International Journal of Emerging Engineering Research and Technology
Volume 4, Issue 7, July 2016, PP 6-13
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Travel Pattern Recognition using Smart Card Data in Public Transit

Chang YU[1, 2, 3], Zhao-Cheng HE[1, 2, 3]

[1]*Guangdong Provincial Key Lab of Intelligent Transportation System*
[2]*Research Centre of Intelligent Transportation System*
[3]*School of Engineering, SUN YAT-SEN University, Guangzhou 510006, China*

ABSTRACT

As the basic travel service for urban transit, bus services carry the majority of urban passengers. A better understanding of transit riders' travel characteristics can provide a first-hand reference for the evaluation, management and planning of urban public transport system. Over the past two decades, data from smart cards have become a new source of travel survey data, providing more comprehensive spatial-temporal information about urban public transport trips. In this paper, a methodology for mining smart card data is developed to recognize the travel patterns of transit riders. A smart card dataset is first processed to obtain the trip information. After reconstructing the transit trip chains from the trip information, this paper adopts the density-based spatial clustering of application with noise (DBSCAN) algorithm to mine the historical travel patterns of each transit riders. In addition, a sensitivity analysis is conducted to evaluate the optimum parameters. In case study, the analysis of travel pattern characteristics is conducted focusing on the transit riders of Guangzhou City, China.

Keywords: Smart card data, Transit rider, Travel pattern, Trip chain, DBSCAN.

INTRODUCTION

As an effective and energy-efficient travel mode, public transport plays an increasingly important role in urban transportation. Public transit trip information provides the basis for urban public transit planning and operational decision-making, and offers a significant data base for bus dispatching and network planning. Planners and managers have been endeavouring to acquire increasing information on public transit trips to characterize the transit patterns of urban residents and to facilitate scientific planning and decision-making in public transit[1].

Travelling to and from work, school and shops are the main reasons for urban travel, and constitute the primary demand for public transport services. The spatial-temporal properties of such commuting behaviour tend to be regular and periodic. That means, most public transit riders usually have several regular travel patterns. With a better understanding of the travel patterns of transit riders, transit authorities will be able to evaluate their current services to reveal how best to adjust their marketing strategies to encourage higher usage [2]. The majority of exploratory travel pattern studies use travel survey data as their principal data source. However, travel survey data generally have limitations, such as small sample sizes, high cost, low response rates and inaccurate travel behaviour information [3, 4]. Therefore, alternative data sources are needed to more accurately and comprehensively understand the spatial-temporal characteristics of travel patterns.

Over the past two decades, smart card has gradually become the most popular transaction mode in urban public transit systems. For example, in Guangzhou, over 80% of public transit transactions are conducted with smart cards and over 8,000,000 transactions are recorded per day. Accordingly, smart card data have become a new source of data for examining public transport trips. The information collected by smart card systems provides a valuable data source for investigating travel behaviour, as the systems store detailed information on individual public transport trips [5].

Address for correspondence:

342013992@qq.com

A number of studies have used smart card data to examine passengers' travel behaviour in public transit systems around the world. For example, Barry, et al. [6] conducted a detailed study of the New York MetroCard system, and processed the passenger flow data to obtain a passenger flow origin-destination (OD) matrix. Zhao, et al. [7] developed a method for inferring rail passenger trip OD matrices from an origin-only automatic fare collection system. Trépanier, et al. [8] proposed a method for estimating the alighting point of a trip in a system in which the users only validate when boarding. Munizaga and Palma [9] developed a method of estimating the boarding and alighting points in multimodal public transport systems and constructed public transport OD matrices. Jun and Dongyuan [10] identified the properties of travel regions considering the travel frequency and spatial-temporal characteristics of the passengers, and then acquired the transit commuters' OD distribution. For a more detailed literature review, see Pelletier, et al. [11].

Most of the above-mentioned research based on smart card data extracted travel behaviour information macroscopically rather than by analysing individual transit riders' travel patterns. Chu and Chapleau [4] applied the association rule and clustering algorithms to measure transit riders' regularity, and conducted an individual travel behaviour analysis using both temporal and spatial methods. However, their analysis was based on high quality data with complete information and their method was not optimized for a large dataset. Ma, et al. [12] identified the patterns of passengers travel behaviour by spatial clustering based on the temporal and spatial characteristics of smart card transaction data. Deficiently, the parameter setting of clustering is subjective.

In this paper, a comprehensive and effective data mining method is proposed to extract individual transit riders' travel patterns from a massive dataset with incomplete information. Firstly, an estimation method is utilized for data pre-processing to obtain the individual trip information. The trip chains of each riders are then generated based on individual trip information. A DBSCAN clustering algorithm is adopted to mine the travel pattern from each transit riders' historical trip chains and to identify the regular OD and time that the rider usually travel. The clustering parameters are determined by Sensitivity Analysis. The travel patterns clustering is respectively conducted on the scale of "Spatial-temporal regular", "Spatial regular" and "Temporal regular". By doing so, this study develops an effective data mining method to extract individual riders' travel patterns, which have practical application value for the planning and management of urban public transport systems.

METHODOLOGY

Individual Trip Information Obtained

The dataset used in this study is consist of smart card data and Intelligent Public Transport System data (bus GPS data, bus dispatching data and bus network data). The smart card data records the time information and vehicle information of passengers' transit trip, but it doesn't record the location information of trips. Moreover, single ticket system is wildly applied in most public transit system. The public transport system has boarding validation only. That means the smart card data is unable to record the passengers' alighting information. In this situation, the individual trip information is obtained by utilized a method of estimating the boarding and alighting points. The estimation methods have been previously studied [7-9]. The method is based on the assumption of a closed trip chain and infers the alighting point of the current trip according to the boarding point of the next trip. The detailed methodology is provided in Yu and He [13]. The obtained individual trip information include smart card ID, transaction time, route ID, vehicle ID, boarding time, boarding point, alighting time and alighting point, as shown in Table1.

Table1. *Individual Transit Trip Information*

Smart card ID	Transaction time	Route ID	Vehicle ID	Boarding time	Boarding point	Alighting time	Alighting point
510****4837	2013-9-3 8:33:02	80180	****G229	2013-9-3 8:32:14	HCL stop	2013-9-3 9:00:03	SNGKG stop
510****4837	2013-9-3 15:04:26	80190	****G235	2013-9-3 15:04:09	SNGKG stop	2013-9-3 15:30:41	TY stop
510****4837	2013-9-4 8:37:53	80180	****G230	2013-9-4 8:37:38	HCL stop	0001-1-1	Null
510****4837	2013-9-4 19:12:13	80060	****G246	2013-9-4 19:11:47	HJHY stop	2013-9-4 19:37:31	HCL stop
…	…	…	…	…	…	…	…

Trip Chain Generation

Actually, the above obtained trip information is the origin-destination (OD) information based on smart card transactions. There is a need to merge the trip information with same travel motivation (such as passenger transfers) to translate individual trip information into activity-based trip chain, which is significant for travel pattern recognition. A fixed temporal threshold is used in this study to link several smart card transaction records into a trip chain. The fixed temporal threshold is set as 20min based on the statistic of travel interval and the time interval of bus dispatching. If the transaction time difference between two consecutive smart card records is less than the temporal threshold, and the previous alighting point is adjacent to the following boarding point, then they are taken to represent a transfer activity between two routes or two transportation modes (bus and subway) and translated into a trip chain.

Individual Travel Pattern Clustering

Most individual transit riders are likely to show a certain travel pattern during a multi-day period. Once the individual trip chain information has been constructed, the travel pattern for each transit rider is further investigated through clustering the trip chains. To retrieve these hidden and repeated travel patterns in an efficient manner, the density-based spatial clustering of application with noise (DBSCAN) algorithm is therefore adopted because of the following reasons.

(1) The DBSCAN algorithms identify clusters of high density and noise of low density. In this study, noise is an unusual travel pattern, in other words, trips that are randomly made. Our goal is to find the clusters (regular pattern) and differentiate it with the unusual pattern.

(2) The DBSCAN algorithms can identify a cluster of any shape and size. A travel pattern could also form any shape and size due to its nature of human behaviour pattern.

(3) The DBSCAN algorithms do not require the predetermination of initial cores or the number of clusters. It is also essential for travel pattern analysis because the number of patterns from an individual passenger is unknown.

The DBSCAN algorithm defines clusters as dense regions, which are separated by regions of a lower point density. The algorithm has two global parameters: the maximum density reach distance ε and the minimum number of points $MinPts$. A point can be considered a "core point" i_c if it has at least $MinPts$ (density) within a radius ε, as expressed in:

$$\left|N_{\varepsilon(\mathrm{i}_c)}\right| \geq MinPts \tag{1}$$

Where $N_{\varepsilon(\mathrm{i}_c)}$ is the number of points in the data set that has a distance to i_c define as $d(i_c, i)$ less than ε. The most common distance metric used is the Euclidean distance.

For a more detailed description of DBSCAN algorithms, see [14]. A transit rider may begin their repeated trips in both the spatial and temporal domains. In this study, in order to mine the spatial-temporal patterns of passengers' travel behaviour, the maximum density reach distance ε is represented as two parameters: the distance from stop to stop ΔS (spatial), and the trip time difference ΔT (temporal). In the multi-day trips of passengers, if the frequent boarding/alighting stops along the recurring routes are adjacent to each other, these stops may be considered as a regular origin/destination. Similarly, the habitual trip time may be considered as a regular time. According to different consideration of selected parameters, a multi-scale travel pattern clustering can be conducted: spatial-temporal regular (ΔS and ΔT are simultaneously considered); spatial regular (only ΔS are considered); temporal regular (only ΔT are considered).

All passengers' trip chain information is clustered by DBSCAN algorithms. Each cluster represents a typical travel pattern of a passenger. And the anomaly trip records are flagged as noise.

Sensitivity Analysis of Parameters

There is an obviously subjectivity in the parameter setting of DBSCAN algorithms. It is needed to conduct a sensitivity analysis for parameter setting to determine the parameters scientifically, and obtain more reliable clustering results. The application of DBSCAN algorithms requires three important parameters: the distance from stop to stop ΔS (spatial), the trip time difference ΔT (temporal) and the minimum number of boarding $MinPts$ (differentiate noise). Different combinations of parameters are used for clustering, the sensitivity analysis is conducted according to the variability of the percentage of regular trips.

For the spatial travel pattern recognition, the maximum density reach distance ε denotes the walking distance ΔS of the passenger from one to another stop of the same travel pattern. If ΔS increase, the algorithm would define more stops as regular. The parameter ΔS should not be too large since more boarding stops might be generally clustered into same travel pattern, which cause the fineness of travel pattern recognition is not enough.

Figure 1 illustrates the ΔS sensitivity analysis results. The slope of contours in the figure can reflect the effect of ΔS on the percentage of regular trips. The smaller the slope of contours, the greater effect of ΔS . It is indicated the slope of contours is the smallest when ΔS increases from 750 to 1000 m. The variability of the percentage of regular trips is greatest because most passengers' walking distances from stop to stop are in this range. In this study, the parameter setting need to retain as more as possible regular trips (the parameter should be large enough), at the same time, ensure an enough high fineness. On the other hand, according to the Transit Capacity and Quality of Service Manual (TCQSM) [15], 1000m fits in with the acceptable walking distance of passengers and the layout principle of bus stops in city. Therefore, the value ΔS is chosen as 1000 m.

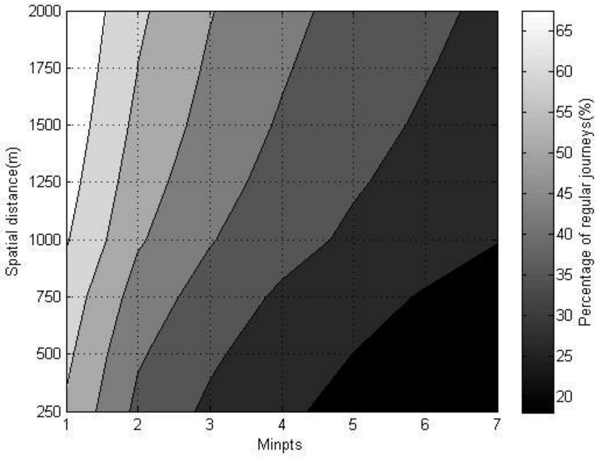

Figure1. ΔS *Sensitivity analysis*

The parameters for the temporal travel pattern recognition have been chosen by a similar approach. Figure 2 shows the ΔT sensitivity analysis results. For the temporal travel pattern recognition, the maximum density reach distance ε denotes the variability of trip times within the same travel pattern. When ΔT increases from 15 to 30 min, the slope of contours is the smallest. Most passengers' variabilities of trip times are in this range. The value ΔT has been chosen as 30 min to allow some variation in passengers' daily schedule.

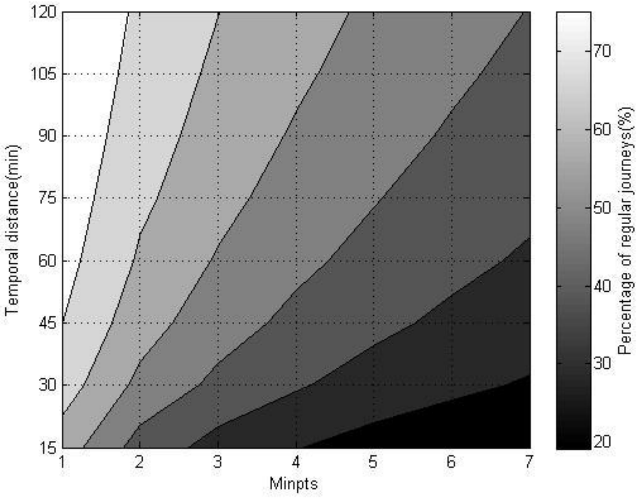

Figure2. ΔT *Sensitivity analysis*

The value of $MinPts$ is chosen to maximize the proportion of the regular travel pattern, but conversely, it should minimize the proportion of anomaly trips because these behaviours could be unreliable. Figure 3 and Figure 4 respectively show the clustering results with different value of $MinPts$ and ΔS / ΔT . It is indicated the distance between the curves " $MinPts$ =2" and " $MinPts$ =3" is the maximum, no matter controlling ΔS or ΔT . That is to say, the parameter is the most sensitive in this range. Accordingly, the value of $MinPts$ is chosen as 3.

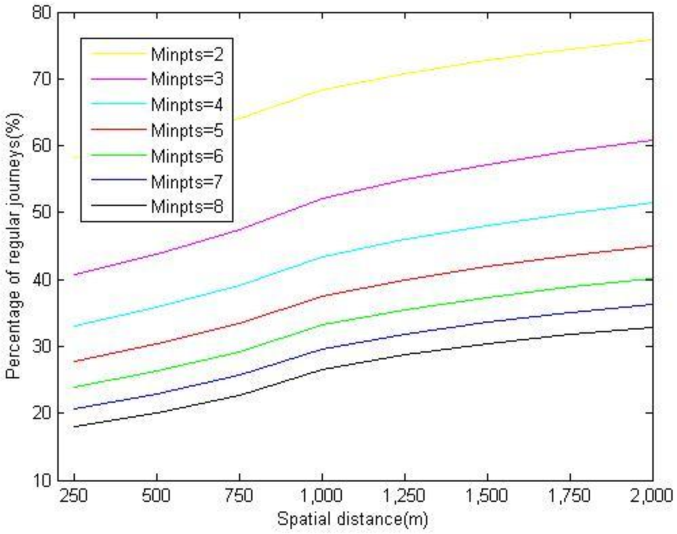

Figure3. *Clustering results with* $MinPts$ *and* ΔS

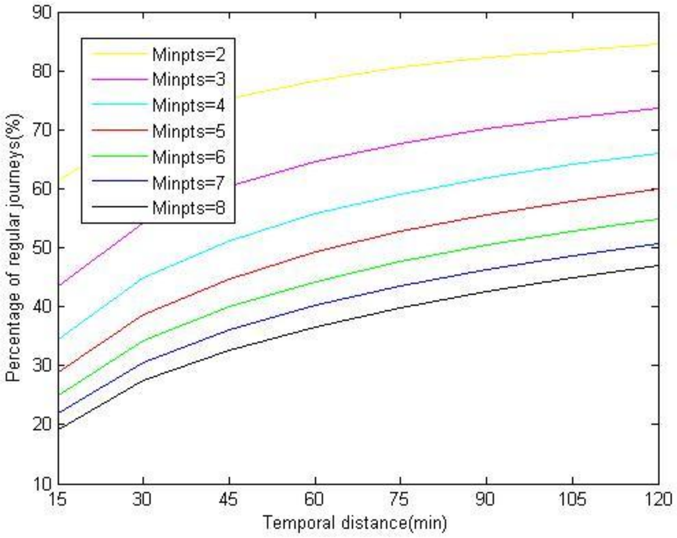

Figure4. *Clustering results with* $MinPts$ *and* ΔT

Case Study

The case study draws on smart card and Intelligent Public Transport System data of a typical month (September, 2014) provided by the Guangzhou transit agency. Under the premise of ensuring the sample representativity, about 32534 smart cards and 1,340,000 transaction records are sampled by Proportional stratified sampling method. This dataset is processed by the proposed method to analysis the travel pattern characteristic of the transit riders in Guangzhou city.

In this study, the travel pattern clustering is conducted on the scale of spatial-temporal regular, spatial regular and temporal regular. By doing so, four types of travel patterns are recognized: spatial-temporal regular (regular OD and habitual time), spatial regular (regular OD), temporal regular (habitual time) and irregular (both of OD and time are irregular). The clustering results are shown as Table 2.

Table2. *Clustering Results of Multi-level Travel Patterns*

Type of patterns	Number of trips	Percentage of trips	Number of patterns
spatial-temporal regular	702810	52%	112689
temporal regular	1146434	83%	152897
spatial regular	1126800	85%	116431
irregular	106067	7.9%	88757

Figure 5 shows the distribution of the number of passengers' travel patterns. As a result: (1) the majority of transit riders have less than 5 kinds of travel patterns, mostly have 2 or 3 kinds. (2) The number of spatial regular of temporal regular travel patterns is generally more than spatial-temporal regular, due to the looser criterion on this scale. (3) The distribution of the number of temporal regular patterns is evener and broader than the distribution of spatial regular, which illustrates passengers have more diverse choice in space than time. Transit riders tend to choose various times to travel in fixed location.

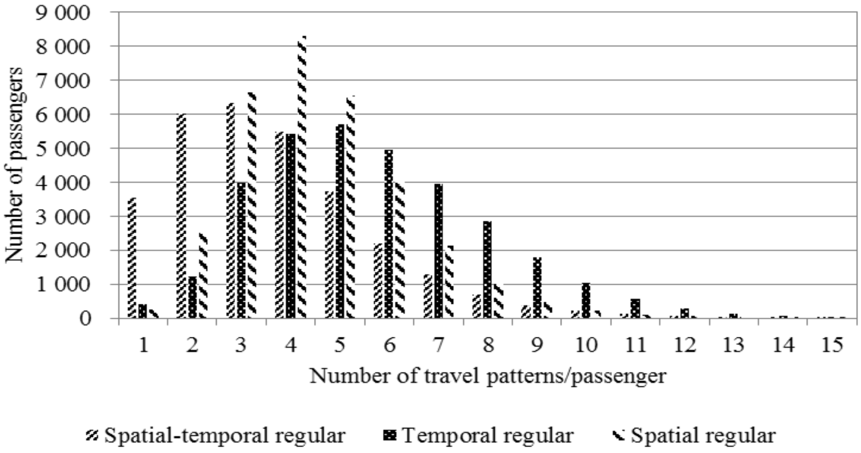

Figure5. *Distribution of the number of passengers' travel patterns*

In order to exploit the usage pattern of each pattern type to understand the daily usage of the transit network. Figure 6 illustrates the trip number of per pattern at different times of the day. It is indicated that the trips of spatial-temporal patterns mainly travelled during morning and evening peak hours, whereas the irregular patterns travelled any time but generally started later in the morning and finished earlier than other patterns. The distributions of spatial regular and temporal regular are similar, but the trips during off-peak hours are more than spatial-temporal regular pattern. A trip purpose assumption can be made that the spatial-temporal patterns mostly are commuting patterns such as school or work-based trips, and the irregular patterns mostly represent less tightly scheduled trips such as leisure or shopping activities. The spatial regular and temporal regular patterns might contain more trips which are flexible in space of time.

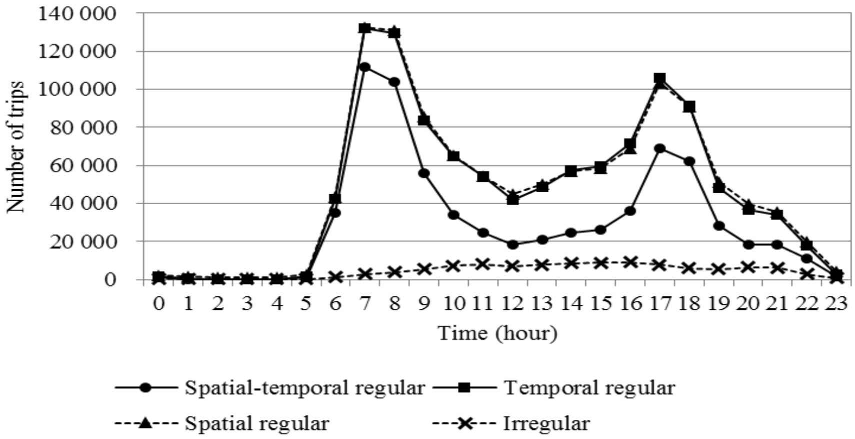

Figure6. *Daily usage characteristics of travel patterns*

There are mainly four types of smart cards in Guangzhou city, respectively correspond to four passenger classes (adult, student, senior, pensioner). Figure 7 compares the proportion of each pattern type under the four passenger classes. The figure shows that adults are the largest contributor in all passenger types. Most of the spatial-temporal regular trips (work-based commuting) are made by adults, who are currently charged with the highest ticket fare. However, due to the large size and complex composition of this class, a large proportion of adults' trips are irregular. The majority trips of students are spatial-temporal regular trips (school-based commuting) because students have tight daily schedule and lack of other travel activities. It is essentially that the transit system is safe and reliable for student travelling by public transport. Relatively, the major contributor in the senior and pensioner classes is irregular trips because passengers from these classes are flexible in space-time and have no mobility needs for work or study.

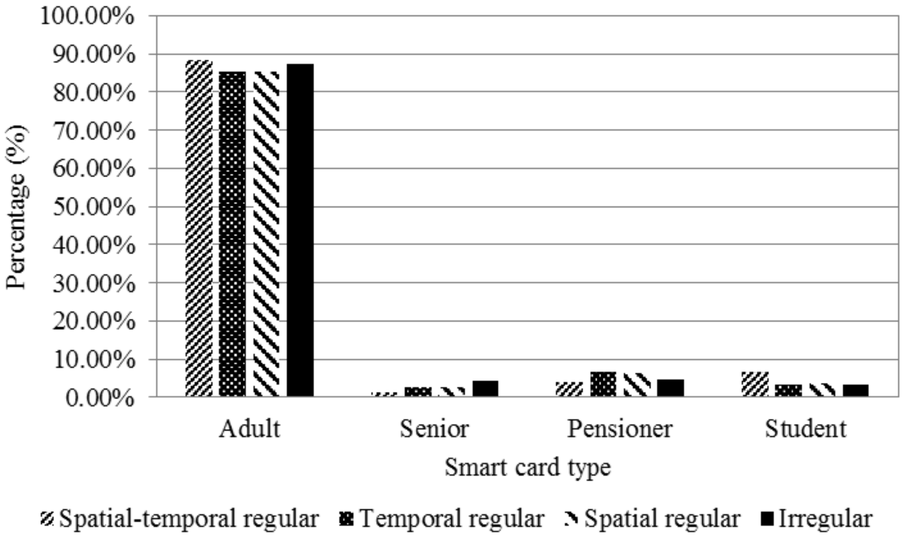

Figure7. *Proportion of each pattern type over different passenger class*

CONCLUSION

This study proposes a series of efficient and effective data mining approaches that is capable of identifying travel patterns for individual transit rider using a large smart card dataset. The DBSCAN algorithm is utilized to successfully detect each passengers' historical travel pattern using the identified trip chains, which exhibits good computational performance. In addition, sensitivity analysis is introduced for determining the optimum parameters and obtain reliable clustering results. A multi-scale clustering procedure is conducted to make comprehensive travel pattern analysis. The travel patterns of transit riders are significant information for transportation researchers seeking to understand day-to-day urban travel behaviour variability. And they also offer substantial benefits for transit agency to support transit market analysis and improve their transit service.

ACKNOWLEDGEMENTS

This work was supported by Science and Technology Planning Project of Guangdong Province, China (No. 2014B010118002). And the authors are grateful to the Traffic Commission of Guangzhou Municipality for data availability.

REFERENCE

[1] C. Daraio, M. Diana, F. Di Costa, C. Leporelli, G. Matteucci, and A. Nastasi, "Efficiency and effectiveness in the urban public transport sector: A critical review with directions for future research," *European Journal of Operational Research*, vol. 248, pp. 1-20, 2016.

[2] D. K. Boyle, P. J. Foote, and K. H. Karash, "Public Transportation Marketing and Fare Policy," *Transportation in the New Millennium*, 2000.

[3] P. Rickwood and G. Glazebrook, "Urban Structure and Commuting in Australian Cities," *Urban Policy and Research*, vol. 27, pp. 171-188, 2009.

[4] K. Chu and R. Chapleau, "Augmenting Transit Trip Characterization and Travel Behavior Comprehension," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2183, pp. 29-40, 2010.

[5] M. Bagchi and P. R. White, "The potential of public transport smart card data," *Transport Policy*, vol. 12, pp. 464-474, 2005.

[6] J. Barry, R. Newhouser, A. Rahbee, and S. Sayeda, "Origin and Destination Estimation in New York City with Automated Fare System Data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1817, pp. 183-187, 2002.

[7] J. Zhao, A. Rahbee, and N. H. M. Wilson, "Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, pp. 376-387, 2007.

[8] 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*, vol. 11, pp. 1-14, 2007.

[9] 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.

[10] C. Jun and Y. Dongyuan, "Estimating Smart Card Commuters Origin-Destination Distribution Based on APTS Data," *Journal of Transportation Systems Engineering and Information Technology*, vol. 13, pp. 47-53, 2013.

[11] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, pp. 557-568, 2011.

[12] 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.

[13] C. Yu and Z.-c. He, "Passenger Flow Estimation Based on Smart Card Data in Public Transit," in *CICTP* 2014, pp. 658-670.

[14] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, "A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases," *International Journal of Computer Applications*, vol. 3, pp. 1-4, 2010.

[15] I. TCRP and Kittelson & Associates, "TCRP Report 100: Transit Capacity and Quality of Service Manual, 2nd ed.," T. R. B. o. t. N. Academies, Ed., ed. Washington, D.C., 2004.