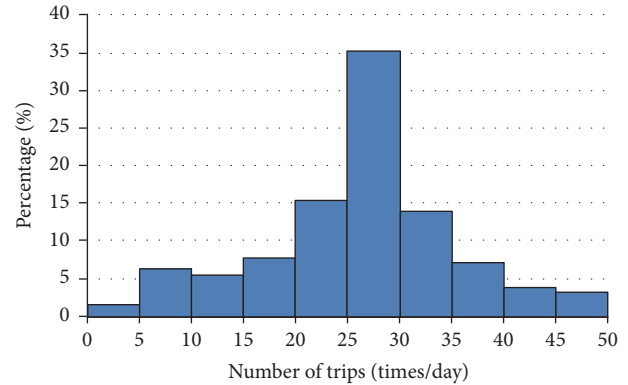


FIGURE 2: The distribution of number of trips for taxi in Shenzhen.

On account of these features, taxis are removed from the dataset to make sure the analysis in this paper is more specific on noncommercial vehicles and the future research will focus on the travel behavior of taxi.

Figure 2 shows the distribution of the number of taxi trips in Shenzhen. There are over 70% of vehicles traveling 20–35 trips per day, and only 7% vehicles are traveling less than 10 trips. Meanwhile, from the clustering result, the travel frequency of nontaxis in a day is no more than 10 trips per day. As a result, we removed vehicles whose travel frequency exceeded 10 trips per day. Under such a definition, there may be two inaccurate results: (1) nontaxis traveling more than 10 times a day were removed and (2) taxis traveling less than 10 times were still retained.

However, in the light of Shenzhen Statistical Yearbook in 2013, the number of taxis is around 17,000 in total, in which less than 50% were detected by the LPR system. Thus, the amount of these two kinds of vehicles will be no more than a thousand, which appears insignificant when compared with tens of thousands ordinary vehicles.

On the basis of the rule proposed above, almost 6,000 taxis for one day are removed from the dataset and when taxis are removed, there are around 122,000 vehicles for each day and 854,000 vehicles in a week.

### 3. Methodology

*3.1. Clustering Methods.* Clustering methods encompass several techniques and algorithms used to group observations based on similar qualitative or quantitative characteristics. They are usually divided into supervised and unsupervised clustering. Supervised methods require a training sample which contains previously known information on each group membership [34]. In accordance with the form of data in this



TABLE 3: Advantages and disadvantages of some unsupervised algorithms.

Clustering algorithm	Representational algorithm	Advantage	Disadvantage
Hierarchical algorithms	BIRCH	(1) No input parameters are required (2) High scalability	(1) high computational complexity and cost; (2) low efficiency in dealing with large-scale data
	CURE		
	Chameleon		
Partition algorithms	$K$ -means	(1) High efficiency in dealing with large-scale data (2) Fast calculation speed	(1) dependent on initial center selection; (2) uncertainty of category number
	CLARANS		

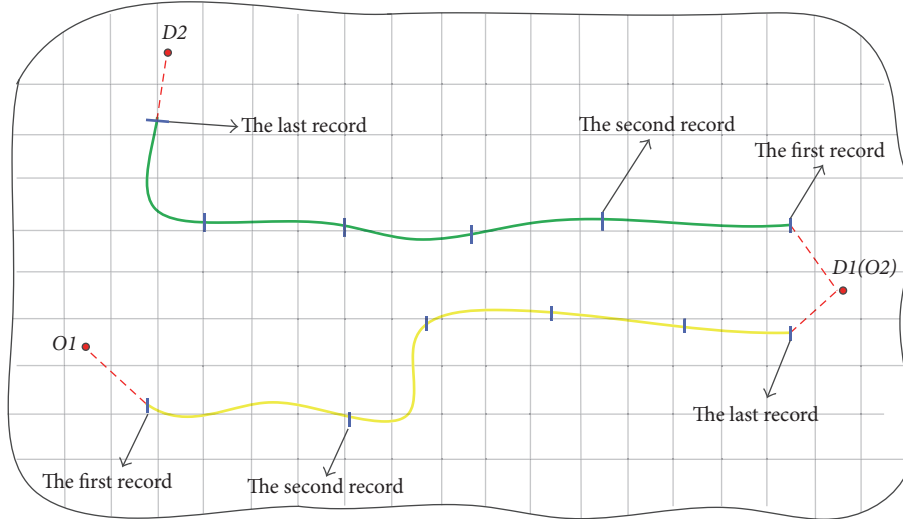


FIGURE 3: The relationship between the travel trajectory and the detection points.

study, the training sample is not available and there are no previously known classes; unsupervised clustering method is the best option. Unsupervised clustering methods aim at categorizing the data objects without a training sample; the goal is to find clusters based on similarities of the input data. There are two main types of unsupervised clustering, the hierarchical algorithms and the partition algorithms. Table 3 discusses the advantages and disadvantages of some unsupervised algorithms [35].

**3.2.  $K$ -Means Algorithm.** As shown in Table 3, hierarchical algorithms have been criticized for low robustness and high sensitivity to noise and outliers. Since the assignment of an object to a cluster is not iterative, hierarchical algorithms are not able to correct potential misclassifications. On the contrast, partition algorithms optimize either a locally or a globally defined objective function to generate groups of observations so they are preferred in studies involving large-scale dataset.

$K$ -means is chosen for this study as a computationally efficient method, which is suitable for situations where all variables are quantitative. It is easy to understand and apply and thus is popular in dealing with the clustering problems. The time complexity of  $K$ -means algorithm is close to linear, is simultaneously suitable for mining large-scale data sets, and is scalable. In this study, the variables used for clustering are all quantitative and we have a large amount of data. So,

$K$ -means is chosen for this study. Nevertheless, the only disadvantage is the difficulty of choosing the number of clusters and their dependency on the initialization scenario. For the first drawback, it can be adjusted by repeated iterations to find the optimal result. For the second one, we have tried several cluster numbers and applied Davies-Bouldin index (DBI) and Silhouette Coefficient (SC) to find the optimal cluster number.

**3.3. Criteria for One Trip.** For the sake of turning the raw data into the form of vehicle trips and the value of its corresponding variables, the criteria for one trip should be given firstly. Due to the inherent limitation of the LPR data, only partial trajectory points of a vehicle can be obtained. As a result, the realistic starting and ending points of a trip cannot be speculated.

Figure 3 shows the travel trajectory of a vehicle in a brief network, where the yellow curve represents the first trip of the vehicle and the green one displays its second trip. In addition, the blue short lines show the detecting points. It is definite that the true trip starting time in origin 1 ( $O1$ ) is earlier than the time of the first trip record, and the trip ending time in destination 1 ( $D1$ ) is later than the time of the last record, as well as the second trip or other trips of the vehicle.

Hence, deviation will exist in the value of some variables inevitably. The average starting time of the first trip will be a little later, and the average ending time of the last trip will be

TABLE 4: The average number of trips for different thresholds.

Threshold (min)	Number of trips	Threshold (min)	Number of trips
20	4.52	50	1.96
25	3.86	55	<b>1.83</b>
30	3.24	60	<b>1.82</b>
35	2.90	65	<b>1.82</b>
40	2.41	70	<b>1.81</b>
45	2.13	75	1.78
50	1.96	80	1.73

a little earlier. The whole activity duration will be longer and the travel distance will be shorter. However, the main purpose of our study is to extract the travel characteristics of vehicles instead of the estimation of the *OD* matrix; these errors are offset in one direction for all vehicles; thus it may not have a critical impact on the clustering result. From this point of view, the definition of one trip is applicable. When applying these values of variables in realistic transportation planning, the deviation should be taken into account.

As mentioned, the “segmentation” refers to the interval between two trips that is the interval of the last record of the first trip and the first record of the second trip, which is different from the vehicle’s accumulated travel time. In order to find the optimal value of the segmentation, we have tested the threshold.

Set *AR* to be the true threshold, *BR* to be the true number of trips, *A* to be the threshold that we will apply, and *B* to be the number of trips that we will calculate. If  $A \leq AR$ , then  $B \geq BR$ ; if  $A \geq AR$ , then  $B \leq BR$ ; only when  $A = AR$ , then  $B = BR$ . Different thresholds ranging from 20 min to 80 min have been tested, and the average number of trips under all circumstance is calculated. The result was illustrated as Table 4.

When the threshold spans from 50 min to 80 min, the value of number of trips has been moving towards stabilization. It implies that the probability of trips to be not detected in this interval is relatively small. Also, the interval of two trips from LPR data is larger than the actual interval. Hence, it is reasonable that one hour is chosen to be the threshold.

## 4. Clustering Variables

*4.1. Spatial and Temporal Variables.* To estimate homogeneous vehicle groups based on their travel patterns using any clustering method, it is necessary to have input information on travel behaviors. Travel patterns can be described by looking at specific variables that together characterize each vehicle’s travel routines [36]. The selected variables must include those vehicles’ characteristics that make their travel patterns distinct [37, 38]. A set of descriptive variables is presented and vehicles are analyzed in weekdays and weekend separately.

*(1) The Proportion of Different Origins/Destinations.* The percentage of different origins/destinations has the potential to be a useful indicator of their mobility patterns. To illustrate,

vehicles with the same starting point for the first trip in a day or the same ending point for the last trip in a day over a week are more likely to be commuters with work or study purposes. This variable is an indicator of spatial travel variability, which could help to infer the vehicle travel predictability. For such vehicles that traveled 3 days in weekdays, the percentage of different origins for the first trip in a day is defined as follows:

*0: The origins of the first trip in a day over the three days are all the same.*

*1/3: There is one difference for the origins of the first trip in a day over the three days.*

*2/3: There are two differences for the origins of the first trip in a day over the three days.*

*1: The origins of the first trip in a day over the three days are all different.*

When the value is 0, the origins for one trip are all the same in the days of travel, suggesting that the behavior of this kind of vehicles has much regularity. In contrast, if the value is 1, the origins for one trip are all different in the days of travel, indicating the irregularity of the travel behaviors.

The calculation for percentage of different destinations for the last trip in a day is defined in the same way, and for vehicles in weekends the dealing method is comparable.

*(2) Travel Distance.* The geometric distance between the origin and destination of one trip can show how accessible activity locations are to a vehicle. Travel distance variability among the trip of a vehicle can also demonstrate travel flexibility and vehicle mobility around the city. The travel distance variables adopted in this study incorporate the maximum/minimum/average travel distance for one trip in the whole week. For the lack of the track points, complete travel trajectory of one trip for a vehicle cannot be obtained. As a result, in this study, the distance of one trip for a vehicle is defined as the exact distance between the start and end points of one trip, which is calculated by the latitude and longitude of the two points.

*(3) Travel Frequency.* The travel frequency of vehicles, that is, trips made over a day/a week (or any other period) incarnates the uncertainty of the travel for vehicles. There are two descriptive variables, number of trips per day, which is the number of complete trips performed on each day of the week and days of travel, which is the number of days within the

period of analysis; a vehicle has at least one trip in a day. For vehicles in weekdays and weekends, the value of their travel days in a week ranges from zero to seven.

(4) *The Trip Start/Finish Time.* The trip start/finish time could give expression to the trip purpose and consistency of trip. Volatility of the start time for the first trip and the finish time for the last trip are crucial aspects when analyzing vehicle travel patterns.

(5) *Total Activity Duration.* Activity refers to all those actions vehicles perform when they are not traveling and in this paper the time interval between the two adjacent trips is defined as the protocol of activity duration. There is a mass of activities purposes, business, work, study, and entertainment, among others. The characteristics of the activity performed at a destination may determine the vehicle's travel decision and the average activity duration of a vehicle in each day varies from weekdays to weekends.

#### 4.2. The Distribution of the Variables for All Vehicles

(1) *Weekdays.* Figure 4 illustrates the distribution of all the temporal and spatial variables in weekdays which is a statistical indicator of the whole vehicles.

In Figure 4(a), there is an obvious peak during the interval of 8:30 am to 9:00 am, representing that the average trip start time of vehicles is mostly focused between 8:30 am and 9:00 am, implying the morning peak hours. Figure 4(b) shows the tendency of the average trip finish time and the majority of the vehicles finish their trip at around 18:00 pm–19:30 pm, which means the afternoon peak hour. Additionally, there is also a large amount of vehicles that start their trip at 12:30 pm–13:30 pm.

For the number of trips per day in Figure 4(c), vehicles traveling 1.5 trips/day occupy a high proportion and vehicles traveling 3.5 trips/day, 2 trips/day, and 4 trips/day followed. The result seems to be confused that vehicles traveling 1.5 trips/day (less than 2 trips/day) conquer such a high rate. Probably, it is because the definition of one trip in the study and the incomplete vehicle detection data.

Figure 4(d) demonstrates days of travel. Vehicles that only travel one day in a week occupy a high rate. The activity duration of most vehicles is within 11 h in Figure 4(e). Figures 4(f), 4(g), and 4(h) reflect the travel distance of vehicles. The maximum travel distance of vehicles for one trip is almost within 60 km, the minimum travel distance is less than 30 km, and the average travel distance is within 40 km. At the same time, we can see that, for the average travel distance of vehicles for one trip, over 68% of trips are within 10 km.

According to Figures 4(i) and 4(j), for the percentage of different starting or ending points, values 0 and 1 seize on a high proportion. Value 0 means the starting/ending points of each trip are identical, and the regularity is high. Analogously, value 1 means that the starting/ending points of each trip are all different, and irregularity is high.

(2) *Weekends.* For vehicles traveling in weekends, the distribution of their temporal indicators is basically similar to the weekdays. For the value of both of the percentages for

different starting and ending points in weekends, value 0 takes up the highest ratio; in other words, these vehicles travel with less regularity. Compared with the weekday vehicles, they travel a relatively short distance; whether it is the maximum travel distance, minimum travel distance, or average travel distance, almost all are within 10 km and relatively concentrated within 5 km.

## 5. Results and Discussions

The values of within-cluster variation and the DBI/SC are shown as functions of the number of clusters in Figures 5(a) and 5(b). A smaller value of DBI and a larger value of SC are better. In Figure 5(a), when the cluster number is six, the value of DBI is the smallest, and when it turns to seven, the value of SC is the largest. The value of SC of seven groups is just a little better than six groups but the value of DBI of six groups is much better than seven groups. As a result, "six" is a relatively better choice. In Figure 5(b) when the cluster number is three, both values of SC and DBI are optimal; there is a lowest point of DBI and a highest point of SC. So, the cluster number for weekdays and weekends is selected as six and three, respectively. The *K*-means clustering method provides not only information about each cluster's core characteristics but also information about the average characteristics of each cluster. Tables 5 and 6 display the average values of each index for each category in weekdays and weekends.

*For Vehicles in Weekdays, Six Groups Are Clustered.* The last column of Table 5 illustrates the proportion of the total number of each category. The smallest cluster contains 4.1% of the vehicles in the sample, and the largest one accounts for 33.7%. Groups 1 to 6 are identified as follows, long travel distance vehicles, commuting vehicles, noon travel vehicles with short travel distance, off-peak hour travel vehicles, midnight travel vehicles, and peak-hour travel with short activity duration vehicles, respectively.

Group 1 is inferred as long travel distance vehicle that travels 1.82 days in a week and makes 2.13 daily trips. On average, the first trip starting time of Group 1 is 10:14 am and the last trip ending time is 19:02 pm. Additionally, the travel behavior of this group is irregular because the trip origins and destinations are all different. Besides, the total activity duration of this group is about 7.41 hours, and the travel distance of this group of vehicles is relatively long. The maximum travel distance for one trip is 78.1 km on average.

Group 2 may be commuting vehicle, which travels 5.94 days of the week on average and makes 2.18 trips per day. The first trip of the day starts at approximately 8:42 am and the last trip of the day ends at 18:18 pm. The activity duration lasts 8.67 hours on average. Furthermore, the distance between the origin and destination of their trips varies from 6.9 km to 59.2 km, and their average travel distance is about 17.5 km for one trip. The proportion of different starting and ending points for Group 2 is 0.12 and 0.09, representing a high regularity in the daily origins and destinations. All of these features support the speculation of Group 2 to be commuting vehicles.

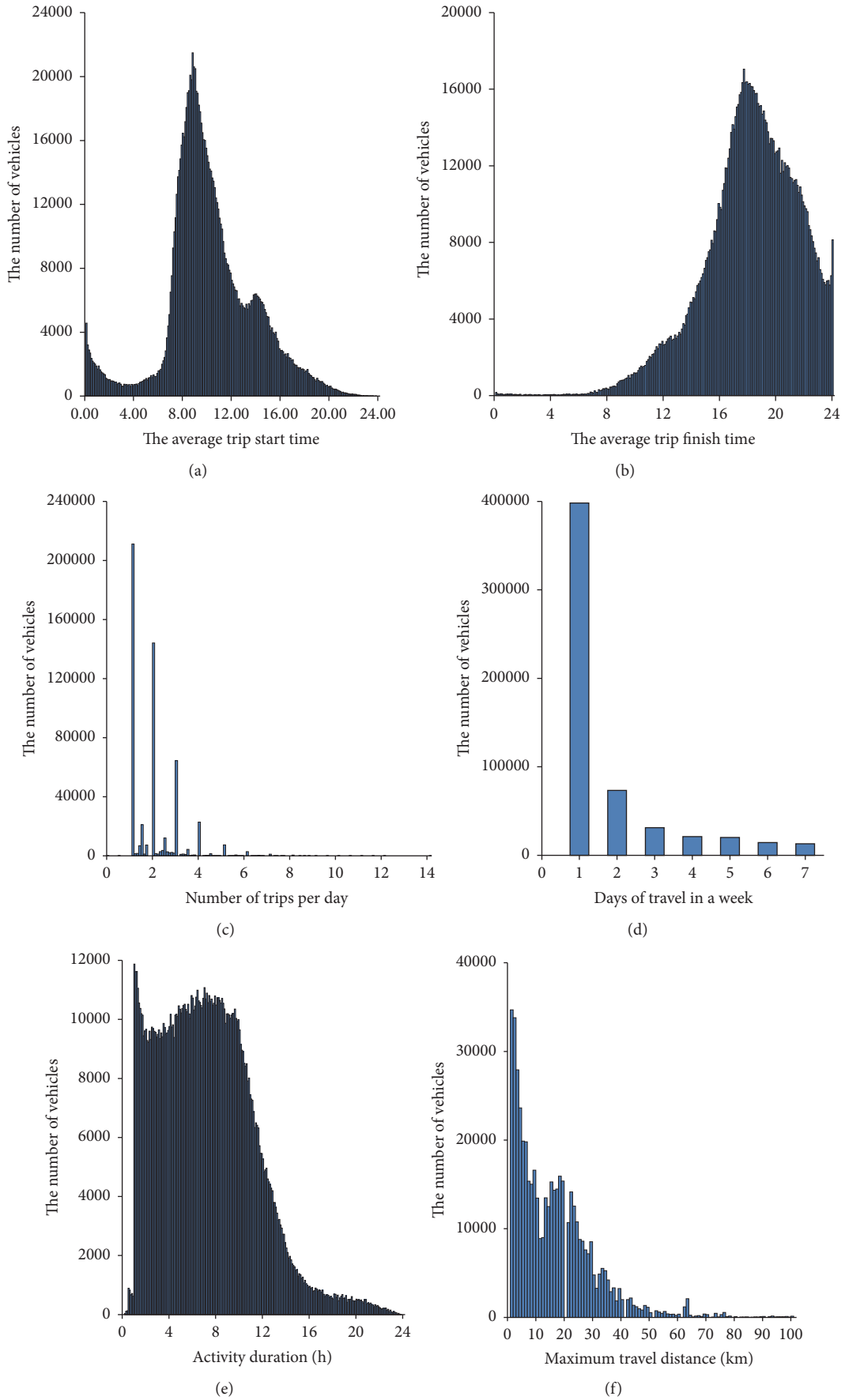


FIGURE 4: Continued.

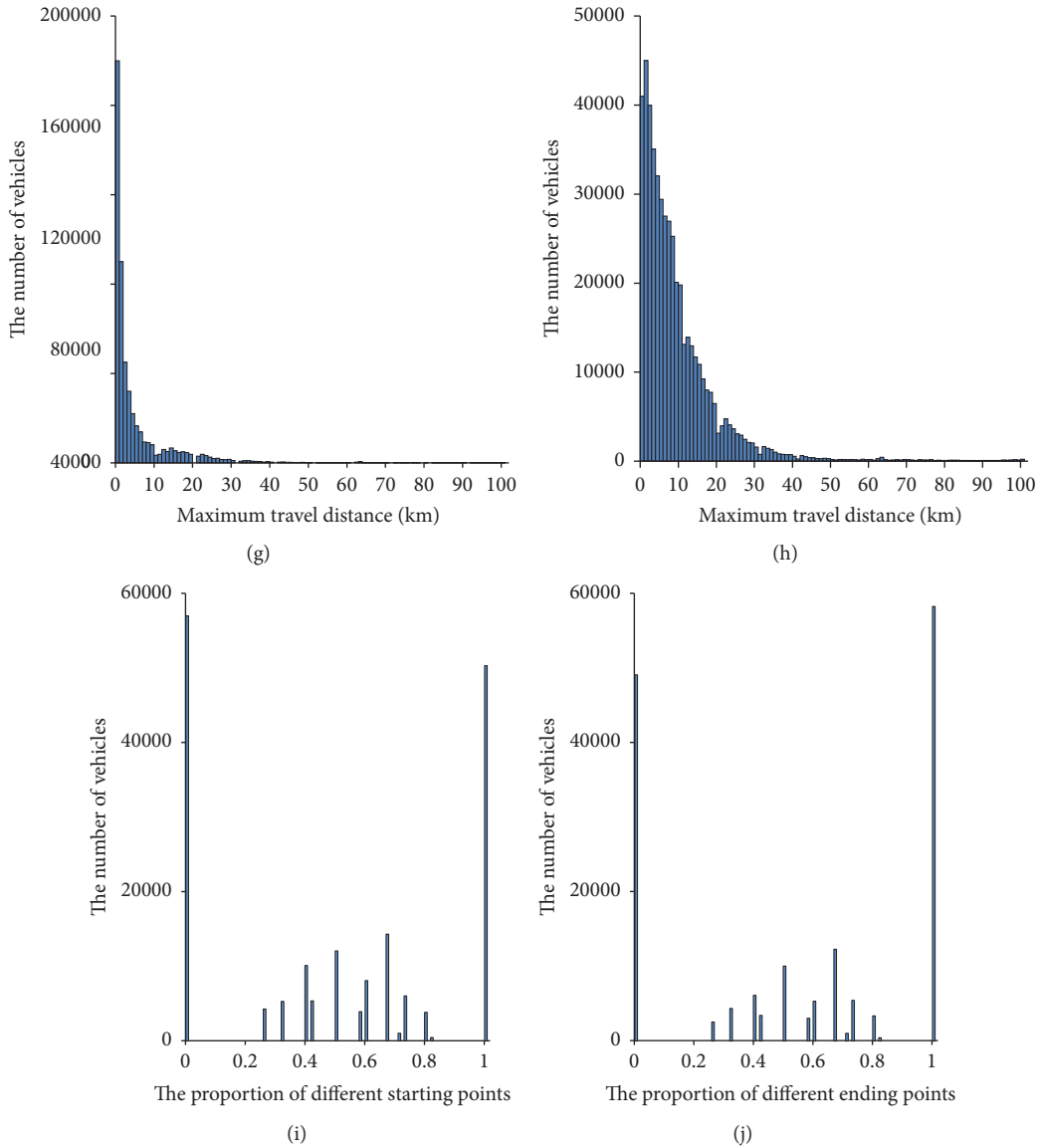


FIGURE 4: The distribution of variables in weekdays. (a) The average trip start time. (b) The average trip finish time. (c) Number of trips per day. (d) Days of travel. (e) Total activity duration. (f) Maximum travel distance. (g) Minimum travel distance. (h) Average travel distance. (i) The proportion of different starting points. (j) The proportion of different ending points.

Group 3 is defined as noon travel vehicle with short travel distance. The first travel starts at 10:07 am and the last travel ends at 15:14 pm; it only travels at noon. Moreover, Group 3 travels only 1.08 days in a week and 1.82 trips in a day, and the activity duration is also short, only 4.02 hours on average. The travel distance varies between 1.9 km and 3.5 km, dropping in a short range and the travel origins and destinations are almost different.

Group 4 is concluded to be off-peak hour travel vehicle; the first trip of the day starts at 10:20 am and the last trip of the day ends at 19:48 pm, which staggers the peak hours. There are 1.82 days of travel in a week and 1.63 trips in a day and the travel distance of Group 4 is similar to that of Group 3. In

particular, the maximum travel distance is only 2.9 km and in accordance with the percentage of different starting and ending points, the travel for Group 4 is not so regular too.

Unlike other groups, Group 5 may be midnight travel vehicle, which has the most distinguish feature. Vehicles start their travel at 0:40 am and the activity duration is around 17.14 hours. Besides, the number of travel times per day is 2.99, which is also higher than others and the travel distance varies between 4.8 km and 32.5 km. The origins and destinations also have a certain degree of randomness.

Group 6 is defined as peak-hour travel with short activity duration vehicle. It starts the first travel at 8:55 am and finishes at 6:11 pm. They travel 1.82 days in a week and 2.71 trips in a

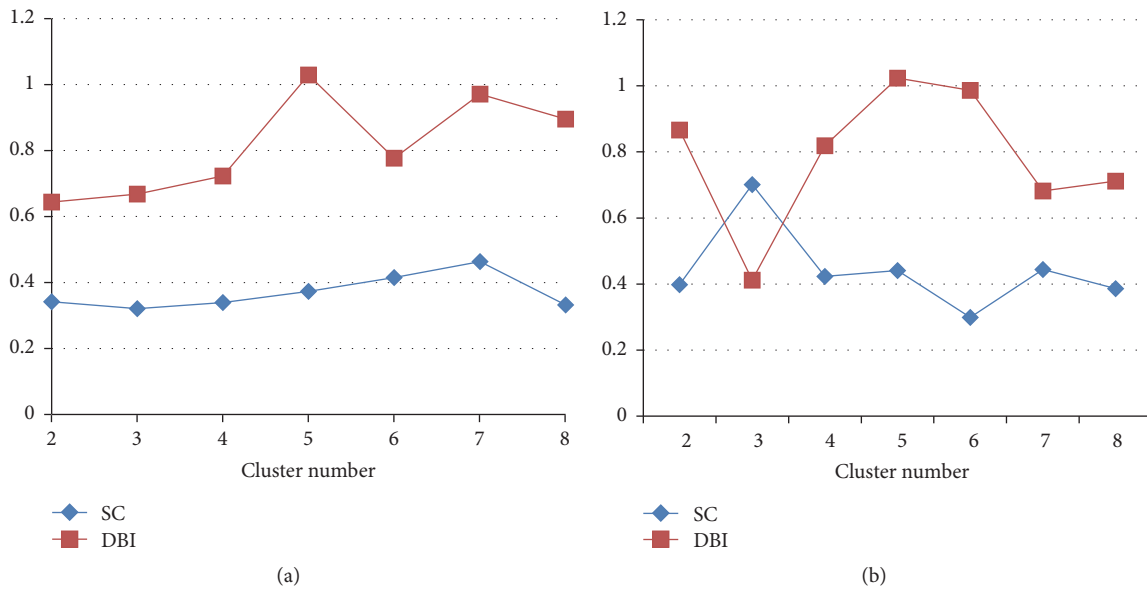


FIGURE 5: DBI and SC of weekdays and weekends. (a) Weekdays. (b) Weekends.

day and they have short activity duration. The travel origins and destinations are not regular and they travel for 28.1 km in average.

In general, the start time of the first trip and the end time of the last trip for Group 2 are similar to those of Group 6, both in the peak hour. Even so, the days of Group 6 traveling in a week are less and its travel distance is much longer. Comparing the characteristics of Group 2 with Group 6, we can conjecture that Group 2 is commuting vehicles traveling twice everyday and Group 6 may be vehicles commuting only in part of the days in a week and consistent with activities for leisure, recreational, or sporadic work in the rest days. Moreover, Groups 2, 3, and 4 are the main composition of traffic flow, taking up 79.1% of the whole vehicle population. Group 4 is off-peak hour travel vehicle and there is no clear travel purpose that could be inferred using only these travel behavior characteristics. These clusters could be composed of leisure travelers, visitors, or sporadic vehicles. They may be vehicles coming out to pick up child or shopping nearby. Group 5 has distinguishing features from others; they travel only in the midnight; it is similar to taxi or online hailing vehicles (i.e., Uber); the travel time, and travel purposes are random and not sure.

*For Vehicles in Weekends, Three Groups Are Clustered.* The characteristics of each group are shown in Table 6.

Group 1 is deduced as off-peak hour travel, where the starting time of the first travel is 10:11 am and the trip ending time is 19:30 pm. They travel 1.87 days in a week and 2.02 trips in a day. In addition, the average travel distance is about 30.8 km and the similarity of the travel origins and destinations is high. Combining with the travel frequency, travel time, and travel distance of these vehicles, they may

live in the city center for work in the weekdays and during weekends they may visit their parents or relatives in the suburbs or have picnics to relax.

Group 2 is defined as afternoon travel with short activity duration vehicle, which travels 1.27 trips per day and 1.66 days in a week. It travels in off-peak hour, which is 12:30 am and 16:18 pm, the travel distance is not long and the activity duration is about 3 hours. Additionally, the origins and destinations are relatively stable. Combined with all of these features, group 2 tends to be vehicles going shopping or leisure on weekends.

Group 3 may be peak-hour travel vehicle, the average start time of the first trip is 7:42 am and the average finish time of last trip is 18:50 pm and it only travels 2.11 days in a week. The travel distance is as short as Groups 3 and 4 in weekdays. Vehicles in this group resemble commuting vehicles in weekdays. This kind of vehicles may work only in weekends, for example, people working for cram schools and the like.

## 6. Conclusions

This paper shows that it is possible to analyze the travel characteristics of vehicles and identify vehicle groups with similar travel behavior using LPR data. The main contribution of this paper is summarized as follows:

- (i) Six vehicle groups with similar travel characteristics in weekdays and three groups in weekends are identified and the detailed behavior of each cluster is presented.
- (ii) Travel characteristics are studied by analyzing the distribution of these variables and the values of each



TABLE 5: Average values of variables for each category in weekdays.

Type	Days of travel per week (day)	Number of trips per day	Average start time of first trip	Average finish time of last trip	Total activity duration (h)	Maximum/minimum/average travel distance (km)	The proportion of different start/end points	Percentage (%)
(1) Long travel distance	1.82	2.13	10:14	19:02	7.41	78.1/24.5/30.3	0.92/0.91	9.5
(2) Commuting	<b>5.94</b>	<b>2.18</b>	<b>8:42</b>	<b>18:18</b>	<b>8.67</b>	<b>59.2/6.9/17.5</b>	<b>0.12/0.09</b>	<b>33.7</b>
(3) Noon travel	1.08	1.82	10:07	<b>15:14</b>	4.02	3.5/1.9/2.2	0.67/0.82	29.6
(4) Off-peak hour travel	1.82	1.63	10:20	19:48	9.09	2.9/1.2/2.4	0.77/0.73	15.8
(5) Midnight travel	3.13	2.99	<b>0:40</b>	21:30	<b>17.14</b>	32.5/4.8/14.3	0.66/0.57	7.3
(6) Peak-hour travel	1.82	2.71	<b>8:55</b>	<b>18:11</b>	9.07	68.0/12.2/28.1	0.97/0.98	4.1

TABLE 6: Average values of variables for each category in weekends.

Type	Days of travel per week (day)	Number of trips per day	Average start time of first trip	Average finish time of last trip	Total activity duration (h)	Maximum/minimum/average travel distance (km)	The proportion of different start/end points	Percentage (%)
(1) Off-peak hour travel	1.87	2.02	10:11	19:30	12.48	46.8/21.6/30.8	0.19/0.21	28.9
(2) Afternoon travel with short activity duration	1.66	1.27	12:30	16:18	2.98	10.2/4.8/7.9	0.34/0.27	39.5
(3) Peak-hour travel	2.11	1.85	7:42	18:50	6.81	4.6/1.9/3.7	0.24/0.31	31.6

variable for each category. In addition, we defined vehicle type for each group of vehicle, to identify the commuting vehicle and other ordinary leisure travel vehicles, and the clustering result can be used in several aspects, such as

- (1) policy making in vehicle classification management;
- (2) transport planning and vehicle travel forecasting;
- (3) urban traffic simulation and monitoring.

For example, with the clustering, we can effectively extract the commuting travel vehicles which provide better decision information for developing urban traffic demand and managing policy by analyzing the spatial and temporal distribution of its travel behavior. In addition, summarizing the clustering result, there are almost 46% (type 3 and type 4) off-peak hour travel vehicles traveling in short distance (less than 3.5 km) in weekdays. Considering that the detectors are mainly installed on expressways, we can guide these vehicles to take arterial roads instead of expressways by implementing some traffic management schemes during off-peak hour to improve the level of services of arterial roads and finally release the traffic pressure of off-peak hours on expressways.

In general, firstly, this study has shown that it is possible to analyze the travel characteristic of vehicles and identify vehicle groups with similar travel behavior using LPR data. Besides, a study of the vehicles' travel pattern can be performed based on this study results and this information can be used to preferably understand how the behavior of the different groups affects the road system, the travel patterns, and travel modes.

Secondly, from the standpoint of transportation planning, clustering vehicle travel patterns allow the analysis of possible differences in level of service experienced by different vehicle segments and the identification of potential biases. It can also provide better understanding of how changes in level of service affect different vehicles and how they respond to those changes. Knowing the main differences between groups can contribute to a better understanding of the effect of disruptions on travel behavior.

Finally, the method displayed in this study is innovative and practical which can be applied in several similar problems and researches. It highlights the potential of using LPR data to mine underlying information of vehicles and the study also reveals the importance of clustering vehicles based on their characteristics.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This research was sponsored by National Natural Science Foundation of China (71171147) and Fundamental Research Funds for the Central Universities.

## References

- [1] W. Wenjing and G. Hongcheng, "Car and rail travel mode choice behavior analysis," *Urban Transportation of China*, vol. 3, pp. 11–14, 2010.
- [2] A. Chakirov and A. Erath, "Use of public transport smart card fare payment data for travel behaviour analysis in Singapore," in *Proceedings of the 16th International Conference of Hong Kong Society for Transportation Studies*, Hong Kong, 2011.
- [3] L. Liu, A. Hou, A. Biderman, C. Ratti, and J. Chen, "Understanding individual and collective mobility patterns from smart card records: a case study in Shenzhen," in *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems (ITSC '09)*, pp. 842–847, St. Louis, Mo, USA, October 2009.
- [4] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, "Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 197–211, 2017.
- [5] E. Chung, "Classification of traffic pattern," in *Proceedings of the 10th World Congress on Intelligent Transport Systems*, vol. 11, pp. 16–20, Madrid, Spain, 2003.
- [6] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, Edward Arnold, London, UK, 2001.
- [7] S. Hanson and J. Huff, "Classification issues in the analysis of complex travel behavior," *Transportation*, vol. 13, no. 3, pp. 271–293, 1986.
- [8] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.
- [9] H. Ling and W. Lingda, "Summary of clustering algorithms in data mining," *Application Research of Computers*, vol. 1, pp. 10–13, 2007.
- [10] W. Weijermars and E. Van Berkum, "Analyzing highway flow patterns using cluster analysis," in *Intelligent Transportation Systems*, pp. 308–313, 2005.
- [11] T. Li, T. Pei, Y. C. Yuan et al., "A summary for the classification of pattern and application of human trajectory," *Progress in Geography*, vol. 33, no. 7, pp. 938–948, 2014.
- [12] B. Zhang, *Research on Taxi Trajectory Data Mining Based on Cloud Computing*, Xidian University, Xi'an, China, 2014.
- [13] L. Gong, H. Sato, T. Yamamoto, T. Miwa, and T. Morikawa, "Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines," *Journal of Modern Transportation*, vol. 23, no. 3, pp. 202–213, 2015.
- [14] L. Gong, H. Sato, T. Morikawa et al., "Activity stop and non-activity stop identification in GPS trajectories utilizing density-based clustering method and support vector machines," in *Proceedings of the Transportation Research Board Annual Meeting*, 2015.
- [15] X. Zhang and W. Guang, "Study on urban traffic road segmentation based on cluster analysis," *Intelligent Transportation Systems and Information Technology*, vol. 9, no. 3, pp. 36–41, 2009.
- [16] H. R. Arkian, R. E. Atani, and S. Kamali, "Cluster-based traffic information generalization in vehicular ad-hoc networks," in *Proceedings of the IEEE International Symposium on Telecommunications*, pp. 197–207, 2014.
- [17] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data,"

- Transportation Research Part B: Methodological*, vol. 53, no. 4, pp. 64–81, 2013.
- [18] N. H. M. Wilson, J. Zhao, and A. Rahbee, “The potential impact of automated data collection systems on urban public transport planning,” in *Schedule-Based Modeling of Transportation Networks*, vol. 46, pp. 1–5, 2009.
- [19] G. Leduc, “Road traffic data: collection methods and applications,” JRC Technical Notes, Working Papers on Energy, Transport and, Climate Change 1, 2008.
- [20] F. Öztürk and F. Özen, “A new license plate recognition system based on probabilistic neural networks,” *Procedia Technology*, vol. 1, pp. 124–128, 2012.
- [21] M. A. Massoud, M. Sabee, M. Gergais, and R. Bakhit, “Automated new license plate recognition in Egypt,” *Alexandria Engineering Journal*, vol. 52, no. 3, pp. 319–326, 2013.
- [22] R. Camus, G. Longo, and C. Macorini, “Estimation of transit reliability level-of-service based on automatic vehicle location data,” *Transportation Research Record*, vol. 1927, pp. 277–286, 2005.
- [23] S. Lee and M. Hickman, “Trip purpose inference using automated fare collection data,” in *Proceedings of the 4th Transportation Research Board Conference on Innovations in Travel Modeling*, Tampa, Fla, USA, 2012.
- [24] X. Zhan, R. Li, and S. V. Ukkusuri, “Lane-based real-time queue length estimation using license plate recognition data,” *Transportation Research Part C: Emerging Technologies*, vol. 57, pp. 85–102, 2015.
- [25] C. Hong, X. Xiangchun, Y. Feng, and J. Zhou, “A novel method of trip route estimation based on vehicle license plate recognition system,” in *Proceedings of the 13th COTA International Conference of Transportation Professionals (CICTP '13)*, November 2013.
- [26] M. P. Dixon and L. R. Rilett, “Population origin-destination estimation using automatic vehicle identification and volume data,” *Journal of Transportation Engineering*, vol. 131, no. 2, pp. 75–82, 2005.
- [27] M. A. Munizaga and C. Palma, “Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile,” *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.
- [28] H. Zhang, X. Liu, X. Duan, D. Miao, and H. Ma, “A comparative study of clustering algorithms in data mining,” *Computer Applications and Software*, vol. 20, no. 2, pp. 5–6, 2003.
- [29] A. Al-Wakeel and J. Wu, “K-means based cluster analysis of residential smart meter measurements,” *Energy Procedia*, vol. 88, pp. 754–760, 2016.
- [30] A. E. Raftery and N. Dean, “Variable selection for model-based clustering,” *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 168–178, 2006.
- [31] J. Yuqing and G. Dunwei, “Fast genetic clustering algorithm,” in *Proceedings of the China Intelligent Automation Conference*, vol. 38, pp. 186–190, 2007.
- [32] L. Zhu, B. Ma, and X. Zhao, “Effectiveness analysis based on clustering coefficient contour,” *Journal of Computer Applications*, vol. 30, pp. 139–141, 2010.
- [33] H. Cai, X. Zhan, J. Zhu, X. Jia, A. S. F. Chiu, and M. Xu, “Understanding taxi travel patterns,” *Physica A: Statistical Mechanics and Its Applications*, vol. 457, pp. 590–597, 2016.
- [34] C. Fraley and A. E. Raftery, “How many clusters? Which clustering method? Answers via model-based cluster analysis,” *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.
- [35] S. Jigui and J. Jie, “Research on clustering algorithm,” *Journal of Software*, vol. 19, pp. 48–61, 2009.
- [36] H. Nishiuchi, J. King, and T. Todoroki, “Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data,” *International Journal of Intelligent Transportation Systems Research*, vol. 11, no. 1, pp. 1–10, 2013.
- [37] M. A. Ortega-Tong, *Classification of London's public transport users US smart card data [Bachelor of Science in Civil Engineering]*, University of Chile, 2007.
- [38] P. Jones and M. Clarke, “The significance and measurement of variability in travel behaviour,” *Transportation*, vol. 15, no. 1-2, pp. 65–87, 1988.

## Research Article

# Impact of Vehicular Countdown Signals on Driving Psychologies and Behaviors: Taking China as an Example

Fuquan Pan,<sup>1</sup> Lixia Zhang,<sup>1</sup> Changxi Ma,<sup>2</sup> Haiyuan Li,<sup>3</sup> Jinshun Yang,<sup>1</sup>  
Tao Liu,<sup>1</sup> Fengyuan Wang,<sup>1</sup> and Shushan Chai<sup>1</sup>

<sup>1</sup>School of Automobile and Transportation, Qingdao University of Technology, Qingdao, Shandong 26650, China

<sup>2</sup>School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou, Gansu 730070, China

<sup>3</sup>Nevada Department of Transportation, Carson City, NV 89712, USA

Correspondence should be addressed to Lixia Zhang; [zlxzhanglixia@163.com](mailto:zlxzhanglixia@163.com)

Received 31 January 2017; Accepted 8 March 2017; Published 10 April 2017

Academic Editor: Xiaolei Ma

Copyright © 2017 Fuquan Pan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Countdown signal control is a relatively new control mode that can inform a driver in advance about the remaining time to pass through intersections or the time needed to wait for other drivers and pedestrians. At present, few countries apply vehicular countdown signals. However, in China, some cities have applied vehicular countdown signals for years, though it is unclear how and how much such signals influence driving psychologies and behaviors compared with non-countdown signal controls. The present work aims to clarify the impact of vehicular countdown signals on driving psychologies and behaviors on the cognitive level. A questionnaire survey with 32 questions about driving psychologies and behaviors was designed, and an online survey was conducted. A total of 1051 valid questionnaires were received. The survey data were analyzed, and the main results indicate that most of the surveyed drivers prefer countdown signal controls and think that such controls can improve not only traffic safety but also traffic operational efficiency. The surveyed drivers also think that countdown signal controls have an impact on driving psychologies and behaviors and the survey results have demonstrated that the driving behaviors of female drivers surveyed are not conservative under the clear conditions of green countdown signal control. Further studies and methods concerning the effects of countdown signals on driving psychologies and behaviors are discussed.

## 1. Introduction

While intersections are part of the road system, they are far more complex than the segments connecting them [1, 2]. Driving psychologies and behaviors at road intersections are considerably different compared to those at road sections. According to the control mode, there are two kinds of road intersections: unsignalized and signalized. In recent years, countdown signalized intersections have appeared in some countries or areas. Compared to traditional non-countdown signals, countdown signals can inform a driver in advance about the remaining time to pass through intersections or the time needed to wait for other drivers and pedestrians. Countdown signals can be divided into vehicular countdown signals and pedestrian countdown signals. Vehicular countdown signals further include green, red, and yellow countdown signals.

Out of more than 200 countries, only a few use countdown signal controls at road intersections, such as China, Thailand, India, Singapore, Malaysia, the United States, and the United Kingdom. Of these countries, only a small number allow pedestrian countdown signals to be used at road intersections, such as the United States and Britain. In other words, few countries currently use vehicular countdown signals. The object of this study is to focus on vehicular countdown signals.

China, as one of the pioneer countries, is playing a leading role in the use of vehicular countdown signals. However, no specific manuals or standards are available for guiding the usage of vehicular countdown signals in China [3–5]. Therefore, Chinese local governments encounter difficulties in making a clear decision on extending the application of vehicular countdown signals. Traffic engineers and researchers

have different opinions on the application of vehicular countdown signals. Some support the use of countdown signal controls and consider such controls as capable of improving traffic operational efficiency by taking full advantage of the signal time. Others are opposed to the application of countdown signal controls and are concerned about the possibility of such controls increasing traffic crashes and having a negative impact on traffic safety [6].

In theory, vehicular countdown signals are different compared to non-countdown signals and should have an impact on driving psychologies and behaviors. However, to date, it is unclear how and how much vehicular countdown signals influence driving psychologies and behaviors compared with non-countdown signal controls. To answer these questions, the present study explored the influence of vehicular countdown signals from two aspects of driving psychologies and behaviors based on a questionnaire survey. The survey consisted of 32 questions related to vehicular countdown signals.

## 2. Literature Review

This paper focuses on vehicular countdown signals (hereafter referred to as countdown signals in the following text unless specified otherwise). To date, countdown signal control, as a new traffic control mode of signalized intersections, is applied in China, Singapore, Thailand, Malaysia, India, and other countries. Lum and Halim conducted a before-and-after study by observing driver reactions at a signalized intersection with a green signal countdown display (GSCD) installed [7]. The finding revealed that the number of red-light running violations is significantly mitigated at the initial period after installing the GSCD. However, the effectiveness of the device tends to dissipate over time, with the number of violations bouncing back to almost the same level as before GSCD. Ibrahim et al. introduced countdown timers installed at some intersections in Kuala Lumpur, Malaysia [8]. The impact of countdown timers on driving behaviors and intersection approach headways was studied by comparing three intersections with countdown timers with three intersections lacking countdown timers. The result indicated that the countdown timers have a significant impact on headways but a little impact on the initial delay. Limanond et al. studied how countdown timers affect the queue discharge characteristics of through movements during the green phase at a signalized intersection in Bangkok, Thailand [9]. They pointed out that the countdown timers had a significant impact on the start-up lost time at the intersection under study, but the effect on the saturation headway was negligible. Chiou and Chang investigated the impact of GSCD and red signal countdown display (RSCD) on driver behaviors in Taiwan [10]. The results showed that GSCD can reduce the late-stopping ratio, but it increases the likelihood of rear-end crashes. Although RSCD can effectively reduce the start-up delay, saturated headway, and cumulative start-up delay, it cannot significantly improve intersection safety in the long term. Sharma et al. presented the usage of countdown timers at signalized intersections in India [11]. The study conducted a before-and-after analysis by comparing predata with postdata collected at a selected intersection in Chennai. The results reflected that the time

information provided at the beginning of the green light (end of the red light) can enhance efficiency and reduce start-up lost time but increase red-light running violations. Papaioannou and Politis found that the percentage of early start violations at the intersection with SCD was 24%, where the percentage for intersections without SCD was less than 1% [12]. Devalla et al. found that GSCD is linked with fewer red light violations (RLVs) cycles, a lower mean number of RLVs per RLV cycle, higher vehicular speeds during the phase transition at different locations upstream to the stop line, a higher number of speeding cars, and higher stop line crossing speeds during amber [13]. Islam et al. found a reduction in start-up lost time at signalized intersections when a red signal countdown timer is present [14].

In China, Wang and Yang conducted a preliminary analysis on the traffic signal countdown by conducting a survey of 337 drivers regarding driver attitudes and behaviors on the green signal countdown in Longyan City, Fujian province [15]. The study advised that the countdown signal should be set cautiously. Wu et al. (2009) focused on the driver's decision-making process at countdown signalized intersections [16]. A logistic model was adopted to build the model of behavior decision at countdown signalized intersections based on vehicle types and speed. Zhang et al. conducted a survey on the countdown signal and collected 200 questionnaires from drivers and pedestrians at four intersections in Wanzhuang, Beijing City [17]. The results showed that the drivers and pedestrians sampled had a preference for the countdown signal. Qian and Han preliminarily studied the influence of green signal countdown on traffic safety through a questionnaire investigation of 390 drivers [18]. The finding indicated that the green signal countdown is good for neither traffic safety nor traffic operational efficiency. Thus, the green signal countdown should be used cautiously. Ma et al. conducted a field observation to obtain critical parameters related to drivers and vehicles at two similar intersections, one with GSCDs and the other without GSCD, in Shanghai City, China [19]. They found that GSCD increased the traffic capacity at the sampled intersection and significantly reduced the number of red-light running violations. Qian carried out an eight-question driver behavior survey of 390 drivers regarding the red signal countdown to analyze driver behaviors [20]. Long et al. studied the impact of the countdown timer on driver behaviors after the yellow onset and found that the countdown timer influences drivers stopping or passing through the intersection [21]. Additionally, a correlation exists between the countdown timer and red-light running violations. Huang et al. found that although GSCD stimulates the drivers in dilemma zones to choose to cross the intersection during amber, which produces a higher RLR risk compared with SCD and GSCD, the intersection with GSCD has the lowest RLR violations due to its strong positive effect in cutting down the range of dilemma zones [22]. Li et al. comparatively analyzed drivers' perception-reaction time (PRT) with and without a countdown timer based on the RGB color model and found that the drivers' PRT was decreased from 2.12 s to 1.48 s with countdown signals [23]. Pan et al. attempted to find effects of the end of a green signal countdown on drivers' behaviors when they drive vehicles through



intersections based on the data of vehicle position, time, and speed at the entrance of intersections [24]. Fu et al. characterized and modeled driver's brake perception-reaction time (BPRT) to yellow signal at signalized intersections with and without countdown timer and found an increase in driver's BPRT because countdown timer may induce risky driving behaviors [25]. Pan et al. did an interesting study demonstrating that the value of a driver's car has influence on driving behaviors at countdown signalized intersections [26].

As introduced above, three methods are commonly used to study the effects of countdown signals on drivers, including survey, video, and observation in the field. Evaluating the existing studies through questionnaire survey method revealed that the questionnaires are not designed comprehensively enough, because these only involve a part of driving behaviors and do not investigate driving behaviors from the psychological aspect.

### 3. Method

*3.1. Design of Questionnaire.* Considering the complexity of driving psychologies and behaviors and the new mode of countdown signal controls, the questionnaire was designed to fully reflect the common and individual characteristics of the driving psychologies and behaviors involving the aspects listed as follows:

- (1) Gender: male and female drivers may have significantly different attitudes toward countdown signal controls
- (2) Age: age refers to the length of time that one has existed. It is an important indicator to reflect the differences in driving psychologies and behaviors of different age groups
- (3) Driving experience: driving experience is used to examine the differences of driving psychologies and behaviors for drivers with different driving experiences
- (4) Specific questions associated with driving psychologies and behaviors on the countdown signal controls

*3.2. Online Questionnaire Survey.* The designed questionnaire was released and conducted by a professional survey website in China. A total of 1051 valid questionnaires were received.

### 4. Analysis of Survey Results

The survey results were classified and analyzed to indicate the general understanding of drivers and their psychological characteristics, as well as the behavioral characteristics of different types of drivers surveyed.

*4.1. Basic Characteristics of Drivers Surveyed.* The characteristics of gender, age, and driving experience for the surveyed drivers are shown in Table 1.

Table 1 shows that the majority of respondents were young drivers and that more male drivers than female drivers

TABLE 1: Characteristics of drivers surveyed.

Questions	Options	Proportion (%)
(1) Gender	(A) Male	62.32
	(B) Female	37.68
(2) Age	(A) Younger than 25 years old	16.94
	(B) 25–30 years old	37.68
	(C) 31–40 years old	22.07
	(D) 41–50 years old	16.46
	(E) 51–60 years old	5.80
	(F) More than 60 years of age	1.05
(3) Driving experience	(A) 0–3 years	36.06
	(B) 4–5 years	26.17
	(C) 6–10 years	20.36
	(D) More than 10 years	17.41

responded to the survey. The drivers surveyed below the age of 40 accounted for 76.69%, indicating that young drivers use the Internet more widely in China. For driving age, the drivers with no more than three years of driving experience accounted for 36.06%, which is consistent with the rapid increase of the number of Chinese drivers in the last three years.

*4.2. Attitudes and Understanding of Drivers on Countdown Signal Controls.* Six questions, questions (4) to (9), were designed to evaluate the attitudes and understanding of drivers on countdown signal controls. The survey data were analyzed, and the results are summarized in Table 2.

Table 2 shows that the majority of the surveyed drivers felt easier driving vehicles on roadway sections than at intersections; they especially felt nervous at unsignalized intersections. Most drivers also supported setting up countdown signal controls, which they considered to be beneficial to driving behavior decisions. Even more drivers thought that countdown signal controls can be conducive to improve traffic safety and traffic operational efficiency. In addition, the proportion of aggressive drivers is not high from the drivers surveyed.

*4.3. Attitudes and Understanding of Drivers on Green Countdown Signal Controls.* Questions (10)–(15) were designed to investigate driver attitudes on green countdown signal controls and to contrast the behavior of “race against time” at the end of the green light at countdown signal and non-countdown signal control intersections. The analysis results for questions (10)–(13) are shown in Table 3.

According to Table 3, the surveyed drivers who supported the setup of green countdown signals accounted for the majority of respondents. The proportion of drivers who regarded the green countdown signal as having an impact on driving behaviors reaches up to 92.30%. With regard to the display modes of the green countdown, 55.19% of the surveyed drivers selected the partial countdown.

TABLE 2: Attitudes and understanding of surveyed drivers on countdown signal controls.

Questions	Options	Proportion (%)
(4) Which kind of city road conditions do you find the least stressful? [single choice]	(A) Roadway section	80.21
	(B) Signalized intersection	13.42
	(C) Unsignalized intersection	6.37
(5) What are your attitudes regarding the setting of countdown signal controls at intersections? [single choice]	(A) Support	81.83
	(B) Do not support	12.18
	(C) Does not matter	5.99
(6) If countdown signals are set up at intersections, do you think it will help drivers make behavioral decisions? [single choice]	(A) Helpful	84.87
	(B) Not helpful	11.61
	(C) Does not matter	3.52
(7) Do you think what kind of signal controls more conducive to traffic safety? [single choice]	(A) Countdown signal controls	76.88
	(B) Non-countdown signal controls	23.12
(8) Do you think what kind of signal controls more conducive to traffic operational efficiency? [single choice]	(A) Countdown signal controls	88.39
	(B) Non-countdown signal controls	11.61
(9) You think your own driving behavior tends to [single choice]	(A) Aggressive	15.43
	(B) Conservative	55.90
	(C) Neutral	28.67

TABLE 3: Attitudes of surveyed drivers on green countdown signal controls.

Questions	Options	Proportion (%)
(10) Do you support setting up a green countdown signal at intersections? [single choice]	(A) Support.	80.02
	(B) Do not support.	17.89
	(C) Does not matter.	2.09
	(D) Others.	0.00
(11) If the green countdown signal control is set up at intersections, what kind of impact will this have on you? [single choice]	(A) Through the countdown signal, I can see the green time decreasing. Thus, I accelerate to pass through the intersection as fast as possible in this phase, thereby increasing my chances of red-light running.	23.98
	(B) Through the countdown signal, I can see the green time decreasing. Thus, I can control vehicle speed better, which reduces my chances of red-light running.	68.32
	(C) Neither the countdown signal nor the non-countdown signal has an impact on me.	6.76
	(D) Others.	0.95
(12) Which of the following control modes of green countdown is better? [single choice]	(A) The countdown from the beginning to the end of the overall period of green light is better.	44.81
	(B) Beginning to show the countdown 30 s before the end of the green light is better.	10.37
	(C) Beginning to show the countdown 20 s before the end of the green light is better.	19.31
	(D) Beginning to show the countdown 10 s before the end of the green light is better.	25.50
(13) When a green light changes to a red light, which transition do you think is more reasonable? [single choice]	(A) Non-countdown green → yellow → red.	13.99
	(B) Countdown green → yellow → red.	47.86
	(C) Non-countdown green → countdown green → yellow → red.	28.83
	(D) Non-countdown green → countdown green → red.	9.32

TABLE 4: Behaviors of surveyed drivers on the green signal coming to an end in the two cases.

Questions	Options
(14) At countdown signalized intersections, what will you do when you approach the intersection stop line, see the green countdown coming to an end and the yellow light starting at once? [single choice]	(A) Accelerate and pass the stop line before the end of the green countdown or before the end of the yellow light. (B) Decelerate and make sure to stop before the stop line prior to the end of the yellow light. (C) Maintain the original speed; if I cannot stop safely in front of the stop line, then I will pass the stop line before the end of the yellow light. (D) Others.
(15) At non-countdown signalized intersections, when you approach the intersection stop line, see the green light is changing yellow light, how do you do? [single choice]	(A) Accelerate and pass the stop line before the end of the yellow light. (B) Decelerate and make sure to stop before the stop line prior to the end of the yellow light. (C) Maintain the original speed; if I cannot safely stop in front of the stop line, and then pass the stop line before the end of the yellow light. (D) Others.

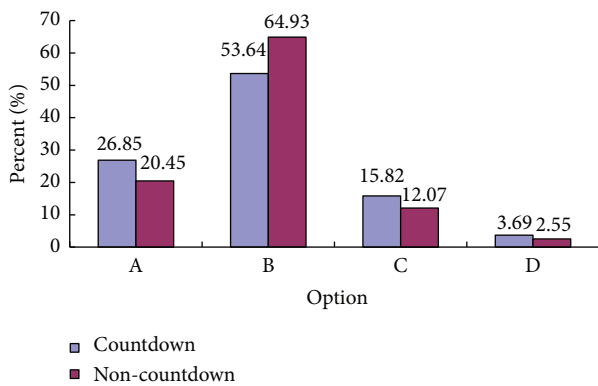


FIGURE 1: Comparison of driving behaviors before the end of the green light at two types of intersections.

Questions (14)-(15) and their options are described in Table 4, and the comparative analysis is shown in Figure 1.

In comparison with non-countdown signalized intersections (Figure 1), more drivers would like to accelerate passing through intersections while the green light time shifts to the yellow light time at green countdown signalized intersections. However, the difference is not significant. More drivers surveyed agreed with option B at non-countdown signalized intersections than at countdown signalized intersections. The result shows that driving behaviors are more adventurous at countdown signalized intersections than at non-countdown signalized intersections.

**4.4. Driver Attitudes and Behaviors on Red Countdown Signal Controls.** Questions (16)–(24) were designed to investigate the attitudes and possible driving behaviors of drivers on red countdown signal controls. The analysis results for questions (16)–(19) are shown in Tables 5 and 6. Table 5 summarizes the questions on attitudes and behaviors on red countdown

signal controls. Table 6 presents the questions on turning off engines while waiting for the green signal.

Table 5 reveals that most of the drivers surveyed are supportive of red countdown signal controls, and 66.51% of the surveyed drivers considered red countdown signal controls as having an impact on driving behaviors. The proportion of drivers (52.05%) who preferred overall countdowns is close to the proportion of drivers (47.95%) who selected partial countdowns. For question (19), 19.79% of the drivers would accelerate passing through the intersection, which is a very dangerous behavior.

At signalized intersections, the reckless behavior of a driver is largely constrained by other drivers or vehicles. Therefore, the behavior of the first driver in a certain lane is focused on. In Table 6, for question (20), 79.82% of the surveyed drivers would engage the engine gear in advance and accelerate to start once the green light changes. This behavior may cause traffic accidents in conflict directions with vehicles or delayed pedestrians. However, from another angle, it can improve traffic operational efficiency. Regarding question (21), 77.38% of the surveyed drivers who would turn off their engines while waiting would start the engines in advance and then accelerate to move while changing to the green light. For question (22), 58.33% of the drivers surveyed would turn off their engines while waiting when the waiting time is longer than 30 s, which reflects the driver’s awareness of conserving energy and reducing exhaust emissions. Question (23) shows that the main causes for not turning off engines at red countdown signalized intersections are feelings of inconvenience and concerns about fuel consumption when restarting the engine. Question (24) indicates that the main causes for not turning off engines are feelings of inconvenience and having no idea of how soon the red light will be over. Note that some questions have a plurality of possible causes or choices; therefore, the percentage sum of multiple-choice questions may be greater than 100%. The calculation principle is the

TABLE 5: Attitudes and behaviors of surveyed drivers on red countdown signal controls.

Questions	Options	Proportion (%)
(16) Do you support setting up a red countdown signal at intersections? [single choice]	(A) Support.	79.64
	(B) Do not support.	14.65
	(C) Does not matter.	4.85
	(D) Others.	0.86
(17) If red countdown signal controls are set up at intersections, what kind of impact will this have on you? [single choice]	(A) Through the countdown signal, I can know the remaining time of the red light, be ready to drive, and accelerate to pass through the intersection.	36.44
	(B) Through the countdown signal, I know the remaining time of the red light and can decide whether or not to turn off the engine according to the remaining time, thereby reducing fuel consumption.	30.07
	(C) Through the countdown signal, I can know the remaining time of the red light, and thus reduce the anxiety of waiting for a red light.	29.97
	(D) Either countdown or non-countdown has no impact on me.	3.52
(18) Do you think which of the following control modes of red countdown is better? [single choice]	(A) The countdown from the beginning to the end of the overall period of red light is better.	52.05
	(B) Beginning to show the countdown 30 s before the end of the red light is better.	12.94
	(C) Beginning to show the countdown 20 s before the end of the red light is better.	10.47
	(D) Beginning to show the countdown 10 s before the end of the red light is better.	24.55
(19) If you happen to approach an intersection at normal speed and no other vehicles are in front of your vehicle, that is to say, your vehicle is the first vehicle in a certain lane, if you find the red countdown is coming to end, what will you do? [single choice]	(A) Accelerate into the intersection.	19.79
	(B) Decelerate into the intersection.	56.99
	(C) Maintain the original speed and enter the intersection.	21.03
	(D) Others.	2.19

number of times an item is selected divided by the number of drivers surveyed.

4.5. Analysis of Red-Light Running Behaviors and Other Risk Behaviors. Questions (25)–(29) were designed to investigate driving behaviors on red-light running, sudden acceleration, sudden braking, and so on. Question (25) and its options are described in Table 7, and the results are shown in Figure 2.

Question (25) indicates that the proportion of drivers who would intentionally run a red light at countdown signalized intersections is 12.33% (Figure 2). Their main reason for running a red light is to rush through intersections within a very short time, even though the drivers may be well aware of the remaining green light time. In addition, 24.63% of the surveyed drivers revealed good driving behaviors by not running a red light.

Question (26) was designed to correspond to question (25). Its options are described in Table 8, and the results are shown in Figure 3.

Question (26) reflects that cases of red-light running at non-countdown intersections caused by not knowing the time the yellow light will appear and by the sudden transition of yellow light account for a large proportion (63.6%) of the

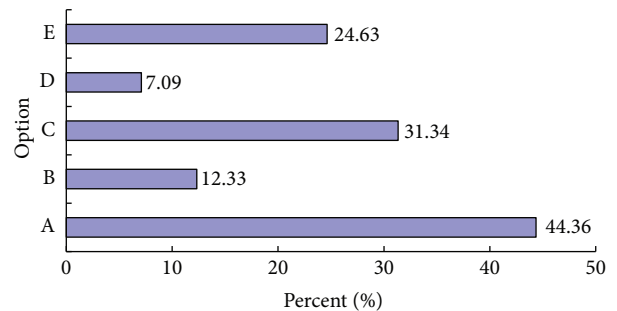


FIGURE 2: Distribution of reasons for running a red light at countdown signalized intersections.

drivers surveyed (Figure 3). Some drivers even attributed their red-light running to the lack of countdown signals. Red-light running that resulted from inattention still accounts for approximately one-third of the total (29.83%).

Question (27) and its options are described in Table 9, and the statistical results are illustrated in Figure 4. Effective countermeasures to reduce red-light running are installing

TABLE 6: Behaviors on whether engines should be turned off while waiting for the green light.

Questions	Options	Proportion (%)
(20) At countdown signalized intersections, if you do not turn off the engine while waiting, assuming your vehicle is the first vehicle in a certain lane, when you see the red countdown is coming to end, what will you do? [single choice]	(A) Engage the engine gear in advance, and then accelerate to pass through the intersection when the green light begins.	44.62
	(B) Engage the engine gear in advance, slowly glide, and then accelerate to pass through the intersection when the green light begins.	35.20
	(C) Continue to wait until the green light begins, and then engage the engine gear to pass through the intersection.	18.93
	(D) Others.	1.24
(21) At countdown signalized intersections, if you turn off the engine while waiting, assuming your vehicle is the first vehicle in a certain lane, when you see the red countdown is coming to end, what will you do? [single choice]	(A) Start the engine in advance, engage the engine gear beforehand, and then accelerate to pass through the intersection when the green light begins.	46.01
	(B) Start the engine in advance, engage the engine gear beforehand, slowly glide, and then accelerate to pass through the intersection when the green light begins.	31.37
	(C) Start the engine in advance, continue to wait until the green light begins, and then engage the engine gear to pass through the intersection.	20.53
	(D) Continue to wait until the end of the red light, start the engine, and then engage the engine gear to pass through the intersection.	2.09
	(E) Others.	0
(22) At countdown signalized intersections, you will turn off the engine to wait at how many remaining seconds of the red countdown? [single choice]	(A) 30–39 s.	5.33
	(B) 40–49 s.	5.71
	(C) 50–59 s.	6.95
	(D) 60–69 s.	20.36
	(E) >70 s.	19.98
	(F) Never turn off the engine.	41.67
(23) At red countdown signalized intersections, when you meet a red light but do not turn off the engine to wait, what are the causes? (Drivers who will turn off their engines do not need to reply to the question.) [multiple choices]	(A) Turning off and restarting the engine is inconvenient.	65.75
	(B) Turning off and restarting the engine will consume more fuel.	48.72
	(C) Turning off and restarting the engine will increase the wear of the vehicle.	27.69
	(D) Others.	7.33
(24) At non-countdown signalized intersections, when you meet a red light but do not turn off the engine to wait, what are the causes? (Drivers who will turn off their engines do not need to reply to the question.) [multiple choices]	(A) Turning off and restarting the engine is inconvenient.	57.66
	(B) Not knowing how much time of the red light is left.	54.52
	(C) Turning off and restarting the engine will increase the wear of the vehicle.	29.59
	(D) Others.	7.71

TABLE 7: Reasons for running a red light at countdown signalized intersections.

Questions	Options
(25) At countdown signalized intersections, what are your reasons for running a red light? [multiple choices]	(A) Inattention, not intentionally running a red light.
	(B) In a hurry, intentionally running a red light.
	(C) Intending to pass through the intersection before the end of the green countdown or the end of the yellow light, but runs a red light.
	(D) Others.
	(E) Never runs a red light.



TABLE 8: Reasons for running a red light at non-countdown signalized intersections.

Questions	Options
(26) At non-countdown signalized intersections, what are your reasons for running a red light? [multiple choices]	(A) Inattention, not intentionally running a red light.
	(B) Not knowing the time that the yellow light will appear, not intentionally running a red light.
	(C) Transition of the yellow light is too sudden, and I cannot brake to stop before the stop line, not intentionally running a red light.
	(D) In a hurry, intentionally running a red light.
	(E) Others.
	(F) Never runs a red light.

TABLE 9: Measures for reducing the chances of running a red light.

Questions	Options
(27) Which of the following measures do you think can reduce the chances of running a red light? [multiple choices]	(A) Install automatic facilities to capture the behavior of red-light running (e.g., electronic police, camera, etc.).
	(B) Install countdown signal lights.
	(C) Strengthen education on traffic safety, and raise awareness on traffic safety.
	(D) Others.

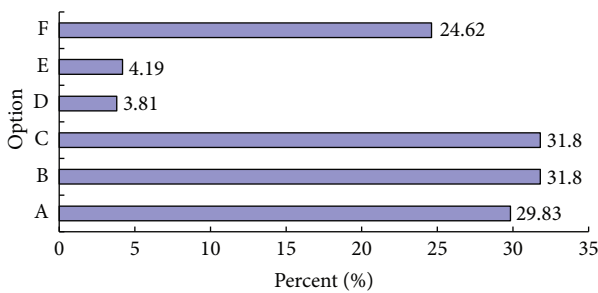


FIGURE 3: Distribution of reasons for running a red light at non-countdown signalized intersections.

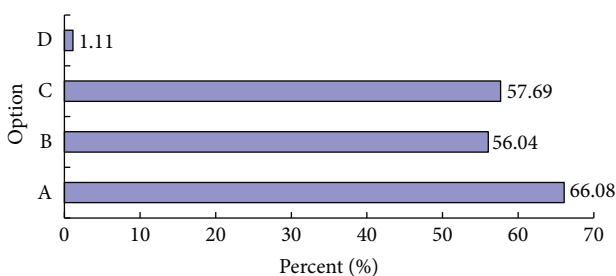


FIGURE 4: Distribution of measures to reduce the chances of running a red light.

automatic-capture facilities, enhancing education, and setting up countdown signal lights based on the degree of influence.

Questions (28)-(29) and their statistical results are shown in Table 10. The proportion of acceleration at the end of the green light at countdown signalized intersections is 26.07% more than that at non-countdown signalized intersections. The key factor is the existence of the countdown signal.

For sudden braking, the possibility of occurrence at non-countdown signalized intersections is higher than that at countdown signalized intersections. Regardless of whether a driver is at countdown signalized intersections or at non-countdown signalized intersections, sudden braking will increase while the yellow light starts.

4.6. *Attitudes and Understanding of Drivers on Display Modes of Countdown Signals.* Questions (30)-(32) were designed to investigate the attitudes of drivers on the display modes of countdown signals. The statistical results are shown in Table 11.

According to Table 11, most of the surveyed drivers selected the mode of countdown display with overall lights (red, green, and yellow). Nearly half of the surveyed drivers considered the red countdown light to be beneficial in improving traffic operational efficiency.

4.7. *Cross-Analysis of Typical Questions.* The psychological and behavioral characteristics of drivers are closely related to the driver's gender, age, and driving experience. The cross-analysis between gender and attitudes toward countdown signals is carried out based on questions (1) and (7) of the survey data. The analysis results shown in Table 12 indicate that the majority of male and female drivers supported the setting up of countdown signal controls, and male drivers are more inclined to support countdown signal controls.

The cross-analysis between gender and behaviors before the end of the green countdown is conducted based on questions (1) and (14), as listed in Table 10. Table 13 shows that the proportions of male and female drivers are very close in decelerating to stop by the end of the green countdown. Compared with the male drivers, the surveyed female drivers were also more likely to accelerate passing through the stop line by the end of the green countdown or by the end of

TABLE 10: Risky driving behaviors at different intersections.

Questions	Options	Proportion (%)
(28) What kind of signal controls make you inclined to accelerate to pass through an intersection? [single choice]	(A) Countdown signal control, the green countdown is coming to an end.	61.75
	(B) Non-countdown signal control, the green light is flashing.	35.68
	(C) Others.	2.57
(29) What kind of signal controls that make you incline to urgently decelerate? [single choice]	(A) Countdown signal control, the green countdown is coming to an end.	19.31
	(B) Countdown signal control, the yellow countdown begins to counting.	27.40
	(C) Non-countdown signal control, the green light is flashing.	21.50
	(D) Non-countdown signal control, the yellow light is starting.	30.73
	(E) Others.	1.05

TABLE 11: Attitudes and understanding of surveyed drivers on the display modes of countdown signals.

Questions	Options	Proportion (%)
(30) Do you think it is reasonable to show only the red countdown, but not the yellow and green countdowns? [single choice]	(A) Reasonable.	23.69
	(B) Unreasonable.	65.37
	(C) Does not matter.	10.94
(31) If you answered “Reasonable” in question (30), what are the reasons? [multiple choice]	(A) It is not easy to lead to red-light running at the end of the green light.	85.82
	(B) It allows the drivers waiting for the red light to be prepared to start in advance, which is conducive to traffic operational efficiency.	47.76
	(C) Others.	5.33
(32) If you answered “Unreasonable” in question (30), what are the reasons? [multiple choice]	(A) It is easy to lead to red-light running at the end of the green light.	67.46
	(B) It allows the drivers waiting for the red light to be prepared to start in advance, which may result in a traffic accident, especially with no vehicle waiting for the red light in a certain lane.	46.81
	(C) Others.	3.62

TABLE 12: Cross-analysis between gender and attitudes toward countdown signals.

X	Y				Total
	(A) Support	(B) Do not support	(C) Does not matter	(D) Others	
(A) Male	87.75%	6.84%	0.85%	4.56%	100%
(B) Female	76.12%	16.42%	3.48%	3.98%	100%

TABLE 13: Cross-analysis between gender and behaviors before the end of the green countdown.

X	Y				Total
	(A) Accelerate and pass the stop line before the end of the green countdown or before the end of the yellow light	(B) Decelerate and make sure to stop before the stop line before the end of the yellow light	(C) Maintain the original speed; if I cannot safely stop in front of the stop line, then pass stop line before the end of the yellow light	(D) Others	
(A) Male	18.23%	62.96%	14.25%	4.56%	100%
(B) Female	20.40%	61.69%	10.95%	6.97%	100%

the yellow light, indicating that female drivers may not be conservative about the behavior in question.

The cross-analysis between gender and behaviors while the green light is turning into the yellow light at non-countdown intersections is conducted based on questions (1)

and (15), as shown in Table 14. According to the results, at non-countdown intersections, for the option “accelerate and pass the stop line before the end of the yellow light,” male drivers are more aggressive than female drivers. In comparison with question (14), the female drivers surveyed were more

TABLE 14: Cross-analysis between gender and behavior before the end of the green light at non-countdown signalized intersections.

X	Y				Total
	(A) Accelerate and pass the stop line before the end of the yellow light	(B) Decelerate and make sure to stop before the stop line before the end of the yellow light	(C) Maintain the original speed; if I cannot safely stop in front of the stop line, then pass the stop line before the end of the yellow light	(D) Others	
(A) Male	19.09%	67.24%	10.83%	2.85%	100%
(B) Female	4.98%	75.62%	13.93%	5.47%	100%

conservative in the condition of uncertain remaining time than male drivers.

## 5. Conclusion and Discussion

According to the analysis results of the survey, several conclusions can be drawn. (1) Most of the surveyed drivers preferred countdown signalized intersections and tended to select the mode of countdown display of overall lights (red, green, and yellow). (2) Most of the drivers considered countdown signal controls as capable of improving not only traffic safety but also traffic operational efficiency, which is consistent with the findings from earlier studies [17, 19] but is contrary to the studies in [18, 20]. (3) Regardless of whether green countdown or red countdown controls are set up, most of the drivers considered countdown signal controls as having an impact on driving psychologies and behaviors. However, the impact may not be conducive to improve traffic safety. (4) The proportion of drivers intentionally running red lights is relatively small at countdown signalized intersections or non-countdown signalized intersections. However, the time by the end of the green signal and at the onset of the yellow signal is the key time of red-light running at both types of intersections. According to the survey results, the installation of an automatic-capture system to catch traffic violations is conducive to reduce the occurrence of red-light running. (5) Female drivers are traditionally viewed as having more conservative driving behaviors compared with male drivers. However, the analysis results indicate that the driving behaviors of female drivers surveyed are not conservative under clear green countdown conditions. Nevertheless, female drivers are very conservative under non-countdown conditions, which confirms the general psychological characteristics indicating that males are more adventurous than females under unknown conditions.

Driving psychologies and behaviors are complex phenomena. To further study the effects of countdown signals on driving psychologies and behaviors, several ways may be recommended: (1) using professional equipment to collect indicator parameters of driving psychologies and behaviors at countdown and non-countdown signalized intersections with actual traffic conditions and then analyzing the data; (2) at countdown signalized or non-countdown signalized intersections, observing or photographing driving behaviors and then analyzing the behaviors; and (3) using a more scientific

comparison and analysis of data obtained by different methods to draw reasonable conclusions.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

The authors acknowledge the support of the Natural Science Foundation of Shandong Province, China (ZR2016EEM14 and ZR2012EEM05), the National Natural Science Foundation of China (51408288 and 51505244), and the Humanities and Social Science Fund Project under the Ministry of Education of the People's Republic of China (12YJCZH162).

## References

- [1] F. Pan, L. Zhang, J. Lu, Q. Xiang, and L. Lu, "Application of access management techniques in safety improvement at highway intersections," *Journal of Beijing University of Technology*, vol. 37, no. 2, pp. 237–242, 2011.
- [2] F. Pan, L. Zhang, J. Lu, J. J. Zhao, and F. Wang, "A method for determining the number of traffic conflict points between vehicles at major-minor highway intersections," *Traffic Injury Prevention*, vol. 14, no. 4, pp. 424–433, 2013.
- [3] Standardization Technical Committee on Transportation in Ministry of Public Security of China, *Road Traffic Counting Down Display Unit (GA508-2004)*, The Ministry of Public Security of the People's Republic of China, 2004.
- [4] Standardization Administration of China, *Specification for Setting and Installation of Road Traffic Signals (GB14886-2006)*, Standards Press of China, Beijing, China, 2006.
- [5] Standardization Administration of China, *Road Traffic Signal Lamps (GB14887-2011)*, Standards Press of China, 2011.
- [6] Y. Wang, "Why not adopt the red light countdown devices at the signalized intersections in Shanghai city," *Traffic and Transportation*, vol. 2, no. 17, 1999.
- [7] K. M. Lum and H. Halim, "A before-and-after study on green signal countdown device installation," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 9, no. 1, pp. 29–41, 2006.
- [8] M. R. Ibrahim, M. R. Karim, and F. A. Kidwai, "The effect of digital count-down display on signalized junction performance," *American Journal of Applied Sciences*, vol. 5, no. 5, pp. 479–482, 2008.

- [9] T. Limanond, S. Chookerd, and N. Roubtonglang, "Effects of countdown timers on queue discharge characteristics of through movement at a signalized intersection," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 6, pp. 662–671, 2009.
- [10] Y.-C. Chiou and C.-H. Chang, "Driver responses to green and red vehicular signal countdown displays: safety and efficiency aspects," *Accident Analysis and Prevention*, vol. 42, no. 4, pp. 1057–1065, 2010.
- [11] A. Sharma, L. Vanajakshi, V. Girish, and M. S. Harshitha, "Impact of signal timing information on safety and efficiency of signalized intersections," *Journal of Transportation Engineering*, vol. 138, no. 4, pp. 467–478, 2012.
- [12] P. Papaioannou and I. Politis, "Preliminary impact analysis of countdown signal timer installations at two intersections in Greece," *Procedia Engineering*, vol. 84, pp. 634–647, 2014.
- [13] J. Devalla, S. Biswas, and I. Ghosh, "The effect of countdown timer on the approach speed at signalised intersections," *Procedia Computer Science*, vol. 52, pp. 920–925, 2015.
- [14] M. R. Islam, D. S. Hurwitz, and K. L. Macuga, "Improved driver responses at intersections with red signal countdown timers," *Transportation Research Part C: Emerging Technologies*, vol. 63, pp. 207–221, 2016.
- [15] Y. Wang and X. Yang, "Discussion on setting traffic signals with counting down display unit at intersection based on traffic safety," *China Safety Science Journal*, vol. 16, no. 3, pp. 55–70, 2006.
- [16] W.-J. Wu, Z.-C. Juan, and H.-F. Jia, "Drivers' behavioral decision-making at signalized intersection with countdown display unit," *Systems Engineering—Theory and Practice*, vol. 29, no. 7, pp. 160–165, 2009.
- [17] J. Zhang, Y. He, X. Sun, and X. Liu, "Effects of countdown signal at urban intersection on driving behaviors," *Journal of Transport Information and Safety*, vol. 27, no. 5, pp. 99–101, 2009.
- [18] H. Qian and H. Han, "Influence of countdown of green signal on traffic safety at crossing," *China Safety Science Journal*, vol. 20, no. 3, pp. 9–13, 2010.
- [19] W. Ma, Y. Liu, and X. Yang, "Investigating the impacts of green signal countdown devices: empirical approach and case study in China," *Journal of Transportation Engineering*, vol. 136, no. 11, pp. 1049–1055, 2010.
- [20] H. Qian, "Influence of red signal countdown on traffic safety and efficiency," *Journal of Transport Information and Safety*, vol. 29, pp. 65–68, 2011.
- [21] K. Long, L. D. Han, and Q. Yang, "Effects of countdown timers on driver behavior after the yellow onset at Chinese intersections," *Traffic Injury Prevention*, vol. 12, no. 5, pp. 538–544, 2011.
- [22] H. Huang, D. Wang, L. Zheng, and X. Li, "Evaluating time-reminder strategies before amber: common signal, green flashing and green countdown," *Accident Analysis and Prevention*, vol. 71, pp. 248–260, 2014.
- [23] Z. Li, J. Zhang, J. Rong, J. Ma, and Z. Guo, "Measurement and comparative analysis of driver's perception–reaction time to green phase at the intersections with and without a countdown timer," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 22, pp. 50–62, 2014.
- [24] F. Pan, D. Yun, L. Zhang, Y. Ma, Y. Meng, and H. Tang, "Analysis and modeling of behavior of catching green signal at countdown signalized intersections," *China Safety Science Journal*, vol. 25, no. 7, pp. 147–152, 2015.
- [25] C. Fu, Y. Zhang, Y. Bie, and L. Hu, "Comparative analysis of driver's brake perception–reaction time at signalized intersections with and without countdown timer using parametric duration models," *Accident Analysis and Prevention*, vol. 95, pp. 448–460, 2016.
- [26] F.-Q. Pan, L.-X. Zhang, T. Liu, G.-X. Kang, M. Li, and F.-Y. Wang, "Modeling of driving behaviors at countdown signalized intersections considering the value of car," *Journal of Transportation Systems Engineering and Information Technology*, vol. 16, no. 2, pp. 64–69, 2016.

## Research Article

# Large-Scale Demand Driven Design of a Customized Bus Network: A Methodological Framework and Beijing Case Study

Jihui Ma,<sup>1</sup> Yang Yang,<sup>1</sup> Wei Guan,<sup>1</sup> Fei Wang,<sup>1</sup> Tao Liu,<sup>2</sup> Wenyuan Tu,<sup>1</sup> and Cuiying Song<sup>1</sup>

<sup>1</sup>MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing 100044, China

<sup>2</sup>Department of Civil and Environmental Engineering, The University of Auckland, Auckland 1142, New Zealand

Correspondence should be addressed to Yang Yang; 11114218@bjtu.edu.cn

Received 10 January 2017; Revised 10 March 2017; Accepted 16 March 2017; Published 29 March 2017

Academic Editor: Xiaolei Ma

Copyright © 2017 Jihui Ma et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, an innovative public transportation (PT) mode known as the customized bus (CB) has been proposed and implemented in many cities in China to efficiently and effectively shift private car users to PT to alleviate traffic congestion and traffic-related environmental pollution. The route network design activity plays an important role in the CB operation planning process because it serves as the basis for other operation planning activities, for example, timetable development, vehicle scheduling, and crew scheduling. In this paper, according to the demand characteristics and operational purpose, a methodological framework that includes the elements of large-scale travel demand data processing and analysis, hierarchical clustering-based route origin-destination (OD) region division, route OD region pairing, and a route selection model is proposed for CB network design. Considering the operating cost and social benefits, a route selection model is proposed and a branch-and-bound-based solution method is developed. In addition, a computer-aided program is developed to analyze a real-world Beijing CB route network design problem. The results of the case study demonstrate that the current CB network of Beijing can be significantly improved, thus demonstrating the effectiveness of the proposed methodology.

## 1. Introduction

With the rapid increase in urbanization, many Chinese cities are facing problems associated with urban environments, such as increased traffic congestion, serious environmental pollution, and extreme energy deficiencies. Furthermore, with the gradual increase in the standard of living of urban residents, existing public transportation (PT) has not satisfied the travel demands of passengers, is unable to encourage private car passengers to switch travel modes, and has been unable to improve PT at attractive rates. Therefore, customized buses (CBs) have been established across China as an innovative mode of PT services. Network planning is a major component of CB systems. The rationality and proportionality of network planning play a vital role in the entire CB operational system. Therefore, scientific and systematic research on CB network planning must be conducted. Scientific and reasonable network planning can maximize the use of CB resources, satisfy the travel demands of most

passengers, and improve the service quality of CBs while reducing operating costs for CB operators and improving the attractiveness of CBs. Therefore, the study reported in this paper has great significance for effectively encouraging private car owners to change travel modes, ease traffic congestion, and mitigate the problems of air pollution.

Since the CB concept was first introduced, scholars have conducted a considerable amount of research, the majority of which remains theoretical in nature. In addition, less specific methods have been used for CB research. Kirby and Bhatt [1] analyzed ten specific CB cases and discussed the potential impact of CBs, including easing urban traffic congestion, environmental pollution, and energy consumption. They also developed guidelines for CB passenger recruitment, network planning, operation scheduling, and fares. In addition, the authors in [2] discussed the seven main features of CBs that assured successful operation, such as having more than 50 same points and ends on the long-distance lines, an organization for operations management, constantly adjusting the



lines and scheduling to meet demands, and guaranteeing seats with personalized service. At the time, the customized services were provided by small, private organizations, which faced such problems as increased numbers of commuters, difficulties in the management of operations, and a lack of security and funds for expansion. Therefore, better services could not be provided to passengers. For those reasons, Bautz [3] proposed CBs as a part of urban PT. They compared the cost of CBs to those of different operating modes, and concluded that, compared with private operation organizations, the encouragement and coordination by governmental operations organizations can maximize the benefits of customized PT. McCall [4] analyzed commuter buses that had 47 lines and provided service to more than 2,000 commuters in Ventura, Los Angeles, and Orange Counties. The commuter bus network was a business model whereby the local private sector did not accept subsidies and had only slightly longer travel times than travelling by private car but had considerably lower travel costs than private cars. The successful operation of the commuter buses played a critical role in reducing environmental degradation, but the majority of the PT system required high subsidies for normal operations. McKnight and Paaswell [5] comprehensively analyzed the Chicago CB network. Their results showed that the scope of the Chicago CB service was small, and the main role was to ease peak rail transport travel. A series of steps to expand the PT market were proposed by them. First, the cause of CB travel demand was determined. Then, methods to meet the demands of commuter travel were established using a chart analysis and market research. At the same time, the paper changed rail operations, which were scheduled to support the development of the CBs. Shaheen et al. [6] reviewed the CB development process after the car-sharing concept was proposed and summarized the advantages of CBs: passengers could save costs, and society could reduce the demand for parking spaces and wasted resources. They also suggested that the future development trend of CBs was a constant expansion of the scope of services and the use of more advanced booking technology. Chang and Schonfeld [7] constructed mathematical models for custom transit fixed-route buses and conventional flexible lines. The vehicle size and service space were based on an optimization of the decision variables and total system cost. The objective was to minimize the combination of operational costs and user costs. Martin and Shaheen [8] suggested that passengers encouraged to participate in a car-sharing service could generate significant traffic, land use, and social and environmental benefits, including a reduction in the total number of kilometers of vehicular travel and carbon dioxide emissions. Their study focused on the impact of car sharing on greenhouse gas emissions through theoretical and methodological systems. Potts et al. [9], who studied CB services in the United States and Canada for nearly a decade, proposed different CB modes that were applicable to large, medium, and small cities and rural areas. His research provided guidance on whether and how to open CB lines, combined with the practical situations of CB operators in different areas. Duncan [10] stated that car-sharing services have been vigorously developed in the United States in recent years. Such services enable sustainable

development for an urban transportation system but also could bring greater benefits to cities. The most effective way to increase car-sharing functions was cost savings. Based on the quantification and comparison of the potential cost savings of different travel modes, the car-sharing services are the most cost-effective ones. El Fassi et al. [11] noted that CB operator organizations must continue to improve reticle layouts and increase station capacities to meet growing passenger demands. In addition, policy makers often saw a loss of resources, time, and market share if they made decisions based solely on experience. Thus, providing decision support to policy makers based on a discrete simulation event was proposed. Such decision support could optimize the network, maximize passenger satisfaction, and minimize the number of buses. De Lorimier and El-Geneidy [12] used a multilevel regression analysis method to determine the factors that influenced the effectiveness of decisions to use vehicles based on the CB system in Montreal, Canada. That method provided a reference for building or expanding PT networks for CB operators. Nair and Miller-Hooks [13] constructed a balanced network model to determine the optimal configuration of a CB system. Passengers could take the nearest bus according to their demands. The operator must determine the optimal location of the station, the number of buses, and the station capacity to maximize benefits. Le Vine et al. [14] studied two CB modes: point to point and round trip. This paper suggests that the number of prospective subscribers to a point-to-point CBs in London is between three and four times as large as the comparable number for round-trip CBs. Point-to-point CBs could be used as an alternative to PT, and round-trip CBs could be used as a complement to a point-to-point CB. Liu and Ceder [15] studied the developmental background of Chinese CBs, analyzed China CB network planning and operation processes, and summarized the advantages and disadvantages of CBs and trends in its development in China. In addition, their studies provided a reference for CB operators for policy formulation and academic research.

Compared with the number of studies focusing on the theory of CB, there have been relatively few CB network planning studies. However, there has been increasing research on conventional bus network planning theory and methods. Lampkim and Saalmans [16] studied a specific case. That paper better optimized the entire PT system and achieved more efficient resource utilization by redesigning the network, determined departure frequencies, and developed better timetables and vehicle scheduling than the previous network plan. Ceder and Wilson [17] summarized different bus network planning methods and proposed an easier implementation method in combination with the advantages of a previous method. That paper considered the interests of passengers and operators and presented the design of a new transit network algorithm. Baaj and Mahmassani [18] combined artificial intelligence methods and a genetic algorithm to solve the problem of bus network planning. Tom and Mohan [19] used the operators' cost and total passenger travel time as a total system cost and objective function. First, a series of candidate paths were established. A genetic algorithm was then used to select the line with

minimum cost and determine the departure frequency. Jerby and Ceder [20] proposed that the reasonable planning of a subway PT network could attract passengers to PT. Their method was based on a situation in which a large number of passengers used private cars to travel to a subway transfer station, which caused traffic congestion and a subway station parking space overload. Based on the Rome PT network, Cipriani et al. [21] used a parallel genetic algorithm to solve a complex network topological structure, multimode PT system, large-city PT network planning problem with the characteristics of multi-to-multi bus travel demands. Nikolić and Teodorović [22] used the artificial bee colony algorithm to design a PT network that minimized the total travel time of passengers, maximized passenger satisfaction, and minimized total passenger transfer times as the objective function. Badia et al. [23] studied the PT network planning method of a radial network city. The size of the center area and the departure interval were decision variables. In addition, the operating cost and minimum passenger cost were the objective functions for the preferred line.

In the research of bus data processing methods, there have been relatively few CB data processing methods studies. However, there has been increasing research on conventional bus data processing methods. Trépanier et al. [24] presented a model to estimate the destination location for each individual boarding a bus with a smart card. Experiments carried out with a database programming approach showed that the data must be thoroughly validated and corrected prior to the estimation process. Li [25] investigated statistical inference for a transit route O-D matrix using on-off counts of passengers, created a Markov chain model, and inferred the unknown parameters of the Markov model using Bayesian analysis. After that, Ma et al. [26] developed a Markov chain based Bayesian decision tree algorithm to extract passengers' origin data from recorded SC transaction information. Using the time invariance property of the Markov chain, the algorithm was further optimized and simplified to have a linear computational complexity. Munizaga and Palma [27] presented a methodology for estimating a public transport OD matrix from smartcard and GPS data for Santiago of Chile, obtained detailed information about the time and position of boarding public transportation, and generated an estimation of time and position of alighting for over 80% of the boarding transactions. Ma et al. [28] proposed an efficient and effective data-mining procedure that models the travel patterns of transit riders in Beijing of China and identified the transit riders' trip chains based on the temporal and spatial characteristics of their smart card transaction data. In addition, Ma and Wang [29] attempted to develop a data-driven platform for online transit performance monitoring and Ma et al. [30] developed a series of data-mining methods to identify the spatiotemporal commuting patterns of Beijing public transit riders.

A review of the related literature on CB research demonstrates that CB research has placed more emphasis on the theoretical and practical significance of CBs. Relatively less research has been performed on CB network planning, but there are more references to conventional bus network planning theory. However, as an innovative PT mode, CB has its

own characteristics. This paper proposes an area clustering algorithm based on the travel demands of passengers because CBs currently lack network planning and the CB resource allocation efficiencies are not high. A multiobjective integer programming model is established in combination with the background and significance of CB and comprehensively considering CB operating costs, environmental costs, and traffic congestion costs. Based on the actual CB travel demands in Beijing, this paper reports a case study and demonstrates the rationality and effectiveness of the proposed method. The results of this research provide a new method for CB network planning that will be useful for guiding practical efforts.

The remainder of this paper is organized as follows. In Section 2, the details of the proposed CB network planning method based on area division are proposed. The specific content of the method is described in Section 3. The method is demonstrated through a case study in Section 4. The summary of this paper, limitations, and next steps are discussed in Section 5.

## 2. Details of the Proposed CB Network Planning Method

CB network planning differs from conventional bus network planning in that it is a bottom-up activity based on travel demand. A travel demand survey of passengers is the most important data source for CB network planning, which provides the most important data base for CB network planning. CB travel demand surveys are typically questionnaires provided on the Internet. The contents of the questionnaire are mainly focused on passengers' trip ODs, travel times, and travel purpose. CB provides travel services for passengers based on their travel demands. Therefore, an analysis of passenger travel demand is key to the operation of CBs.

In this paper, considering CB network planning and using concept from a point-to-line layout into a network, the original (O) and destination (D) areas of a line are first divided according to the travel demands of passengers. The operating line scheme is then determined according to the model solution, all operation lines are laid out, and the final CB network is determined.

In this paper, in combination with the conventional bus and existing CB network planning experience, a CB network planning method is proposed based on area division. The specifics of this method are as follows.

(1) *Large-Scale Travel Demand Data Processing.* A travel demand survey was the basis of the transportation planning and provided comprehensive and accurate data for transportation planning. CB travel demand surveys are typically questionnaires provided on the Internet. Because CB operations are based on the actual travel demands of passengers, it is important to determine passengers' trip ODs, travel times, and travel purpose. The passengers' travel demand data were analyzed, and the collected data were quantified to provide data support for network planning.

(2) *CB Line OD Area Division.* CB operators do not use bus stops. The line OD area division is a component of the line and is an important part of network planning. The number of area divisions decreases with increases in the radius of

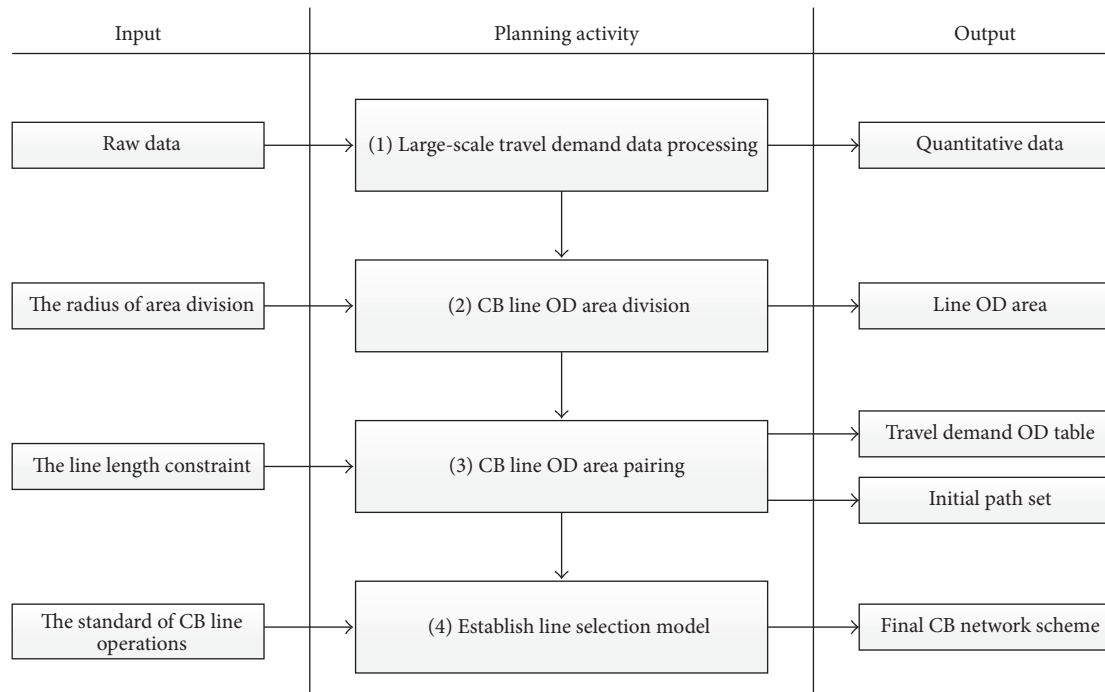


FIGURE 1: CB network planning process.

the line OD area division, which can effectively reduce the operating costs; however, it is difficult to attract passengers because passengers' walking distance is increased. On the contrary, reductions in the radius of the line OD area division can increase operating costs and lead to losses. These two conditions are not conducive to the long-term development of CBs. This paper divides the CB line OD area according to the passengers' position distribution using an area division algorithm, which places similar demands on an area and establishes a reasonable radius that the line OD area must cover. Passengers board the bus at the origin area stops and exit at the destination area stops, which can prevent halfway stops. This scheme will satisfy the demands of large numbers of passengers, make full use of resources, and effectively reduce companies' operating costs.

(3) *CB Line OD Area Pairing.* After the line OD area is determined, different CB lines can be established by pairing the OD areas. Using travel demand data processing, travel demand OD tables can be developed based on the line OD area. By deleting lines that do not conform to the distance constraint, the initial path set can be determined and the OD table can be updated.

(4) *Establish the Line Selection Model.* The objective of CB operations is to provide a comfortable and rapid riding environment for passengers, ease urban traffic congestion, and solve the problem of air pollution. These factors are regarded as the standard of CB line operations. The operating lines and CB network scheme were determined in this study by constructing an objective function for minimizing the sum of the operating, environmental, and traffic congestion costs and using the linear programming method to obtain

a solution. The CB network planning process based on area division is shown in Figure 1.

### 3. CB Network Planning Method

3.1. *Large-Scale Travel Demand Data Processing.* Quantitative data related to CB network planning can be obtained by sorting and calculating a series of raw data collected from the demand survey; these data can provide data support for the planning method. First, according to the passenger OD survey, the latitudes and longitudes of the passenger ODs can be determined. The latitude and longitude coordinates are then translated into planar coordinates using software so that the Euclidean distance between the origins and destinations of each demand can be calculated. Finally, the coordinates are marked on the diagram.

3.2. *CB Line OD Area Division.* Area division algorithms primarily include hierarchical clustering and K-means clustering, and the choice of algorithm is based on the purpose and application of the data analysis [31]. The K-means cluster method needs to set a parameter in advance, indicating the number of classes. Besides, the result is sensitive to the initial core of the data. In the clustering process of CB travel demand, the number of classes is unknown, and the initial core is selected randomly. Agglomerative hierarchical clustering is based on the bottom-up strategy, it aggregates a given data set according to the distance measurement criteria between categories until a certain condition is satisfied. Thus, the agglomerative hierarchical clustering method is more suitable for the CB travel demand clustering.

Agglomerative hierarchical clustering is based on the bottom-up strategy. First, each object is treated as a category, the distance between the objects is calculated, and the initial distance matrix is obtained. Two categories of minimum distances between categories are merged into one category, and the distances between the new category and all other categories are recalculated. The previous step is repeated, and the categories become increasingly larger until all objects are merged into a single category or a certain end condition is satisfied. According to the different distance measurement criteria between categories, hierarchical clustering can be divided into 4 types: (1) the average distance method, (2) the minimum distance method, (3) the maximum distance method, and (4) the barycenter method. Considering the actual definition of the distance between categories, the maximum distance method was selected as the distance measurement criteria for the hierarchical clustering. The distance between the two points that are most distant in the two categories is expressed as the distance between the two areas, which is formulated in (1). When the distance between the two areas is less than the maximum distance, the two categories are combined into a single category.

$$D(A, B) = \max_{i \in A, j \in B} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad (1)$$

where  $A, B$  are the sets of points that belong to categories  $A$  and  $B$ , respectively,  $x_i, y_i$  are the horizontal and vertical coordinates of the points that belong to category  $A$ , respectively,  $x_j, y_j$  are the horizontal and vertical coordinates of the points that belong to category  $B$ , respectively, and  $D(A, B)$  is the distance between categories  $A$  and  $B$ .

In this paper, the data set was initially divided using hierarchical clustering. Then, a hierarchical clustering tree was generated, the maximum distances between the categories were used to determine the number of categories  $m$ , the data points were divided into  $m$  categories, and the area division scheme was determined. Additional details are provided below.

(1) By collecting the travel demand data, the origin address set and the destination address set can be obtained. Based on the plane coordinate set of the origins and destinations determined by processing the travel demand data in Section 3.1, the points of the origin and destination sets are marked on the diagram.

(2) The distance between all the points in the origin set and destination set is separately calculated, and the two data sets based on the maximum distance measurement criteria are then separately classified. Two categories that are separated by a distance less than the maximum distance are combined into a large category until all points are merged into a single category, and the hierarchical clustering tree is generated.

(3) According to expert opinion or practical experience, the area division radius can be determined. In addition, the maximum distance between the categories also can be determined. In this way, the number of categories  $m$  is determined according to the maximum distances between the categories.

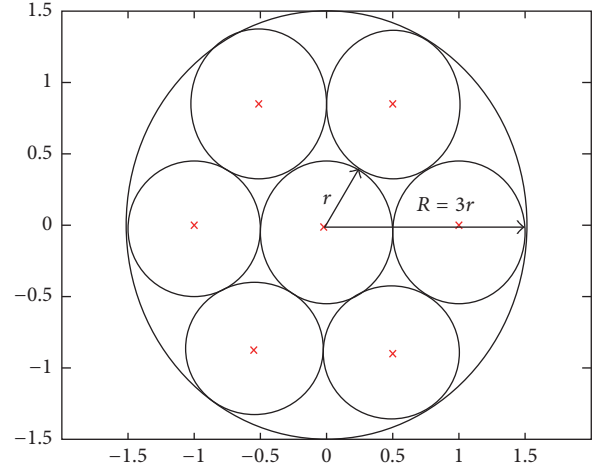


FIGURE 2: Diagram of the method for determining the area division radius.

(4) The two data sets are classified separately according to the number of categories  $m$  in the last step. Each category represents an origin area or destination area. The centroid coordinates of each origin area and each destination area can be calculated using (2). The centroid is then marked on the diagram, and the final area division scheme is determined.

$$(X_m, Y_m) = \left( \frac{\sum_{i=1}^n x_i}{n}, \frac{\sum_{i=1}^n y_i}{n} \right), \quad (2)$$

where  $X_m, Y_m$  are the horizontal and vertical coordinates of the centroid of category  $m$ , respectively,  $x_i, y_i$  are the horizontal and vertical coordinates of a certain point belonging to category  $m$ , respectively, and  $n$  is the total number of points in category  $m$ .

If there are an excessive number of stops in the CB OD areas, the time spent at the stops will be excessively long, which will increase the passenger travel time. Therefore, considering the passengers' walking distances, the radius of the area division cannot be overly large. However, operating costs will increase if the radius of area division is overly small. Therefore, the radius of the CB area division should be reasonably selected. The method for determining the area division radius is shown in Figure 2. The large circle represents a category that is an OD area, whereas the small circle represents the coverage of a stop in the OD area. The radius  $r$  of the small circle indicates the passengers' walking distances, which are between 500 and 1,000 m. No more than 7 stops in an area are appropriate. If the 7 stops are distributed in the area and guarantee the maximum areal coverage, an area coverage radius  $R$  of  $3r$  can be obtained. Therefore, the area coverage radius is 1.5 to 3 km. Then, the range of the maximum category distance is the diameter of the circle, which is 3 to 6 km. The reasonable value of the maximum category distance is typically determined based on the actual situation, which includes the city scale, number of buses, operating scale, and number of commuters.

3.3. *CB Line OD Area Pairing.* After the line OD area is determined, different CB lines can be constituted by pairing



the OD areas. Assuming that there are  $m$  origin areas and  $n$  destination areas,  $m \times n$  bus lines can be connected by pairing. Each path from the origin area to the destination area is represented as a bus line. The travel demand data corresponding to those  $m \times n$  lines are processed. By taking the number of people from the  $i$  origin area to the  $j$  destination area as the OD quantity of this line, the travel demand OD table can be determined by integration. Each cell in the demand OD table represents the travel demand of a line from an origin area to a destination area.

CBs can reduce passengers' commuting transfer and travel times and reduce travel costs to attract private car passengers. If the passenger travel distance is short, traditional buses can satisfy the passenger travel demand at a low price and high non-bus stop rate. If the passenger travel distance is long, the CB can effectively reduce transfer times. In Beijing, the round-trip travel cost of a 20 km indicates that CB costs are 30% of the cost of travelling by private car and 15% of the cost of taking a taxi. Thus, CBs can effectively reduce the travel costs of private car passengers. Therefore, the distance from the origin area to the destination area, which is also the line, should not be overly short. In this paper, the distance between the origin area clustering center and the destination area clustering center is represented as the line length. The length of each line  $l_b$  is given by a logistic function:

$$l_b = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}, \quad (3)$$

where  $l_b$  is the length of each line,  $b = \{1, 2, \dots, m \times n\}$ ,  $X_i, Y_i$  are the horizontal and vertical coordinates of the origin area centroid, respectively,  $i = \{1, 2, \dots, m\}$ ,  $X_j, Y_j$  are the horizontal and vertical coordinates of the destination area centroid, respectively, and  $j = \{1, 2, \dots, n\}$ .

$l_b$  should not be less than the minimum line length constraint:

$$l_b \geq l_{\min}, \quad (4)$$

where  $l_{\min}$  is the minimum line length.

In this paper, the minimum line length was set to  $l_{\min} = 8$  km. An initial path set  $L_0 = \{l_1, l_2, \dots, l_k\}$  is formed by all  $k$  lines that satisfy the minimum line length constraint. After determining the initial path set, the travel demand OD table should be updated. If the line is not included in the initial path set, the value of the line travel demand is then changed to 0.

**3.4. Establishing the Line Selection Model.** The main considerations of CB operations are passenger comfort, economical operation, and social benefits. Because the aforementioned area division algorithm has been designed to ensure that passenger walking distances are not overly long, the line selection model in this section mainly considers the remaining two aspects: economical operation and social benefits. The operation of a CB will increase the operating costs of a company, but with the CB operations, the company can attract more private car passengers who choose CB, which will reduce vehicle pollution emissions and ease traffic congestion. Therefore, there is a certain contradiction between operating costs and

social benefits in CB operations. In this section, considering those two aspects, a CB line selection model that is based on total cost is established. By solving the model, the directions of the CB lines and the number of buses used in each line that minimize the total cost of a single line can be determined.

The model is established based on the following assumptions:

- (1) Passengers travel by either bus or car.
- (2) The linear distance between the origin area clustering center and destination area clustering center is represented by the CB line length.
- (3) Each line uses the same CB vehicle.

The target function presented in this paper consists of three parts. The first is the company's operating cost, including the CB operating cost per kilometer and fixed cost (including drivers' and conductors' salaries, management fees, and maintenance fees). The second is the environmental pollution cost, including the pollution cost of CBs and car emissions. The third is road congestion cost, including the additional time cost to passengers who travel by CB and by car because of traffic congestion.

The company's operating cost is determined using the following logistic function:

$$Z_1 = c_{Gm} \times \left[ \frac{n_{Gi}}{\alpha_G} \right] \times l_i + c_{Go} \times \left[ \frac{n_{Gi}}{\alpha_G} \right], \quad (5)$$

where  $Z_1$  is the total operating cost,  $c_{Gm}$  is the fuel cost per kilometer per vehicle,  $c_{Go}$  is the fixed operating cost per vehicle per day,  $l_i$  is the mileage per line,  $i = \{1, 2, \dots, k\}$ ,  $n_{Gi}$  is the number of people travelling by CBs on each line,  $\alpha_G$  is the maximum load capacity per CB vehicle, and  $\lceil n_{Gi}/\alpha_G \rceil$  is the number of CB vehicles on each line calculated by rounding up.

The environmental cost is given by the following logistic function:

$$Z_2 = c_a \left( W_{Ga} \times \left[ \frac{n_{Gi}}{\alpha_G} \right] \times l_i + W_{Ca} \times \left[ \frac{n_{Ci}}{\alpha_C} \right] \times l_i \right), \quad (6)$$

where  $Z_2$  is the total environmental pollution cost,  $c_a$  is the environmental pollution cost per unit of pollution,  $W_{Ga}$  is the pollutant emissions per bus per kilometer,  $W_{Ca}$  is the pollutant emissions per car per kilometer,  $n_{Ci}$  is the number of people travelling by car on each line,  $\alpha_C$  is the average load capacity per car, and  $\lceil n_{Ci}/\alpha_C \rceil$  is the number of cars on each line calculated by rounding up.

The traffic congestion cost is the additional time cost incurred by all passengers because of the change in the number of buses and cars on the road, which is formulated as

$$Z_3 = c_s \left( \left( \frac{l_i}{v_{Gs}} - \frac{l_i}{v_G} \right) \times n_{Gi} + \left( \frac{l_i}{v_{Cs}} - \frac{l_i}{v_C} \right) \times n_{Ci} \right), \quad (7)$$

where  $Z_3$  is the total congestion cost,  $c_s$  is the value per unit of time,  $v_{Gs}$  is the average running speed of the bus in the case of exclusive bus lanes and traffic congestion,  $v_G$  is the average



running speed of buses during normal running,  $v_{Cs}$  is the average running speed of cars in the case of traffic congestion, and  $v_C$  is the average running speed of cars during normal running.

In summary, the objective function  $Z$  of the line operating standard model can be determined using the linear weighted sum of the three parts as

$$\min Z = \omega_1 Z_1 + \omega_2 Z_2 + \omega_3 Z_3, \quad (8)$$

where  $Z$  is the total cost of a single line and  $\omega_1, \omega_2, \omega_3$  are the operating, environmental, and traffic congestion cost weights, respectively.

The decision variables of the line selection model in this paper are the number of commuters who travel by customized bus and the number of commuters who take a private car, respectively, represented by  $n_{Gi}$  and  $n_{Ci}$ . Actually, in this paper, the total travel demand on each line is known; thus  $n_{Gi}$  and  $n_{Ci}$  should satisfy the following constraints:

$$n_{Gi} + n_{Ci} = n_i, \quad (9)$$

$$n_{Gi}, n_{Ci} \in N, \quad (10)$$

where  $n_i$  is the travel demand on each line and  $N$  is the natural number set.

Equation (9) specifies that the sum of the numbers of people who are travelling by CBs and by car is equal to the travel demand on each line. Equation (10) specifies that the number of people who are travelling by CBs and by car is a natural number.

**3.5. Model Solution.** Solving the CB line selection model is essentially an integer programming problem and a discrete optimization problem. The purpose of the problem is to find a solution that conforms to the objective function from a limited number of possible scenarios. The optimal solution can be obtained by comparing the objective function value and using an enumeration method. However, in practical problems, the solution space is large and consumes large amounts of computational time and memory. Integer programming optimization methods for solving research problems are also gradually maturing with the development of optimization theory. At present, the branch-and-bound method, cutting-plane method, tabu search, and genetic algorithms are commonly used methods. In this paper, the branch-and-bound method [32] was used to solve the CB line selection model, and the final network operation scheme was obtained using a computer.

Integer programming is a branch of linear programming. If the integer constraint is removed, integer programming is transformed into linear programming. This linear programming problem is called the linear programming relaxation problem of integer programming, and all the feasible solutions of the integer programming problem are included in the linear programming relaxation problem. If the solution of the linear programming relaxation problem is expressed as  $Z_0$ , the optimal integer solutions that have been found are expressed by  $Z_i$ , the optimal integer solutions are expressed by  $Z^*$ , the lower bound is expressed by  $Z_l$ ,

and the upper bound is expressed by  $Z_u$ . For the objective function minimization problem, the optimal integer solution must satisfy

$$Z_l = Z_0 \leq Z^* \leq Z_i = Z_u. \quad (11)$$

The branch-and-bound method is based on the above relationship. The method begins by solving the linear programming relaxation problem. The feasible region of the linear relaxation problem is then decomposed into smaller subdomains (branches). The next step is to continuously update the upper and lower bounds by finding better integer solutions, which are obtained by the branches together. This approach can accelerate convergence, simplify the operation, and cause the lower bound to be equal to the upper bound. In this manner, the optimal solution of the integer programming is obtained [33]. The algorithm steps are as follows.

*Step 1.* The linear relaxation problem of integer programming is solved by removing the integer constraints. If there is no feasible solution, then the integer programming problem has no optimal solution, and the procedure is stopped. If an integer solution is obtained, then the solution is the integer programming optimal solution  $Z^*$ , and the procedure is stopped. If a noninteger solution is obtained, then the solution is taken as the lower bound of the integer programming problem, the upper bound is expressed by infinity, and the processing continues to Step 2.

*Step 2.* Choose any variable that does not meet the integer criteria  $n_{Gi}$  or  $n_{Ci}$   $i = 1, 2, \dots, k$  from the solution of the linear relaxation problem. Its value is  $b_i$ . Using  $[b_i]$  represents the largest integer one that does not exceed  $b_i$ . Next, two mutually exclusive inequality constraints,  $n_{Gi} \geq [b_i] + 1$  and  $n_{Gi} \leq [b_i]$  ( $n_{Ci} \geq [b_i] + 1$  and  $n_{Ci} \leq [b_i]$ ), are constructed. The two constraints are added to the linear relaxation problem, two subproblems can be obtained, and the optimal solution can be sought. The smallest result can then be found from the obtained optimal solutions and is used as a new lower bound  $Z_l$ . Finally, the greatest value of the objective function from each subproblem consistent with integer conditions is found and used as a new upper bound  $Z_u$ .

*Step 3.* For the subproblems with nonfeasible solutions and integer solutions, the downward branch is not continued downward and the node is closed. For the subproblems with noninteger optimal solutions, determine whether the optimal value is greater than the upper bound of  $Z_u$ . If the optimal value is greater than or equal to  $Z_u$ , stop the branch; otherwise, branch out and repeat Step 2 until all the nodes are closed. Optimal integer solutions can be obtained at this time with  $n_{Gi}^*$  and  $n_{Ci}^*$  ( $i = 1, 2, \dots, k$ ), and the optimal value is  $Z^* = Z_u$  [34].

The flowchart of branch-and-bound method is shown in Figure 3.

Using the branch-and-bound method to solve the CB line selection model, the numbers of people travelling by CB and car on each line can be obtained. If the number of people travelling by CB is zero, the line is not running; otherwise,

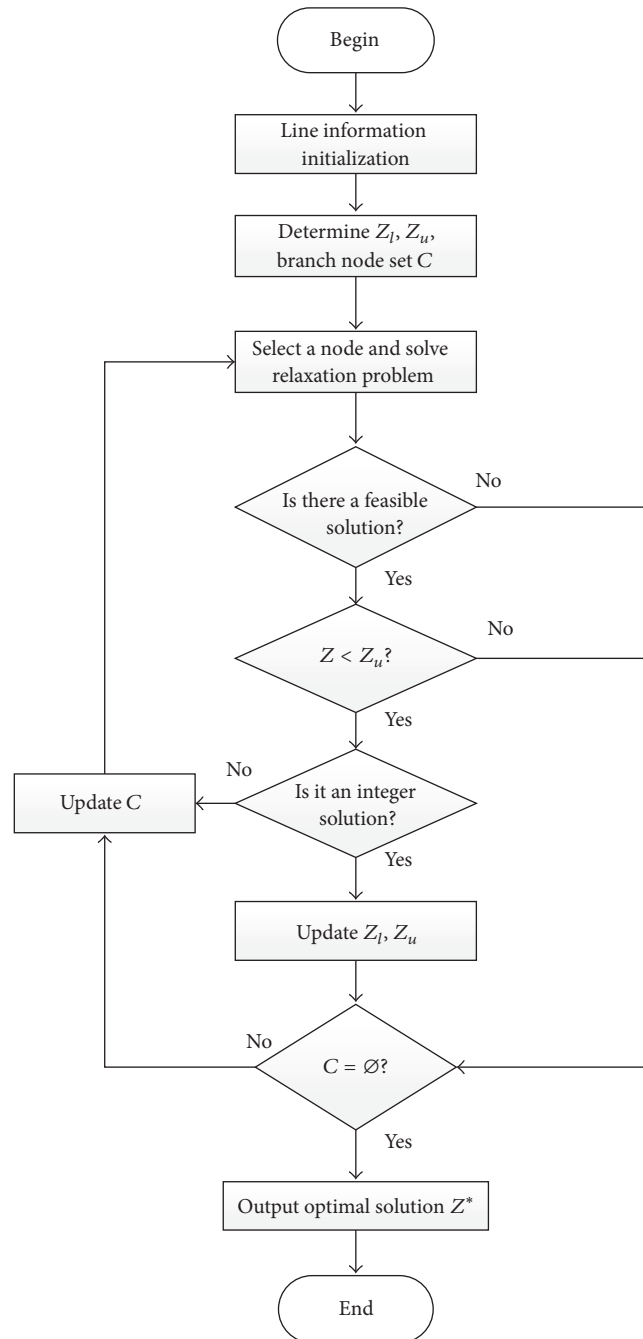


FIGURE 3: The flowchart of branch-and-bound method.

the line is operating, and the required number of vehicles is computed according to the travel demand of passengers.

#### 4. Case Study

In this section, Beijing CB network planning is employed as a case study. The optimization scheme for Beijing CB network planning can be determined using the CB network planning method introduced in this paper. The CB network planning method introduced in this paper was proven to be applicable

and effective compared with the current status of Beijing CB network planning.

*4.1. Large-Scale Travel Demand Data Processing.* The travel demands of passengers were collected using an Internet survey. A total of 15,000 morning peak travel passenger demands for CBs were collected from August 2015 to November 2015. Each travel demand includes a passenger's trip origin and destination. According to the passenger travel demand address data, the origin address and destination

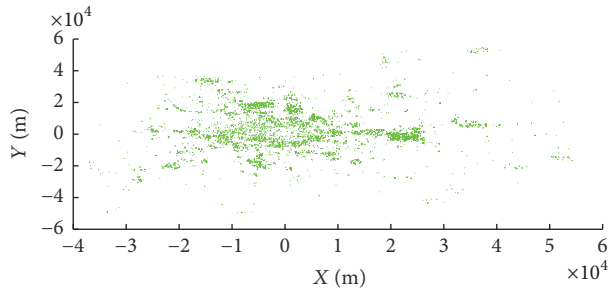


FIGURE 4: Scatterplot of the passenger trip origins.

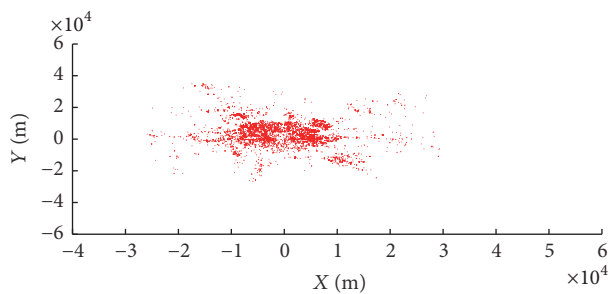


FIGURE 5: Scatterplot of the passenger trip destinations.

address of every passenger were stored as text documents or EXCEL documents. The documents then were imported into the XGeocoding software program and saved, and parsing was initiated. Parsing batch queried the longitude and latitude coordinates of every passenger OD, and the latitude and longitude coordinates of the passengers' locations were stored in text format. Subsequently, the text documents were imported into the COORD coordinate conversion software program, which can transform the latitude and longitude coordinates into plane coordinates that can be easily used to calculate the point-to-point distances. Because the passenger OD coordinate values are large numbers, for convenience, the coordinates of Tiananmen were used as the origin of the coordinates, the Tiananmen coordinates were subtracted from all the coordinates, and the origin and destination demands of the passengers then were marked separately using MATLAB software. The passenger trip origins are shown by green points in Figure 4. The passenger trip destinations are shown by red points in Figure 5.

In Figure 4, the range in the distribution of the CB passenger trip origins was wide and concentrated in a large residential zone. However, the passenger trip destinations were mainly concentrated in a large business zone centered on Tiananmen with a radius of 10 km, as shown in Figure 5. The characteristics of the morning peak travel demand in Beijing City is an agglomeration from rural areas to the urban area, and the travel distances were long. Therefore, the operation of the CB in Beijing City can effectively ease traffic congestion and reduce travel time.

**4.2. Line OD Area Division Based on Hierarchical Clustering.** From the processing results of the passenger travel

demand data in Section 4.1, all ODs were classified using hierarchical clustering according to the maximum distance measurement criteria. In this paper, considering the walking cost of passengers, the passenger walking distance should not exceed 800 m. The area coverage radius was 2.5 km, and the maximum category distance was 5 km. Thus, when the maximum distance between two categories was less than 5 km, the two categories were combined into one category. The final results yielded 84 origin categories and 67 destination categories. The hierarchical clustering tree was then built to represent the clustering results vividly. The origin hierarchical clustering tree is shown in Figure 6. The destination hierarchical clustering tree is shown in Figure 7. In the figure, the distance between categories is represented by the vertical coordinates, the category is represented by the horizontal coordinates, and the number on each horizontal coordinate represents the travel demands that are clustered by this category.

According to the results, the origin and destination were classified using hierarchical clustering. The passenger plane coordinate data for each category were stored as an array. Using (2), the centroid coordinates of each category were calculated, and they are marked on the diagram. The origin clustering results are shown in Figure 8. The destination clustering results are shown in Figure 9. In the diagram, the points in different colors represent the passenger travel demand points in different categories. The crosses represent the centroids of the categories. In the origin hierarchical clustering diagram, each category represents an origin area, and each centroid represents the center of an origin area. In the destination hierarchical clustering diagram, each category represents a destination area, and each centroid represents the center of a destination area.

**4.3. Determination of the Initial Path Set.** Based on the hierarchical clustering results of the origin and destination areas presented in Section 4.2, approximately 5,628 CB lines were obtained by pairing the OD areas. A travel OD demand table was obtained by processing the travel demand data corresponding to the 5,628 lines. In addition, the lengths of the lines were the distances between the centroids of the origin and destination areas. In the travel OD demand table, each cell indicates the travel demand of a line from an origin area to a destination area. In the line length table, each cell indicates the length of a line. The names of the origin area and destination area were derived from the centroid coordinates of each area using COORD and an XGeocoding software transformation.

In this paper, the minimum length of the line operation was determined to be  $l_{\min} = 8$  km. According to the line length constraint, a total of 233 lines that did not satisfy the line length constraint were deleted, and the 5,395 lines that satisfied the line length constraint were used as the initial line set. Using the initial path set, the travel demand OD table was updated, a total of 1,973 travel demands were removed, and 13,321 travel demands remained.

**4.4. Model Solution and Scheme Determination.** Based on the travel OD demand and line length of each line, the

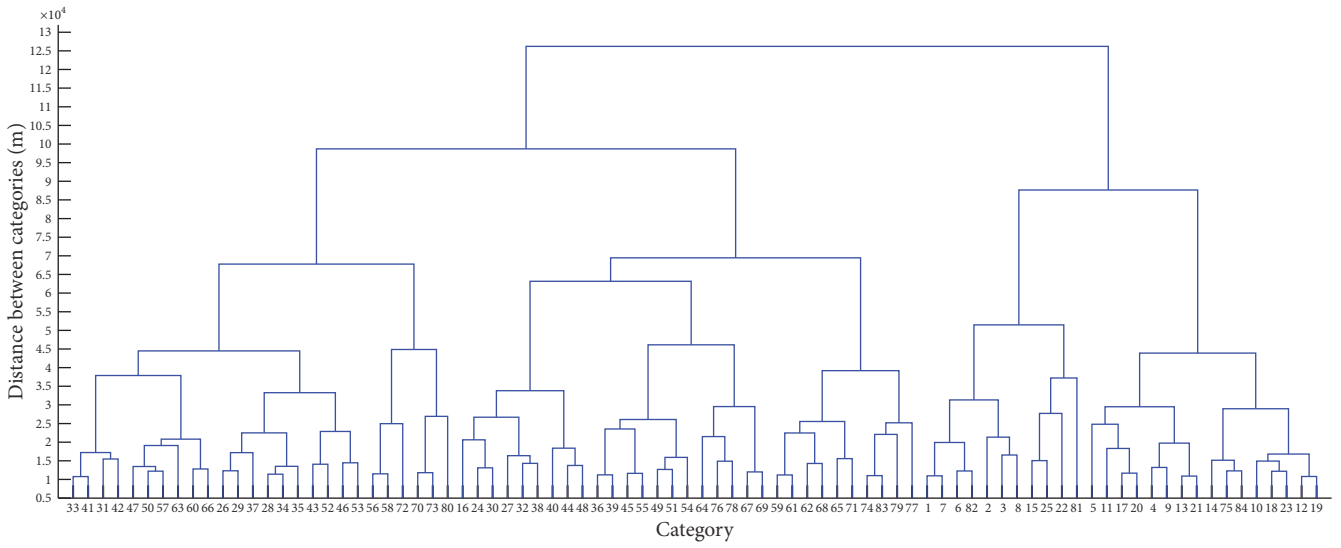


FIGURE 6: Origin hierarchical clustering tree.

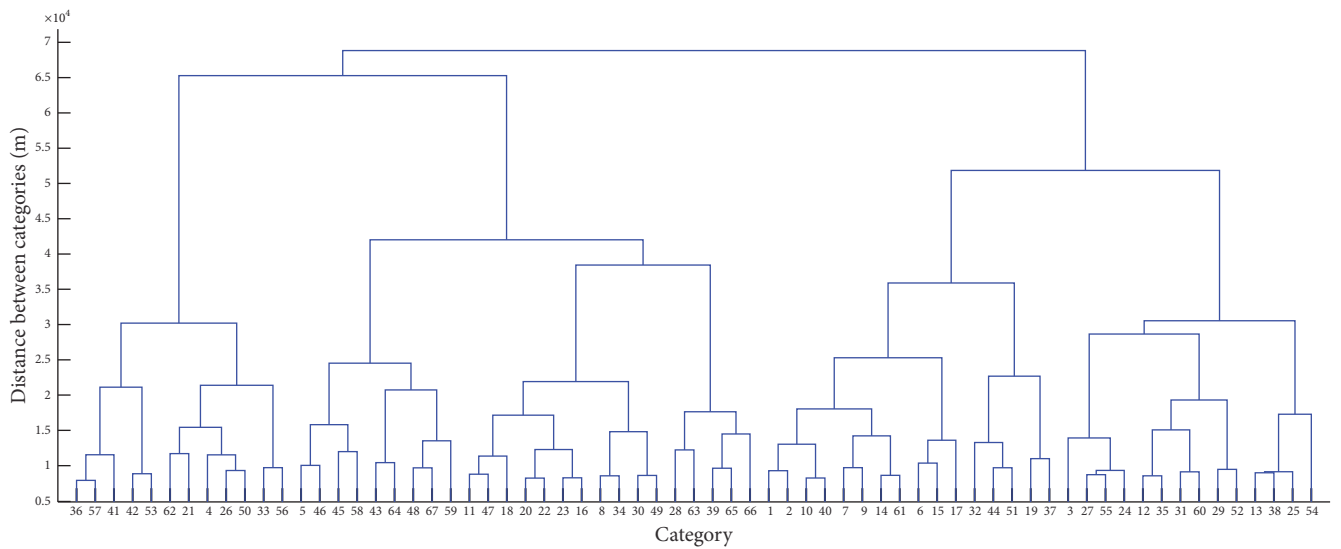


FIGURE 7: Destination hierarchical clustering tree.

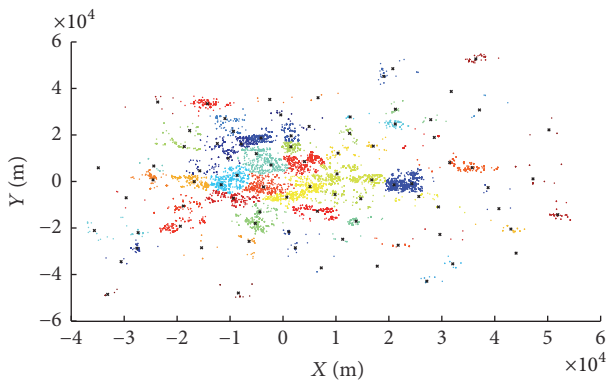


FIGURE 8: Origin hierarchical clustering diagram.

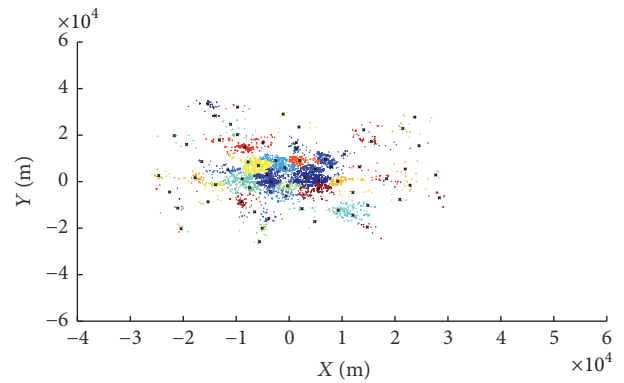


FIGURE 9: Destination hierarchical clustering diagram.

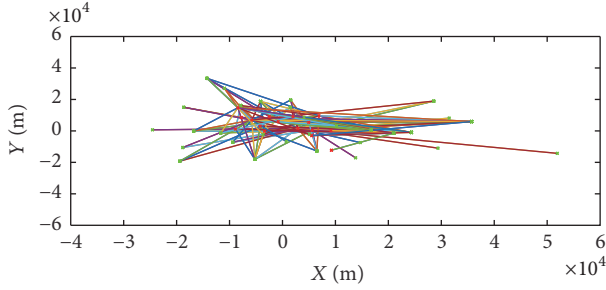


FIGURE 10: Planning line diagram for the Beijing CB network.

final operating scheme for the Beijing CB network was solved using MATLAB according to the line selection model in Section 3.4 and the branch-and-bound method in Section 3.5. The related parameters are as follows: the fuel cost per kilometer per vehicle  $c_{Gm} = 2.1$ , the fixed operating cost per vehicle per day  $c_{Go} = 350$  [35], the environmental pollution cost per unit of pollution  $c_a = 3.3$  [36], the unit time value  $c_s = 36$ , the maximum load capacity per CB vehicle  $\alpha_G = 30$ , the average load capacity per car  $\alpha_C = 2$ , the pollutant emissions per bus per kilometer  $W_{Ga} = 1.2$ , the pollutant emissions per car per kilometer  $W_{Ca} = 0.4$  [37], the average running speed of bus in the case of bus lanes and road congestion  $v_{Gs} = 28$ , the average running speed of buses during normal running  $v_G = 35$ , the average running speed of cars in the case of road congestion  $v_{Cs} = 26$ , the average running speed of cars during normal running  $v_C = 46$  [38], the weight of operating cost  $\omega_1 = 0.3$ , the weight of the environmental cost  $\omega_2 = 0.4$ , and the weight of congestion cost  $\omega_3 = 0.3$ .

According to the 15,000 CB morning peak travel demands of passengers, approximately 123 CB lines were found using the CB network planning method introduced in this paper. In addition, the operating kilometers totaled 2,708.3 km, which required 183 CBs with 30 seats, each of which served 5,009 passengers. The specific line distribution range is shown in Figure 10. In the figure, the green crosses represent the CB origins, the red crosses represent the CB destinations, and the connecting lines between them represent CB operating lines.

**4.5. Comparison and Evaluation.** In reality, using the approximately 100,000 CB morning peak travel demands of passengers from September 2013 to November 2015, the Beijing CB company designed a total of 92 morning peak lines. By comparison, in this paper, a total of 15,000 CB morning peak travel demands of passengers were collected from August 2015 to November 2015, and a total of 123 morning peak lines were designed by using the method proposed in this paper. The OD distributions of the current and planning schemes are compared in Figures 11 and 12. The contrast figure of the CB origin distribution is shown in Figure 11. The contrast figure of the CB destination distribution is shown in Figure 12. In the figure, the black crosses represent the current stations, and the green crosses represent the planning stations. In addition, in combination with the current planning line diagram of the CB and the planning line diagram of the

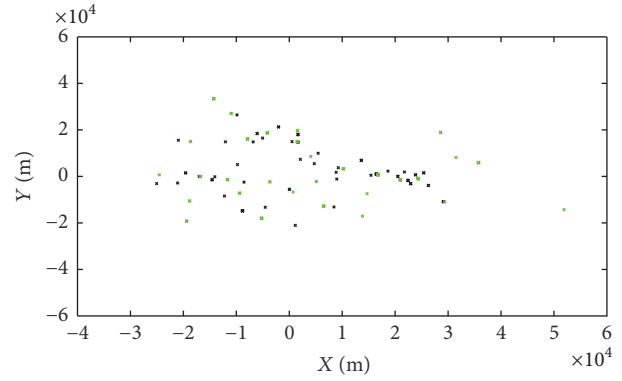


FIGURE 11: Contrast figure of the CB origin distribution.

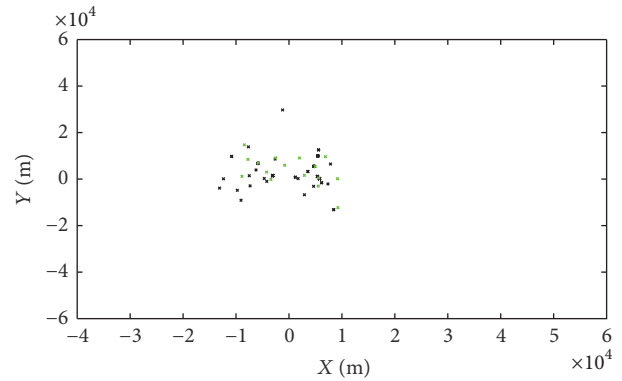


FIGURE 12: Contrast figure of the CB destination distribution.

CB determined using this paper's method, it was found that the planning scheme generated using the method of this paper was basically consistent with the current scheme. This shows that the method proposed in this paper is feasible and effective for solving CB network planning problems.

Figure 11 illustrates that the origins of the planning and current schemes were concentrated in large residential areas, such as Tongzhou, Huilongguan, and Fengtai, whereas the distribution of the planning scheme was more balanced and the stations had a wider coverage. Figure 12 illustrates that the destinations of the planning and current schemes were concentrated over a range that included Tiananmen as its center and had a radius of 10 km, which included the China World Trade Center, Finance Street, and Zhongguancun. In contrast, the station coverage range of the planning scheme was smaller than that of the current scheme.

CB is a new public transportation mode without transfer based on travel demands of passengers. The CB operating line is the shortest line between origin area stops and destination area stops. CB network evaluation differs from conventional bus network evaluation but also has some similar places as conventional bus. In this paper, the following evaluation indexes are established to evaluate the service level of CB network.

(1) The site coverage rate  $\gamma$  is the proportion of CB site areal coverage in the total areal coverage of travel demand,



TABLE 1: Beijing CB network evaluation index contrast.

	Current scheme	Planning scheme
The number of lines	92	123
The total length of lines	1412.44 km	2708.30 km
The average line distance	15.35 km	22.02 km
The number of vehicles	192	241
The site coverage rate	23.38%	32.71%
The average load factor	72.53%	77.05%
The service rate	30.17%	37.60%

that is, the proportion of the areal coverage of all original area in the total areal coverage of travel demand, which is formulated in

$$\gamma = \frac{\sum_{i=1}^m s_i}{S_C} \times 100\%, \quad (12)$$

where  $s_i$  is the areal coverage of each original area,  $i = \{1, 2, \dots, m\}$ , and  $S_C$  is the total areal coverage of travel demand.

(2) The average load factor  $\bar{\lambda}$  is the proportion of the number of people travelling by CBs in the maximum number of passengers provided by CBs, which is formulated in

$$\bar{\lambda} = \frac{\sum_{i=1}^k n_{Gi}}{\sum_{i=1}^k b_i \times \alpha_G} \times 100\%, \quad (13)$$

where  $b_i$  is the number of vehicles of each line,  $i = \{1, 2, \dots, k\}$ .

(3) The service rate  $\varphi$  is the proportion of the number of people travelling by CBs in the total travel demand of passengers, which is formulated in

$$\varphi = \frac{\sum_{i=1}^k n_{Gi}}{N} \times 100\%, \quad (14)$$

where  $N$  is the total travel demand of passengers.

The comparative results of the specific network evaluation index are listed in Table 1. Compared with the current network scheme, the travel demand data used by the planning scheme were smaller. However, the total number of lines, total length of the operating lines, and number of people being served were greater. Thus, the planning scheme placed more emphasis on the passenger service rate. In addition, the average line distance of the current scheme was 15.35 km, and the planning scheme was 22.02 km, demonstrating that the planning scheme lines mainly served passengers who travelled medium and long distances.

From the perspective of the company's operations, 192 vehicles are used in the current scheme, with an average load factor of 72.53%; in contrast, 241 vehicles are used in the planning scheme, with an average load factor of 77.05%. Although the average load factor cannot reflect the actual operational situation, as long as the planning is scientific and reasonable regarding the stops and timetable, the actual average load factor should be equal to the numerical value. Therefore, the

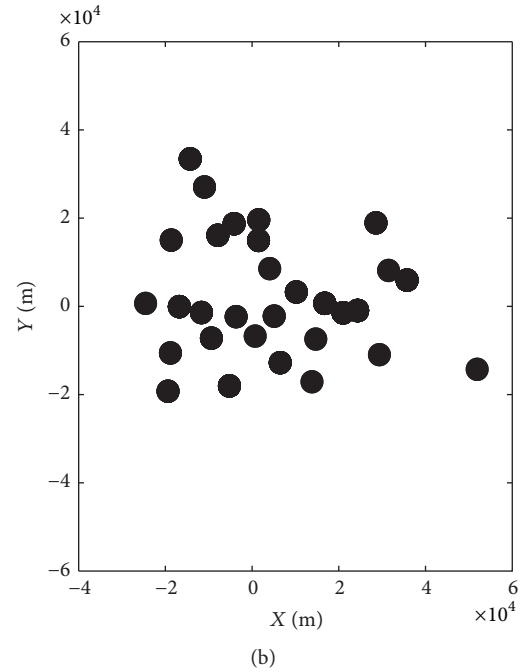
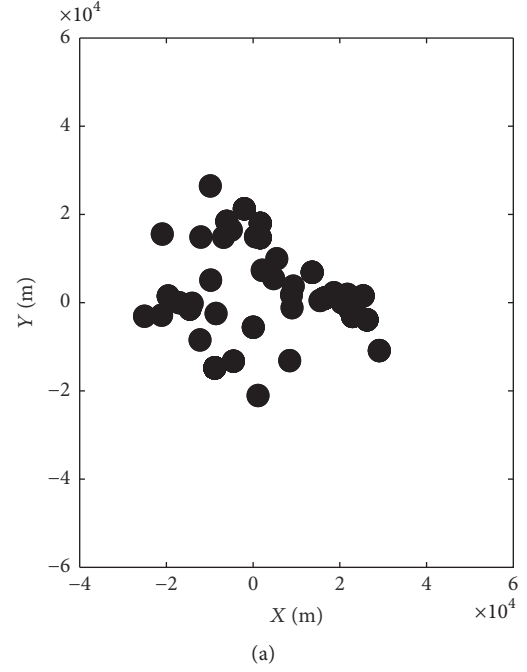


FIGURE 13: Contrast figure of the CB original stations' areal coverage. (a) Current original stations' areal coverage. (b) The planning original stations' areal coverage.

planning scheme can optimize the allocation of resources and effectively reduce the operating company's costs.

The coverage radius for each traffic zone was set to 2.5 km, and the site coverage rate for the original traffic zone was calculated. The contrast in the original stations' areal coverage for the current and planning schemes is shown in Figure 13. Combined with Table 1 and Figure 13, the station distribution

of the planning scheme was more balanced, and the site coverage rate was considerably higher. Therefore, the CV network can more effectively encourage private car owners to change travel modes, ease traffic congestion, and solve the problem of air pollution.

In summary, combined with the actual operational status of the Beijing CB network and the comparison results, a more scientific and reasonable network scheme can be obtained using the method presented in this paper. Compared with the current scheme, the planning scheme can more effectively provide services to passengers, save operating costs, and create positive social benefits.

## 5. Conclusions

This paper presented the status of research on CB network planning methods. On that basis, combined with conventional bus and existing CB network planning experience, a CB network planning method based on area division was proposed. The method uses ideas from a point-to-line layout into a network. First, the line OD area is divided according to the passenger travel demand; O and D of line OD represent the passenger's trip origin and trip destination, respectively. Second, the operating line scheme is determined according to the solution of the model, all operating lines are laid out, and the final CB network is determined. The results of this paper advance the theoretical research on CB network planning and provide a precise and efficient technical method for CB operators who are planning networks. The following aspects should be investigated further to optimize and improve the methods proposed in this paper.

(1) A major feature of CBs is that they have exclusive bus lane rights. When the traffic is highly congested in the morning and evening peak periods, exclusive bus lanes can ensure the effectiveness of the running time and improve the punctuality. In this paper, the line layouts did not consider the specific paths of each line. A straight line between the origin and destination was used as a line, and the important role of exclusive bus lanes in the layout of CB lines was neglected. In future research, the layout of exclusive bus lanes should be considered.

(2) In this paper, a line operation model based on operating cost and social benefits was proposed. The main aspect of the model is the determination of the parameter values of various influencing factors and quantitative methods. By consulting various references to determine the parameter values, the parameters values achieved good results in the study case demonstration. However, whether those parameters values are applicable to all cities and how to assign the weight ratio of each factor should be considered in future research.

(3) In this paper, the vehicle standard was 30 seats. In future research, the vehicle standard should be chosen according to the line travel demand, which can improve the use efficiency of CB vehicles.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

The paper is supported by National High-tech R&D Program of China (863 Program): 2015AA124103.

## References

- [1] R. F. Kirby and K. U. Bhatt, *Guidelines on the Operation of Subscription Bus Services*, National Technical Information Service, Alexandria, Va, USA, 1974.
- [2] R. F. Kirby and K. U. Bhatt, "An analysis of subscription bus experience," *Traffic Quarterly*, vol. 29, no. 3, pp. 403–425, 1975.
- [3] J. A. Bautz, "Subscription service in the United States," *Transportation*, vol. 4, no. 4, pp. 387–402, 1975.
- [4] C. H. J. McCall, *COM-BUS: A Southern California Subscription Bus Service*, National Technical Information Service, Alexandria, Va, USA, 1977.
- [5] C. E. McKnight and R. E. Paaswell, *The Potential of Private Subscription Bus to Reduce Public Transit Subsidies*, Urban Mass Transportation Administration, Washington, DC, USA, 1985.
- [6] S. A. Shaheen, D. Sperling, and C. Wagner, "Carsharing in Europe and North America: past, present and future," *Transportation Quarterly*, vol. 52, no. 3, pp. 35–52, 1998.
- [7] S. K. Chang and P. M. Schonfeld, "Optimization models for comparing conventional and subscription bus feeder services," *Transportation Science*, vol. 25, no. 4, pp. 281–298, 1991.
- [8] E. Martin and S. A. Shaheen, "Assessing greenhouse gas emission impacts from carsharing in north america: theoretical and methodological design," in *Proceedings of the 15th World Congress on Intelligent Transport Systems and ITS America Annual Meeting*, pp. 1183–1194, November 2008.
- [9] J. F. Potts, M. A. Marshall, E. C. Crockett, and J. Washington, "A guide for planning and operating flexible public transportation services," TCRP Report 140, 2010.
- [10] M. Duncan, "The cost saving potential of carsharing in a US context," *Transportation*, vol. 38, no. 2, pp. 363–382, 2011.
- [11] A. El Fassi, A. Awasthi, and M. Viviani, "Evaluation of car-sharing network's growth strategies through discrete event simulation," *Expert Systems with Applications*, vol. 39, no. 8, pp. 6692–6705, 2012.
- [12] A. De Lorimier and A. M. El-Geneidy, "Understanding the factors affecting vehicle usage and availability in carsharing networks: a case study of commuauto carsharing system from Montréal, Canada," *International Journal of Sustainable Transportation*, vol. 7, no. 1, pp. 35–51, 2012.
- [13] R. Nair and E. Miller-Hooks, "Equilibrium network design of shared-vehicle systems," *European Journal of Operational Research*, vol. 235, no. 1, pp. 47–61, 2014.
- [14] S. Le Vine, M. Lee-Gosselin, A. Sivakumar, and J. Polak, "A new approach to predict the market and impacts of round-trip and point-to-point carsharing systems: case study of London," *Transportation Research Part D: Transport and Environment*, vol. 32, pp. 218–229, 2014.
- [15] T. Liu and A. Ceder, "Analysis of a new public-transport-service concept: customized bus in China," *Transport Policy*, vol. 39, pp. 63–76, 2015.
- [16] W. Lampkin and P. D. Saalmans, "The design of routes, service frequencies, and schedules for a municipal bus undertaking: a case study," *OR: Operational Research Quarterly*, vol. 18, no. 4, pp. 375–397, 1967.

- [17] A. Ceder and N. H. M. Wilson, "Bus network design," *Transportation Research Part B: Methodological*, vol. 20, no. 4, pp. 331–344, 1986.
- [18] M. H. Baaj and H. S. Mahmassani, "Hybrid route generation heuristic algorithm for the design of transit networks," *Transportation Research, Part C: Emerging Technologies*, vol. 3, no. 1, pp. 31–50, 1995.
- [19] V. M. Tom and S. Mohan, "Transit route network design using frequency coded genetic algorithm," *Journal of Transportation Engineering*, vol. 129, no. 2, pp. 186–195, 2003.
- [20] S. Jerby and A. Ceder, "Optimal routing design for shuttle bus service," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1971, no. 1, pp. 14–22, 2006.
- [21] E. Cipriani, S. Gori, and M. Petrelli, "Transit network design: a procedure and an application to a large urban area," *Transportation Research Part C: Emerging Technologies*, vol. 20, no. 1, pp. 3–14, 2012.
- [22] M. Nikolić and D. Teodorović, "Transit network design by bee colony optimization," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5945–5955, 2013.
- [23] H. Badia, M. Estrada, and F. Robusté, "Competitive transit network design in cities with radial street patterns," *Transportation Research Part B: Methodological*, vol. 59, pp. 161–181, 2014.
- [24] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 11, no. 1, pp. 1–14, 2007.
- [25] B. Li, "Markov models for Bayesian analysis about transit route origin-destination matrices," *Transportation Research Part B: Methodological*, vol. 43, no. 3, pp. 301–310, 2009.
- [26] X.-L. Ma, Y.-H. Wang, F. Chen, and J.-F. Liu, "Transit smart card data mining for passenger origin information extraction," *Journal of Zhejiang University: Science C*, vol. 13, no. 10, pp. 750–760, 2012.
- [27] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.
- [28] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders' travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 1–12, 2013.
- [29] X. Ma and Y. Wang, "Development of a data-driven platform for transit performance measures using smart card and GPS data," *Journal of Transportation Engineering*, vol. 140, no. 12, Article ID 04014063, 2014.
- [30] X. Ma, C. Liu, H. Wen, Y. Wang, and Y. Wu, "Understanding commuting patterns using transit smart card data," *Journal of Transport Geography*, vol. 58, pp. 135–145, 2017.
- [31] M. Duan, *The Research and Application of Hierarchical Clustering Algorithm*, Central South University, Changsha, China, 2009.
- [32] L. Yupo, *Study on Problems to Select Initial Cluster Centers of the K-means Algorithm*, Lanzhou Jiaotong University, Lanzhou, China, 2012.
- [33] P. Zhao, *Course of Management Operations Research*, Tsinghua University Press, Beijing, China, 2008.
- [34] Q. Pingping, *The Application of Branch and Bound Algorithm in the Model of Operational Research*, Yanshan University, Qinhuangdao, China, 2009.
- [35] L. Dongmei, *The Transit Scheme Research on Customized City Bus Service of Xi'an*, Chang'an University, Xian, China, 2014.
- [36] Y. Chen, *A Comparative Analysis on the Transportation Costs between Public Transportation and Cars*, Nanjing Forestry University, Nanjing, China, 2009.
- [37] H. Cai and S. Xie, "Determination of emission factors from motor vehicles under different emission standards in China," *Journal of Peking University: Natural Science Edition*, vol. 03, pp. 319–326, 2010.
- [38] Q.-B. Wu, F. Chen, Y. Huang, and Y.-Y. Hu, "Calculation and analysis of traffic congestion cost in Beijing," *Journal of Transportation Systems Engineering and Information Technology*, vol. 11, no. 1, pp. 168–172, 2011.

## Research Article

# Compression Algorithm of Road Traffic Spatial Data Based on LZW Encoding

Dong-wei Xu,<sup>1,2</sup> Yong-dong Wang,<sup>1,2</sup> Li-min Jia,<sup>3</sup> Gui-jun Zhang,<sup>1,2</sup> and Hai-feng Guo<sup>1,2</sup>

<sup>1</sup>College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

<sup>2</sup>United Key Laboratory of Embedded System of Zhejiang Province, Hangzhou 310023, China

<sup>3</sup>State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Dong-wei Xu; dongweixu@zjut.edu.cn

Received 20 October 2016; Revised 19 December 2016; Accepted 10 January 2017; Published 16 February 2017

Academic Editor: Xiaolei Ma

Copyright © 2017 Dong-wei Xu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Wide-ranging applications of road traffic detection technology in road traffic state data acquisition have introduced new challenges for transportation and storage of road traffic big data. In this paper, a compression method for road traffic spatial data based on LZW encoding is proposed. First, the spatial correlation of road segments was analyzed by principal component analysis. Then, the road traffic spatial data compression based on LZW encoding is presented. The parameters determination is also discussed. Finally, six typical road segments in Beijing are adopted for case studies. The final results are listed and prove that the road traffic spatial data compression method based on LZW encoding is feasible, and the reconstructed data can achieve high accuracy.

## 1. Introduction

The advent of big data brings unprecedented opportunities as well as challenges, especially in the field of transportation and traffic engineering [1, 2]. With the rapid development of science and technology, the intelligent transportation system (ITS) has developed continuously, and its applications have become wide ranging. The ITS system can accomplish the tasks of road traffic data acquisition, processing, and transportation. Besides, it can complete the job of traffic state analysis, route guidance, and traffic control. As various road traffic detection systems are adopted in the road traffic field, the collected road traffic state data increase and become massive. This serious situation introduces a challenge for real-time transmission, storage, and guidance of massive road traffic data. Thus, it is necessary to find an efficient approach to compress real-time traffic states data which can save much storage space as well as providing some other applications [3]. And the compression method of road traffic states data has deeply promoted the managements for transportation administrators. Besides, useful compression method of road traffic data can also be applied to transportation research fields, and some inspirations may occur to the researchers.

The essence of road traffic state data compression is to represent the signal information with less data. Through effective compression and reconstruction, traffic data transmission and storage can be achieved [4–6].

In recent years, a great many of data compression methods have been explored in traffic and transportation fields. With the popularity of machine learning and data mining study among practitioners and researchers, some road traffic compression methods are presented. Due to the multidimensional and multigranularity characteristics of traffic and transportation big data, PCA method realizes the compression of road traffic states data through reducing the dimensions of original data [7]. As an emerging technology, compression sensing has also been used in data compression due to its superiority. Compression sensing breaks through traditional Nyquist sampling theorem restricts and can collect and compress data simultaneously. Making use of the redundancy characteristics of road traffic states, compression sensing technology achieves the estimation [8] and compression [9–11] of road traffic states data. Since the road traffic states data possess the spatial-temporal correlation and similar characteristics, Xiao et al. presented a spatial-temporal model based on road traffic data compression and



decompression technology of 2D discrete wavelet transformation, realizing the denoising compression of ITS system [12]. Ou et al. proposed road traffic volume data compression based on artificial neural network [13].

Some modified and improved methods also fill the compression gap. The embedded devices in motor vehicles also generate abundant data for researchers to investigate the compression of road traffic states data [14]. Making use of the GPS positioning data produced by the mobile devices of travelers, Ma et al. presented a differential preprocessing method, and a dynamical Huffman algorithm was adopted to compress GPS positioning data [15]. Wang et al. put forward an encoding algorithm with self-adaptive switching mode according to specific format [16]. Hou presented a stop-wave mode based on the concept of the compression factor and its differential equation [17]. Song et al. proposed a hybrid spatial compression algorithm and error bounded temporal compression algorithm to compress the spatial and temporal information of trajectories, respectively [18]. However, many researches do not have a common baseline for their performance analysis and provide the infrastructure to operate on a publicly available dataset.

The existing road traffic data compression methods mainly focus on the compression of road traffic network data. However, in recent years, limited literature has been written on the road traffic spatial data compression methods of different road segments on similar time nodes. Some literatures on predictions are investigated temporally and spatially in recent years. The studies are not only in road traffic field, but also in the field of transportation.

The travel needs and travel routes of traffic participants exhibit certain regularity; thus, the road traffic spatial states of different road segments on similar time nodes represent strong relationships. That is, the changing curve of road traffic spatial state on different road segments on similar time nodes possesses some similarity. The correlation presents great probability for the compression of road traffic spatial data. Thus, based on the spatial correlation characteristics of the road traffic states, the road traffic spatial data on different road segments on similar time nodes are extracted for compression. LZW inherits the merits of LZ77 and LZ78 on compression efficiency and speed. Besides, the method easily achieves good performance. Thus, the LZW encoding is introduced in the study. Based on the spatial correlation of road traffic, a compression method of road traffic spatial data based on LZW encoding is proposed in this paper.

In this study, a compression method of road traffic spatial data based on LZW encoding is proposed to compress the road traffic spatial states data under the same time intervals, realizing efficient transmission and storage as well as display. The useful compression of road traffic states data can be efficiently used into feature extraction and traffic states prediction. Multivariate time series analysis is similar to the proposed method, which can take into consideration both spatial and temporal correlations. In our study, we used the spatial correlation characteristics of road traffic states to compress the states data. The aims of the two studies are different.

Some motivations are explained here. Although the proposed compression method of our study is tested on the road traffic states data, it is also very useful for transportation management as well as transportation prediction. Besides, the compression can be also used for feature extraction, which can be applied to evaluate the traffic running states.

Based on the characteristics of road traffic flow, the PCA method can be used to analyze the correlation of spatial road segments [19, 20]. Then, the spatial road segments are selected to extract the data for compression. The spatial road segments denote the different road segments; the data on these segments are extracted on the spatial road segments at the same time intervals.

The contributions of the proposed algorithm are three-fold:

- (1) The PCA method was introduced to the algorithm to select the road segments with spatial correlation.
- (2) A novel road traffic spatial data compression algorithm based on LZW encoding was proposed to construct the difference data on selected spatial road segments under the same mode.
- (3) The proposed algorithm could determine the optimal parameters in the training process based on spatial historical data and base data on road traffic states.

The rest of this paper is organized as follows. The modeling methodology of the proposed algorithm is discussed in Section 2. In Section 3, parameter determination of the road traffic spatial data compression study based on LZW encoding is presented. The experiment results are shown in Section 4. The conclusion and direction for future studies are discussed in Section 5.

## 2. Compression Algorithm of Road Traffic Spatial Data Based on LZW Encoding

*2.1. Framework of the Algorithm.* The process of compression and reconstruction of road traffic spatial data is shown in Figures 1 and 2, respectively. First, the PCA method was used to select the road segments with the characteristics of spatial correlation. The road traffic spatial data under the same mode on different road segments were acquired to construct the reference sequences of road traffic characteristics. Based on the analysis of spatial correlation, the base road segment was selected and the data on which were regarded as spatial base data. Second, the historical data on other spatial road segments under the same mode was extracted as training data. The optimal threshold of road traffic spatial difference data was determined based on road traffic spatial base data under the same mode. Third, real-time spatial data on other road segments under the same mode were acquired as experimental data and the road traffic spatial difference data were acquired on the basis of road traffic spatial base data under the same mode. Finally, the compression and reconstruction of road traffic spatial difference data were achieved through LZW encoding and decoding technology, respectively.



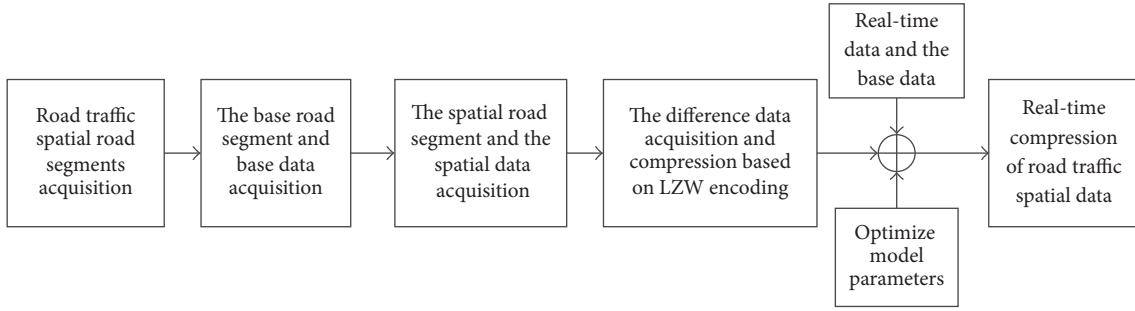


FIGURE 1: The process of road traffic spatial data compression based on LZW encoding.

TABLE 1: Road traffic characteristics reference sequence information chart.

Reference sequence ID	Mode	Road segment ID	Time	Road traffic state parameters
-----------------------	------	-----------------	------	-------------------------------

TABLE 2: Road traffic characteristics reference sequence description chart.

Reference sequence ID	Reference sequence name	Description
-----------------------	-------------------------	-------------

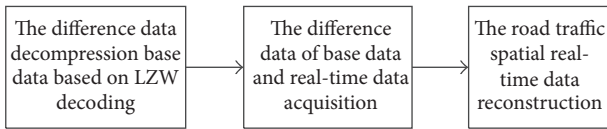


FIGURE 2: The process of road traffic spatial data reconstruction based on LZW decoding.

## 2.2. Acquisition of Road Traffic Spatial Base Data

**2.2.1. Selection of Road Segments with Correlation Based on PCA Method.** Road traffic flows possess the characteristics of periodicity, similarity, correlation, and so on. The road traffic flows of spatial road segments indicate a strong spatial correlation. Thus, the PCA method was used in this study to select the road segments with the characteristics of correlation.

PCA is a multivariate statistical method that eliminates the correlation among the variable indicators.  $n$ -dimensions of road traffic state data can be effectively reduced to two dimensions, which can be illustrated in a 2D figure. Taking advantage of these characteristics, the related road segments can be selected. The process has been described in previous studies [19, 20].

**2.2.2. Division of Road Traffic Running Modes.** The road traffic running modes can be divided into two levels: the road network level and road segments level. Assuming that the running modes division identification of road network level and road segments level can be divided into  $g$  and  $h$

submodes, respectively, the road traffic running modes can be divided into  $g \times h$  modes in total. The modes can be shown as  $M_{ode} = \{M_{11}, M_{12}, \dots, M_{gh}\}$ .  $g$  and  $h$  can be determined by the road traffic running modes division identification. The running modes division identification of road network mainly refers to the impact factors of road traffic running modes on different dates. The road traffic running modes division identification of road segments refers to the influence factors of the road traffic running modes of the specific condition of the road segments, which can be illustrated as in Figure 3.

**2.2.3. Construction Design of Road Traffic Characteristics Reference Sequences.** Assuming the collection period of road traffic state data was  $\Delta t$ , then time format of road traffic information template can be illustrated as in Figure 4. The table format of the road traffic characteristics reference sequence can be described as in Tables 1 and 2.

Let  $p + 1$  denote the total number of selected road segments, which can be described as follows:

$$L = [L_1 \ L_2 \ \dots \ L_{p+1}], \quad (1)$$

where  $p + 1$  is the number of spatial road segments;  $L_i$  ( $1 \leq i \leq p + 1$ ) denotes the  $i$ th road segments;  $L$  represents the set of selected road segments with correlation.

Based on the correlation of road traffic spatial data, the base road segment is acquired to extract the road traffic data as road traffic base data. The road traffic data on other  $p$  spatial road segments are extracted as historical data and real-time data.

**2.3. Optimal Threshold Determination of Road Traffic Difference Data.** The data on other spatial road segments are extracted as training data. Under  $M_{gh}$  mode, the road traffic spatial difference data under the same mode are acquired based on road traffic spatial base data to conduct the threshold processing. Through LZW encoding, the optimal

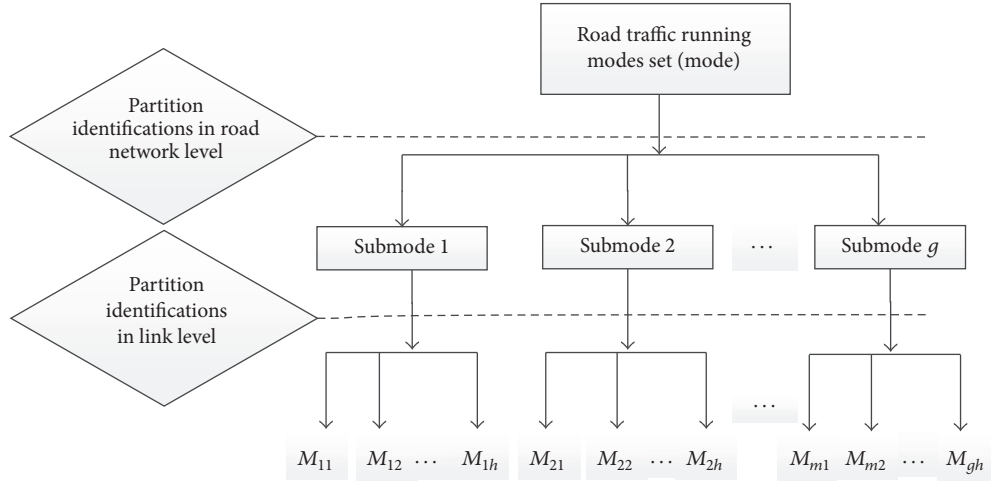


FIGURE 3: The division chart of road traffic running mode.

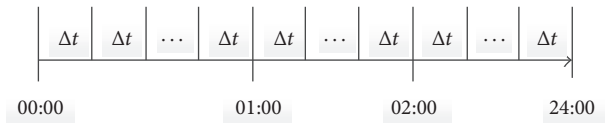


FIGURE 4: The time format of road traffic characteristic reference sequence.

threshold is identified. The main expressions can be described as follows:

$$S_i(m * \Delta t, M_{gh}) = ST_i(m * \Delta t, M_{gh}) - SB(m * \Delta t, M_{gh}),$$

$$e_i(m, M_{gh}) = [S_i(\Delta t, M_{gh}) \ S_i(2 * \Delta t, M_{gh}) \ \dots \ S_i(m * \Delta t, M_{gh})],$$

$$he_i(m, M_{gh})$$

$$= \begin{cases} 0, & e_i(m, M_{gh}) < E_i(m, M_{gh}) \\ e_i(m, M_{gh}), & e_i(m, M_{gh}) > E_i(m, M_{gh}), \end{cases}$$

$$pe_i(n, M_{gh}) = w(he_i(m, M_{gh})),$$

$$pe_i(n, M_{gh}) = [S'_i(1, M_{gh}) \ S'_i(2, M_{gh}) \ S'_i(n, M_{gh})].$$

(2)

The characteristics are described in Table 3.

Based on the formulas of (2), the optimal threshold of difference data can be identified.

#### 2.4. Road Traffic Spatial Data Compression Based on LZW Encoding

**2.4.1. Acquisition of Road Traffic Spatial Difference Data.** The spatial data on other road segments were extracted as real-time data. Under  $M_{gh}$  mode and based on the spatial base data, the road traffic difference data were acquired. The main expressions can be described as follows:

$$MS_j(m * \Delta t, M_{gh}) = SM_j(m * \Delta t, M_{gh}) - SB(m * \Delta t, M_{gh}), \quad (3)$$

$$err_j(m, M_{gh}) = [MS_j(\Delta t, M_{gh}) \ MS_j(2 * \Delta t, M_{gh}) \ \dots \ MS_j(m * \Delta t, M_{gh})]. \quad (4)$$

The characteristics are described in Table 4.

**2.4.2. Road Traffic Spatial Difference Data Compression Based on LZW Encoding.** LZW encoding is a lossless compression method based on dictionary coding. By constructing a string table, the long code word is presented by a shorter code word to realize data compression. The string and code

word are gradually built, and the string table is constructed dynamically. The string table is constantly improved and is greater in comparison with the latter string and string table. The created string table does not need to be stored along with the data. In the decompression process, the same string word can still be reconstructed. Thus, the compression ratio can be improved by another step.

TABLE 3

$\Delta t$	The collection period of road traffic state data
$(m * \Delta t)$	The $m$ th period collection of road traffic state data, $0 \leq m \leq N$
$N$	The number of daily collected road traffic data
$i$	The $i$ th road segment under $M_{gh}$ mode
$ST_i(m * \Delta t, M_{gh})$	The road traffic data on $i$ road segment at $(m * \Delta t)$ moment under $M_{gh}$ mode
$SB(m * \Delta t, M_{gh})$	The road traffic data on base road segment at $(m * \Delta t)$ moment under $M_{gh}$ mode
$S_i(m * \Delta t, M_{gh})$	The road traffic difference data between the training data on $i$ road segment and the base data on base road segment at $(m * \Delta t)$ moment under $M_{gh}$ mode
$e_i(m, M_{gh})$	The road traffic difference data between the training data on $i$ road segment and the base data on base road segment at $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode
$he_i(m, M_{gh})$	The road traffic difference data between the training data on $i$ road segment and the base data on base road segment after threshold processing $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode
$E_i(m, M_{gh})$	The threshold of road traffic difference data at $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode
$pe_i(n, M_{gh})$	The data on the difference data between $i$ road segment and base road segment after LZW encoding at $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode
$S'_i(n, M_{gh})$	The $n$ th data on the difference data between $i$ road segment and base road segment after LZW encoding at $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode
$m$	The number of difference data between $i$ road segment and base road segment before LZW encoding at $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode
$n$	The number of difference data between $i$ road segment and base road segment after LZW encoding at $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode

TABLE 4

$j$	The $j$ th road segment, $1 \leq j \leq p + 1$
$SM_j(m * \Delta t, M_{gh})$	The real-time data on $j$ road segment at $(m * \Delta t)$ moment under $M_{gh}$ mode
$MS_j(m * \Delta t, M_{gh})$	The difference data on $j$ road segment and base road segment at $(m * \Delta t)$ moment under $M_{gh}$ mode
$err_j(m, M_{gh})$	The road traffic difference data between the real-time data on $j$ road segment and the base data on base road segment at $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode

Based on LZW encoding, the road traffic spatial data compression can be achieved. The best threshold of the difference data between  $i$  road segment and base road segment can be introduced into the difference data on the  $j$  road

segment and the base road segment under the same  $M_{gh}$  mode. Combining the LZW encoding, the difference data compression of  $j$  road segment and base road segment can be realized. The main expressions can be described as follows:

$$\begin{aligned}
herr_j(m * \Delta t, M_{gh}) &= \begin{cases} 0, & err_j(m * \Delta t, M_{gh}) < E_{opt}(M_{gh}) \\ err_j(m * \Delta t, M_{gh}), & err_j(m * \Delta t, M_{gh}) > E_{opt}(M_{gh}) \end{cases}, \\
herr_{s_p}(m * \Delta t, M_{gh}) &= [herr_1(m * \Delta t, M_{gh}) \quad herr_2(\Delta t, M_{gh}) \quad \cdots \quad herr_{p'}(m * \Delta t, M_{gh})], \\
perr_{p'}(m * \Delta t, M_{gh}) &= w(herr_{s_p}(m * \Delta t, M_{gh})), \\
perr_p(m * \Delta t, M_{gh}) &= [MS_1(m * \Delta t, M_{gh}) \quad MS_2(m * \Delta t, M_{gh}) \quad \cdots \quad MS_{p'}(m * \Delta t, M_{gh})].
\end{aligned} \tag{5}$$

The characteristics are explained in Table 5.

The compression ratio is  $p/p'$ .

**2.5. Road Traffic Spatial Data Decompression Based on LZW Decoding.** Based on LZW decoding technology, the data reconstruction of difference data between  $p$  road segments and base road segment can be realized. Combining the base

data, the decompression of  $p$  road segments real-time data can be achieved. The main expressions are as follows:

$$\begin{aligned}
dperr_p(m * \Delta t, M_{gh}) &= w'(perr_{p'}(m * \Delta t, M_{gh})), \\
CSM_p(m * \Delta t, M_{gh}) &= SB(m * \Delta t, M_{gh}) + dperr_p(m * \Delta t, M_{gh}),
\end{aligned} \tag{6}$$

TABLE 5

$E_{opt}(M_{gh})$	The optimal training threshold
$herr_j(m * \Delta t, M_{gh})$	The difference data between the real-time data on $j$ road segment and the base data on base road segment after threshold processing at $(m * \Delta t)$ moment under $M_{gh}$ mode
$m$	The number of difference data between $j$ road segment and base road segment before LZW compression at $\Delta t$ to $(m * \Delta t)$ time intervals under $M_{gh}$ mode
$herrs_p(m * \Delta t, M_{gh})$	The road traffic difference data set of $p$ road segments at $(m * \Delta t)$ moment under $M_{gh}$ mode
$Perr_{p'}(m * \Delta t, M_{gh})$	The set of difference data on $p$ road segment after LZW encoding at $(m * \Delta t)$ moment under $M_{gh}$ mode
$p'$	The number after LZW encoding at $(m * \Delta t)$ moment
$MS'_j(m * \Delta t, M_{gh})$	The $j'$ th data on difference data after LZW encoding at $(m * \Delta t)$ time moment under $M_{gh}$ mode

where  $w'$  denotes the LZW decoding;  $dperr_p(m * \Delta t, M_{gh})$  denotes the spatial difference data on  $p$  road segments after LZW decoding at  $(m * \Delta t)$  moment under  $M_{gh}$  mode; and  $CSM_p(m * \Delta t, M_{gh})$  denotes the reconstructed road traffic real-time data on  $p$  road segments at  $(m * \Delta t)$  moment under  $M_{gh}$  mode.

### 3. Parameter Determination

In the process of road traffic spatial data compression based on LZW encoding, the following parameters were involved:  $SB(m * \Delta t)$ ,  $ST_j(m * \Delta t)$ ,  $E_i(m * \Delta t)$ ,  $per$ ,  $err_i(m * \Delta t)$ ,  $n$ , where  $E_i(m * \Delta t)$  can be acquired by  $SB(m * \Delta t)$  and  $per$ ,  $n$ , and  $err_i(m * \Delta t)$  can be acquired by  $SB(m * \Delta t)$ ,  $ST_j(m * \Delta t)$ , and  $E_i(m * \Delta t)$ . Parameter settings here are only concerned with the effect analysis of the road traffic spatial data compression based on LZW encoding. Separately analyzing the effect of each parameter on the accuracy of the algorithm cannot guarantee an optimal algorithm because these parameters influence the accuracy of the algorithm in different ways. All of the parameters in the road traffic spatial data compression results should be considered when conducting the algorithm analysis.

The compression ratios are introduced to measure the effect of parameters on the precision of the algorithm. The main expression can be described as follows:

$$CR_p(m * \Delta t, M_{gh}) = \frac{CM_a(m * \Delta t, M_{gh})}{CM_b(m * \Delta t, M_{gh})}, \quad (7)$$

where  $CR_p(m * \Delta t, M_{gh})$  denotes the compression ratio of  $p$  road segment at  $(m * \Delta t)$  moment under  $M_{gh}$  mode;  $CM_a(m * \Delta t, M_{gh})$  denotes the number of road traffic data before compression at  $(m * \Delta t)$  moment under  $M_{gh}$  mode; and  $CM_b(m * \Delta t, M_{gh})$  denotes the number of road traffic data after compression at  $(m * \Delta t)$  moment under  $M_{gh}$  mode.

Different  $(SB(m * \Delta t, M_{gh}), ST_j(m * \Delta t, M_{gh}), per)$  corresponds to different  $NMAE$ . Thus, the following expression is reasonable:

$$CR_p(m * \Delta t, M_{gh}) = f(SB(m * \Delta t, M_{gh}), ST_j(m * \Delta t, M_{gh}), per). \quad (8)$$

That is, a certain distribution relationship  $f$  exists between  $(SB(m * \Delta t, M_{gh}), ST_j(m * \Delta t, M_{gh}), per)$  and

TABLE 6: The road segments information.

Road segment ID	Road segment name
HI3009b	Xiao Jie Bridge East to Bei Xiao Jie Bridge
HI3008b	Bei Xiao Jie Bridge to Yong He Gong Bridge
HI7058b	Yong He Gong Bridge to Capital Library
HI7036b	Capital Library to An Ding Men Bridge
HI7057b	An Ding Men Bridge to Zhong Lou North Bridge
HI7056b	Zhong Lou North Bridge to Gu Lou Bridge

$CR_p(m * \Delta t, M_{gh})$ . The process of finding the maximum  $CR_p(m * \Delta t, M_{gh})$  that corresponds to  $(SB(m * \Delta t, M_{gh}), ST_j(m * \Delta t, M_{gh}), per)$  is training optimal parameters. Thus, the following model can be obtained:

$$\min f(SB(m * \Delta t, M_{gh}), ST_j(m * \Delta t, M_{gh}), per)$$

$$CR_p(m * \Delta t, M_{gh}) = \frac{CM_a(m * \Delta t, M_{gh})}{CM_b(m * \Delta t, M_{gh})}. \quad (9)$$

Finally, the value of  $(SB(m * \Delta t, M_{gh}), ST_j(m * \Delta t, M_{gh}), per)$  can be determined through statistical analysis of the reconstructed results of road traffic state.

## 4. Experiments

### 4.1. Data Acquisition

**4.1.1. Road Segment Acquisition.** The proposed compression algorithm is conducted with the road traffic spatial relevant data; thus, the selected data must exhibit the characteristics of spatial correlation. The road segments will be briefly explained here. The types of the road segments are express ways, the wide of which is similar. First, the volume data on six typical road segments in Beijing were adopted in the present study. The specific road segments were determined in Table 6.

Five days (June 11, 18, 19, 25, and 26 in 2011) of road traffic data were extracted to construct the reference sequences of road traffic characteristics. The road traffic state data collection interval is 2 min. As the correlation of road

TABLE 7: The cross correlation of the road segment.

Road segment	HI3009b	HI3008b	HI7058b	HI7036b	HI7057b	HI7056b
HI3009b	1	0.9218	0.9286	0.9289	0.9087	0.8349
HI3008b	0.9218	1	0.9645	0.9330	0.9011	0.8310
HI7058b	0.9286	0.9645	1	0.9285	0.8915	0.8294
HI7036b	0.9289	0.9330	0.9285	1	0.9107	0.8676
HI7057b	0.9087	0.9011	0.8915	0.9107	1	0.8199
HI7056b	0.8600	0.8616	0.8655	0.8676	0.8398	1

segments mentioned in the literatures [19, 20], the first two principal components can reflect most of the information of road traffic state. Based on PCA method, we can find that four road segments, HI3009b, HI3008b, HI7058b, HI7036b, exhibited strong correlation that can be determined by cross correlation.

The volume data on the six road segments from June 11, 2011, were extracted to determine the spatial correlation. The cross correlation is shown in Table 7. According to the table, the correlation of all road segments can be determined.

As shown in Table 7, the cross correlations between HI3008b and the other three road segments (HI3009b, HI7058, and HI7036b) were greater than 0.9. Thus, the HI3008b road segment served as the base road, and its collected data were considered as the base data. The volume data on the four road segments were selected for the case study to prove the performance of the proposed algorithm. This can be explained by the following reasons.

The change regularity of volume is mainly determined by the regularity of people's origin-destination (OD) travel. But for different date, people travel OD changes randomly. The travel on weekends has a comparative regularity. Thus, four days (June 18, 19, 25, and 26 on 2011) of road traffic data on spatial road segments were extracted to construct the reference sequences of road traffic characteristics.

**4.1.2. Data Instruction.** The collected road traffic data on the HI3009b, HI7058b, and HI7036b road segments from June 11, 2011, were considered as training data to conduct algorithm parameter settings. Under the same mode, the collected road traffic data on the HI3009b, HI7058b, and HI7036b road segments from four other days were regarded as real-time data to validate the proposed algorithm.

**4.2. Results.** The road traffic spatial volume data compression results based on LZW encoding on the HI3009b, HI7058b, and HI7036b road segments are illustrated in Figures 5–16.

The running time is provided here, which can indirectly reflect the calculation speed of the proposed method. Through several times testing, the average running time is approximate to 0.45. From the running time, we can see that the proposed method is simple and practicable.

The statistical reconstructed results of spatial volume data based on LZW encoding on HI3009b and HI7058b road segments from June 18, 19, 25, and 26, 2011, are illustrated in Tables 8 and 9, respectively.  $CR$ ,  $AE$ ,  $marerr$ , and  $\sigma$  denote the compression ratio, mean absolute error, absolute relative

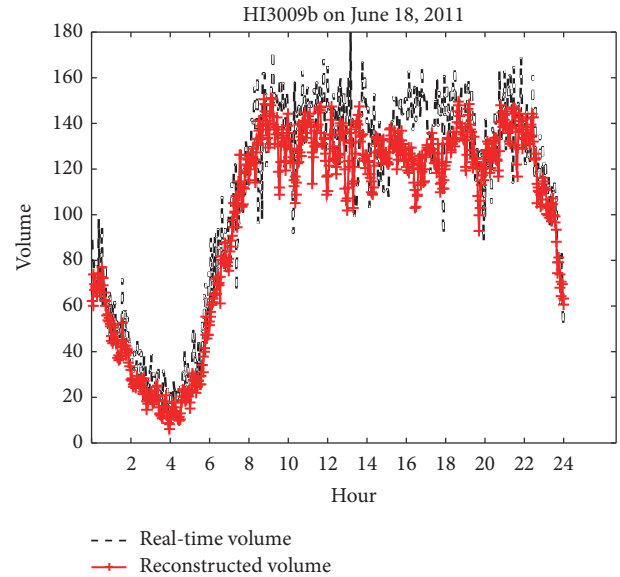


FIGURE 5: HI3009b on June 18, 2011.

TABLE 8: Reconstructed results on HI3009b road segment.

Date	18	19	25	26	Average
$CR$	11.08	9.73	10.14	8.67	9.91
$AE$	11.24	11.28	14.14	11.92	12.15
$marerr$	12.42	14.82	14.09	13.81	13.79
$\sigma$	13.93	13.41	15.12	14.00	14.12

TABLE 9: Reconstructed results on HI7058b road segment.

Date	18	19	25	26	Average
$CR$	16.74	16.00	15.65	11.80	15.05
$AE$	6.65	6.93	6.83	7.41	6.96
$marerr$	6.87	7.67	7.07	8.51	7.53
$\sigma$	8.65	9.33	8.87	9.79	9.16

error percentage, and error standard deviation, respectively. Average denotes the mean value of the four indicators.  $CR$  is described in (7).  $AE$ ,  $marerr$ , and  $\sigma$  can be described as follows:

$$AE = \frac{1}{P} \cdot \sum_p \left| CSM_p(m * \Delta t, M_{gh}) - SM_p(m * \Delta t, M_{gh}) \right|,$$



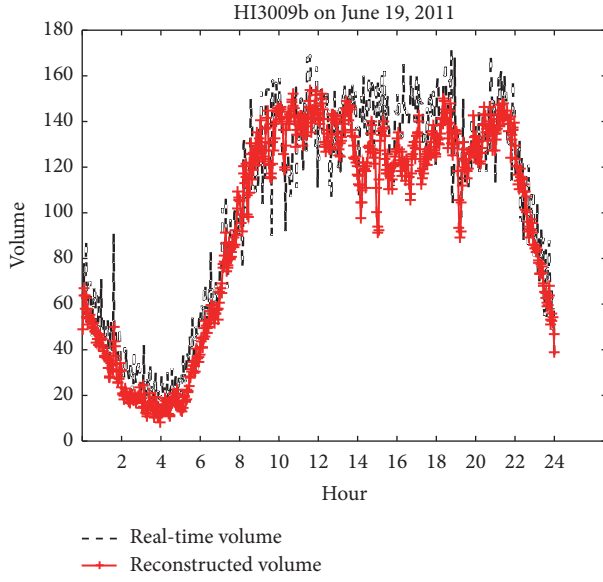


FIGURE 6: HI3009b on June 19, 2011.

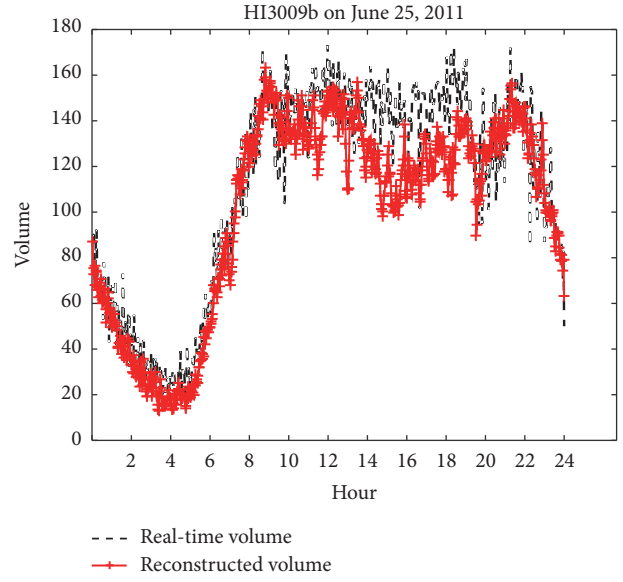


FIGURE 7: HI3009b on June 25, 2011.

$$marerr = \frac{1}{p}$$

$$\cdot \frac{\sum_p |CSM_p(m * \Delta t, M_{gh}) - SM_p(m * \Delta t, M_{gh})|}{SM_p(m * \Delta t, M_{gh})},$$

$$\sigma = \sqrt{\frac{\sum_p (y_p(m * \Delta t, M_{gh}) - \bar{e}_p(m * \Delta t, M_{gh}))^2}{p - 1}}, \quad (10)$$

where

$$y_p(m * \Delta t, M_{gh}) = CSM_p(m * \Delta t, M_{gh}) - SM_p(m * \Delta t, M_{gh}), \quad (11)$$

$$\bar{e}_p(m * \Delta t, M_{gh}) = \frac{1}{p} \sum_p y_p(m * \Delta t, M_{gh}),$$

where  $y_p(m * \Delta t, M_{gh})$  denotes the error data between the original real-time data and the reconstructed real-time data on  $p$  road segments at  $(m * \Delta t)$  moment under  $M_{gh}$  mode;  $\bar{e}_p(m * \Delta t, M_{gh})$  denotes the mean error at  $(m * \Delta t)$  moment under  $M_{gh}$  mode.

**4.3. Sensitive Analysis.** A sensitivity analysis is the study of how the uncertainty in the output of a mathematical model or system (numerical or otherwise) can be apportioned to different sources of uncertainty in its inputs [21]. In Section 4.2, four road segments are selected, and HI3008b is used for training and the others are used for testing. To test the effect of data size on the compression and reconstruction results, a sensitive analysis is urgently needed. Since the proposed algorithm is applicative for big data in road traffic transportation data, a sensitive analysis is also required to test the feasibility for little and medium-size data. The data size

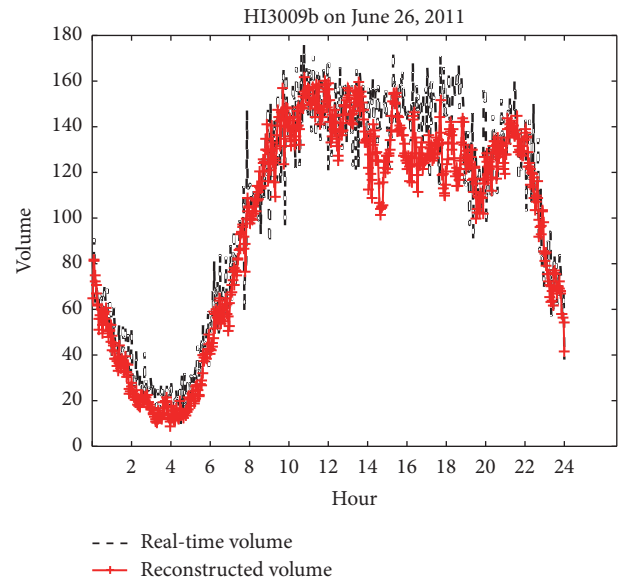


FIGURE 8: HI3009b on June 26, 2011.

can be indicated by the collecting time. Thereby, a sensitive analysis is conducted through testing the compression and reconstruction results indicators under different collecting time, that is,  $CR$ ,  $AE$ ,  $marerr$ , and  $\sigma$ .

The process of parameters determination is performed in Section 3, but the optimal parameter is determined under fixed collecting time. For different collecting time, the optimal parameters will be different. Thus, collecting time is considered as a variable to test compression results. Besides, in this process, we also follow the rule in (9).

Here, a brief data declaration is provided. In Section 4.1, one-day collected data (720) are used for experiment. To test the feasibility of the proposed method, we calculate

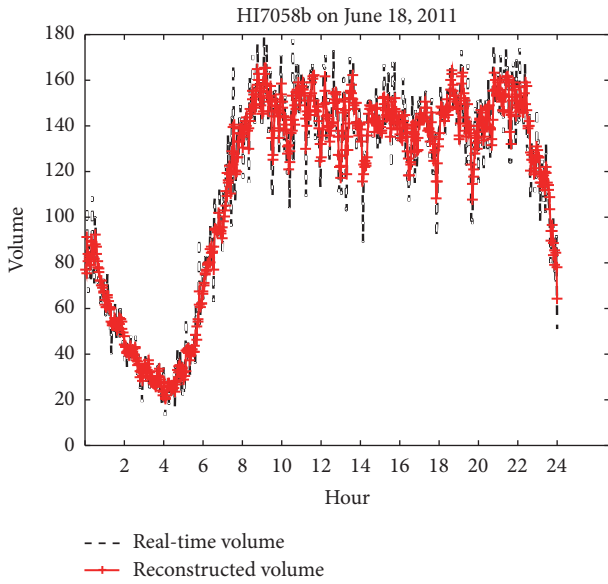


FIGURE 9: HI7058b on June 18, 2011.

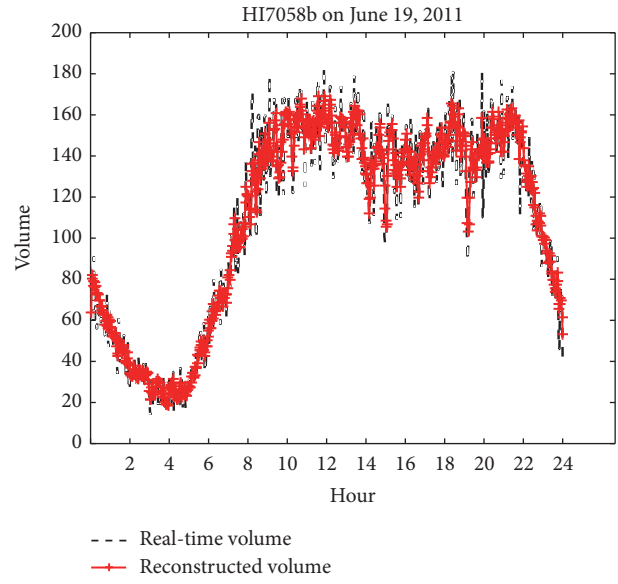


FIGURE 10: HI7058b on June 19, 2011.

the experimental index under different collecting time on HI7058b. This may be regarded as a test bed. The sensitive analysis can be seen in Tables 11–14.

From the sensitive analysis results shown in Tables 11–14, we can see that the compression ratio of big-size data is relatively greater than little and medium-size data. And  $AE$ ,  $marerr$ , and  $\sigma$  are all less than 10. The results show that the proposed algorithm is feasible.

A comparison is also provided here. PCA method is a famous data compression method; thus, we compare the proposed method with PCA method. We compare the reconstruction indicators on on June 19, 2011. The specific results are shown in Table 15.

From Table 15, we can see that the  $CR$  of LZW encoding is dramatically greater than that of PCA. The  $AE$ ,  $marerr$ , and  $\sigma$  of PCA and LZW are very similar. The comparison proves that the performance of the proposed method is comparatively better.

**4.4. Analysis of Experiment Results.** Based on the experiment results conducted in Section 4.2, the following analyses are presented:

- (1) From Tables 8–10, the following results can be obtained:

For the reconstructed volume data, the average compression ratios are 9.91, 15.05, and 5.94 for the HI3009b, HI7058b, and HI7036b road segments, respectively; the average mean absolute error rates are 12.15, 6.96, and 10.32 for the HI3009b, HI7058b, and HI7036b road segments, respectively; the average absolute relative error percentages are 13.79, 7.53, and 12.00 for the HI3009b, HI7058b, and HI7036b road segments, respectively; the average error standard deviations are 14.12, 9.16, and 13.37 for the HI3009b, HI7058b, and HI7036b road segments, respectively.

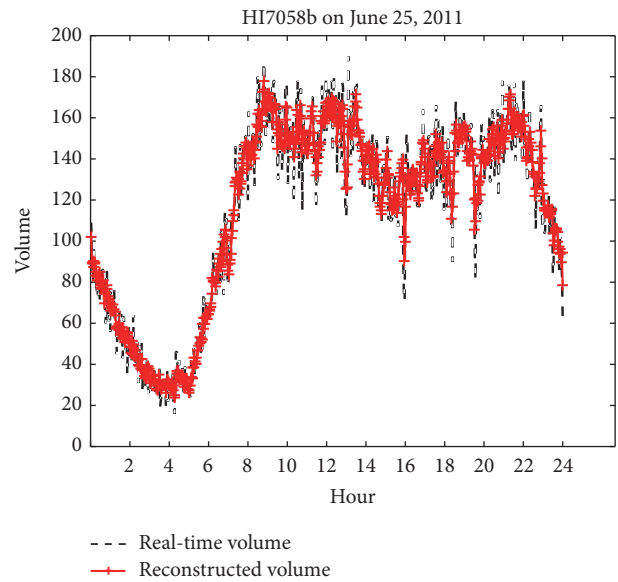


FIGURE 11: HI7058b on June 25, 2011.

TABLE 10: Reconstructed results on HI7036b road segment.

Date	18	19	25	26	Average
$CR$	6.32	5.71	6.43	5.29	5.94
$AE$	9.61	9.76	10.73	11.19	10.32
$marerr$	11.10	12.11	12.02	13.25	12.00
$\sigma$	12.33	12.65	13.85	14.64	13.37

As the statistical data show, we can find that the performance of the HI7058 road segment is better than that of the HI3009b and HI7036b road segments.

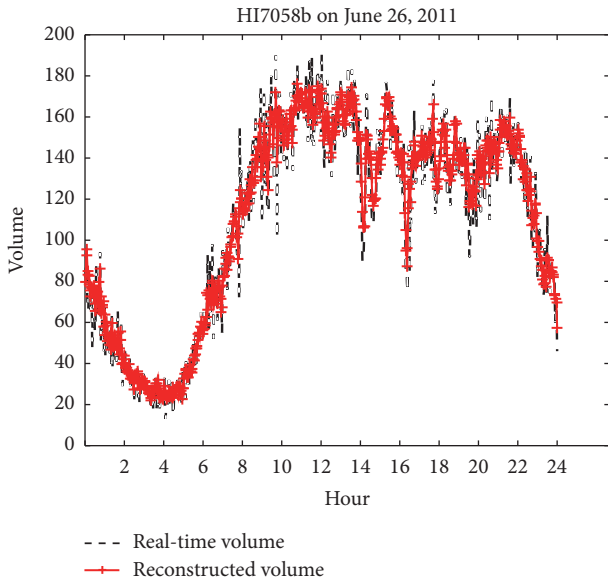


FIGURE 12: HI7058b on June 26, 2011.

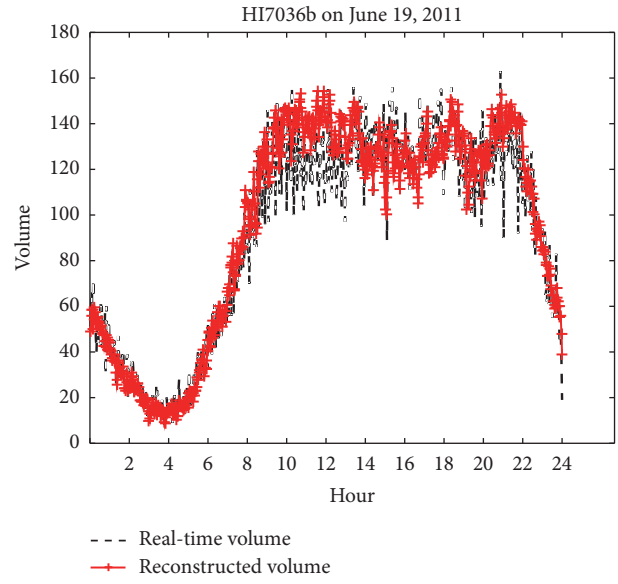


FIGURE 14: HI7036b on June 19, 2011.

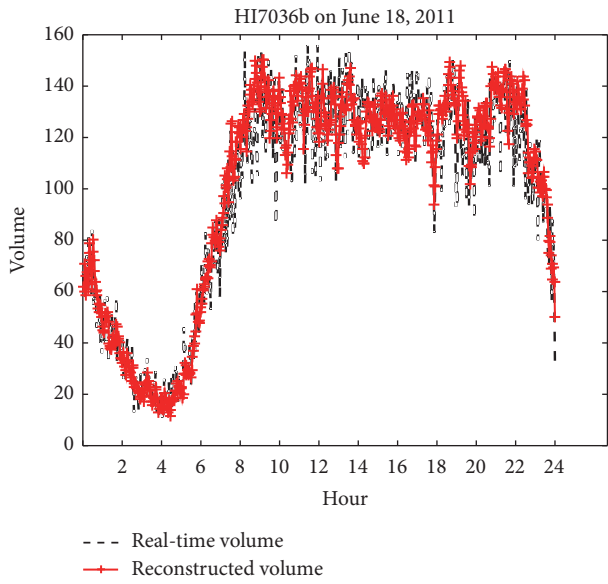


FIGURE 13: HI7036b on June 18, 2011.

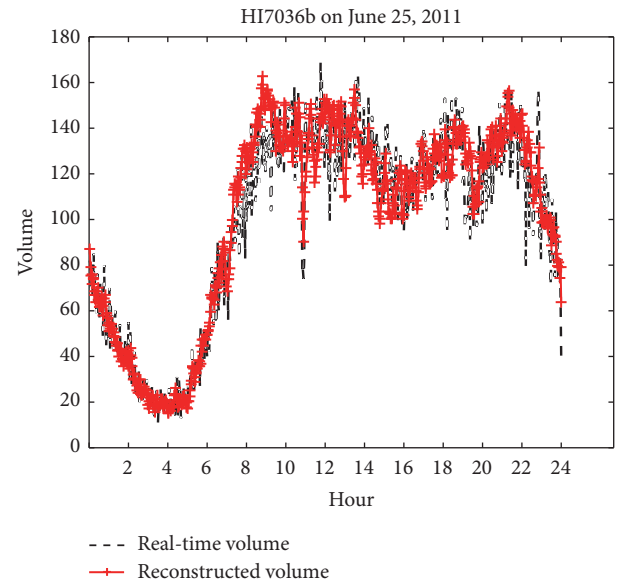


FIGURE 15: HI7036b on June 25, 2011.

- (2) The road traffic spatial volume compression ratio for the HI7058b road segment is higher than that for the HI3009b and HI7036b road segments.

The main reason is that the cross correlation of road traffic spatial volume for HI7058b is higher than that for HI3009b and HI7036b. From Table 7, a similar conclusion can be reached. Consequently, the volume compression ratio for the HI7058b road segment is higher than that for the other two road segments.

- (3) The precision of the reconstructed results of volume for the HI7058b road segment is higher than that for the other two road segments.

From Figures 5–16, we can get that the precision and stability of the reconstructed volume data for the HI7058b road segment is higher than those for the HI3009b and HI7036b road segments based on LZW encoding. The phenomenon is mainly caused by the cross correlation between the base volume data and real-time volume data on base road segment and other spatial road segments, respectively. From Table 7, similar conclusions can be reached.

- (4) Some peak points are missing and the phenomenon can be described by the following reason.

TABLE 11: The sensitive analysis results on June 18 on HI7058b.

Data size	180 (0:00–6:00)	360 (0:00–12:00)	540 (0:00–18:00)	720 (0:00–24:00)
CR	5.45	10.29	14.21	16.74
AE	4.09	5.92	6.35	6.65
marerr (%)	9.92	8.13	7.24	6.87
$\sigma$	5.43	7.90	8.40	8.65

TABLE 12: The sensitive analysis results on June 19 on HI7058b.

Data size	180 (0:00–6:00)	360 (6:00–12:00)	540 (12:00–18:00)	720 (0:00–24:00)
CR	3.00	10.59	15.00	16.00
AE	3.84	6.16	6.45	6.93
marerr (%)	10.98	9.14	7.86	7.67
$\sigma$	4.79	8.51	8.74	9.33

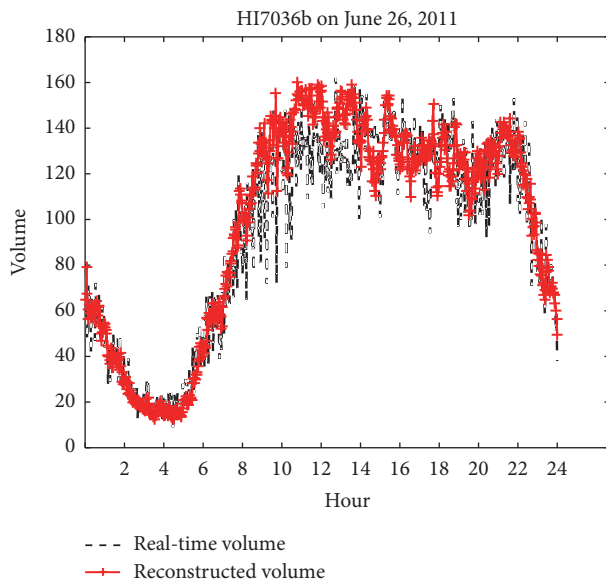


FIGURE 16: HI7036b on June 26, 2011.

The road traffic data at the peak points has a sudden change compared to the data on the base road segment. In the feature extraction process, the features are extracted based on the threshold processing of the difference data. If the features needed to be retained, we can shift down the threshold. In the reconstruction process, the peak points can be retained at the cost of compression ratio. From the reconstructed results shown above, the peak points are sustainable.

- (5) Several errors occur in the road traffic state reconstruction of this algorithm.

The errors are mainly caused by the following two reasons:

- (1) Obtaining the corresponding road traffic spatial states with a perfect match based on LZW encoding is difficult because of the limitations of the road traffic running characteristics.

- (2) The parameters exhibit a certain deviation. Determining the optimal parameters is irregular because they vary for different road traffic state datasets. The selected optimal parameters are determined based on the historical road traffic state data. Therefore, the current optimal parameters are approximately different from the historical optimal parameters.

## 5. Conclusions

An effective road traffic data compression algorithm can boost the data transportation and storage effectiveness of a road traffic system. The PCA method can be used to select the road traffic segments with strong correlation. Based on the spatial correlation of the road traffic spatial data, this study proposes a road traffic spatial data compression algorithm that uses LZW encoding. The contributions of this study can be effectively used for the road traffic spatial data compression of different road segments. Besides, the high spatial correlation roads are selected by PCA, which can also be used in transportation research. Further, the compression method can motivate some interesting ideas in transportation research field as well.

For the road segments with high spatial correlation, the proposed algorithm performs effectively. According to the reconstructed results of the HI3009b and HI7036b road segments, the algorithm is sensitive for correlation. The stronger the correlation is, the better the performance of the algorithm is. Thus, to ensure improved performance, the cross correlation should be greater than 0.95. Then, the expected compression ratio can be obtained.

Considering the remarkable performance of the proposed algorithm, we will explore the traffic state compression based on the spatial-temporal correlations in our next study.

## Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

TABLE 13: The sensitive analysis results on June 25 on HI7058b.

Data size	180 (0:00–6:00)	360 (0:00–12:00)	540 (0:00–18:00)	720 (0:00–24:00)
CR	3.27	10.59	12.86	15.65
AE	4.19	5.92	6.43	6.83
marerr (%)	10.00	8.16	7.22	7.07
$\sigma$	5.26	7.76	8.39	8.87

TABLE 14: The sensitive analysis results on June 26 on HI7058b.

Data size	180 (0:00–6:00)	360 (0:00–12:00)	540 (0:00–18:00)	720 (0:00–24:00)
CR	2.09	4.86	15	11.80
AE	4.29	6.51	7.32	7.41
marerr (%)	11.70	9.60	9.03	8.51
$\sigma$	5.60	9.14	9.84	7.79

TABLE 15: The indicators results of PCA and LZW encoding on June 19, 2011.

Road segment	CR		AE		marerr		$\sigma$	
	PCA	LZW	PCA	LZW	PCA	LZW	PCA	LZW
HI3009b	2.01	9.73	7.43	11.28	8.10	14.82	9.94	13.43
HI7036b	2.01	5.71	9.43	9.76	13.31	12.11	11.82	12.65
HI7058b	2.01	16.00	7.15	6.93	7.30	7.67	9.23	9.33

## Acknowledgments

This work was supported by the Zhejiang Provincial Natural Science Foundation (Grant no. LQ16E080012), the National Natural Science Foundation of China (Grant no. 6157331), and Open Fund for a Key-Key Discipline of Zhejiang Province (2015001).

## References

- [1] E. I. Vlahogianni, B. B. Park, and J. W. C. Van Lint, "Big data in transportation and traffic engineering," *Transportation Research Part C: Emerging Technologies*, vol. 58, p. 161, 2015.
- [2] Q. Shi and M. Abdel-Aty, "Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 380–394, 2015.
- [3] Z. Zhang, Q. He, H. Tong, J. Gou, and X. Li, "Spatial-temporal traffic flow pattern identification and anomaly detection with dictionary-based compression theory in a large-scale urban network," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 284–302, 2016.
- [4] K. Sayood, *Introduction to Data Compression*, Elsevier Press, Newnes, 4th edition, 2012.
- [5] G.-H. Ahn, Y.-K. Ki, and E.-J. Kim, "Real-time estimation of travel speed using urban traffic information system and filtering algorithm," *IET Intelligent Transport Systems*, vol. 8, no. 2, pp. 145–154, 2014.
- [6] X.-H. Yao, F. B. Zhan, Y.-M. Lu, and M.-H. Yang, "Effects of real-time traffic information systems on traffic performance under different network structures," *Journal of Central South University of Technology (English Edition)*, vol. 19, no. 2, pp. 586–592, 2012.
- [7] Z. Q. Zhao, Y. Zhang, J. M. Hu et al., "Comparative study of PCA and ICA based traffic flow compression," *Journal of Highway and Transportation Research and Development*, vol. 25, no. 11, pp. 109–113, 2008.
- [8] D. W. Xu, H. H. Dong, H. J. Li, L. M. Jia, and Y. J. Feng, "The estimation of road traffic states based on compressive sensing," *Transportmetrica B*, vol. 3, no. 2, pp. 131–152, 2015.
- [9] Q.-Q. Li, Y. Zhou, Y. Yue, and A. G.-O. Yeh, "Compression method of traffic flow data based on compressed sensing," *Journal of Traffic and Transportation Engineering*, vol. 12, no. 3, pp. 113–126, 2012.
- [10] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: from theory to applications," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4053–4085, 2011.
- [11] B. Li, J. Z. Xie, and B. L. Wang, "Signal reconstruction based on compressed sensing," *Computer Technology and Development*, vol. 19, no. 5, pp. 23–25, 2009.
- [12] Y. Xiao, L.-Y. Lu, S. Gao, Y.-M. Xie, and D.-Y. Xu, "Traffic data denoising compression for intelligent traffic systems based on 2-D discrete wavelet transformation," *Beifang Jiaotong Daxue Xuebao/Journal of Northern Jiaotong University*, vol. 28, no. 5, pp. 1–5, 2004.
- [13] X. L. Ou, J. T. Ren, and Y. Zhang, "A neural network mode for urban volumes compression," *Cybernetics and Informations*, vol. 1, no. 4, 2003.
- [14] G. J. Xu and H. Wang, "Implementation of locating data real-time compression of embedded GPS system with car," *Information Technology*, vol. 4, p. 14, 2006.
- [15] Q.-L. Ma, W.-N. Liu, and D.-H. Sun, "A high-speed compression scheme for vast quantities of GPS data," *Journal of Sichuan University*, vol. 43, no. 1, pp. 123–128, 2011.
- [16] Q.-Z. Wang, K. Wang, and Z.-S. Yang, "Coding algorithm of traffic flow in intelligence guidance system based on adaptive



- switching mode,” *China Journal of Highway and Transport*, vol. 6, article no. 14, 2009.
- [17] M. Hou, *QoS Management with Differentiated Services IP over the Internet*, Kingston, Ontario, Canada, 1999.
- [18] R. Song, W. Sun, B. Zheng et al., “RESS: a novel framework of trajectory compression in road networks,” *Proceedings of the VLDB Endowment*, vol. 7, no. 9, pp. 661–672, 2014.
- [19] J. Dong, Y. Zhang, Z. Zhang, and X.-X. Kuang, “Principal component analysis of dependency of urban intersections,” *Journal of Southwest Jiaotong University*, vol. 38, no. 6, pp. 619–622, 2004.
- [20] Y. Pan and J. Zhang, “Traffic-flow prediction based on dependency analysis of urban intersections,” *Transportation and computer*, vol. 23, no. 1, pp. 31–34, 2005.
- [21] A. Saltelli, “Sensitivity analysis for importance assessment,” *Risk Analysis*, vol. 22, no. 3, pp. 579–590, 2002.

## Research Article

# Dynamic Route Choice Prediction Model Based on Connected Vehicle Guidance Characteristics

Jiangfeng Wang, Jiarun Lv, Chao Wang, and Zhiqi Zhang

MOE Key Laboratory for Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Jiangfeng Wang; wangjiangfeng@bjtu.edu.cn

Received 26 December 2016; Accepted 24 January 2017; Published 14 February 2017

Academic Editor: Xiaolei Ma

Copyright © 2017 Jiangfeng Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A route choice prediction model is proposed considering the connected vehicle guidance characteristics. This model is proposed to prevent the delay in the release of guidance information and route planning due to inaccurate timing predictions of the traditional guidance systems. Based on the analysis of the impact of different connected vehicle (CV) guidance strategies on traffic flow, an indexes system for CV guidance characteristics is presented. Selecting five characteristic indexes, a route choice prediction model is designed using the logistic model. A simulation scenario is established by programming different agents for controlling the flow of vehicles and for information acquisition and transmission. The prediction model is validated using the simulation scenario, and the simulation results indicate that the characteristic indexes have a significant influence on the probability of choosing a particular route. The average root mean square error (RMSE) of the prediction model is 3.19%, which indicates that the calibration model shows a good prediction performance. In the implementation of CV guidance, the penetration rate can be considered an optional index in the adjustment of the guidance effect.

## 1. Introduction

The connected vehicle (CV) guidance system is a new type of guidance system. This system realizes dynamic vehicle guidance by utilizing connected vehicle technologies. Based on the vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), and vehicle-to-smart terminal (V2T) technologies, the CV guidance system facilitates dynamic guidance for road network flow using real-time traffic information. Traditional guidance systems have several shortcomings such as the delay in the release of guidance information and route planning due to inaccurate timing predictions. The CV guidance overcomes these shortcomings and enhances the spatial-temporal guidance efficiency in road networks.

The United States (US), the European Union (EU), and Japan have conducted a study on dynamic route guidance using connected vehicle (CV) technologies; these countries have launched their own application projects, including Connected Vehicle [1], AERIS [2], DRIVE C2X [3], and Smartway [4]. These projects have helped improve the travelling efficiency of the road network, propose many road guidance theories and methods, and conduct corresponding

field applications [5–8] by applying the CV technologies to dynamic route guidance. Recently, Beijing, Shenzhen, and Chongqing also have applied the concept of wireless communication to dynamic route guidance [9, 10].

Over the years, many researchers have paid attention to the theory and algorithm of route guidance and proposed various optimization algorithms to analyze their impact on the road network flow [11]. Some intelligent algorithms for route guidance have been proposed, such as the Dijkstra algorithm [12], Floyd algorithm [13], A\* algorithm [14], genetic algorithm [15], neural network algorithm [16], and ant colony optimization [17]. Su et al. [18] proposed a multiobjective and multipath optimization selection method based on the genetic algorithm. The algorithm can provide several alternative routes and satisfies drivers' varying preferences. Furthermore, travelling fitness functions were designed to provide better multiroute selection. A dynamic route guidance method based on the real-time forecast of traffic information was proposed to solve the problem of seeking answers ineffectively in a route guidance algorithm. This method combined the neural network algorithm and the genetic algorithm, and the proposed method improved

the computation efficiency and solution quality [19]. Yang [20] analyzed the major factors that influence the choice of optimal route and then provided an improved K-optimal chaos ant colony algorithm. The results of a simulation experiment showed that the algorithm has much higher capacity for global optimization and can use the basic ant colony algorithm to optimize the route choice. Introducing the idea of depth parameter, Lee and Kim [21] combined the Dijkstra algorithm and A\* algorithm to propose a hybrid route guidance algorithm. Experimental results indicated that the algorithm reduced the computation cost considerably compared to the costs involved in traditional searching algorithms.

Nowadays, scholars have begun to study route guidance in a connected vehicle (CV) environment. Tian et al. [22] presented a real-time route guidance system based on CV technologies, and simulation results showed that better routes are found using the V2V and V2I technologies. Paikari et al. [23] realized CV guidance by developing a V2V and V2I application interface (API). Experimental results demonstrated that the extended simulation system can handle the load of urban freeways and reduce crash risks. Chim et al. [24] proposed a navigation scheme in the CV environment using anonymous credentials and limited jurisdiction, and they addressed the security requirements associated with CV guidance. Vreeswijk et al. [25] proposed a CV guidance strategy based on travelling bounded rationality degree. The results showed that the proposed strategy can effectively reduce the total travel time and realize the goal of system optimal guidance. Genders and Razavi [26] used V2V communication to share warning information about the work zone to nearby vehicles, and a dynamic route guidance algorithm was proposed. The results showed that the CV penetration rate of less than 40% contributes to a safer traffic network for vehicles in the work zone. In general, these studies cannot comprehensively reflect the impact of the CV guidance characteristics on route choice, although the penetration rate was considered.

Existing route guidance research mainly focuses on route guidance algorithms [16, 17, 27–29], guidance strategy [30–32], and system design [33] using traditional data detectors. CV technologies allow for some innovative means for data acquisition, and many scholars have attempted to study the dynamic route guidance models in a CV environment [32, 34, 35]. Several studies have even considered the influence of CV characteristics, such as penetration rate [24, 26, 36–38]. However, existing research does not provide a detailed analysis of the CV guidance characteristics. Therefore, this paper presents a CV guidance algorithm based on the proposed CV guidance characteristic indexes. A simulation experiment is conducted to analyze the prediction accuracy.

## 2. Prediction Model

In a CV environment, travellers receive traffic guidance information using V2V and V2I technologies and select the optimum route according to the guidance strategy. Different guidance strategies have different effects on the flow distribution of a road network. Figure 1 shows the guidance effect of

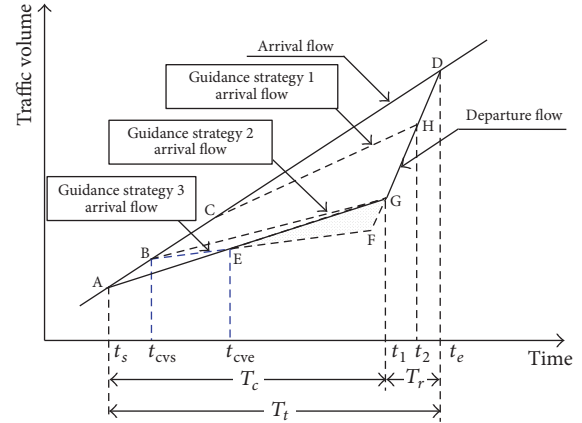


FIGURE 1: Impact of CV guidance strategy on traffic flow.  $t_s$ —start time of traffic congestion;  $t_{cvs}$ —start time of CV guidance strategy 3;  $t_{cve}$ —end time of CV guidance strategy 3;  $t_1$ —end time of CV guidance strategy 2;  $t_2$ —end time of guidance strategy 1;  $t_e$ —end time of congestion.

different guidance strategies in a CV guidance environment. The guidance information from the variable message signs (VMS) is received during traffic congestion on a route to reduce the arrival flow. The arrival flow is shown by the line representing guidance strategy 1 in Figure 1. Considering the large delay in information release, the guidance effect has not been exhibited well. Using the V2V technology, the guidance information is transmitted in time, and the arrival flow is shown by the line representing guidance strategy 3 in Figure 1. However, guidance strategy 3 might generate surplus capacity, as indicated by the area of the triangle EFG. Hence, this guidance strategy is not a good option and the optimal option would be guidance strategy 2. Therefore, a reasonable guidance strategy can be obtained by adjusting the CV guidance characteristics, such as the penetration rate.

The CV guidance characteristics have a direct impact on the guidance effect. The characteristic indexes for CV guidance features are shown in Figure 2.

In this study, five characteristic indexes were chosen as the variables for analysis: compliance rate (CR), following rate (FR), penetration rate (PR), release delay time (DT), and congestion level (CL). CR is the ratio of vehicles with CV ability to comply route adjustments to all vehicles with CV ability. FR refers to the ratio of vehicles without CV ability following to adjust the route because of the influence of the leading vehicle changing the route to vehicles without CV ability. PR refers to the proportion of vehicles with CV ability to all vehicles. DT refers to the interval from the generation of traffic information to the reception of traffic information by the vehicle. CL is an important index for evaluating the flow operation states of the road network, and the duration of congestion caused by an accident vehicle is used as an alternative variable in this paper.

To study the influence of the CV guidance characteristic indexes on route choice behavior, the route choice model was established using the logistic model. Utility refers to a metric that should be maximized to satisfy travellers' demands, such

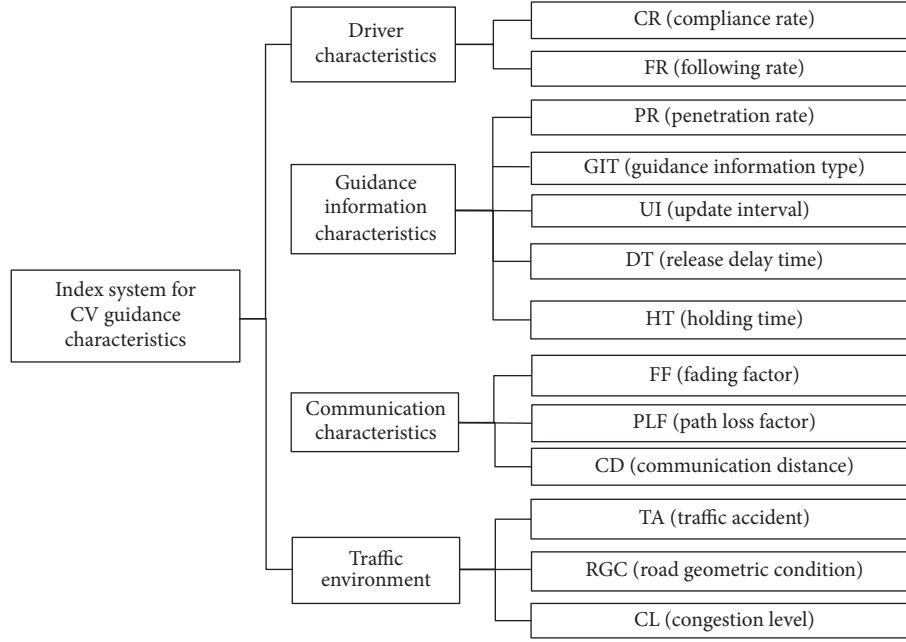


FIGURE 2: Indexes system for CV guidance characteristics.

as travel time. The logistic model is based on the utility maximization theory, which states that all vehicles will always choose the route with maximum utility during the route selection process. The utility is represented by the following equation:

$$U_j = \sum_{n=1}^N \beta_{nj} x_{nj} + \beta_0, \quad (1)$$

where  $U_j$  is the utility of the choice route  $j$ ,  $x_{jn}$  is the explanatory variable for the characteristics index  $n$  of the choice route  $j$ ,  $\beta_{nj}$  is the coefficient of  $x_{jn}$ , and  $\beta_0$  is a constant.

In a CV environment, origin-destination (od) pairs have  $r$  ( $r \geq 2$ ) routes. The route choice model is

$$G_{L_j^{\text{od}}} = \ln \left( \frac{P_j^{\text{od}}}{P_j^{\text{od}}} \right) = \beta_j^{\text{od}} + \sum_n \beta_{jn}^{\text{od}} x_{jn}, \quad (2)$$

$$j = 1, 2, \dots, r-1,$$

where  $G_{L_j^{\text{od}}}$  is the utility of the choice route  $L_j^{\text{od}}$  relative to the reference route  $L_j^{\text{od}}$  ( $j \neq J$ );  $P_j^{\text{od}}$  and  $P_j^{\text{od}}$  are the probabilities of the choice routes  $L_j^{\text{od}}$  and  $L_j^{\text{od}}$ , respectively.  $\beta_j^{\text{od}}$  is a constant;  $\beta_{jn}^{\text{od}}$  is the coefficient of the explanatory variable  $x_{jn}$ ;  $x_{jn}$  is the explanatory variable for the characteristic index  $n$  of the choice route  $j$  in a CV environment.

Based on (2), assuming  $P_j^{\text{od}}/P_j^{\text{od}} = e^{G_{L_j^{\text{od}}}}$  and  $\sum_{j=1}^{r-1} e^{G_{L_j^{\text{od}}}} = \sum_{j=1}^{r-1} P_j^{\text{od}}/P_j^{\text{od}} = y$ , the route choice probability in a CV environment is as follows:

$$P_j^{\text{od}} = \frac{e^{G_{L_j^{\text{od}}}}}{1 + y}. \quad (3)$$

### 3. Model Validation

To verify the effectiveness of the route choice model, a simple road network scenario is designed, which includes three routes and four 3-way intersections, as shown in Figure 3. In the initialization stage, vehicles with connected vehicle ability (CVs) and vehicles without connected vehicle ability (non-CVs) are arranged in the road network. Car-Agent is used to control the vehicles' traffic behaviors, such as car following and lane changing by programming the agents using the EstiNet tool. The Roadside Unit (RSU) at an intersection is used to collect the traffic volume entering the intersection and transmit the volume to the Central-Roadside Unit (C-RSU) in real time. The C-RSU is responsible for receiving the volume from the downstream intersections; it calculates the shortest route using the Bureau of Public Roads (BPR) impedance function. To generate the traffic congestion, a broken vehicle (BV) was set on route 1. During the simulation stage, all vehicles, including CV and Non-CV, choose the shortest route to travel depending on the real-time traffic information obtained through V2V and V2I communication. When the simulation starts, the BV sends the accident information to the C-RSU. Meanwhile, the C-RSU broadcasts the real-time traffic information to all vehicles. CVs will choose the shortest route to travel according to the CV guidance information, and non-CVs will choose their routes to travel based on their own reasonable judgment. For example, these non-CVs may change their initial route following the leading vehicles depending on the FR.

The basic parameters of the simulation are listed in Table 1. The total number of vehicles including CVs and non-CVs is 100 vehicles, which is represented by the parameter of number of vehicles.

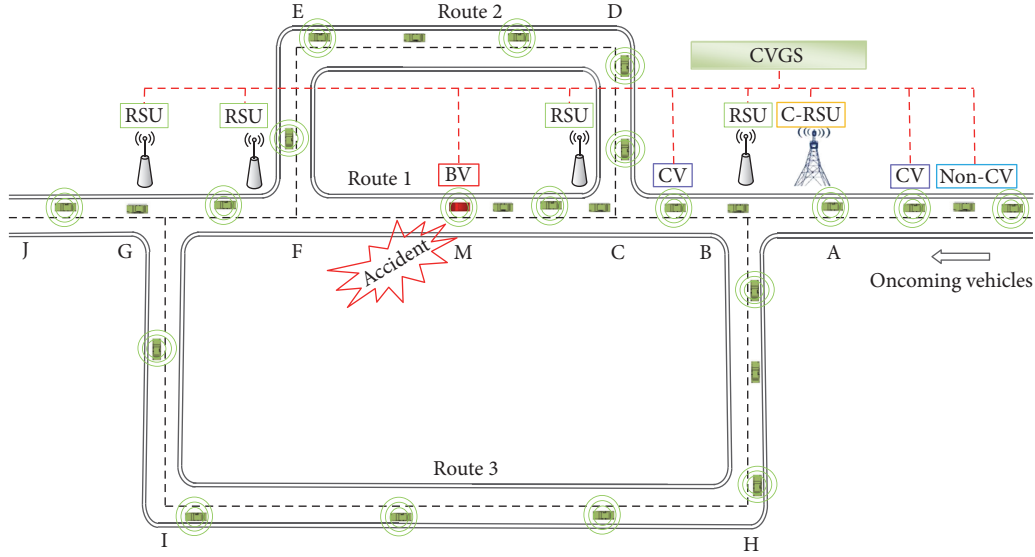


FIGURE 3: CV guidance simulation scenario.

TABLE 1: Basic parameters of the simulation.

Parameters	Values	Parameters	Values
Simulation time (s)	1800	Number of intersections	4
Number of vehicles	100	Number of road sections	12
Maximum velocity (m/s)	18	MAC communication protocol	IEEE802.11p
Acceleration velocity (m/s <sup>2</sup> )	-4 to 1	CL (s)	300/500
Transit power (m)	1000	PR (%)	25/50/75/100
Number of lanes	2	CR (%)	25/50/75/100
Lane width (m)	3.5	FR (%)	0/10/20/30
Vehicle type	Passenger car unit	DT (s)	0/120/180

Using the CV guidance scenario, the five characteristic indexes with their corresponding values were selected for simulation. 312 simulation experiments were carried out, and the experimental samples were divided into 200 calibration samples and 112 test samples. Using the calibration samples and considering route 3 as the reference choice route, the multinomial logit model is used to obtain the coefficients of variables. The calibrated results are as follows:

$$\begin{aligned}
 G_{L_1}(x) &= 1.608PR + 2.006CR + 1.881FR \\
 &\quad - 0.085CL(0) + 0.280DT(0) \\
 &\quad + 0.114DT(1) - 3.582, \\
 G_{L_2}(x) &= 3.143PR + 3.494CR + 3.737FR \\
 &\quad - 0.413CL(0) + 0.487DT(0) \\
 &\quad + 0.126DT(1) - 7.046.
 \end{aligned} \tag{4}$$

In the next section, the test samples are used to evaluate the performance of the proposed route choice model and analyze the impact of the five characteristic indexes on the route choice.

#### 4. Results Analysis

To verify the effects of CV guidance, the results of sample 1 (without guidance) and sample 2 (with guidance) were compared, and the results are listed in Table 2. With CV guidance, the average travel time on the entire road network decreased by 20.31%. The volume distribution ratios in route 2 and route 3 gradually tend to balance one another in the CV guidance environment, as shown in Figure 4.

The PR is one of the most important characteristic indexes, which determines the proportion of CVs. Figure 5 illustrates the impact of the PR on the probability of the route choice in a CV guidance environment. The figure also illustrates that a higher proportion of CV vehicles will choose route 2 and route 3 as the PR increases. Finally, the distribution of volume will even out when the PR is approximately 100%. This may be because more CVs will follow the guidance information to choose the optimal route 2 to travel when route 1 is blocked owing to an increase in the PR. With an increase in the impedance of route 2, the CV will choose route 3 to travel.

Figures 6(a) and 6(b) show the impact of different CR and FR values on the choice probability of route 2 and route 3 in a CV guidance environment. With an increase in the FR, the



TABLE 2: Distribution of volume and average travel time for different routes.

	Flow distribution ratio (%)			Travel time (s)			
	Route 1	Route 2	Route 3	Route 1	Route 2	Route 3	Mean
Sample 1	87	8	5	1325.19	1076.73	1004.92	1289.30
Sample 2	11	43	46	1006.22	1055.50	1006.37	1027.48
Savings	87.36% ↓	81.40% ↑	89.13% ↑	24.07% ↓	1.97% ↓	0.14% ↑	20.31% ↓

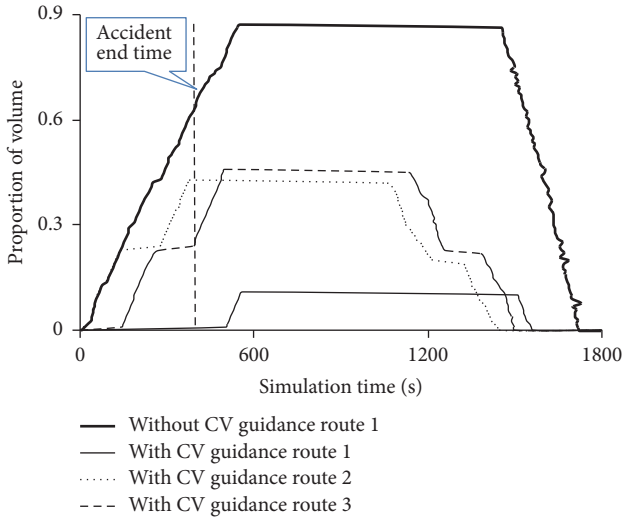


FIGURE 4: Volume distribution of each route in CV guidance environment.

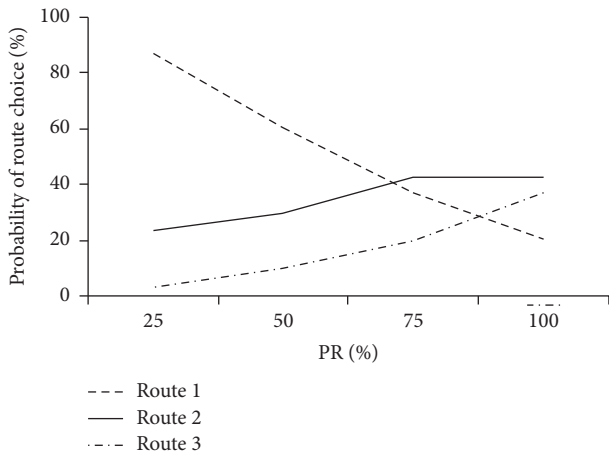


FIGURE 5: Impact of different PR values on probability of route choice.

choice probability of route 2 and route 3 increases steadily. On the other hand, the index CR has a more obvious impact on the probability of route choice. Overall, an increasing number of vehicles will choose a detour as the FR and CR increase when route 1 is blocked, and the traffic impedance values of route 2 and route 3 gradually reach a state of equilibrium. Figure 6 indicates that the indexes FR and CR have a significant influence on the probability of route choice, which is consistent with the theoretical expectations.

Using the test samples, a calibration model is employed to predict the vehicles' route choice considering the impact of the five characteristic indexes, and the prediction accuracy is analyzed using the root mean square error (RMSE). The prediction results of the calibration model are presented in Table 3. It is shown in the table that the prediction accuracy of the route choice ranges from 2.40% to 4.52% for different values of the five characteristic indexes. The prediction accuracy of the calibration model is high for all PR values, while the index of the CR has a significant influence on the prediction accuracy of the calibration model, which shows relatively big fluctuation. Overall, the average RMSE of the calibration model is 3.19%, which indicates that the calibration model shows a good prediction performance.

The following section will analyze the impact of the five characteristic indexes on the prediction accuracy. Figure 7 shows the influence of the indexes FR, CR, and FR on the prediction accuracy of the calibration model. It is concluded from Figure 7 that the prediction accuracy decreases initially and then increases with an increase in the values of FR, CR, and FR. A reasonable prediction accuracy is obtained when the values of PR, CR, and FR are 50%, 50%, and 10%, respectively.

Figure 8 shows the influence of the indexes CL and DT on the prediction accuracy. The RMSE of the calibration model decreases with an increase in the value of the CL. This might be explained by a more stable route choice behavior in congestion states, and the calibration model has a higher prediction accuracy of route choice. When the value of the CL is 500 s, the model has a smaller RMSE, which indicates that the calibration model exhibits better prediction in a congested environment. The RMSE of the calibration model increases with an increase in the value of the DT, which shows that the delay in the release of traffic information has a negative effect on the prediction accuracy. This might be explained by a more confused route choice behavior for a bigger release delay time, and the calibration model has a lower prediction accuracy of route choice. The real-time release of the CV guidance information helps the vehicles choose a suitable route to avoid traffic congestion.

### 5. Conclusion

A route choice model was proposed considering the characteristics of CV guidance; this model was validated using the EstiNet simulation tool. The effect on CV guidance was statistically analyzed, and the impact of the five characteristic indexes on the prediction accuracy of the calibration model was studied. The simulation results showed that the indexes

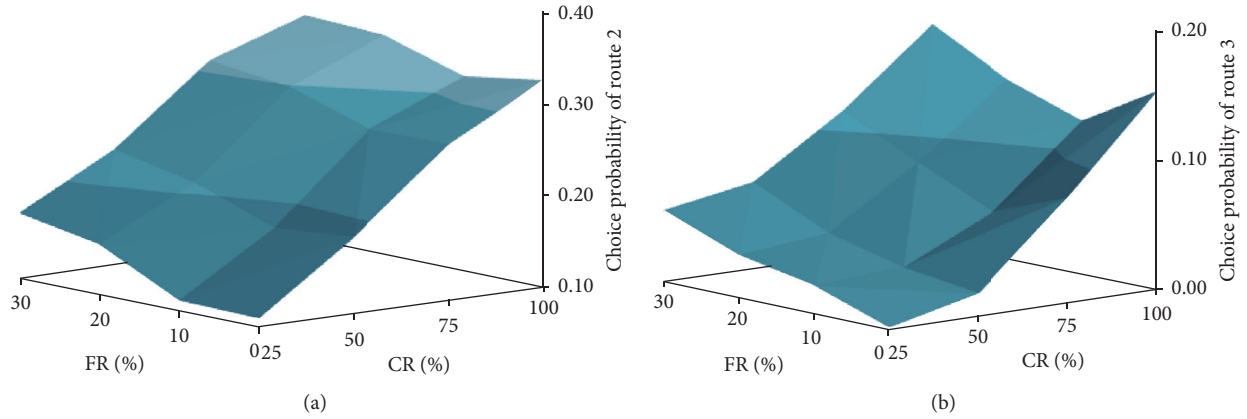


FIGURE 6: Influence of different FR and CR values on route choice probability of routes 2 and 3.

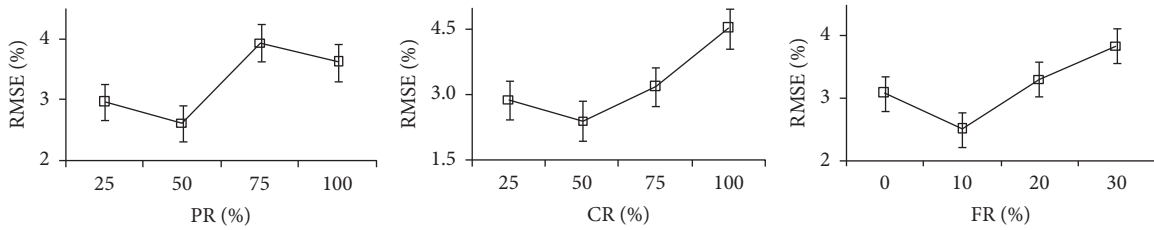


FIGURE 7: Influence of PR, CR, and FR on RMSE of calibration model.

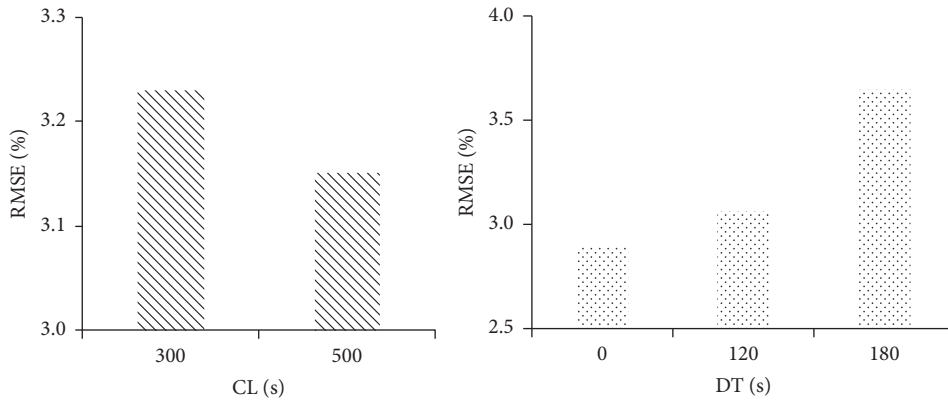


FIGURE 8: Influence of CL and DT on RMSE of calibration model.

TABLE 3: Influence of CV characteristics indexes on prediction accuracy.

CL (s)	300	500	—	—
RMSE	3.23%	3.15%	—	—
DT (s)	0	120	180	—
RMSE	2.89%	3.06%	3.64%	—
PR (%)	25	50	75	100
RMSE	2.96%	2.62%	3.93%	3.61%
CR (%)	25	50	75	100
RMSE	2.88%	2.40%	3.18%	4.52%
FR (%)	10	20	30	40
RMSE	3.08%	2.49%	3.30%	3.84%

PR, FR, and CR had a significant influence on the probability of route choice, which was consistent with the theoretical expectations. Overall, the average RMSE of the calibration model was 3.19%, which indicates that the calibration model exhibits a good prediction performance. In the implementation of CV guidance, the PR can be considered an optional index to adjust the guidance effect.

There are several considerations for future research works. First, the route choice model will be calibrated and validated through a field experiment to provide better understanding of the benefit of CV guidance. Second, more characteristic indexes will be considered in the design of the route choice model in a CV environment.

## Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

## Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (Grant no. 61473028, 71621001), Beijing Municipal Natural Science Foundation (Grant no. 8162031), and the National High Technology Research and Development Program of China ("863" Program) (Grant no. 2015AA124103).

## References

- [1] US Department of Transportation, Connected Vehicle Research 2014 [http://www.its.dot.gov/connected\\_vehicle/connected\\_vehicle.htm](http://www.its.dot.gov/connected_vehicle/connected_vehicle.htm).
- [2] H. A. Rakha, C. E. Via, and R. K. Kamalanathsharma, "AERIS: Eco-Vehicle Speed Control at Signalized Intersections using I2V Communication," *Speed Control*, 2012.
- [3] European Commission DG CONNECT, DRIVE C2X 2015 <http://www.drive-c2x.eu/project>.
- [4] National Institute for Land and Infrastructure Management, *Smartway 2007*, 2007, [http://www.nilim.go.jp/japanese/its/3paper/pdf/071128ITSWC\\_ss.pdf](http://www.nilim.go.jp/japanese/its/3paper/pdf/071128ITSWC_ss.pdf).
- [5] K. J. Malakorn and B. Park, "Assessment of mobility, energy, and environment impacts of intellidrive-based Cooperative Adaptive Cruise Control and Intelligent Traffic Signal control," in *Proceedings of the IEEE International Symposium on Sustainable Systems and Technology (ISSST '10)*, pp. 1–6, IEEE, Arlington, Va, USA, May 2010.
- [6] H. Lee, W. Lee, and Y.-K. Lim, "The effect of eco-driving system towards sustainable driving behavior," in *Proceedings of the 28th Annual CHI Conference on Human Factors in Computing Systems (CHI '10)*, pp. 4255–4260, Atlanta, Ga, USA, April 2010.
- [7] X. L. Ma, Y. J. Wu, and Y. H. Wang, "Drive Net: an E-science of transportation platform for data sharing, visualization, modeling, and analysis," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2215, pp. 37–49, 2011.
- [8] M. Zhou, X. Qu, and S. Jin, "On the impact of cooperative autonomous vehicles in improving freeway merging: a modified intelligent driver model-based approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, no. 99, pp. 1–7, 2016.
- [9] Z.-S. Yang and L.-Y. Chu, "Study on the development of the Dynamic Route Guidance Systems (DRGS)," *Journal of Highway and Transportation Research and Development*, vol. 17, no. 1, pp. 34–38, 2000.
- [10] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transportation Research Part C: Emerging Technologies*, vol. 54, pp. 187–197, 2015.
- [11] S. Wang and X. Qu, "Station choice for Australian commuter rail lines: equilibrium and optimal fare design," *European Journal of Operational Research*, vol. 258, no. 1, pp. 144–154, 2017.
- [12] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.
- [13] R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.
- [14] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.
- [15] D. J. Bertsimas and D. Simchi-Levi, "A new generation of vehicle routing research: robust algorithms, addressing uncertainty," *Operations Research*, vol. 44, no. 2, pp. 286–304, 1996.
- [16] C. L. Giles and M. W. Goudreau, "Routing in optical multistage interconnection networks: a neural network solution," *Journal of Lightwave Technology*, vol. 13, no. 6, pp. 1111–1115, 1995.
- [17] M. Dorigo, G. Di Caro, and L. M. Gambardella, "Ant algorithms for discrete optimization," *Artificial Life*, vol. 5, no. 2, pp. 137–172, 1999.
- [18] H. B. Su, Y. L. Shi, and Z. Z. Hou, "Multiobjective and multi-path optimization selection methods based on genetic algorithms," *Microelectronics & Computer*, vol. 23, no. 10, pp. 41–43, 2006.
- [19] C. H. D. Wu, L. Y. Yang, and K. Xu, "Application of neural network and genetic algorithm in dynamic route guidance," *Application Research of Computers*, vol. 23, no. 5, pp. 177–179, 2006.
- [20] H. Yang, "K-optimal chaos ant colony algorithm and its application on dynamic route guidance system," in *Computer Engineering and Networking*, pp. 227–234, Springer International, Berlin, Germany, 2013.
- [21] Y. Lee and S. Kim, "A hybrid search method of a\* and dijkstra algorithms to find minimal path lengths for navigation route planning," *Journal of the Institute of Electronics and Information Engineers*, vol. 51, no. 10, pp. 109–117, 2014.
- [22] D. Tian, Y. Yuan, J. Zhou, Y. Wang, G. Lu, and H. Xia, "Real-time vehicle route guidance based on connected vehicles," in *Proceedings of the IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing (GreenCom-iThings-CPSCOM '13)*, pp. 1512–1517, August 2013.
- [23] E. Paikari, L. Kattan, S. Tahmasseby, and B. H. Far, "Modeling and simulation of advisory speed and re-routing strategies in connected vehicles systems for crash risk and travel time reduction," in *Proceedings of the 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE '13)*, Saskatchewan, Canada, May 2013.
- [24] T. W. Chim, S. M. Yiu, L. C. Hui, and V. Li, "VSPN: VANET-based secure and privacy-preserving navigation," *IEEE Transactions on Computers*, vol. 63, no. 2, pp. 510–524, 2014.
- [25] J. D. Vreeswijk, R. L. Landman, E. C. Van Berkum, A. Hegyi, S. P. Hoogendoorn, and B. Van Arem, "Improving the road network performance with dynamic route guidance by considering

- the indifference band of road users,” *IET Intelligent Transport Systems*, vol. 9, no. 10, pp. 897–906, 2015.
- [26] W. Genders and S. N. Razavi, “Impact of connected vehicle on work zone network safety through dynamic route guidance,” *Journal of Computing in Civil Engineering*, vol. 30, no. 2, Article ID 04015020, 2016.
- [27] D. Zhao, C.-F. Shao, J.-L. Wang, J. Li, and B.-B. Wang, “Modelling combined mode choice behavior of commute trip chain under multi-modal guidance,” *Journal of Jilin University*, vol. 45, no. 6, pp. 1763–1770, 2015.
- [28] M. Yildirimoglu, M. Ramezani, and N. Geroliminis, “Equilibrium analysis and route guidance in large-scale networks with MFD dynamics,” *Transportation Research Part C: Emerging Technologies*, vol. 59, pp. 404–420, 2015.
- [29] M. Heydar, J. Yu, Y. Liu, and M. E. Petering, “Strategic evacuation planning with pedestrian guidance and bus routing: a mixed integer programming model and heuristic solution,” *Journal of Advanced Transportation*, vol. 50, no. 7, pp. 1314–1335, 2016.
- [30] B. Y. Chen, W. H. Lam, and Q. Li, “Efficient solution algorithm for finding spatially dependent reliable shortest path in road networks,” *Journal of Advanced Transportation*, vol. 50, no. 7, pp. 1413–1431, 2016.
- [31] Y.-F. Xu, H.-Y. Yu, B. Su, and H.-L. Zhang, “Traffic flow distribution models based on time and path preference and inducement strategy,” *System Engineering Theory and Practice*, vol. 32, no. 10, pp. 2306–2314, 2012.
- [32] L. Wu, *Modeling and optimization of dynamic route guidance system under vehicular ad-hoc networks [dissertation]*, Shandong University, Jinan, China, 2014.
- [33] Y. Z. Fan, F. C. Jiang, R. Mao En, and G. H. Wang, “Analysis and design of distributed dynamic route guidance system,” *Computer Engineering and Design*, vol. 15, pp. 3737–3739, 2007.
- [34] Q. Y. Xie, *Study of vehicle dynamic route guidance under intellidrive [M.S. thesis]*, South China University of Technology, Guangzhou, China, 2012.
- [35] Z. C. Li and H. J. Huang, “Determination of equilibrium market penetration under multi-criteria route guidance systems,” *Systems Engineering—Theory & Practice*, no. 9, pp. 125–130, 2004.
- [36] R. Ding, *Research on dynamic path selection based on VANET [thesis]*, Jilin University, Changchun, China, 2013.
- [37] L. D. Baskar, B. De Schutter, and H. Hellendoorn, “Optimal routing for automated highway systems,” *Transportation Research Part C: Emerging Technologies*, vol. 30, pp. 1–22, 2013.
- [38] J. Sun, Y. Yang, and K. Li, “Integrated coupling of road traffic and network simulation for realistic emulation of connected vehicle applications,” *Simulation*, vol. 92, no. 5, pp. 447–457, 2016.

## Research Article

# A Novel Trip Coverage Index for Transit Accessibility Assessment Using Mobile Phone Data

**Zhengyi Cai, Dianhai Wang, and Xiqun (Michael) Chen**

*College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China*

Correspondence should be addressed to Xiqun (Michael) Chen; [chenxiqun@zju.edu.cn](mailto:chenxiqun@zju.edu.cn)

Received 30 November 2016; Accepted 18 January 2017; Published 13 February 2017

Academic Editor: Xiaoyue Liu

Copyright © 2017 Zhengyi Cai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Transit accessibility is an important measure on the service performance of transit systems. To assess whether the public transit service is well accessible for trips of specific origins, destinations, and origin-destination (OD) pairs, a novel measure, the Trip Coverage Index (TCI), is proposed in this paper. TCI considers both the transit trip coverage and spatial distribution of individual travel demands. Massive trips between cellular base stations are estimated by using over four-million mobile phone users. An easy-to-implement method is also developed to extract the transit information and driving routes for millions of requests. Then the trip coverage of each OD pair is calculated. For demonstrative purposes, TCI is applied to the transit network of Hangzhou, China. The results show that TCI represents the better transit trip coverage and provides a more powerful assessment tool of transit quality of service. Since the calculation is based on trips of all modes, but not only the transit trips, TCI offers an overall accessibility for the transit system performance. It enables decision makers to assess transit accessibility in a finer-grained manner on the individual trip level and can be well transformed to measure transit services of other cities.

## 1. Introduction

Public transportation plays an important role in solving traffic problems in urban cities. It is well recognized among transportation planners that transit accessibility is an important measure of the service performance. The Transit Capacity and Quality of Service Manual summarized spatial, temporal, information, and capacity availability factors of public transit systems [1]. A major concern in the public transit sector has been the adequate assessment of access to transit services. Measures of transit accessibility are important in assessing existing transit services, allocating investments, and making decisions on the land development [2]. Transit accessibility has been one of the key indicators of transit planning, performance evaluation, and quantification of the level of service. Transit system planners design the layout of transit lines and stops to improve accessibility and enhance the transit attractiveness. One of the recent research concerns is the extent to which public transit systems enable the less privileged population of privately nonmotorized travelers to access the systems more conveniently, efficiently, and comfortably [3]. To achieve this objective, the first question

we need to answer is: *How to assess accessibility of public transit for trips in a network with spatially and temporally nonuniform travel demands?*

As shown in Table 1, modeling public transit accessibility has attracted numerous research efforts in the past decades, primarily including the evaluation of the spatial coverage [4–6], temporal coverage [7–10], and trip coverage [11–13]. Since passengers' departure time varied according to their own need, Owen and Levinson calculated continuously in time for the evaluation of transit systems rather than at a single of a few departure times [14]. On the other hand, more and more research combined the spatial coverage and temporal coverage to assess transit accessibility. For instance, Mavoa et al. combined public transit, walking accessibility index, and transit frequency to calculate accessibility at the parcel level [15]. Mamun et al. combined the spatial coverage, temporal coverage, and trip coverage to measure public transit performance [16]. El-Geneidy et al. incorporated both travel time and transit fares to determine whether people residing in socially disadvantaged neighborhoods [17].

Based on the literature review it is evident that (1) most assessment models belong to the physical location-based



TABLE 1: Summary of transit accessibility measures.

Category	Measure description	Application	Reference
Physical access to transit	Proximity to transit stops in time or distance	Measuring accessibility for local transit operators in London	Hillman and Pool [5]
	Quarter-mile buffers around transit routes	Transit coverage in the Queen Anne Community of Seattle	Nyerges [6]
	Pedestrian average and maximum walking distance to transit stops	Three neighborhood plans for a 23.3 ha site	Aultman-Hall et al. [4]
Accessibility to destination	Travel-impedance measurements (e.g., travel distance, time, or cost)	Mass/light rapid transit systems in Singapore	Liu and Zhu [11]
	Public transport relative accessibility percentage (transit catchment area and population by transit within 60 min)	90 sites in the south east of England	Gent and Symonds [12]
	Transit accessibility index (TAI) and transit dependence index (TDI)	TransCAD-based transit accessibility measure (TAM) software tool	Bhat et al. [13]
Temporal accessibility	Span and headway of transit service and time-of-day distribution of travel demand	Numerical illustration	Polzin et al. [7]
	Space-time accessibility measures with opportunities and human activity-travel behavior	Commercial and industrial land parcels of Portland Metropolitan Region, Oregon	Kim and Kwan [8]
	Dynamic activity opportunities that can be reached within a prespecified time limit with known transit schedules	Southern California Association of Governments megaregion	Lei et al. [9]
	Rate of access poverty among population	Regional transportation plan scenarios from the San Francisco Bay Area	Golub and Martens [10]

proximity analysis, while few studies take into account the transit coverage of individual travels in a real-world large-scale network; (2) most of the existing transit accessibility measures account for the spatial and temporal coverage, while very few studies consider the trip coverage; (3) the accessibility metrics produced by most existing tools are therefore static in the sense that they describe the transit system but consider less the temporal variability of individual travel demands. To assess whether the public transit service is well accessible for trips of specific origins, destinations, and origin-destination (OD) pairs, a public transit accessibility measure coping with the trip coverage is needed to provide a more reasonable assessment of transit quality of service.

There are some reasons for the gap of the previous studies. In the past, multiple sources of data required to evaluate transit accessibility considering individual travel demands are difficult to collect and consequently extensive efforts are required in order to obtain the useful data. In particular,

it is difficult to measure real-world travel demands due to the small amount of household survey data in the past. In addition, many surveys are zone based and unable to describe individual travel behavior. In some cases, public transit operational data (e.g., stops, routes, schedules, frequencies, and hours of operation) may be hard to access and data fusion could become uneasy due to their inconsistent formats in the time scale and data particle size [18].

Fortunately, the recent advent of data collection technologies, for example, mobile phone signaling data and automated vehicle location, has shifted a data-poor environment to a data-rich environment and offered opportunities to conduct comprehensive transit system performance evaluation. For example, cell phone signaling data have emerged to be a widely used resource to measure both individual travel behavior and network demand, for example, individual human mobility patterns [19, 20], estimation of OD matrices [21], and OD trip purposes [22]. On the other hand, more

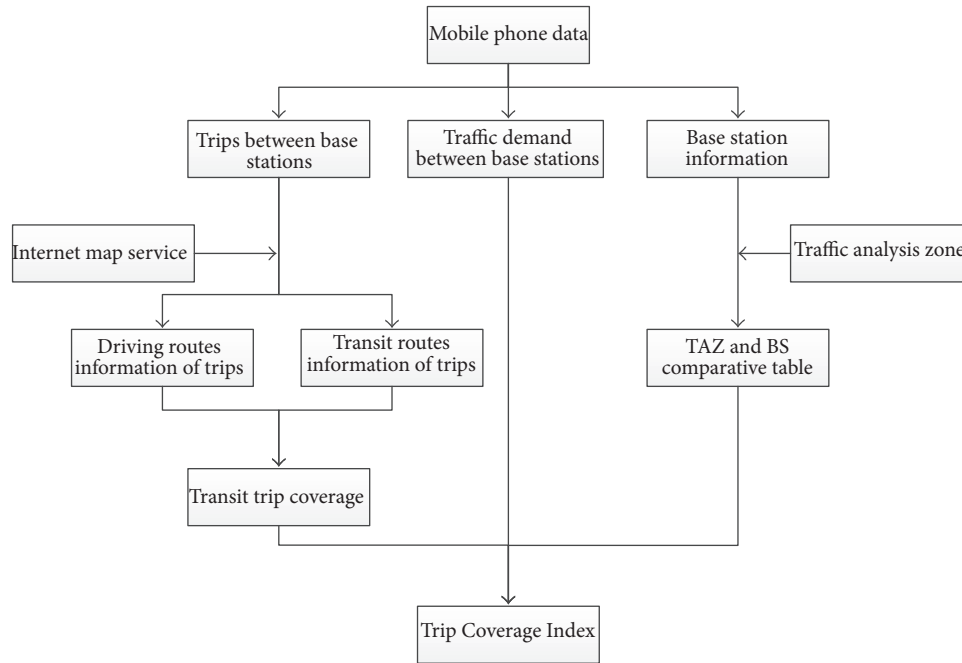


FIGURE 1: Estimation framework of TCI.

and more web map service data become readily available for public use, for example, Google Maps APIs [23], Baidu Map API [24], and AMAP Open Platform [25], which can provide massive on-demand transit trip planning services in real time. Ma and Wang developed a data-driven platform for online transit performance monitoring using automated fare collection and automated vehicle location [26]. Ma et al. developed a series of data mining methods to identify the spatiotemporal commuting patterns of public transit riders using one-month transit smart card data [27]. Therefore, new accessibility indicators taking into account individual trips will definitely provide a more powerful tool.

This paper is aimed at presenting a new public transit quality of service measure, the Trip Coverage Index (TCI), which takes into account both the trip coverage for transit systems and the spatial distribution of heterogeneous and dynamic individual travel demands. The TCI provides a quantitative measure of transit accessibility on the basis of massive trips collected from mobile phone data. The transit accessibility information is extracted from the Baidu Map with the Python code implementation, for example, the access to transit facilities, transit routes (shortest in time/length and alternatives), transit on-vehicle time, and OD connectivity. The novel measure of transit service performance fills the research gap that the conventional spatial coverage index does not consider the coverage to individual trips or the percentage of travel demands that can be served by the transit systems.

The rest of the paper is organized as follows: Section 2 presents the methodological approach to the trip coverage analysis, which is different from the conventional spatial coverage of transit services. In Section 3, an illustrative and

tractable numerical example is employed to present how to calculate TCI and compare it with the conventional measure of spatial coverage. Section 4 shows the field data utilized in this paper and presents results of a real-world city-wide case study that applies TCI to the transit network of Hangzhou, China. Finally, Section 5 concludes the paper and outlooks the future research.

## 2. Methodology

In this part, we first propose a new method to acquire the transit route information for millions of trips determined from the mobile phone data automatically based on online map and programming; then a new public transit quality of service measure (TCI) is proposed considering the access to transit facilities, transit routes information, driving routes information, and OD connectivity. The development of the proposed TCI requires several steps and the framework is shown in Figure 1. The first step is to acquire travel flows between cellular base stations using mobile phone data. Second, the information of transit services and driving routes between each OD pair is then extracted by accessing the online map service for millions of times. Third, the transit trip coverage from base station  $m$  to base station  $n$  is estimated based on the data retrieved from the online map. Fourth, TCI from Zone  $i$  to Zone  $j$  is estimated using the transit trip coverage and travel demand between the base stations. Each of the key procedures of the transit accessibility assessment will be presented in the following sections.

**2.1. Trip Estimation.** In this section, we introduce the mobile phone data and present the methods used to determine trips from the mobile phone data.

**2.1.1. Mobile Phone Signaling Data.** The dataset used in this study consists of two tables in the database: one is the base station table and the other is the anonymous table of mobile phone records. The mobile phone record is generated when a device connects to the cellular network in any of the following cases:

- (i) when the phone makes or receives a call;
- (ii) when the phone sends or receives a message;
- (iii) when the phone is switched on or off;
- (iv) when the user moves from one base station to another; or
- (v) when the system sends the periodic location update request on the phone, for example, 2 h.

The mobile phone signaling data contain Call Details Records (CDR), which were previously utilized to estimate OD demands in numerous related studies [19–22]. Each record of the mobile phone signaling data contains an anonymous user ID, base station ID, and timestamp at the instance of the phone communication with the base station. The base station table contains the base station ID, longitude, and latitude. There are more than 52 thousand of base stations in the urban area of Hangzhou, China. The average covering radius of each base station is less than 100 m.

**2.1.2. Determining Trips.** In order to infer trips from the mobile phone signaling data, the first step is to filter out noise resulting from one base station to another. The call balancing is conducted by the mobile service provider, which creates the appearance of false movements, and distinguishes users' stay locations. Once the stay locations are determined, we evaluate the trips as paths between a user's consecutive locations. To achieve this, we estimate the trips by employing the method of using mobile phone traces data [20]. The estimation is carried out as follows:

- (i) Each mobile phone signaling record  $R_i(k)$  is characterized by a position  $p_i(k)$  expressed by latitude, longitude, and a timestamp  $t_i(k)$  for the  $k$ th observation of a given anonymous user  $i$ .
- (ii) Then the signaling records are connected into a sequence of records  $\{R_i(1), R_i(2), \dots, R_i(n)\}$  according to their time series.
- (iii) If the signaling record series  $\{R_i(q), R_i(q+1), \dots, R_i(z)\}$  satisfy the criteria, (I)  $t_i(z) - t_i(q) > \Delta T$ ; (II)  $\max\{p_i(j), p_i(k)\} < \Delta S, \forall q \leq j, k \leq z$ , which mean a user should stay in an area with the radius  $\Delta S$  (set as 200 m) over a certain time interval  $\Delta T$  (set as 30 min). Then the points  $\{R_i(q), R_i(q+1), \dots, R_i(z)\}$  are fused together by selecting the point with the longest stay time as the stay location.
- (iv) We evaluate paths between a user's stay locations at consecutive points, and the stay locations are assumed to be trip origins or destinations.

## 2.2. Extracting Transit Routes from an Online Map

**2.2.1. Online Map Service.** Calculating the trip coverage indicators requires a database with transit data such as the transit network, road network, operational transit information, and bus stops. Based on those data, we know how many transit lines serve the trips from base station  $m$  to  $n$  and how large the distance is from base station  $m$  to  $n$  by transit line  $l$  and other associated information. Some literatures collected data based on Google Transit or GTFS (General Transit Feed Specification), a supplemental service to Google maps [3, 9, 18]. In those studies, the public transit network in a GIS format and the road network data were required, however, which were difficult to acquire as the data might be from different sources and difficult to use since the data must share the same coordinate, scale, context, and so forth [18].

More and more online map services provide path navigation in China, for example, Baidu Map and AMAP. If the user selects the transportation mode, enters the origin and destination, and chooses a departure time on the map website, it will return the route planning information including the trip distance, trip time, and suggested routes from the origin to destination. Some online map services provide open resources to developers, which are mostly in the form of the Application Programming Interface (API). The API is a set of predefined user applications and the operating system's function, by means of which programmers can easily achieve the underlying operating system feature development or packages. Launched in April, 2010, Baidu Map API [24] not only includes the basic interface to build maps, but also provides information such as local search, route planning, and other data services, through which we can acquire the route information of trips. However, since it needs to search the transit route information for millions of trips completed by over 4 million mobile phone users for a week in Hangzhou, it is impossible to manually acquire such huge information.

As shown in Figure 2, we propose a new method to acquire the transit route information for millions of trips automatically. The trip database stores millions of trips on a local computer server (each trip includes origin/destination geolocations and departure and arrival time). The Python code extracts one piece of trip data and makes an API request to the Baidu Map server via HTTP for the transit route data and the server will return data in the form of XML or JSON after it queries the back-end database. After that we call JSON parsing functions [28] and store the result to the database.

**2.2.2. Transit Routes Data.** The response of the transit route information from the Baidu Map API contains fruitful information and we just extract the useful information for assessing transit accessibility, for example, the taxi route information and bus route information. An example of response is shown in Table 2. The status code indicates whether the online service returns valid results, 0 means a correct record, and 1 means invalid information. The taxi route information includes the taxi distance of the trip, taxi travel time, and monetary costs. The bus route information is much more complicated. There may be several suggested bus schemes per trip and several segments per bus scheme.

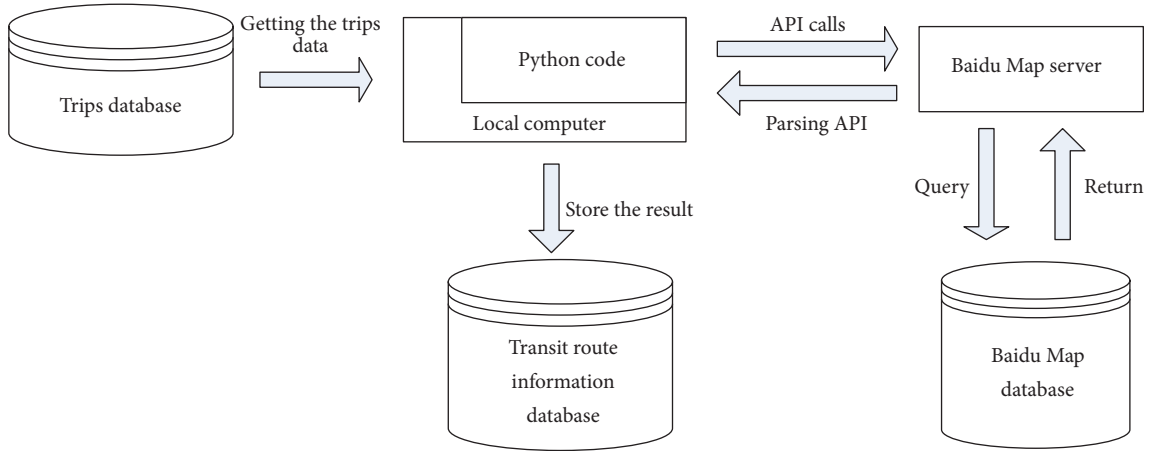


FIGURE 2: Web scraping of transit information.

TABLE 2: A sample of the transit route information from the Baidu Map API.

Status	0	0: correct record 1: error record
Taxi	Taxi_distance (m)	4,540
	Taxi_travel time (s)	503
	Total distance (m)	4,584
	Total travel time (s)	2,276
Bus scheme 1	Scheme_type	1
	Segment1_type	5
	Segment1_distance (m)	311
	Segment2_type	3
	Segment2_distance (m)	3523
	Segment3_type	5
Bus scheme 2	Segment3_distance (m)	750
	...	...
		0: correct record 1: error record

In addition, the segment\_type code means the travel model (e.g., 5 for walking, 3 for bus).

Generally, there are three segments per scheme in a direct transit route without transfer, which means that the trip distance  $d_{m,n,l}$  from origin base station  $m$  to destination base station  $n$  by taking transit line  $l$  consists of the access distance by walking, in-vehicle distance by bus, and egress distance by walking, given by

$$d_{m,n,l} = d_{m,n,l}^a + d_{in-vehicle} + d_{m,n,l}^e \quad (1)$$

2.3. Trip Coverage Index. The conventional evaluation criterion for the transit service includes the transit spatial

coverage area, which is usually estimated using the buffer area covered by the transit route or by the area within a walking distance threshold of a transit stop or transit route [1]. The walking distance threshold is modified for various features, for example, the percent elderly in the population and street connectivity [29]. It is commonly accepted by transit planners and researchers that bus transit users are willing to walk up to 1/4 mile (400 m) to reach their nearest transit stop [30–33]. The government agencies and researchers of China use 500 m as the buffer radius to evaluate the transit serving area [34, 35]. The sensitivity of the walking distance threshold will be analyzed in Section 4. In the context of transit, a traveler may transfer from one bus route to another and continue to reach

his/her destination. According to Modesti and Sciomachen [36], more than two times of transfers in a transit trip are generally intolerable for transit users, such that two transfers can be chosen as the maximum value allowed per trip.

Based on the aforementioned idea, this paper presents the binary connectivity parameter ( $\gamma_{m,n,l}$ ) to indicate whether two regions are connected by transit services. Here we only consider walking to reach the transit station and other options such as bike, park, and ride have not been considered. For any trip from base station  $m$  to  $n$ , if (1) there exists a transit line  $l$  that connects the two regions; (2) both the access and egress distances are smaller than the preselected walking distance threshold; and (3) the transfer count  $k_{m,n,l}$  is less than the transfer tolerance threshold  $N$ , then the binary connectivity parameter  $\gamma_{m,n,l}$  is 1; and 0, otherwise. The binary connectivity parameter is given by

$$\gamma_{m,n,l} = \begin{cases} 1, & d_{m,n,l}^a \leq d_0, \quad d_{m,n,l}^e \leq d_0, \quad k_{m,n,l} \leq N, \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

$\forall l, m, n.$

The research concern is the extent of the efficiency and attractiveness of the public transport system compared to private cars. As is known to all, there are many factors that may influence the travel mode choice such as travel time, transit distance, transit fare car parking fare, and weather. However, this paper is not concerned with individual travel mode choice behavior but provides the trip-level assessment of transit accessibility, so we just take into account the travel time and travel distance. On the other hand, the developed index (i.e., TCI) is applied to assess accessibility of public transit for all trips, not only those currently or likely to be using transit and no matter whether he/she owns cars or not. Instead, they can rent a car or take a taxi to reach the destination for those have no access to personal car. Therefore, the trip coverage from base station  $m$  to  $n$  served by the transit line  $l$ , that is,  $TC_{m,n,l}$ , is measured by the ratio of the driving distance to the transit distance from base station  $m$  to  $n$  by transit line  $l$  and the ratio of the total travel time by driving to the total travel time by transit line  $l$ , given by

$$TC_{m,n,l} = \alpha \frac{d_{m,n}}{d_{m,n,l}} + (1 - \alpha) \frac{T_{m,n}}{T_{m,n,l}}, \quad 0 \leq \alpha \leq 1, \quad (3)$$

where the weighting factor  $\alpha$  can be determined according to the preference on travel distance or travel time by decision makers. The default value is 0.5.

Considering there may be several transit lines or no transit lines serving the trip, we select the maximum value between 0 and  $TC_{m,n,l}$  multiplied by the binary connectivity parameter  $\gamma_{m,n,l}$  as the trip coverage score  $TC_{m,n}$  for the OD pair in the network, given by

$$TC_{m,n} = \max \{ \gamma_{m,n,l} \cdot TC_{m,n,l}, 0 \}, \quad (4)$$

where the trip coverage score  $TC_{m,n}$  takes into account the OD pairwise transit accessibility which is rarely considered by previous studies.

There are some short trips, of which the distance is shorter than the walking distance threshold. In other words, it is unnecessary to take bus for this trip. So when calculating TCI, we only consider the trips with a distance longer than twice of the service access distance threshold, that is, 1,000 m.

The spatial relationship between TAZs and base stations (BSs) can be obtained by the ArcGIS spatial analysis toolbox. The TAZ-BS membership table can be obtained, which is the foundation of calculating TCI from TAZ  $i$  to  $j$ . TCI from TAZ  $i$  to  $j$  is defined as the weighted average trip coverage score ( $TC_{m,n}$ ) by the travel demand, given by

$$TCI_{i,j} = \frac{\sum_{P_{m,n} \in \{d_{m,n} > 2d_0\}} TC_{m,n} \cdot P_{m,n}}{\sum_{P_{m,n} \in \{d_{m,n} > 2d_0\}} P_{m,n}}. \quad (5)$$

TCI can be used to quantify the coverages of origin TAZ  $i$  and destination TAZ  $j$ , given by

$$TCI_i = \frac{\sum_j TCI_{i,j} \cdot P_{i,j}}{\sum_j P_{i,j}},$$

$$TCI_j = \frac{\sum_i TCI_{i,j} \cdot P_{i,j}}{\sum_i P_{i,j}}, \quad (6)$$

$$TCI = \frac{\sum_i \sum_j TCI_{i,j} \cdot P_{i,j}}{\sum_i \sum_j P_{i,j}}.$$

### 3. An Illustrative Numerical Example

This section provides a tractable numerical example to illustrate the application of TCI to the assessment of transit accessibility. As shown in Figure 3, we consider a road network of four zones served by three transit lines, and each line has four stops. The dashed circles represent 500-meter buffers around each transit stop.

Table 3 shows the travel demands between base stations and the trip coverage results for the trips. Columns 4 and 5 of Table 3 show which TAZs base stations  $m$  and  $n$  belong to, respectively. Column 6 provides the number of transit lines serving the OD pair. Columns 7–13 present the transit route information which can be obtained from the online map for a real-world network. The binary connectivity parameter for each transit line estimated by (2) is shown in Column 14.

The illustrative numerical example helps understand the difference between the proposed measure and the conventional spatial coverage measure. There are two lines and two bus stops serving BS1, while there is only one transit line and one bus stop serving BS2. At the same time, there are two bus lines serving trips of BS1–5 and only one bus line serving trips of BS2–5. It is reasonable to expect a higher level of transit coverage for BS1–5 than that of BS2–5. However, for trips of BS1–5,  $d_{m,n,l}^e$  of both line 1 and line 2 (see Column 11 in bold in Table 3) exceed the preselected distance threshold (500 m), and the binary connectivity parameters for both lines are zero, which means no buses can offer services to trips of BS1–5 given the stop buffer distance threshold.

Column 15 shows the trip coverage of different OD pairs, that is,  $TC_{m,n}$  estimated by (3)–(4). For BS1–6,  $TC_{1,6}$  is



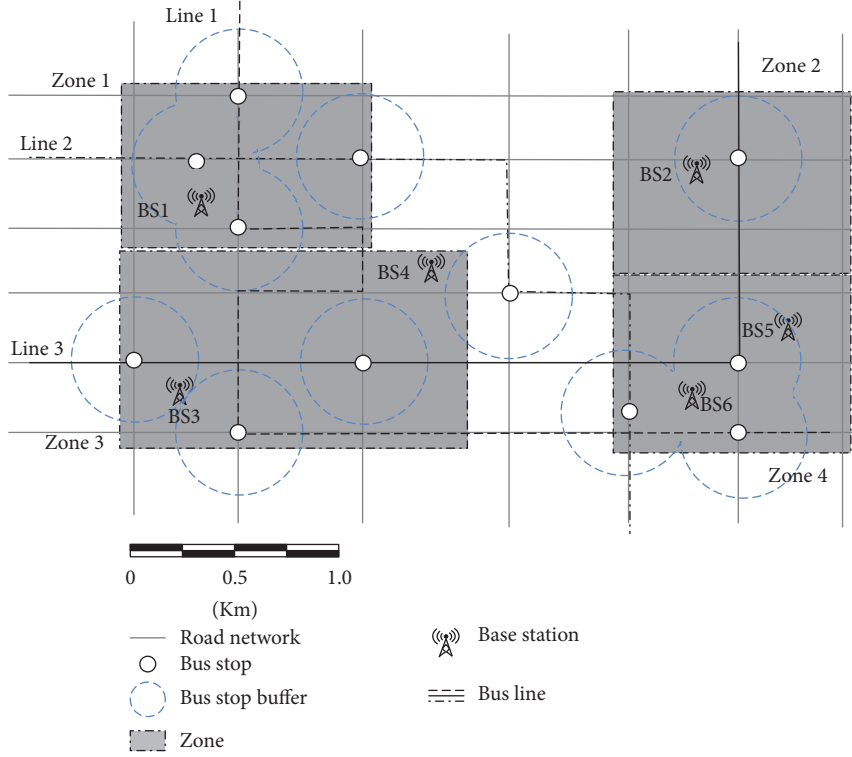


FIGURE 3: An illustrative road and transit networks.

TABLE 3: Trip coverage calculation.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$BS_m$	$BS_n$	$P_{m,n}$	Zone $i$	Zone $j$	Bus line	$d_{m,n}$	$T_{m,n}$	$d_{m,n,l}^a$	$T_{m,n,l}$	$d_{m,n,l}^e$	$d_{in.veh}$	$d_{m,n,l}$	$\gamma_{m,n,l}$	$TC_{m,n}$
1	2	0	1	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
<b>1</b>	<b>3</b>	4	1	3	1	<b>1400</b>	613	300	1225	400	2100	<b>2800</b>	1	<b>0.500</b>
<b>1*</b>	<b>5*</b>	<b>8*</b>	<b>1*</b>	<b>4*</b>	1/2*	3700	1245/1353	300/250	2125/2178	<b>700/1100</b>	4500/3300	5500/4650	0/0	0.000
<b>1*</b>	<b>6*</b>	<b>3*</b>	<b>1*</b>	<b>4*</b>	1/2*	3300	1566/1227	300/200	1875/1470	450/450	4500/3300	5250/3950	1	0.835
2	1	1	2	1	NA	2600	NA	NA	NA	NA	NA	NA	0	0.000
2	3	12	2	3	3	3400	1198	300	1550	300	3800	4400	1	0.773
2	5	8	2	4	3	<b>1400</b>	827	300	975	450	900	<b>1650</b>	1	<b>0.848</b>
3	1	7	3	1	1	1400	613	400	1225	300	2100	2800	1	0.500
3	2	15	3	2	3	3400	1198	300	1550	300	3800	4400	1	0.773
3	5	2	3	4	3	3200	1293	300	1475	450	2900	3650	1	0.877
<b>5*</b>	<b>1*</b>	<b>3*</b>	<b>4*</b>	<b>1*</b>	1/2*	3700	1873/1695	1100/700	2279/1800	250/300	4500/3300	5850/4300	0/0	0.000
<b>6*</b>	<b>1*</b>	<b>8*</b>	<b>4*</b>	<b>1*</b>	1/2*	3300	1580/1311	450/450	1893/1571	300/200	4500/3300	5250/3950	1	0.835
6	2	8	4	2	3	1600	945	450	975	300	900	1650	1	0.970
6	3	3	4	3	3	2600	1236	450	1475	300	2900	3650	1	0.712

Note: NA = not applicable.

calculated as follows and other OD pairs can be calculated in the same way:

$$\begin{aligned}
 TC_{1,6} &= \max \{ \gamma_{1,6,1} \cdot TC_{1,6,1}, \gamma_{1,6,2} \cdot TC_{1,6,2}, 0 \} \\
 &= \max \left\{ 1 \times \left( 0.5 \times \frac{3300}{5250} + 0.5 \times \frac{1566}{1875} \right), 1 \right. \\
 &\quad \left. \times \left( 0.5 \times \frac{3300}{3950} + 0.5 \times \frac{1227}{1470} \right), 0 \right\} = 0.835.
 \end{aligned} \tag{7}$$

Table 3 shows that driving distances of BS1–3 and BS2–5 have the same value of 1400 m, but the transit route distance of BS1–3 is longer than that of BS2–5, which means the transit route of BS1–3 makes a detour and the connectivity level is lower than that of BS2–5 (see Columns 7, 13, and 15 in bold in Table 3). This situation is embodied in the calculation of  $TC_{m,n}$ .

Finally, TCI from TAZ  $i$  to  $j$  should incorporate the trip coverage with the travel demand of the trip. TCI for TAZ 1 to

TABLE 4: TCI and travel demand (in brackets).

Origin	Destination				TCI <sub>i</sub>
	1	2	3	4	
1	NA	0.000 (0)	0.500 (4)	<b>0.228</b> (11)	0.300 (15)
2	0.000 (1)	NA	0.773 (12)	0.848 (8)	0.765 (21)
3	0.500 (7)	0.773 (15)	NA	0.877 (2)	0.702 (24)
4	<b>0.607</b> (11)	0.970 (8)	0.712 (3)	NA	0.753 (22)
TCI <sub>j</sub>	0.536 (19)	0.841 (23)	0.706 (19)	0.526 (21)	0.658 (82)

Note: NA = not applicable.

4 can be calculated according to (5) and the results are shown in Table 4.

$$\begin{aligned} TCI_{1,4} &= \frac{TC_{1,5} \cdot p_{1,5} + TC_{1,6} \cdot p_{1,6}}{p_{1,5} + p_{1,6}} \\ &= \frac{0 \times 8 + 0.835 \times 3}{8 + 3} = 0.228. \end{aligned} \quad (8)$$

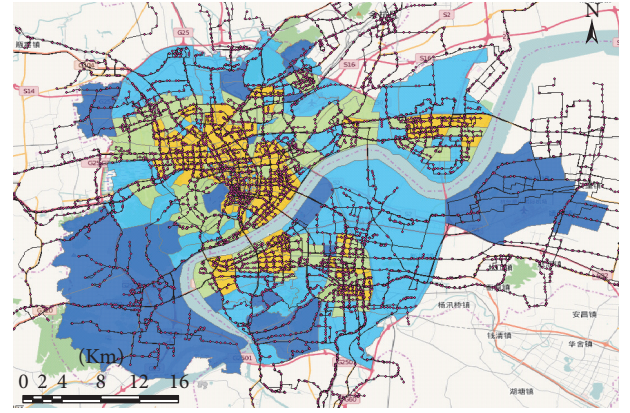
The TCI<sub>i</sub> for TAZ 1 as the origin can be calculated by

$$\begin{aligned} TCI_1 &= \frac{\sum_j TCI_{1,j} \cdot p_{1,j}}{\sum_j p_{1,j}} = \frac{0.5 \times 4 + 0.228 \times 11}{4 + 11} \\ &= 0.300. \end{aligned} \quad (9)$$

Similarly, other TCI<sub>i</sub> and TCI<sub>j</sub> can be obtained. Results of the trip coverage as well as the travel demand of each OD pair are shown in Table 4. It has been realized that the trips in the opposite direction have the same trip coverage scores as Table 3, for example, trips of BS1–6 and BS6–1 (see italic rows in Table 3) and trips of BS1–3 and BS3–1. This is because the bus lines are set two ways in this numerical example. However, the zone-to-zone TCIs in the opposite directions show different scores, for example, TCI<sub>1,4</sub> and TCI<sub>4,1</sub> highlighted in Table 4. Recalling the contents indicated by asterisk of Table 3, we find that the demands from Zone 1 to Zone 4 and from Zone 4 to Zone 1 are different in opposite directions, which means the transit system covering more travel demands has a higher value of TCI.

TCI also offers a way to quantify the transit service level of OD pairs that require a transfer between transit lines. Equation (3) can be improved by considering the transfer distance  $d_{transfer}$  and transfer travel time  $T_{transfer}$ , given by

$$\begin{aligned} TC_{m,n,l} &= \alpha \frac{d_{m,n}}{\sum_l d_{m,n,l} + d_{transfer}} \\ &+ (1 - \alpha) \frac{T_{m,n}}{\sum_l T_{m,n,l} + T_{transfer}}, \quad 0 \leq \alpha \leq 1. \end{aligned} \quad (10)$$



Coverage radius meter



FIGURE 4: Layout of transit and the average base station coverage radius in terms of TAZs in Hangzhou, China.

The spatial coverage is the proportion of the area served by transit stops, which can be calculated by the Transit Capacity and Quality of Service Manual [1]. This method uses a buffer (set as 500 m) around each stop to define the spatial coverage of bus services. Table 5 shows the zonal data of the spatial coverage calculations and the corresponding TCI results. The buffer area for each stop is calculated using the ArcGIS toolbox and the overlapped buffers are calculated only once. The results show that the spatial coverage of Zone 1 is much higher than that of Zone 2, while the TCI of Zone 1 is lower than that of Zone 2, which are highlighted in Table 5.

## 4. Case Study

4.1. Study Area and Data. In this section, TCI is applied to a case study in Hangzhou, China, to assess the transit

TABLE 5: Zonal data of spatial coverage and TCI.

Zone, $i$	Zone area (km <sup>2</sup> )	Bus stop buffer (km <sup>2</sup> )	Spatial coverage	TCI <sub><math>i</math></sub>	TCI <sub><math>j</math></sub>
1	0.880	0.680	<b>0.773</b>	<b>0.300</b>	<b>0.536</b>
2	1.026	0.283	<b>0.275</b>	<b>0.765</b>	<b>0.841</b>
3	1.650	0.807	0.489	0.702	0.706
4	1.026	0.503	0.490	0.753	0.526

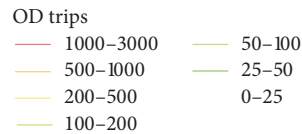
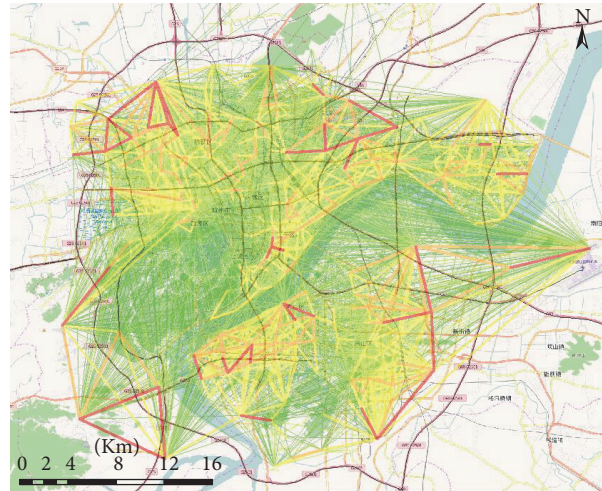


FIGURE 5: Desire lines of OD trips during the AM peak hours.

accessibility. Hangzhou is the capital and most populous city of Zhejiang Province, China. As shown in Figure 4, the study area contains the Shangcheng District, Xiacheng District, Jianggan District, Binjiang District, Xihu District, Gongshu District, and part of Xiaoshan District. The study area is 955 km<sup>2</sup>, and it contains 540 TAZs with 4.43 million residents. As shown in Figure 4, there are 912 transit lines and 18,508 transit stops in the transit network of Hangzhou.

The mobile phone signaling data used in this study consist of two tables, that is, the base station table and the anonymous mobile phone records table collected from 4.17 million mobile phone users in Hangzhou over one month between August and September, 2015. The position accuracy of a trip is determined by the coverage radius of base stations. There are 41,823 base stations in the 540 TAZs, and the average BS coverage radius for each TAZ is shown in Figure 4. Results show that the average coverage radius of 90% base stations is less than 100 m. The remaining 10% are distributed in less populated areas such as the mountainous and wetland areas shown in Figure 4.

The study time periods are AM peak hours (7:00–9:00) and PM peak hours (17:00–19:00). After processing the mobile phone signaling data using the method proposed in

Section 2.1, we obtain 2,816,910 trips in AM peak hours and 2,756,187 trips in PM peak hours on September 8, 2015, a regular working day. The desire lines of trips are shown in Figure 5. The average trip distance is 5.68 km and more than 50% of the trips are less than 3 km.

We also obtain spatial and temporal distributions of the population density using the trip information, for example, origin, destination, and timestamp. As shown in Figure 6, the population distribution is dispersed before 7:00 and after 20:00 and is aggregating during working hours. These observations are consistent with the daily experience.

**4.2. Results.** Combining both the mobile phone data and transit information extracted from the online map service, we are able to calculate the TCI for different time of day. The analytical results are as follows: the distribution of TCI <sub>$i$</sub>  in the AM peak hours and PM peak hours are shown in Figures 7(a)-7(b). We can see a totally different picture of TCI <sub>$i$</sub>  in the AM peak hours as compared to that in PM peak hours, which indicates that some of the TCI <sub>$i$</sub>  vary during different periods while the transit routes and departure interval are the same, which is similar to the findings of [37]. Based on those pictures, we should have different principles and



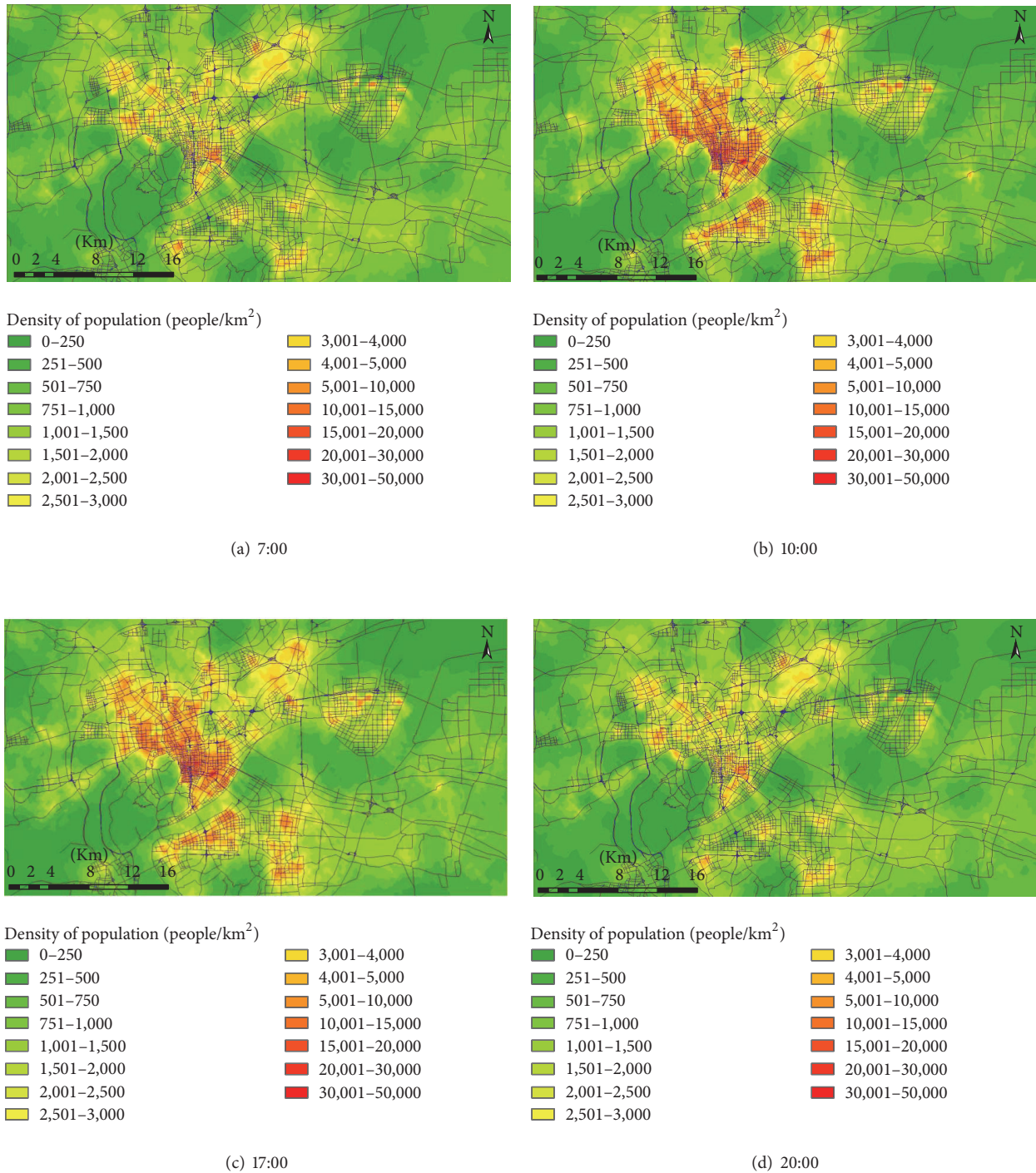
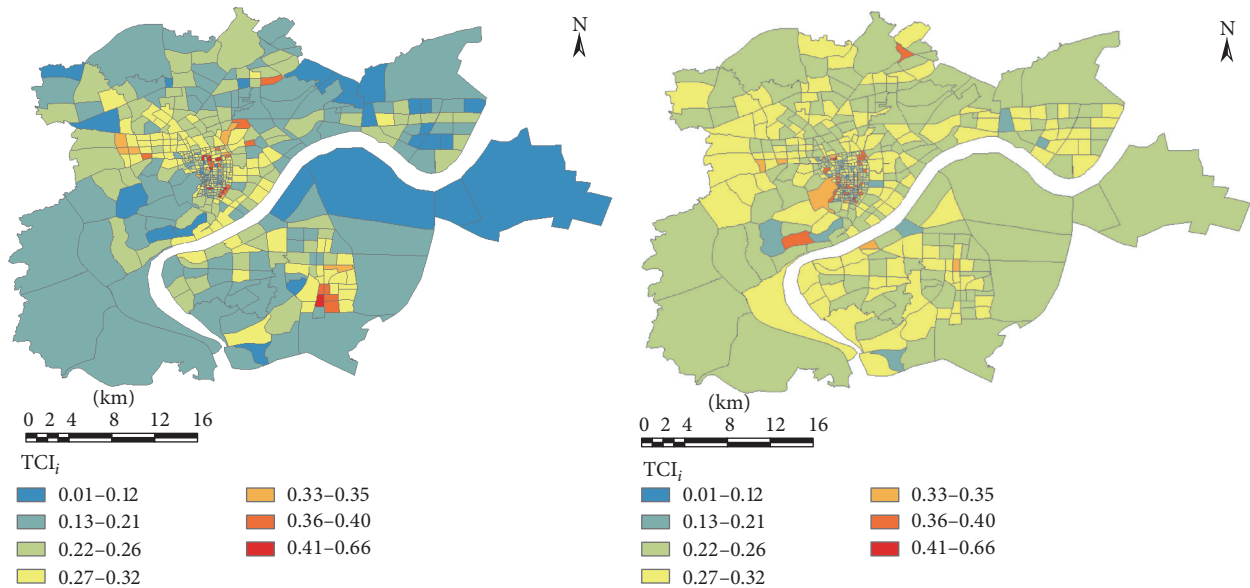


FIGURE 6: Spatiotemporal distributions of population density using mobile phone signaling data (September 8, 2015).

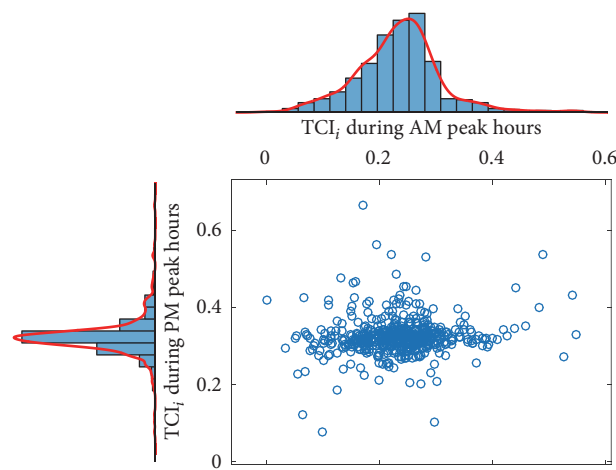
strategies in terms of deploying our bus-related resources and services. Most of the  $TCI_i$  in the AM peak hours are higher than those in the PM hours according to Figure 7(c) and the outliers show that  $TCI_i$  of some TAZs significantly vary with time periods and different travel demands. Transit operators should reschedule transit routes in a dynamic manner to be more consistent with travel demands.

As shown in Figure 8, results are compared with the spatial transit coverage during the AM and PM peak hours, respectively.  $TCI_i$  of TAZs is a skewed distribution, and most of the  $TCI_i$  is lower than 0.5, which means the transit system provides a poor coverage for the travel demand, while the spatial transit coverage calculated by the buffer method shows higher scores in both peak hours, which means most of



(a) Distribution of  $TCI_i$  of the TAZs in the AM peak hours

(b) Distribution of  $TCI_i$  of the TAZs in the PM peak hours



(c) Comparison of  $TCI_i$  of TAZs between AM and PM peaks

FIGURE 7: Distributions of TCI of TAZs during peak hours ( $d = 1000$  m,  $\alpha = 0.5$ ).

the population can access to the bus stops in the walking threshold. The comparison between  $TCI_i$  and the spatial transit coverage shows the bus stops which can be accessed in the walking threshold may not lead travelers to their destinations by transit services.

In order to further explore the sensitivity of the walking distance threshold, acceptable times of transfer, and the weighting factor  $\alpha$ , we summarize the statistics of TCI in terms of these factors in Figure 9. As shown in Figure 9(a), as the walking distance threshold rises from 500 m to 900 m, the TCI of whole network rises significantly; then it slows down after the walking distance threshold is longer than 900 m,

which means that transit operators could gain more trip coverage by improving bus services for those travel demands that the walking distance is under 900 m.

As the acceptable transfer times increase from 0 to 2, the TCI increases both during the AM peak hours and PM peak hours, which is comparable with experience. The results suggest that increasing the transit route crossings would provide a better transit service.

We also explore the interaction between the weighting factor  $\alpha$  and TCI. The larger value of  $\alpha$  is, the more weight of travel distance is considered in the TCI. Figure 9(c) shows that TCI increases with  $\alpha$  in all transfer scenarios, which



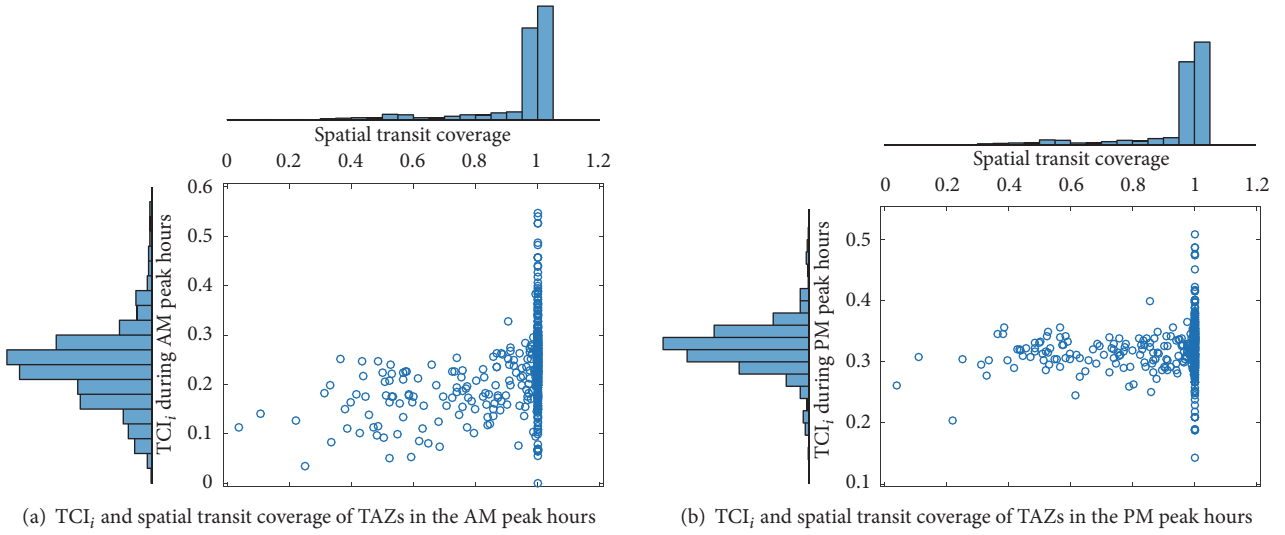


FIGURE 8: Comparison of  $TCI_i$  with the spatial transit coverage during AM and PM peak hours ( $d = 1000$  m,  $\alpha = 0.5$ ).

means TCI is sensitive to ratio of the bus travel time and the driving time. So we should promote the travel time reliability of buses to provide a better transit service.

## 5. Conclusions

In this paper, the novel TCI is proposed for measuring transit connectivity and accessibility. It is built on the existing transit service measures and allows us to analyze the transit connectivity and accessibility for massive trips between the origin and destination, as well as the transit coverage from or to a TAZ. This paper is among the first attempts considering the connectivity of trips from point to point and real-world complicated travel demand in a large-scale urban area. The TCI developed in this paper provides the capability to quantify the level of accessibility of the transit system and vary the assessment of transit accessibility with the temporal and spatial change of travel demands.

This paper also presents an easy-to-implement method to acquire the transit route information for millions of trips based on the online map. Since the data is acquired automatically using computer programming, it is possible to easily construct the data repository and analyze large public transit networks.

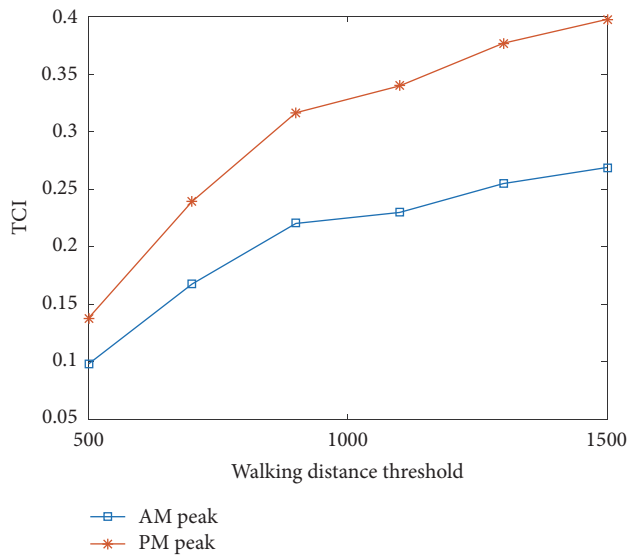
TCI can be applied to all trips, not only those currently or likely to be using transit, such that TCI is demonstrated as an overall measure of transit accessibility and can be used to measure how the transit system reaches its target, which is to provide services for more potential users.

Through the case study of Hangzhou, we find that fluctuations in the travel demand in different time periods make TCI distributing diversely across the city, which means transit operators should reschedule transit routes in a dynamic way to be consistent with travel demands. The sensitivity analysis is performed to determine how the walking distance threshold, times of transfer, and the weighting factor would impact

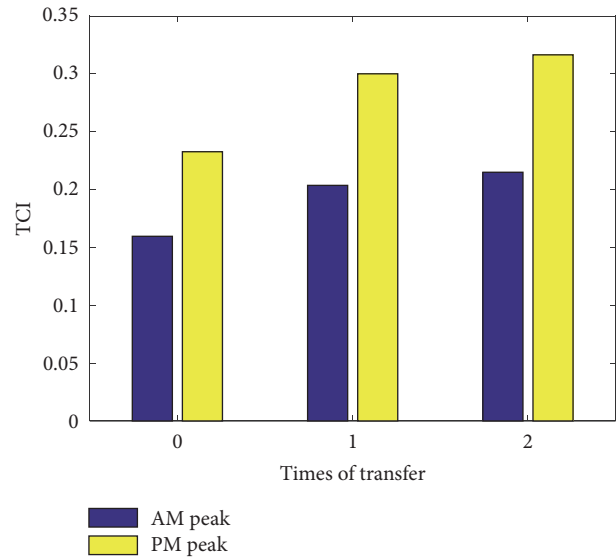
the network-wide TCI. The results can provide operators targeted measures to improve transit services.

## Notations

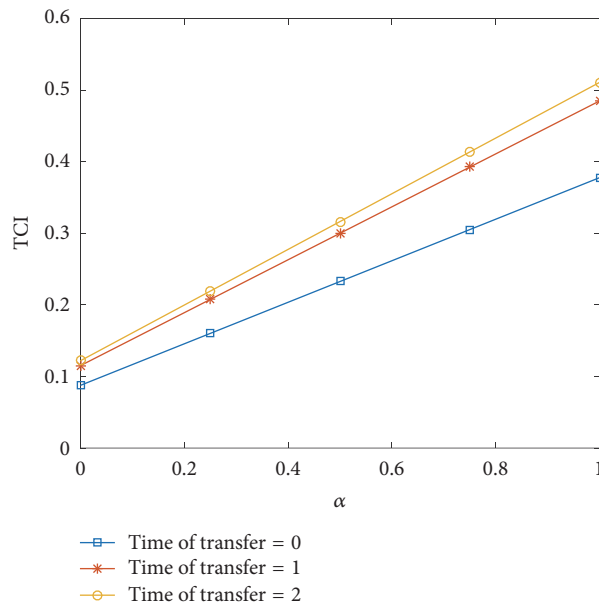
$d_0$ :	Service access distance threshold
$d_{m,n}$ :	Distance from base station $m$ to $n$ by car
$T_{m,n}$ :	Travel time from base station $m$ to $n$ by car
$d_{m,n,l}$ :	Total distance from base station $m$ to $n$ by transit $l$
$k_{m,n,l}$ :	Transfer count from base station $m$ to $n$ by transit $l$
$T_{m,n,l}$ :	Total travel time from base station $m$ to $n$ by transit $l$
$d_{m,n,l}^a$ :	Access distance from base station $m$ to $n$ by transit $l$
$d_{m,n,l}^e$ :	Egress distance from base station $m$ to $n$ by transit $l$
$i$ :	Origin traffic analysis zone (TAZ)
$j$ :	Destination TAZ
$l$ :	Transit line
$m$ :	Origin base station
$n$ :	Destination base station
$p_{m,n}$ :	Travel demand from base station $m$ to $n$
$p_{i,j}$ :	Travel demand from TAZ $i$ to $j$
$TC_{m,n,l}$ :	Trip coverage from base station $m$ to $n$ by transit $l$
$TC_{m,n}$ :	Trip coverage from base station $m$ to $n$
$TCI_{i,j}$ :	Trip coverage index from TAZ $i$ to $j$
$TCI_i$ :	Trip coverage index of origin TAZ $i$
$TCI_j$ :	Trip coverage index of destination TAZ $j$
TCI:	Trip coverage index of a transit network
$\gamma_{m,n,l}$ :	Binary connectivity parameter, 1 if a transit line $l$ connects the trip from $m$ to $n$ with $d_{m,n,l}^a$ and $d_{m,n,l}^e$ smaller than $d_0$ ; 0 otherwise.



(a) Transit network TCI with respect to the walking distance threshold ( $\alpha = 0.5$ )



(b) Transit network TCI with respect to transfers ( $d = 1000$  m,  $\alpha = 0.5$ )



(c) Transit network TCI with respect to  $\alpha$  ( $d = 1000$  m)

FIGURE 9: Sensitivity analysis for transit network.

### Competing Interests

The authors declare that they have no competing interests.

### Acknowledgments

This research is financially supported by Zhejiang Provincial Natural Science Foundation of China under Grant no. LR17E080002, National Natural Science Foundation of China under Grant nos. 51508505, 51338008, and 51278454, and Hangzhou Municipal Science and Technology Commission under Grant no. 20142013A57. Mr. Yanlei Cui helped in

processing some of the data used in this paper, and his assistance is gratefully acknowledged.

### References

- [1] KFH Group, *Transit Capacity and Quality of Service Manual*, Report 100, Transit Cooperative Research Program, Washington, DC, USA, 2nd edition, 2003.
- [2] S. Al Mamun and N. E. Lowmes, "Measuring service gaps: accessibility-based transit need index," *Transportation Research Record*, vol. 2217, pp. 153–161, 2011.

- [3] K. Fransen, T. Neutens, S. Farber, P. De Maeyer, G. Deruyter, and F. Witlox, "Identifying public transport gaps using time-dependent accessibility levels," *Journal of Transport Geography*, vol. 48, pp. 176–187, 2015.
- [4] L. Aultman-Hall, M. Roorda, and B. W. Baetz, "Using GIS for evaluation of neighborhood pedestrian accessibility," *Journal of Urban Planning and Development*, vol. 123, no. 1, pp. 10–17, 1997.
- [5] R. Hillman and G. Pool, "GIS-based innovations for modelling public transport accessibility," *Traffic Engineering and Control*, vol. 38, no. 10, pp. 554–559, 1997.
- [6] T. Nyerges, "Geographic information system support for urban/regional transportation analysis," in *Geography of Urban Transportation*, S. Hanson, Ed., pp. 240–265, The Guilford Press, New York, NY, USA, 1995.
- [7] S. E. Polzin, R. M. Pendyala, and S. Navari, "Development of time-of-day-based transit accessibility analysis tool," *Transportation Research Record*, no. 1799, pp. 35–41, 2002.
- [8] H.-M. Kim and M.-P. Kwan, "Space-time accessibility measures: a geocomputational algorithm with a focus on the feasible opportunity set and possible activity duration," *Journal of Geographical Systems*, vol. 5, no. 1, pp. 71–91, 2003.
- [9] T. Lei, Y. Chen, and K. Goulias, "Opportunity-based dynamic transit accessibility in Southern California: measurement, findings, and comparison with automobile accessibility," *Transportation Research Record*, no. 2276, pp. 26–37, 2012.
- [10] A. Golub and K. Martens, "Using principles of justice to assess the modal equity of regional transportation plans," *Journal of Transport Geography*, vol. 41, pp. 10–20, 2014.
- [11] S. Liu and X. Zhu, "Accessibility analyst: an integrated GIS tool for accessibility analysis in urban transportation planning," *Environment and Planning B: Planning and Design*, vol. 31, no. 1, pp. 105–124, 2004.
- [12] C. Gent and G. Symonds, "Advances in public transport accessibility assessments for development control—a proposed methodology," in *Planning and Transport, Research and Computation (PTRC) Annual Transport Practitioners' Meeting*, 2005.
- [13] C. R. Bhat, S. Bricka, J. La Mondia, A. Kapur, J. Y. Guo, and S. Sen, "Metropolitan area transit accessibility analysis tool," Tech. Rep. 0-5178-P3, Texas Department of Transportation, 2006.
- [14] A. Owen and D. M. Levinson, "Modeling the commute mode share of transit using continuous accessibility to jobs," *Transportation Research Part A: Policy and Practice*, vol. 74, pp. 110–122, 2015.
- [15] S. Mavoia, K. Witten, T. McCreanor, and D. O'Sullivan, "GIS based destination accessibility via public transit and walking in Auckland, New Zealand," *Journal of Transport Geography*, vol. 20, no. 1, pp. 15–22, 2012.
- [16] S. A. Mamun, N. E. Lownes, J. P. Osleeb, and K. Bertolaccini, "A method to define public transit opportunity space," *Journal of Transport Geography*, vol. 28, pp. 144–154, 2013.
- [17] A. El-Geneidy, D. Levinson, E. Diab et al., "The cost of equity: assessing transit accessibility and social disparity using total travel cost," in *Proceedings of the Transportation Research Board 95th Annual Meeting*, 16-3715, Washington, DC, USA, 2016.
- [18] Y. Hadas, "Assessing public transport systems connectivity based on Google Transit data," *Journal of Transport Geography*, vol. 33, pp. 105–116, 2013.
- [19] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.
- [20] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: a mobile phone trace example," *Transportation Research Part C*, vol. 26, pp. 301–313, 2013.
- [21] M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin-destination matrices using mobile phone call data," *Transportation Research Part C: Emerging Technologies*, vol. 40, pp. 63–74, 2014.
- [22] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin-destination trips by purpose and time of day inferred from mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 240–250, 2015.
- [23] "Google Maps APIs," 2016, <https://developers.google.com/maps>.
- [24] "Baidu Map Open Platform," 2016, <http://lbsyun.baidu.com>.
- [25] AMAP, <http://lbs.amap.com>.
- [26] X. Ma and Y. Wang, "Development of a data-driven platform for transit performance measures using smart card and GPS data," *Journal of Transportation Engineering*, vol. 140, no. 12, Article ID 04014063, 2014.
- [27] X. Ma, C. Liu, H. Wen, Y. Wang, and Y. Wu, "Understanding commuting patterns using transit smart card data," *Journal of Transport Geography*, vol. 58, pp. 135–145, 2017.
- [28] R. Lawson, *Web Scraping with Python*, Packt Publishing Ltd, 2015.
- [29] S. Biba, K. M. Curtin, and G. Manca, "A new method for determining the population with walking access to transit," *International Journal of Geographical Information Science*, vol. 24, no. 3, pp. 347–364, 2010.
- [30] A. T. Murray, "Strategic analysis of public transport coverage," *Socio-Economic Planning Sciences*, vol. 35, no. 3, pp. 175–188, 2001.
- [31] F. Zhao, L.-F. Chow, M.-T. Li, I. Ubaka, and A. Gan, "Forecasting transit walk accessibility: regression model alternative to buffer method," *Transportation Research Record*, no. 1835, pp. 34–41, 2003.
- [32] J. Gutiérrez, O. D. Cardozo, and J. C. García-Palomares, "Transit ridership forecasting at station level: an approach based on distance-decay weighted regression," *Journal of Transport Geography*, vol. 19, no. 6, pp. 1081–1092, 2011.
- [33] Q. Chen and J. Pan, "Coverage rate about the bus stop serving area," *Urban Public Transport*, vol. 2, pp. 17–18, 2002.
- [34] Y. Wang and Q. Li, "Evaluation of the public bus network in Beijing," *Journal of Transportation Systems Engineering and Information Technology*, vol. 7, no. 5, pp. 135–141, 2007.
- [35] Ministry of Construction of the People's Republic of China, *Code for Transport Planning on Urban Road*, Ministry of Construction of the People's Republic of China, Beijing, China, 1995.
- [36] P. Modesti and A. Sciomachen, "A utility measure for finding multiobjective shortest paths in urban multimodal transportation networks," *European Journal of Operational Research*, vol. 111, no. 3, pp. 495–508, 1998.
- [37] W. A. Xu, Y. Ding, J. Zhou, and Y. Li, "Transit accessibility measures incorporating the temporal dimension," *Cities*, vol. 46, pp. 55–66, 2015.