IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 26, 2019, accepted December 8, 2019, date of publication December 16, 2019,
date of current version December 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2959907

Mining Customized Bus Demand Spots Based on Smart Card Data: A Case Study of the Beijing Public Transit System

YIYI CHENG[1], AILING HUANG[2], GEQI QI[2], AND BEI ZHANG[3]

[1]Intelligent Transport System Research Center, Southeast University, Nanjing 211189, China
[2]Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Ministry of Transport, Beijing Jiaotong University, Beijing 100044, China
[3]China International Engineer Consulting Corporation, Beijing 100048, China

Corresponding author: Ailing Huang (alhuang@bjtu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1601200, and in part by the Foundation for Innovative Research Groups of the National Nature Science Foundation of China under Grant 71621001.

ABSTRACT In recent years, to fix the shortcomings of traditional bus service and meet the diversified needs of passengers, a new type of transit system, the customized bus (CB), has been proposed. However, how to define and mine the CB's demand is still less being addressed. Since the data of bus smart cards can provide more travel information, it makes the mining of potential CB's demand spots more possible, which can be helpful in CB service design. In order to mine the demand spots more scientifically, this paper, for the first time, quantitatively defines the CB demand characteristics and criteria of selecting potential area, and develops a demand hotspots extraction methodology for CB. The methodology solves two issues primarily. One is how to organize massive smart card data and obtain the space-time pattern and mobility of passenger efficiently; the other is how to mix the CB demand characteristics into the method. This demand spots extraction method can generate multi-style maps, including the heat and origin-destination maps, for spatial cluster of CB's demand spots in rational areas in terms of the CB demand characteristics based on geographic information system. By using the bus smart card data in Beijing, China, this paper carries out a case study to validate the method. The empirical data mining analysis shows that our proposed method can define demand spots ideally. Our work can provide a valuable reference for decision makers to design CB system.

INDEX TERMS Bus smart card data, customized bus, potential demand area, geographic information system, spatial clustering analysis.

I. INTRODUCTION

Customized Bus (CB) is a new and innovative demand-responsive transportation system (DRTS) that provides advanced, user-oriented transportation services to specific customers, especially commuters, by using online information platforms to aggregate similar travel demand patterns [1]. The CB system is more reliable, comfortable and convenient than traditional public transportation (PT) systems, and more efficient, cost-effective and environment friendly than private cars, which makes it become a more competitive PT system in metropolis [2]. Accordingly, CB have their unique advantages in reducing traffic congestion, reducing pollutant emissions, and serving PT, especially in commuting travel. Therefore, CB is increasingly valued by the PT administration departments of major cities. However, many cities lack complete thinking about the planning and management of CB system currently. The design of CB sites and routes is mainly determined by experience, so it is difficult to find out the huge potential demand of CB and thus maximize its due advantages. Existing studies are widely focused on the optimization of CB service patterns, bus network optimization, pricing strategies and so on [3]–[5], but the research on the mining demand of CB has not been discussed. At present, with the accelerating development of communication and information technology, more and more traffic data are available. How to

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

scientifically and effectively mine CB demand hotspots based on existing traffic data, especially bus smart card/intelligent card (IC) data, to provide data support for follow-up CB service system design has become an urgent problem to be studied.

A. SOURCE OF MINING CB'S DEMAND

To mine CB demand, bus smart card data are an important source. With the continuous development of related technologies, the travel information stored by bus smart cards is becoming more and more abundant. Bagchi and White [6] summarized the information that the smart card can contain and summarized it as the linking of data, the volume and scope of the data, and continuous information. They pointed out that the analysis of smart card data can collect more long-term information and provide more accurate predictions than traditional sampling survey methods.

Since smart card stores so much information, privacy issues need to be considered first. Although the Automatic Fare Collection (AFC) system itself is helpful to improve the security of the card holder's travel information [7], [8], the generated data set contains potential privacy issues. The research by Dempsey and Stephen and Pelletier *et al.* [8], [9] suggests that perfecting relevance laws and statutes and placing smart card data under the common supervision of PT enterprises and governments is an effective means to protect privacy. The recommendation was supported and the privacy of smart card holders was strictly protected throughout this paper.

Pelletier *et al.* [9] summed up the previous scholars' research and divided the study of smart cards into the level of strategy, tactics and operation. The strategic level of research includes long-term network planning, passenger behavior analysis and demand forecasting. It also illustrates the important position of smart card data in the analysis of urban PT services. Therefore, more and more studies have begun to pay attention to the analysis of bus smart card data. Ma *et al.* and Zhong *et al.* [10], [11] separately paid attention to the distribution characteristics of smart cards in time and space. The temporality of PT travel is more likely to be obtained from its long-term regularity, and the spatial distribution reflects the relationship between passenger travel and land use. Both have positive significance for bus service planners and even city planners to adjust PT services perfectly. Therefore, the temporality and spatiality of smart card data had best be studied together. When analyzing the spatial distribution of housing prices and housing (including new home purchase and personnel relocation), Gao *et al.* [12] analyzed the relationship between travel and housing prices through data analysis of smart cards, and then conducted short-term forecasts of spatial residential distribution after housing prices rose. Ingvardson *et al.* [13] used the bus smart card data to study the time characteristics of passengers' waiting for bus, and based on this, made a reasonable timetable, thus shortening the waiting time of passengers and improving the service quality of the bus [14]. Smart cards data can also be combined with other data to explore the intrinsic link between urban PT and other urban systems. Wei *et al.* [15] combined the bus smart card data with the buses' GPS data to obtain the link between passengers and population, land use, and transportation factors. Qi *et al.* [16] used smart card data and points of interest (POI) data to analyze and predict passenger regional mobility patterns. However, similar to previous studies, this study did not pay enough attention to the demand characteristics of the passenger.

B. POTENTIAL PROBLEMS OF USING BUS SMART CARD DATA AND SOLUTIONS

It is undeniable that at present, the smart card being the source of CB's demand still has its limitations, which can be divided into three types: (1) the problems in the operation of the AFC system, including potential fare evasion and the erroneous data (missing data, illogical values and duplicate transactions) generated by the AFC; (2) the adoption the entry-only charging system, that is, only taping the smart card once during the whole ride; (3) the usage volume/rate of the smart card. The existence of these problems will affect the study of subsequent bus cards, resulting in analytical errors. The first two will cause errors in the data set itself, and when the smart card usage volume/rate is low, the analyzed results will be one-sided and cannot provide effective guidance to the PT manager. Therefore, if smart card data are expected to be used as the source of mining CB's demand, these problems need to be carefully considered.

According to Delbosc and Currie [17], fare evasion is a problem that cannot be neglected by transportation agencies. Reddy *et al.* and Barabino and Salis [18], [19] analyzed the main scenes of fare evasion and proposed solutions such as video monitoring equipment and an appropriate number of inspectors. Although AFC cannot avoid the occurrence of fare evasion, it can help drivers or inspectors reduce the chances of fare evasion, and improve the efficiency of charging efficiency [20]. The study by Guarda *et al.* [21] also shows that the probability of fare evasion can be reduced with the improved AFC system such as Back-door entry system, the supervision of drivers and reasonable bus design. During data acquisition, AFC generates three types erroneous data, including missing data, illogical values and duplicate transactions due to software or hardware errors [22]. These problems usually can be found through comparisons between data, such as a record lacking the necessary attributes, two records at the same time and place from one card. Barabino *et al.* [23] created a model that used the individual and the entire route tap in& out records as constraints to detect data anomalies. Considering that the amount of data from smart cards is large and the proportion of erroneous data is very small usually, the main processing method is directly eliminating [23]–[26]. And some researchers use historical data and similar data to make estimation supplements [27]–[29].

Currently, entry-only charging systems are adopted by most cities' AFC, such as New York, America, Chicago, America, San Diego, America and Guangzhou, China.

And the entry-exit charging system requires the passengers to tap in& out their smart cards when they take a ride. The South East Queensland, Australia, Seoul, South Korea and Beijing, China adopt this system in their AFC [30]. The entry-only charging system can only accurately provide the passenger's pick-up location. This is obviously not conducive to passenger origin-destination (OD) extraction, which is one of the main goals of mining smart card data [23]. Therefore, how to estimate the passenger's possible drop-off location has become the focus of many scholars' study. Nunes *et al.*, Trépanier *et al.* and Munizaga *et al.* [27], [29], [31], [32] designed models to estimate the possible drop-off position of passenger by using multi-day card data, and both models' accuracy reached 80%. At this stage, the estimation model needs to be improved to improve the accuracy rate. Or, entry-exit charging system could be used to get accurate passenger's drop-off position directly.

The smart card usage volume/rate is critical for demand mining, and even has bad impact on the passenger drop-off position estimation for entry-only charging system with low usage volume/rate [29]. According to the study from Li *et al.* [30], the sample size below 10000 is low smart card usage volume. Although there is no clear correlation between card usage volume/rate and accuracy of demand mining, it is still worth noting that many scholars acknowledged that smart cards offer convenience for commuters, and they identify commuters by analyzing smart card data sets [9], [10], [12], [25]. Fayyaz *et al.* [33] conducted an interesting study. They analyzed the bus dwell time at station to obtain the proportion of different payment methods and achieved good results. Considering that the use of AFC is the overall development trend of PT [7], the feasibility of using large smart card data as the source of mining CB's demand is becoming higher and higher. But nonetheless, for the low smart card usage volume/rate cases, auxiliary investigations are still recommended to conduct to verify the results of mining CB's demand.

In summary, it is feasible to find CB demand spot by mining a large number of smart card data, especially for the AFC which adopting entry-exit charging system, with the premise that the three problems have been properly solved.

C. MINING METHODOLOGY OF TRAFFIC DEMAND

Data mining is a common method used in traffic system analysis, such as road accidents analysis, identifying congestion events, and searching internal relations between various traffic data [34]–[36]. The main methods include statistics, genetic algorithms, artificial intelligence (AI) algorithms, and visualization [37]. Bus travel data characteristics in large cities presents complex, mass and spatial-temporal. And the results generated by a single mining method are difficult to achieve sufficient validity and accuracy. In order to reconcile the data patterns and demand characteristics, the visualization and spatial clustering methods are selected as the methods of demand mining in this paper. Because the former can make complex and spatio-temporal data easy to be understood and applied, while the latter can take into account the point-to-point service, which is one of the CB service characteristics [38], [39].

The temporality and spatiality of bus smart card data make the heat map and the OD map be selected as the visualization method. A heat map (or heatmap) is a graphical representation of data, which is represented as colors according to its value [40]. The heat map can evaluate a series of indicators such as the demand distribution of urban PT, vehicle operating conditions, and the rationality of station location according to passenger flow volume, bus speed and so on, thus providing positive guidance for the operation of urban PT system [41], [42]. Similarly, it can also be used to present bus passenger pattern [10], [16]. But using only one style, heat map, is not sufficient for passenger pattern analysis.

The OD map provides the other side of geographic data, showing the interaction between different regions, and studying the spatio-temporal patterns and trends of large-scale passenger mobility [43], [44]. So, it can be a powerful complement to the heat map. Geographic Information System (GIS) is a computer-aided system for capturing, storing, retrieving, analyzing, managing and displaying spatial or geographic data [45]. It is a common tool for data visualization and has been widely applied in the field of PT. GIS can generate isochronous lines of PT trips (lines with equal travel time) to show the status of urban PT, and is used to evaluate the accessibility of urban PT to enhance the attractiveness of PT, thus providing city manager or policymaker with optimized plan for integrated land use and transportation system design [46]–[48]. Agrawal and Nagrath [49] used GIS to construct heat maps when exploring the autonomous route allocation of urban bus. The area is gridded and the color of different grid is dyed according to the population, and the bus route will be allocated according to the change of grid color to serve the most demanding area. Domènech and Gutiérrez [50] used GIS to present the proportion of population near the station and the connectivity of different areas when evaluating the coverage and utility of PT systems in tourist areas, so as to determine the bus plan in different seasons. Lee and Miller [51] used GIS to outline the spatio-temporal accessibility maps of different bus network combinations when evaluating the urban redesigned PT system and the new rapid transit system, CMAX. This improved the efficiency of PT assessment. At present, the most mature GIS tools include ArcGIS, TransCAD, Google Earth, etc.

Cluster analysis can be used to group objects with similar degrees of similarity [52], while spatial clustering can be used to process data with temporality and spatiality, which has been used in study of hotspot extraction on traffic. For example, K-method or spatial clustering method are used to extract congestion hotspots, accident hotspots, or some vehicle's demand hotspots, like taxis [53]–[55]. The main way to locate bus sites is to establish a certain cost function (like company's cost, convenience, site spacing), then use different methods to gradually approach the optimal solution to obtain the optimal site location [56]–[58]. The extraction of the

TABLE 1. Comparison of existing researches.

References	Data	Research Level	Visualization Plan	Research Focus	Application	Demand Characteristics of Passenger are discussed?
[9]	Bus AFC	City	Heat map & Sankey diagram	Bus spatial-temporal distribution and variability	Lay the foundation for the follow-up research	No
[10]	Bus AFC		Heat map	Spatial-temporal distribution of research object	Provide better bus service for commuter	
[41]	Bus AFC& GPS				Provide guidance for the overall design of the bus system	
[49]	Population	Regional				
[50]	Bus GPS	City				Yes
[51]	Bus AFC					No
[42]	Bus GPS	Road			Evaluate site selection	/
[16]	Bus AFC& POI	Regional		Spatial-temporal distribution and mobility of research object	Promote intelligent transportation design	No
[44]	Pubilc bike AFC& Car GPS		Heat& OD map		Discover periodic patterns and mass longer-term mobility trends	Yes
[55]	Car GPS	City	/	/	Optimize taxi dispatch	

demand spot can be used as the initial solution of the existing model, thereby improving the efficiency and accuracy of the solution. Iliopoulou *et al.* [59] used time-space clustering to observe the phenomenon of bus gathering on the line, and pointed out that the bus gathering phenomenon will be more significant during peak hours.

As the research on traffic data mining and analysis is becoming more sufficient, for the sake of better understanding, this paper makes a comparison on the related research fields among some key literatures (shown on Table 1).

In light of the review, though some contribution on mining CB demand has been made recently, several questions still need to be addressed further. Firstly, currently there is still no clear and accepted definition of the characteristics of CB demand at home and abroad, not to mention the criteria of selecting CB demand spot. Secondly, the research level is only at the city or regional level, which may be sufficient for those researches, but it is inflexible to mining spots throughout the whole city. In addition, the visualization plan is simple, often only one or two styles of map. Therefore, many researchers only focus on the spatial-temporal distribution of their research objects, while ignoring the mobility, which is a powerful complement to the former. Finally, some papers' applications are too macroscopic and theorization, resulting in a lack of consideration of passenger demand characteristics. In fact, exploring the demand from the perspective of the passenger can amplify the reliability of the research results.

As the development of CB has become more and more important, there is a lack of research on how to exploit its potential needs through more resources, such as smart card data. At the same time, limitations of previous studies need to be overcame and the CB demand definition and criteria are needed to be quantitatively presented for mining demand accurately. Therefore, aiming at above research limitations, this paper seeks to provide a general method for mining CB demand spots during the peak period based on bus smart card data.

The remaining paper is organized as follows. In Section II, the demand characteristics of CB and how to incorporate them into the mining method are discussed. In Section III, a general methodology, designed for entry-exit charging system, of the extraction of CB demand spots based on smart card data is introduced, which includes data preprocess, data visualization and demand spots extraction. A case study, with bus smart card data from Beijing, China, is carried out by applying this method in Section IV. Lastly, in Section V, the results of the process and analysis are summarized to arrive at pros and cons of the method and to make recommendations.

II. CB DEMAND CHARACTERISTICS AND SELECTION CRITERION OF POTENTIAL CB DEMAND AREA

Although the source of demand analysis has been identified, and the potential problems in AFC operation can be properly tackled, it is still necessary to clarify the demand characteristics of CB, which is the CB's services expected by potential passengers. It can help avoiding this DRTS to failure [60], and work in finding CB's potential demand area (PDA).

A. DEFINITION OF CB DEMAND CHARACTERISTICS

The analysis of the PT demand characteristics is complex, and the demand characteristics can be systematically obtained by using economics demand concepts and comparing other bus service modes. Economics research indicates that PT demand shows significant regular fluctuations, while travel distance, cost and purpose, and other modes of transportation both have an impact on PT demand [61], [62].

The main PT service modes include the traditional PT service, the Bus Rapid Transit (BRT), the flex-route transit services and the community bus system. Compared with traditional PT service, CB is more reliable, comfortable and convenient [1], [2]. Compared to BRT, CB only needs to consider the endpoint site setting, so its route is more flexible, and its site and route construction costs are lower [2], [63]. Flex-route transit service is another type of DRTS, but it's

TABLE 2. Comparison of main PT modes.

PT mode	Traditional PT service	BRT	Flex-route transit service	Community bus	Customized bus
Site setting	Multipoint	Multipoint	Multipoint point	Multipoint	Single-point/ few point
Route setting	Fixed	Fixed	Flexible	Fixed	Flexible
Service mode	Wait in site	Wait in site	Wait in site	Appointment first	Appointment first
Service speed (km/h)	≤25	≤35	≤15	≤15	≤35
Service scope	City region	Arterial road	Limited region	Limited region	City region
Service distance (km)	≤10	≤15	≤5	≤5	≤15
Vehicle capacity (person)	≤100	≤160	≤30	≤30	≤30

TABLE 3. Selection criteria for PDA.

Name	Specification	Recommended Value	
		City size (million population)	Travel distance CR_D (km)
Distance (D)	Passengers origined from the area have the commuting demand to travel in far/long distance/time.	0.5-1.0	5km≤ CR_D ≤10km
		1.0-2.0	5km≤ CR_D ≤11.5km
		Greater than 2.0	5km≤ CR_D ≤15km
Economics (E)	The economic conditions of the area is in a quite high level of affluence.	The personal total cost of housing and transit (CR_E) should be less than or equal to the 45% of PCDI.	
Activity (A)	The sum (CR_A) of passenger production and attraction at the peak period within an area is high enough.	$CR_A \geq C_v/f_{co}/f_{cb}$	
OD Volume (V)	Passenger flow (CR_V) per hour exists between origin and destination areas.	CR_A ≥20	

more focus on the service in low-density areas [64], which is the exactly the opposite of CB. The community bus is mainly applied for the transfer between hub stations and Inter-community roads [65]. That means commuting passengers have to take more than one transfer behavior, which is inconvenient and may increase their commuting time.

Based on the above, Table 2 is generated to show the performance indicators of these PT modes in detail. It should be noted that due to different factors such as city size and traffic conditions in different cities, the recommended value of the indicator's upper limit is defined by referring to *Code for transport planning on urban road* [66] of China. And "Site setting" in Table 2 refers to the location of sites on the route.

Combined with the previous concept of CB, the fluctuations in PT demand characteristics and CB's advantages over other bus modes, the demand characteristics of CB can be summarized. CB demand refers to the requirements of users who generally have regular trip patterns and expect to be served by the point-to-point, transfer-free, time-reliable transit service to achieve longer distance travel, and are with a certain ability to pay, which is especially suitable for commuters, who travel for long-distance travel, require time reliability, and are willing to pay higher fare.

B. CRITERIA OF SELECTING POTENTIAL DEMAND AREA

It will be time-consuming and labor-intensive to mine CB demand spots across the whole city. The proper consideration is to find PDA and to mine demand spots in these PDAs. Based on the work in Section II. A. and the definition of CB demand characteristics, the selection criteria for the PDA are summarized in Table 3 including Distance, Economics condition, Activity and Volume criteria, respectively.

These criteria are used to ensure passengers' demand for reliable long-distance services, while to ensure the benefits of bus operators as well. All 4 criteria need to be quantified to facilitate and prioritize follow-up hot area selection. And recommended values are also presented in Table 3.

1) DISTANCE CRITERION (D)

Metropolises with different sizes and traffic conditions have different standards for measuring "far/long" commuting distance/ time. Therefore, PT planners can define a suitable value based on the distribution of commuting CR_D/ time in residents trip surveys and local code. The recommended values are defined for cities with different sizes by referring to [66] in China.

2) ECONOMICS CRITERION (E)

It means that the potential areas should be in a quite high level of affluence so that there will be potential CB users who can afford higher fare in these areas. Most of the current research on defining high level of economics is qualitative, hence there is also still no implementing quantitative standards to be implemented. Here, the Position Affordability Index proposed by the US Department of Housing and Urban Development and the Department of Transportation is referenced to determine economics criterion. It indicates that the personal total cost of housing and transit (CR_E) accounts for 45% of personal total revenue is "affordable" [67]. As the economic situation varies in different regions, in practice, local PT planners should conduct survey to investigate economic conditions such as the average rent price, per capita disposable income (PCDI) or Gross Domestic Product (GDP) for adjusting the value of CR_E in terms of 45%.

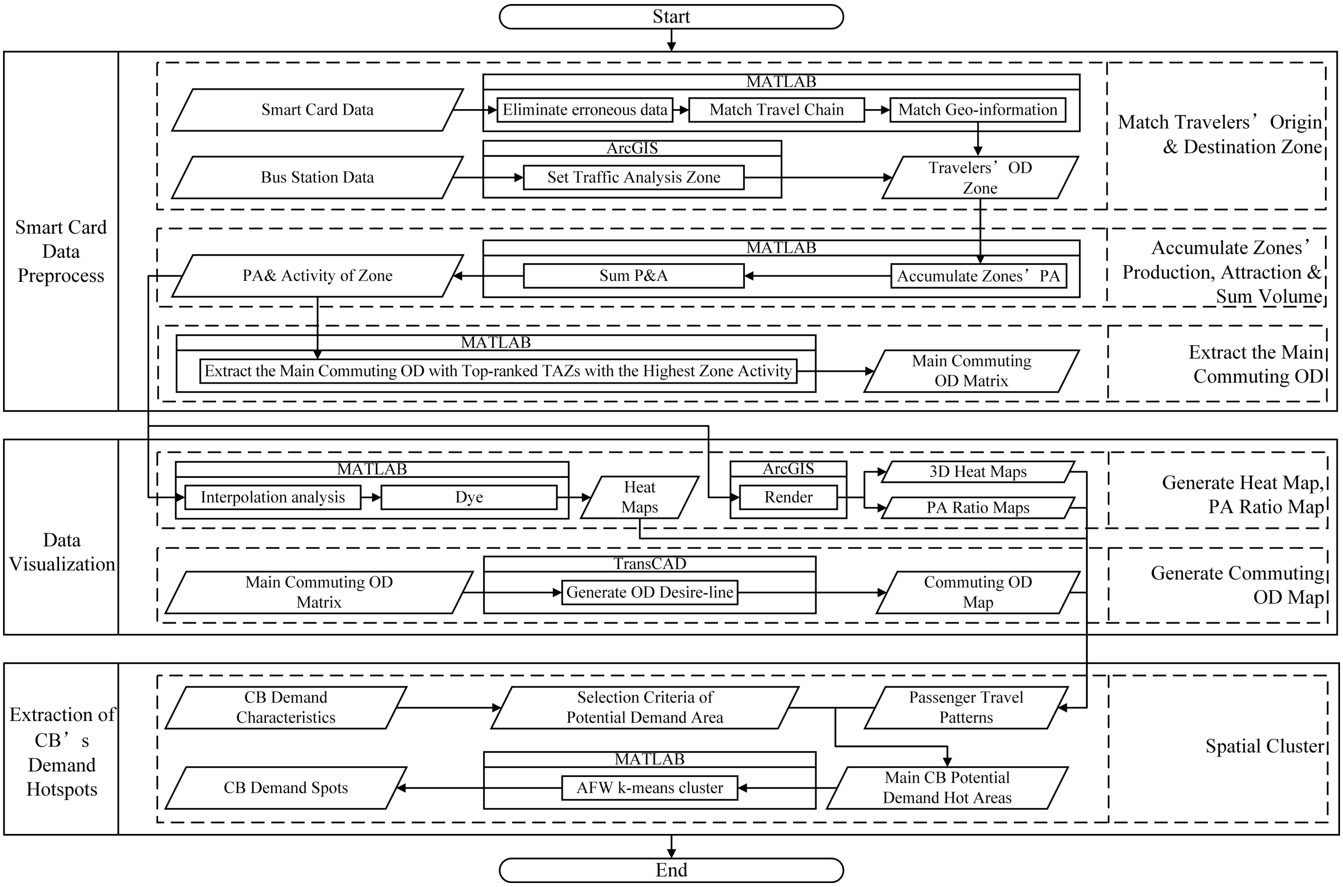

Start
Smart Card Data Preprocess
Smart Card Data
MATLAB
Eliminate erroneous data
Match Travel Chain
Match Geo-information
Bus Station Data
ArcGIS
Set Traffic Analysis Zone
Travelers' OD Zone
Match Travelers' Origin & Destination Zone
PA& Activity of Zone
MATLAB
Sum P&A
Accumulate Zones' PA
Accumulate Zones' Production, Attraction & Sum Volume
MATLAB
Extract the Main Commuting OD with Top-ranked TAZs with the Highest Zone Activity
Main Commuting OD Matrix
Extract the Main Commuting OD
Data Visualization
MATLAB
Interpolation analysis
Dye
Heat Maps
ArcGIS
Render
3D Heat Maps
PA Ratio Maps
Generate Heat Map, PA Ratio Map
Main Commuting OD Matrix
TransCAD
Generate OD Desire-line
Commuting OD Map
Generate Commuting OD Map
Extraction of CB's Demand Hotspots
CB Demand Characteristics
Selection Criteria of Potential Demand Area
Passenger Travel Patterns
CB Demand Spots
MATLAB
AFW k-means cluster
Main CB Potential Demand Hot Areas
Spatial Cluster
End

FIGURE 1. Total process of hotspots extraction.

3) ACTIVITY CRITERION (A)

Activity refers to the sum of passenger production and attraction (PA) at the peak period within this area. It means that the area should be active enough in trips, so there are probably existing potential CB users. Generally, the value of CR_A should be larger than the capacity C_v of a CB vehicle at least. Considering that there may be only a proportion of commuters for all PAs to use CB, we define the activity value as $CR_A \geq C_v/f_{co}/f_{cb}$, here, f_{co} stands for the commuter ratio of PA within this area and f_{cb} is the CB potential CB user ratio of commuter.

4) OD VOLUME CRITERION (V)

Because CB is a type of DRTS, its operation and route settings are flexible, which can be flexibly subscribed, changed and cancelled [1], [2], the passenger flow (CR_V) between origin and destination areas should meet at least one busload. As different size of bus has different capacity C_v, different cities should set different CR_V as the V criterion. For referring to [4], [5] and this Beijing case, the recommended value should be greater than 20 or C_v/f_{co} passenger per hour.

It should be emphasized that if an area has met the V criterion, whether the A criterion is met is not important. Because A criterion cannot guarantee that it can support the existence of CR_V completely, but can help PT planners quickly narrow down the searching scope.

III. METHODOLOGY OF CB'S DEMAND SPOTS EXTRACTION FROM SMART CARD DATA

The research methods proposed in this paper are as follows: (1) First, the smart card data are preprocessed to eliminate the erroneous data, to match the passenger travel, to calculate passenger PA and to extract commuting OD by self-designed preprocessing algorithm. (2) After that, interpolation analysis and dyeing algorithm are applied to generate preprocessed data as heat maps. And the GIS tool is used to visualize and analyze the processed data, and generate the commuting OD map, 3D heat maps and PA ratio maps of the peak period. Through the analysis of the different colors and lines contained in those maps, the passenger demand pattern can be summarized, the hot areas where patterns are matched with the selection criteria of PDA can be identified. And these hot areas are set as PDAs. (3) Finally, the demand spots are positioned in those PDAs by means of a spatial clustering algorithm, the AFW k-means algorithm.

It should be noted that this method is primarily designed for the entry-exit charging system due to the structure of the data samples. For the entry-only charging system, the paper

TABLE 4. Smart card attributes information (partial).

Card ID	Mark time	Mark station	Trade time	Trade station	……
*****833	20170206082344	170473890	20170206085523	170487614	……
*****451	20170206082737	170476393	2.0170206094703	170465055	……
……	……	……	……	……	……

will also give corresponding supplementary instructions. The specific process is shown in Figure 1 below.

A. SMART CARD DATA PREPROCESS

The goal of preprocessing data is to make it available in GIS, thus making it easier for planners to observe the passenger travel patterns. The individual smart card datum has limited attributes and is not directly associated with geographic information. In addition, erroneous data exists in the original data set. Therefore, the overall preprocessing work consists of four parts: the cleaning of erroneous data from AFC, the matching of passenger travel data, the calculation of passenger PA within each traffic analysis zone (TAZ), and the extraction of the main commute OD.

1) THE ELIMINATION OF ERRONEOUS DATA FROM AFC

Data cleaning is mainly to improve the quality of data, thereby improving the accuracy for follow-up analysis. As stated in Section I. B, AFC has two problems in its operation, which are potential fare evasion and the erroneous data generated by the AFC. For the former, the impact on smart card data set is small, as some fare dodgers may not hold smart cards. At the same time, CB's passengers should have good economic capability (Section II. A). And if a person does not expect to pay for traditional bus fare, then he or she will have a lower demand for a more expensive type of bus service. Based on these two points, this method ignores the loss of data caused by the fare evasion.

There are two ways, elimination [23]–[26] and supplement [27]–[29], for the processing of erroneous data. But there is currently no standard available for referring to choose which way, and choice needs to be determined according to specific research objects or conditions. For example, if the cause of erroneous data can be clearly understood, data supplement is necessary [27]. But that is based on a good understanding on the local AFC operation. In fact, with the development of smart card technology and the growing popularity of smart card, the proportion of erroneous data in smart card data is becoming smaller and smaller, hence the feasibility of elimination will become higher and higher. And elimination can also help save calculation cost and avoid the risk of producing new erroneous data in supplement. Therefore, this method recommends the elimination for dealing with the erroneous data.

For the entry-exit charging system, the attributes of card data usually include the unique card ID, boarding site number (Mark station) and time (Mark time), alighting site number (Trade station) and time (Trade time), etc., as shown in Table 4. And for the exit-only system, there are no attributes of Trade station& time. The elimination consists of three parts, (1) missing data, (2) illogical values and (3) duplicate transactions elimination.

(1) Missing data elimination. It refers to the default of one or several attributes in a record. Missing attributes can be found directly by the "find" command, and the records containing these attributes are deleted.

(2) Illogical values elimination. It consists of two types, ① getting on and off at the same site and ② abnormal long-time interval between getting on and off. The former is mainly caused by passenger taping out too early when the vehicle is still in the range of the initial site. Abnormal long-time interval means the length of time exceeding the normal riding time, for example, 24h, which is an impossible time length. The latter may be caused by passengers forgetting to tap out their cards when they get off from the previous bus. For ①, the station number of the Mark station is subtracted with the station number of the Trade station. If the result is 0, the record will be deleted. For ②, the time of Mark and Trade are compared. If the time interval between the two is longer than 3h, the record will be deleted.

This part is mainly designed for the entry-exit charging system. For the exit-only system, it should be adjusted according to the actual case. For example, the changeover between consecutive vehicle trips can cause passengers' abnormal getting on at the terminal station [27], which can be eliminated directly.

(3) Duplicate transactions elimination. Firstly, all records in data set are sorted according to the Card ID, which can gather all the records of each passenger. Secondly, each passenger's records are divided into temporary groups. Thirdly, records in each temporary group which have same Mark and Trade station and time are located and deleted.

2) THE MATCHING OF PASSENGER TRAVEL

The matching of passenger travel data mainly includes two steps of matching, (1) one is to match the travel chain of passengers with transfer behavior. As for passengers without transfer behavior, this step is skipped directly. And (2) the other is to match passenger travel data and geographic information to find passengers' geographical starting point and end point.

(1) Match of passengers' travel chain. The travel chain need to be matched because some passengers have transfer behavior, that is, two or more rides are required to reach the final destination, and the final OD statistics maybe error if the behavior wasn't be identified. As shown in Figure 2, where

TABLE 5. Bus station information table (partial).

Station ID	Line ID	Site name	Latitude	Longitude	……
170473890	5819834	Chaoyang xincheng	116.5493	39.94955	……
170476393	5819834	Donggangzi	116.5654	39.97417	……
……	……	……	……	……	……

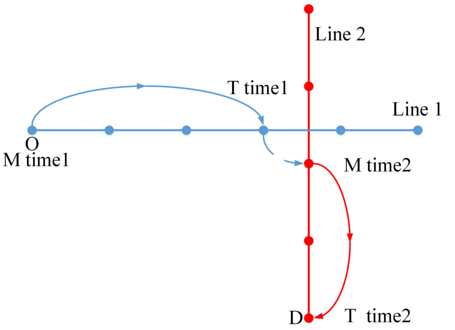


FIGURE 2. Passenger transfer behavior.

M (M hereafter) stands for Mark and T (T hereafter) stands for Trade.

It is worth noting that this method is mainly designed for entry-exit charging system. For entry-only charging system, Reference [27], [29], [31], [32] have proposed some models to estimate the possible drop-off sites, and the follow-up work can be based on these models.

To match the travel chain, the transfer behavior needs to be identified. At present, there are two ideas for screening the transfer behavior, which are to consider the Euclidean distance and time interval between the two rides. For a passenger, if the distance between T station1 and M station2 or the time interval between T time1 and M time2 is lower than the threshold, his or her two rides should be considered as a transfer behavior. And the choice of identifying idea and threshold setting needs to be based on a specific sample.

The specific processing method is as follows: ① the card information table is sorted according to the Card ID; ② starting from the i-th row of the table, create a temporary table with the i-th row as the control row and the i-th to (i+4)-th row (i starts at 1); ③ in the temporary table, the search is based on the control row (i-th row). If the Card ID of a row is found to be the same as the control row, and the distance or time interval is lower than the threshold, the trade time and station of this row will replace the corresponding attribute of the control row, and this row is deleted. There are two considerations for sorting and creating a five-line temporary table. One is to simplify the operation and shorten the calculation time, the other is to consider that few people actually transfer more than five times during the peak period. The flow chart of the travel chain matching is shown in Figure 3.

(2) Matching the travel data with the geographic information. The bus station information table (shown in Table 5) includes the site ID, the line ID, the site name, and the latitude and longitude, etc.

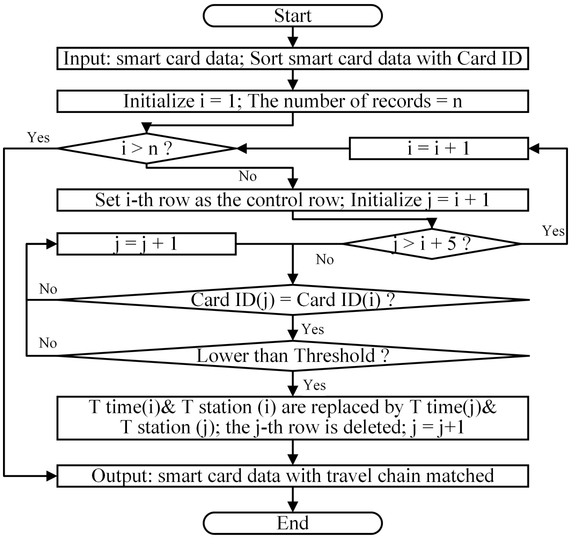

Start
Input: smart card data; Sort smart card data with Card ID
Initialize i = 1; The number of records = n
Yes
i > n ?
i = i + 1
No
Set i-th row as the control row; Initialize j = i + 1
Yes
j = j + 1
j > i + 5 ?
No
No
Card ID(j) = Card ID(i) ?
Yes
No
Lower than Threshold ?
Yes
T time(i)& T station (i) are replaced by T time(j)& T station (j); the j-th row is deleted; j = j+1
Output: smart card data with travel chain matched
End

FIGURE 3. Match of passengers' travel chain.

However, at this step, directly matching is still not available. In order to extract the distribution of passenger demand perfectly, thus to obtain the heat distribution of passengers' OD, it is necessary to aggregate the travel information of all passengers and count their departure and arrival. But if aggregated at the site level, the statistical results will be complicated in the case of so many sites in metropolis. For example, looking for ODs between more than 50,000 sites requires billions of cumulative processes. Further, the results presented on the map are also inconvenient for follow-up processing by the researcher. Therefore, it is considered to select a suitable size range as a TAZ and count the travel OD of each zone. At the same time, the zone needs to be linked to the bus stop to prepare for subsequent matching.

Considering that most passengers will use the bus stop nearby, the bus stop service area is used as the reference for the TAZ. For the service radius (R_s) of the bus station, 400m [68] given in the *Transit Capacity and Quality of Service Manual* is usually used. But in fact, different scholars adopt various R_s. For example, Ingvardson *et al.* [13] used 500m as R_s. Li *et al.* [69] confirmed that it is more reasonable to use 500m as the bus R_s in major cities in China. Alsger *et al.* [26] tested the accuracy of OD extraction with 400m, 800m, 1000m and 1100m as the walking distance. The result indicated that 400m can get very accurate OD, and further increasing the R_s length cannot increase the accuracy a lot. Nunes *et al.* [27] performed sensitivity analysis with 400m, 640m and 1000m walking distance respectively. The results also proved that the increase in R_s length is not significant

TABLE 6. Supplementary bus station information table (partial).

Station ID	Line ID	Site name	Latitude	Longitude	ZID	……
170473890	5819834	Chaoyang xincheng	116.5493	39.94955	12681	……
170476393	5819834	Donggangzi	116.5654	39.97417	10594	……
……	……	……	……	……	……	……

TABLE 7. Matched smart card attributes information (partial).

Card ID	TF	Mark time	Mark station	Trade time	Trade station	Mark zone	Trade zone	……
*****833	0	20170206082344	170473890	20170206085523	170487614	12681	12470	……
*****451	1	20170206082737	170476393	20170206094703	170465055	10594	11224	……
……	……	……	……	……	……	……	……	……

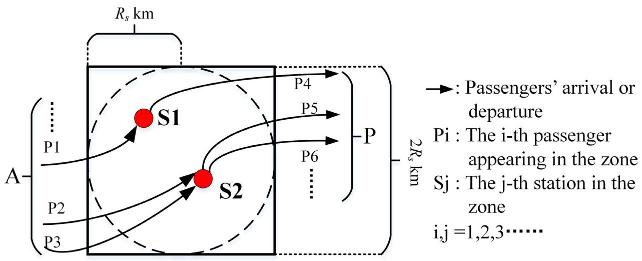

R_s km
P4
P5
P6
S1
S2
P1
P2
P3
A
P
$2R_s$ km
: Passengers' arrival or departure
Pi : The i-th passenger appearing in the zone
Sj : The j-th station in the zone
i,j =1,2,3……

FIGURE 4. PA volume accumulation for every TAZ.

enough for the accuracy of the get-off station estimation, and the accuracy in 400m is high enough. Therefore, the selection of the R_s should be based on the actual situation, and a value near 400m should be selected.

For the convenience of processing, the processing area is divided into $2R_s$km × $2R_s$km lattice area by ArcGIS, and all grids are considered as separate TAZs, that is, the sum of the PA volume of all stations in the grid are taken as the passenger volume of the zone, as shown in Figure 4. This setting can be considered that there is a symbolic station in the middle of each TAZ, which serves the entire TAZ. Since the service radius is R_s, the service area is a circle with a diameter of $2R_s$. And to ensure that the TAZs can cover all the research area, their shape is designed as a square with the length of the diameter of its embedded circle.

All stations are projected into the map by applying GIS tools, and the TAZs to which they belong are confirmed. The TAZ code of each station is added to the station information table (Table 5) as a new attribute. This attribute is taken as ZID and the new table is taken as supplementary station information table. Finally, the passenger travel data and geographic information are matched, by matching the M&T station in the card information table (Table 4) with the same station ID in the supplementary station information table (Table 6). The final supplementary station information table and the matched card information table results are shown in Tables 6 and 7. TF represent the number of transfers.

3) THE CALCULATION OF PASSENGER PA AND ACTIVITY WITHIN EACH TAZ

Based on previews steps, the OD of each TAZ can be accumulated. Through computer programming, firstly, the departure and arrival TAZ of each travel in smart card information table (Table 7) are counted; then according to the TAZ code, the travel PA volume of each TAZ are accumulated. The final processing results are shown in Table 8. Among them, OV and DV are the passenger origin and destination volume of the zone respectively, and the SV is the sum of the first two, which is defined as the activity of the zone, and prepared for the follow-up generation of the commuter OD. Correspondingly, the unit of activity is "passenger".

TABLE 8. OD volume of each zone (partial).

ZID	OV	DV	SV
88	8	3	11
90	27	62	89
……	……	……	……

4) THE EXTRACTION OF THE MAIN COMMUTING OD

As stated in Section II. A, commuters are potential customers of CB. Reliable time and higher comfortability during the travel are important for them. The establishment of the commuting OD matrix helps us to further observe the regular mobility patterns between the TAZs, thus to determine which hot zones meet the CB demand characteristics. This helps to further narrow the range of feasible areas, or PDAs, for CB demand. The OD matrix is generated as follows:

① Merge the morning peak data table (the matched smart card table, ie, Table 7) of one weekdays, retain only three attributes, including Card ID, M ZID, and T ZID. Keep records with different Card ID, and the obtained new data table are used as a passenger table. Compare the daily peak data table with the passenger table, select the passengers that have more than 4 trips in 5 days with same M&T ZID. Finally, all of commuters are selected in this step.

② Select the top-ranked TAZs with the highest zone activity (the SV), then use their zone ID (ZID) as the row and list header of the OD matrix;

③ Accumulate the OD volume of each grid in the OD matrix against the attributes, the M&T ZID, in the commuter table (obtained in step ①).

And the flow chart of generating commuting OD matrix is shown in Figure 5.

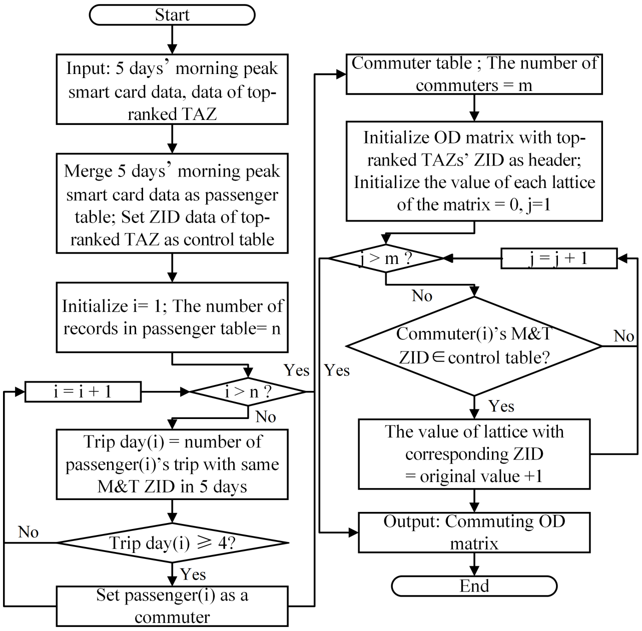

Start
Input: 5 days' morning peak smart card data, data of top-ranked TAZ
Merge 5 days' morning peak smart card data as passenger table; Set ZID data of top-ranked TAZ as control table
Initialize i= 1; The number of records in passenger table= n
i = i + 1
i > n ?
Yes
No
Trip day(i) = number of passenger(i)'s trip with same M&T ZID in 5 days
No
Trip day(i) ⩾ 4?
Yes
Set passenger(i) as a commuter
Yes
Commuter table ; The number of commuters = m
Initialize OD matrix with top-ranked TAZs' ZID as header; Initialize the value of each lattice of the matrix = 0, j=1
j > m ?
j = j + 1
No
Commuter(i)'s M&T ZID∈control table?
No
Yes
The value of lattice with corresponding ZID = original value +1
Output: Commuting OD matrix
End

FIGURE 5. Generation of commuting OD matrix.

The reason why passengers who have more than 4 trips in 5 days with same M&T ZID are regarded as commuters is that, according to the results of reference [10], 75% of commuters have more than 15 trips per month. When generating the OD matrix, it's appropriate to selects only the top-ranked TAZs, rather than all of them, with the highest zone activity as the basic header of the OD matrix, if a city can be divided into a lot of TAZs. The number of selected TAZs is to ensure that the main OD can be outlined while reducing the amount of computation. The structure of the generated OD table is roughly as shown in Table 9 below.

TABLE 9. The main commuters' OD.

ZID	8688	8710	……
8688	**	**	……
8710	**	**	……
……	……	……	……

It should be noted that this method is designed for high card usage volume/rate. As mentioned above (Section I. B), if card usage volume/rate is low (below 10000), the OD extraction results still need to be verified with auxiliary investigations.

B. DATA VISUALIZATION

Using GIS, the geographic data obtained in the previous part can be visualized to generate heat maps and the OD map. The generation method of the heat map is spatial interpolation analysis.

In this paper, the production (OV) and attraction (DV) of each TAZ are taken as the "heat" attribute of the heat map. ① First, generate center point in the middle of each TAZ and assign it "heat" value of the TAZ; ② with the interpolation analysis tool provided by MATLAB, the original "heat" data are supplemented to obtain a new one, so that the discrete heat data become more continuous; ③ arrange and group the original heat data and the new heat data, and dye the corresponding (Red, Green, Blue) RGB color according to its value from small to large; ④ project the heat pattern on the map to generate a heat map.

The purpose of using interpolation analysis is to make the generated image smoother, that is, to supplement the discrete data set as a continuous data set. A smoother map can characterize the spatial distribution of demands better, thus assisting planners in follow-up passenger's pattern analyzing perfectly. This process not only makes the heat map easier to understand, but also preserves the true patterns of the "heat". Because in the follow-up step of confirming which TAZ can be selected into PDA, the heat value of each TAZ's central point is decisive, rather than the points from interpolation.

The generation of OD maps is relatively easy. Because the geo-traffic analysis software, TransCAD, is very mature, the main OD map can be generated by directly importing the OD matrix (Table 9) to it.

C. EXTRACTION OF CB'S DEMAND HOTSPOTS

The overall size of heat map generated by previous step is large, thus a two-step selection is applied to obtain the demand spots, which is the selection of the PDAs and the final demand hotspots.

The selection of the PDA can be performed from analyzing the multi-style maps and selecting hot areas that meet the selection criteria (stated in Table 3). In this step, the heat map is used to judge whether the areas satisfy the D, E & A criterion because it is more convenient and efficient in this respect; and the OD map is more accurate in judging V criterion, so it can be a powerful complement to the heat map as described above (stated in Section I. C).

But the range of potential hot areas selected is still large, and often contains dozens to tens of TAZs, which means that demand hotspots should be explored in areas of dozens to tens km^2. In practical situation, it is feasible to consider road flow, building layout and other factors as demand spot selection criteria, but conducting such research work in such a large area will bring heavy workload for PT planners.

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other clusters [52]. Many sites have already been built in those PDAs. As there are no selection criteria, it is reasonable to use the activity of the station (ie, the SV of Table 8) as the criteria for selection. The calculation of stations' SV is similar to the calculation of TAZs' SV. First, the stations belonging to the PDAs are selected based on Table 6. After that, the number of records in Table 7 which M/T station are same with station ID of selected stations are counted. Then, the OV, DV and SV can be calculated.

Each station in the hot area will be clustered as the analysis object, and the center of each cluster is taken as the final demand hotspots. Because the cluster center is closest to the most active points in the cluster it represents, and it can radiate the entire region relatively uniformly. In other words, the closer to the cluster center, the stronger the CB demand. Therefore, the spatial cluster analysis method is used as the method of demand hotspots extraction in this paper.

In this paper, the k-means method is used for spatial three-dimensional clustering. The three dimensions are the latitude x_i, longitude y_i and activity of the station z_i. The idea is to find k center points in the stations sets, and after repeated iterations, the sum of the cost functions is minimized. The cost function used in this paper is Equation 1.

$$F(C) = \sum_{i=1}^{m} \left[\alpha \times (x_i - c_{xi})^2 + \alpha \times (y_i - c_{yi})^2 + \beta \times (z_i - c_{zi})^2\right] \quad (1)$$

where x_i, y_i, z_i represent the three-dimensional values of the station i, C_{xi}, C_{yi}, C_{zi} represent the three-dimensional values of the nearest cluster center point, m represents the number of station, and α and β respectively represent different weights, which should have $2\alpha + \beta = 1$, $\beta > \alpha$. Since x_i and y_i represent latitude and longitude, they should have the same weight; in actual cases, the more active the station, the higher the probability that it becomes the selected spot, and the weight of the activity should be higher than the former two. However, there are no clear criteria for determining the α and β. And at the data level, there is no clear connection between latitude & longitude and activity itself. Therefore, this paper will use the AFW (Adaptive Feature Weighted) k-means algorithm [70], which can generate reasonable weight values according to the data set itself. The process of assigning weights is as follows:

① Assuming $\alpha = w_x = w_y = \beta = w_z$, the first cluster is executed by Equation 1. K groups are got with $n_1, n_2, \ldots, n_k$ objects in each group;

② Sum of the intraclass distances on the j-th dimension of all groups are calculated by Equation 2. Where j can be x, y or z, and $\bar{j_k}$ is the mean of No.k group on the j-th dimension.

$$d_n = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (j_i - \bar{j_k})^2 \quad (2)$$

③ Sum of the distances between classes on the j-th dimension of all groups are calculated by Equation 3. Where $\bar{j}$ is the mean of all data on the j-th dimension.

$$d_w = \sum_{k=1}^{K} (\bar{j_k} - \bar{j})^2 \quad (3)$$

④ Contribution of the j are calculated by Equation 4.

$$co_j = d_w / d_n \quad (4)$$

The weight of j (w_j) is $co_j / \sum_{j=1}^{3} co_j$. And there must be $w_x = w_y = (1\text{-}w_z)/2$ to keep only one same α.

⑤ The new α and β are put into Equation 1 and the next iteration is executed.

⑥ Repeat ② to ⑤ many times to obtain stable values of α and β. And the stable α and β are put into Equation 1. After many iterations, when there is the smallest F(C), the demand spots are extracted.

IV. CASE STUDY

A. DESCRIPTIONS

Beijing, which is studied in this paper, is an international metropolis with an area of 164.1 million m^2 and a resident population of 21.707 million (2017 dat,[1] provided by the Beijing Municipal Bureau of Statistics). There are currently 53,422 bus stations, mainly located within the 6th Ring Road in Beijing, where is the main research area. Beijing suffers a serious congestion problem, so many studies have attached importance to easing the congestion through the development of PT [10], [16], [35], [45], [71]. It is worth noting that there are many colleges and commercial districts near the 2nd and 3rd Ring Roads in Beijing. And, because Beijing is the capital of China, a large number of government agencies are also widely distributed in this region. The 3rd Ring to the 6th Ring are mainly residential areas. Outside the 6th Ring Road is not urban area, which goes beyond our research. According to the administrative district, Beijing can be divided into 16 districts, including Dongcheng, Xicheng, etc. The six ring roads and administrative regions can be represented by Figure 6.

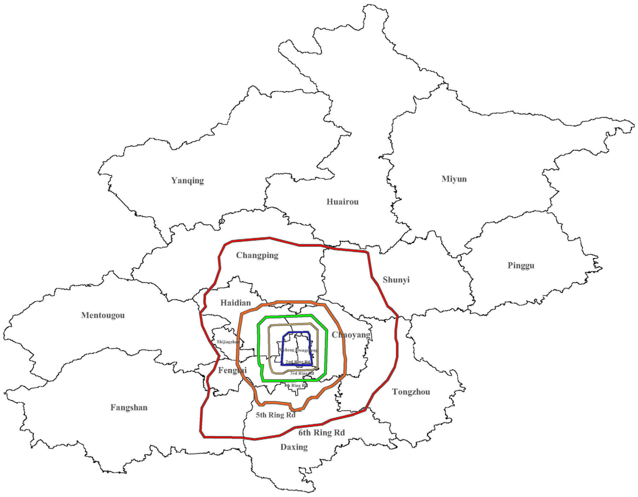
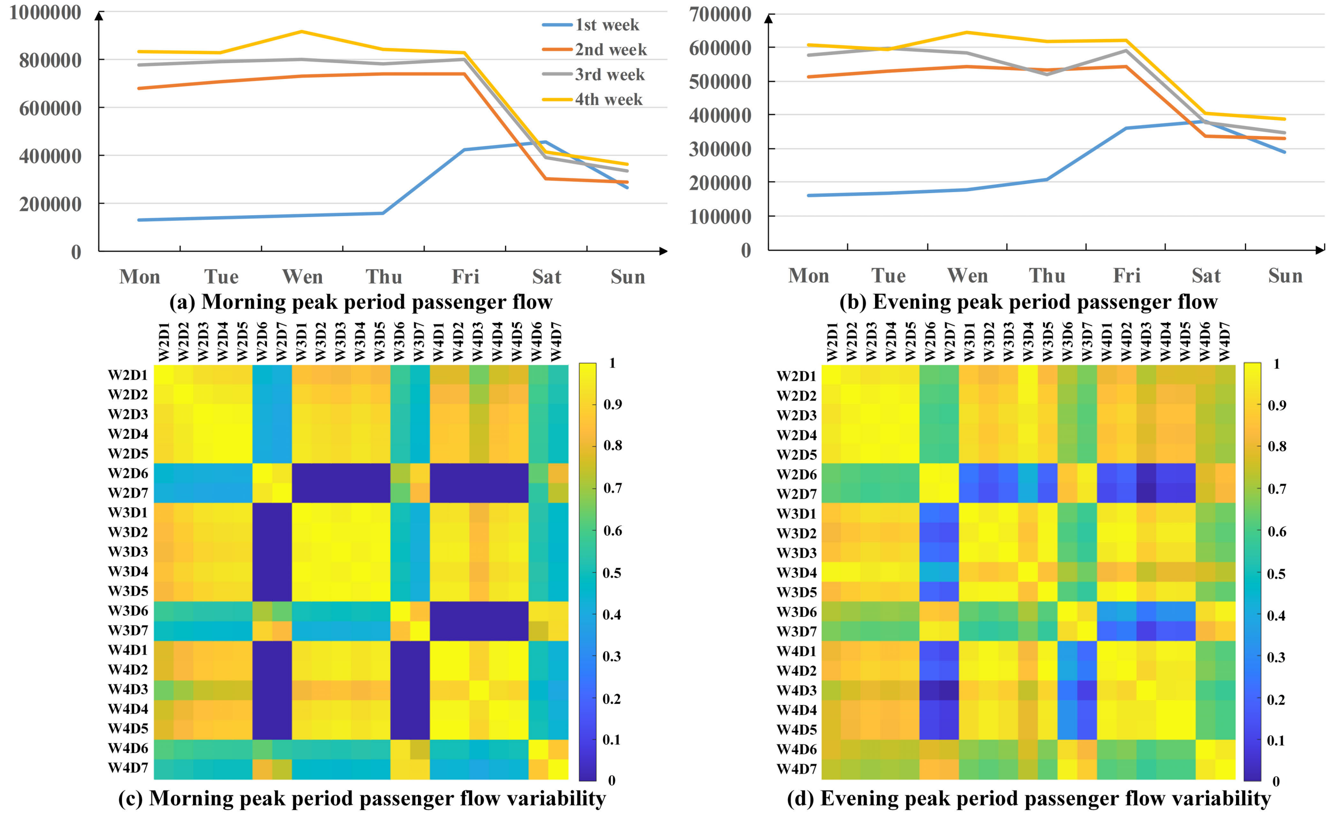


FIGURE 6. Beijing administrative regions and ring roads.

According to the official website, all buses running in Beijing adopt the entry-exit charging system. There are only two payment methods, smart card and cash. Even if the smart card can be set as a season ticket or a monthly ticket, the card holder still needs to tap in and out their card when riding. If a smart card holder does not tap out when getting off the bus, the system will automatically deduct the penalty money at the next ride, which effectively circumvents the fare evasion.

From January 30 to February 26, 2017, the morning and evening peak passenger flow of these four weeks are shown in Figure 7. The travel chain of passengers has been matched. According to Figure 7(a) and (b), except the first week from January 30 to February 2 when most of people are still on vacation because of the most important festival (i.e. Spring Festival) in China, we can see that there are similar travel

[1] http://tjj.beijing.gov.cn



(a) Morning peak period passenger flow

(b) Evening peak period passenger flow

(c) Morning peak period passenger flow variability

(d) Evening peak period passenger flow variability

FIGURE 7. Three week's passenger flow in morning and evening peak.

patterns for other three weeks. To study the commuter's pattern in workdays, the data of 2nd, 3rd and 4th weeks are used to analyze the flow variability further. (c) and (d) show the daily peak flow variability. Variability is the flow rate of change between two days. The calculation method is shown in Equation 5. Flow(i) represents the peak flow on the i-th day, and V(i, j) represents the variability on the i-th day and the j-th day. If Flow(j) is greater than Flow(i), in order to unify the color distribution, their ratio is subtracted by 2 to ensure that the distance of the ratio from 1 is constant.

$$Variability\,(i,j) = \begin{cases} Flow(j)/Flow(i),\; Flow(j) < Flow(i) \\ 2 - Flow(j)/Flow(i),\; Flow(j) \geq Flow(i) \end{cases} \quad (5)$$

The yellower the color of a grid indicates that the two-day flow represented by the horizontal and vertical coordinates is closer. Conversely, if the color is blue, the two-day flow is quite different. As shown in Figure 7(c) and (d), it is clear to find that for workdays, there is little variation in daily flow and the difference between workdays is small. Therefore, the second week's workday data, from February 6 to February 10, are used as research data.

The total number of records we studied is more than 2.5 million, and about 93% of passengers pay with smart card while the rest use the cash, which is enough to support the entire study (Section I. B). Again, the privacy of smart card holders was strictly protected throughout this study. The five days are working days, and there is no rain or snow, and the temperature is not abnormal. Therefore, the demand for PT within these five days is less affected by the external interference, and its regularity better represents the PT demand on the working day.

B. RESULTS AND DISCUSSION

1) PRODUCTION AND ATTRACTION HEAT MAPS FOR MORNING PEAK OF WORKING DAYS

The heat maps can be generated in this section by applying the method given in Section III. However, the paper aims to provide a general framework based on entry-exit charging system, so for a practical case, some settings in mining method need to be adjusted or supplemented accordingly, which includes: (1) Idea of identifying transfer behavior, (2) Bus service radius and (3) the number of top-ranked TAZs.

(1) Idea of identifying transfer behavior (Section III. A. 2). (1)). The amount of every peak period data in case of Beijing is very large, which usually counts above fifty thousand, hence the cumulative calculation time will be long if adopt the idea of Euclidean distance. Therefore, the time interval, that is the interval between T time1 and M time2 of the same card, is chosen to screen travel chain. According to Wang, J. [72], the average bus to bus transfer time is 8.1 minutes and 20% of passenger's total travel time is below 30 minutes. To save the computational time when calculating the time interval, the M&T time (Table 7) need to be rounded to ten. Referring to the morning peak data on February 7, there are 45000 passengers who transfer among 7.1 billion passengers, and 81.35% of their transfer time is within 20 minutes, 87.39% within 30 minutes, and 92.43% within 40 minutes. The differences among the three are small in terms of quantity. So, the time interval is rounded to 30 minutes in this identifying method.

(2) Bus service radius (Section III. A. 2). (2)). Previous studies have shown that a value of around 400m is preferably taken as the R_s (Section III. A. 2). (2)), and since this case using a Chinese city, according to the recommendation of Li *et al.* [69], $R_s = 500$m in this case.

(3) The number of top-ranked TAZs (Section III. A. 3)). The number is set 1000, which has two advantages: on one hand, it can reduce the amount of computation as much as possible, because even if only 1000 zones are selected, there will be one million data in the final matrix; on the other hand, these TAZs almost covers the areas within the 6th Ring Road, which is the main research areas in this case, as shown in Figure 8. The red squares are the top 1000 TAZs, and the blue ones are the other TAZs.

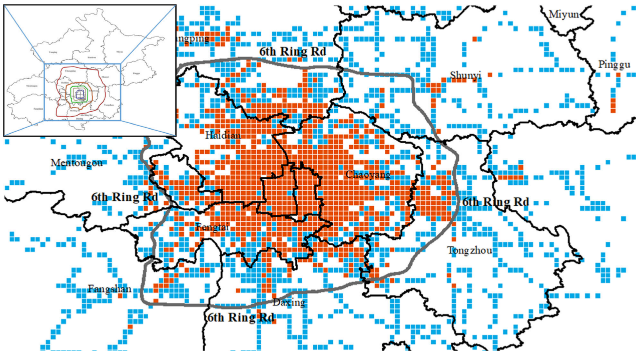


FIGURE 8. The first 1000 TAZs.

The five-day production heat map and the attraction heat map of Beijing in morning peak (7 to 9 a.m.) are plotted, as shown in Figure 9 (a) to (o).

The five figures ((a), (d), (g), (j) and (m)) on the left side of Figure 9 are the PA ratio maps from Monday to Friday, and the figures ((b), (e), (h), (k) and (n)) and figures ((c), (f), (i), (l) and (o)) are production and attraction heat maps of workdays respectively. The x and y axes in the heat maps represent longitude and latitude, respectively. The legend on the lower left side is the legend of the PA ratio maps, and the color bar on the lower right side corresponds to the heat maps. The patterns that can be directly obtained by analyzing the heat maps are as follows:

I. The main area of PT travel is concentrated within the 6th Ring Road of Beijing, and the heat gradually increases as it approaches the city center, but decreases within the 2nd Ring Road;

II. During the peak of the workday, the daily regularity of the passenger travel is basically fixed. In the heat map from Monday to Friday, the overall heat intensity and range of each hot zone does not change much;

III. During the early peak period, the range of the production heat map is larger than the attraction heat map, while the latter is more concentrated.

From these patterns we can summarize some regularities about the travel patterns of bus passenger. For I, the reason why the heat intensity increases as it approaches the city center is because Beijing's urban construction is spread from the interior to the exterior, that is, the closer the area is to the outer ring, the less the population is, and the heat is not as strong as the city center. Naturally, the decrease in heat in the 2nd Ring Road is because those areas are mostly the old city of Beijing. It contains many government agencies and historic sites (such as the Forbidden City, covers 720,000 m^2). There are fewer residents, so commuting travel in this area is not as much as the outer. For II, it reflects that commuting travel accounts for a large proportion of PT trips, because the characteristics, fixed travel time and locations, of commuter determine that the daily variation is not very significant. In addition, the external influence was small, which also led to the stability of PT travel. This phenomenon is also confirmed in the reference [11], that is, the spatial mobility of the smart card user is fixed during the working day. For III, it reflects that the general flow of bus passengers during the morning peak hour is from the external to the internal. In order to further verify this feature, the PA pie charts of the top 1000 activity zone (generate by the method given in Section III. A. 3)) are plotted by ArcGIS, as shown in Figure 9 (a), (d), (g), (j) and (m). As it shows in the new chart sets, the green ratio of the pie chart is larger in the outer zones, while the center zones charts are the opposite. And it reflects bus passengers moving from the exterior to the interior during the morning peak.

The 3D heat map can better show these regularities. By importing the data into ArcGIS and setting a certain angle of view, color and scale, 3D heat maps are generated, as shown in Figure 10. In the figure, the more reddish and the higher the peak, the larger the amount of PA. These two maps illustrate the pattern III described above. The similarity of the heat distribution during the morning peak of the five days is very high. Therefore, Monday heat distribution is chosen as an example.

2) PRODUCTION AND ATTRACTION HEAT MAPS FOR EVENING PEAK OF WORKING DAYS

The evening peak (5 to 7 p.m.) heat maps are generated in the same way, as shown in Figure 11. And Figure 11 is arranged in the same way as Figure 10.

The patterns of the heat maps during the evening peak are as follows:

I. Like the morning peak, the main area of PT is still concentrated within the 6th Ring Road in Beijing, and the

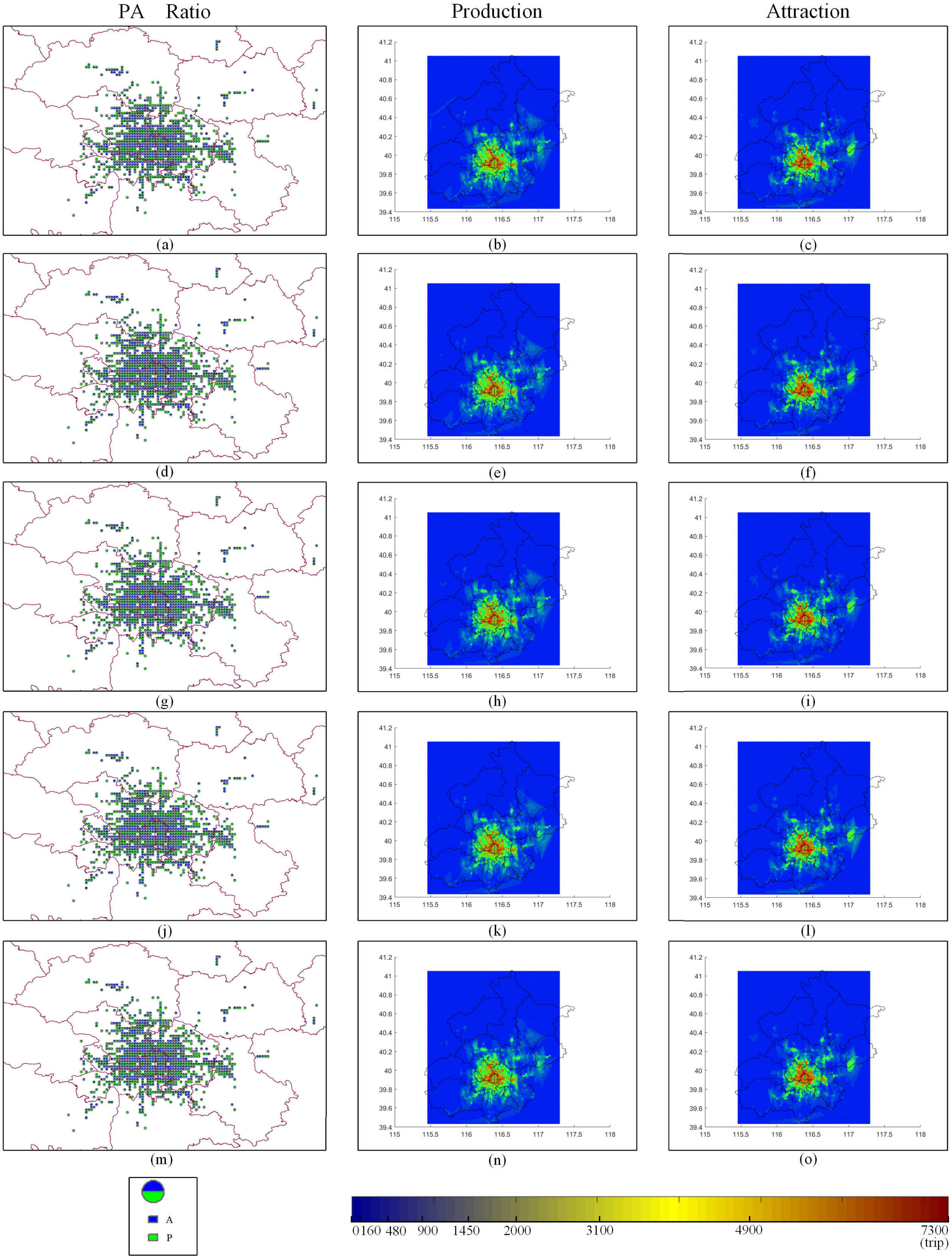
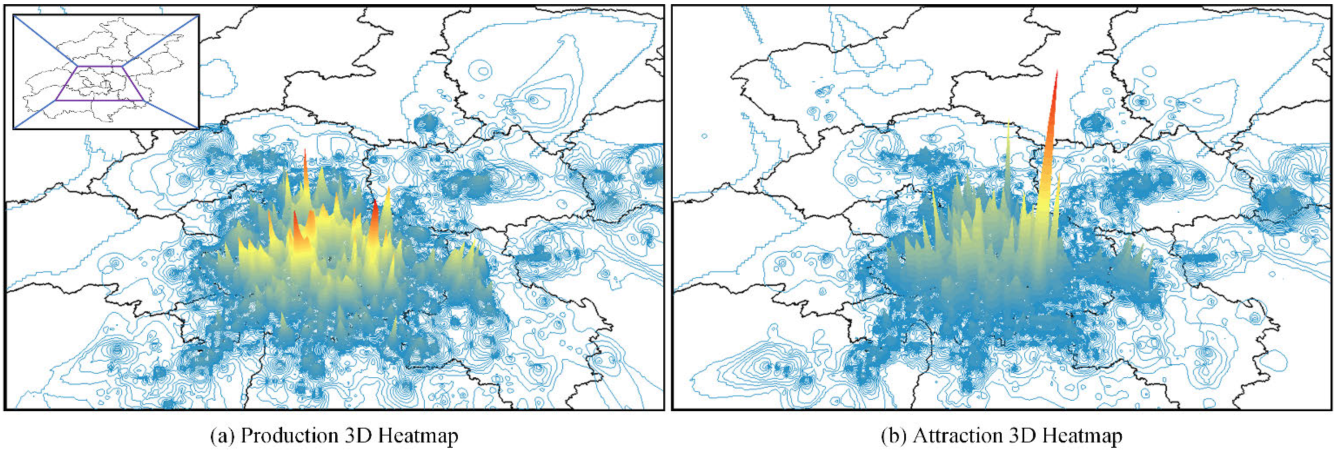


FIGURE 9. Monday-Friday morning peak Production and Attraction heat maps, PA ratio maps of main TAZs.


(a) Production 3D Heatmap

(b) Attraction 3D Heatmap

FIGURE 10. **Morning peak 3D heat map.**

heat gradually increases as it approaches the city center, but decreases within the 2nd Ring Road;

II. The overall intensity and range of travel in the first four days are still small, but the travel intensity on Friday is larger than in previous days;

III. Comparing with the morning peak, during the evening peak period, the range of the attraction heat map is larger than the production heat map. But the difference between production and attraction heat is smaller than the morning peak period.

We can further get more conclusions about the regularities of the passenger patterns. For I, the summary is the same as previously described. For II, considering that there will be a weekend after Friday afternoon, people are more enthusiastic about hanging out and visiting at this time, so heat intensity and scope is bigger than the previous workdays afternoon. For III, it reflects the different patterns of the evening peak and the morning peak. At the morning peak, people need to go to work on time and the travel time is relatively concentrated, so the heat difference between production and attraction is obvious. On the contrary, when people get off work in the afternoon, people no longer care more for going back in time. Travel time, even OD, are no longer concentrated as in the morning. Some people may need to go to other places for shopping, gatherings, etc. These factors have led to a small gap in the distribution of heat intensity between production and attraction during the evening peak. To further demonstrate these patterns, the production and attraction pie charts of the top 1000 activity zone are plotted by ArcGIS, as shown in Figure 11, (a), (d), (g), (j) and (n). From these figures, it can be observed that the green ration in outer zones is no longer as large as they are in the morning peak, and the ratio of blue and green is relatively even. In fact, the blue-green ratio in the central area is more uniform, and only part of zones where blue ratio is larger. This reflects the fact that during the evening peak, the flow of bus passengers is more dispersed, rather than simply the opposite of the morning peak.

Similarly, the 3D heat maps of the evening peak can also be generated, as shown in Figure 12. These figures reconfirm the above-mentioned Pattern III and its related explanation.

3) MAIN POTENTIAL HOT AREAS OF CB DEMAND

The next step is to identify the main CB PDAs for further hot spots searching. The selection criteria of PDA are stated in Table 3. Similarly, 4 criteria need to be quantified with referring to the case.

a: DISTANCE CRITERION (D)

According to the *2018 China Metropolis Commuting Research Report*(provided by AURORA company),[2] the average commuting distance in Beijing is 13.2km, of which 68.2% of passengers have commuting distance greater than 5km. Therefore, for Beijing, setting the "$CR_D \geq 5$km" as D criterion is reasonable, which can guarantee a sufficient length and number of potential passengers.

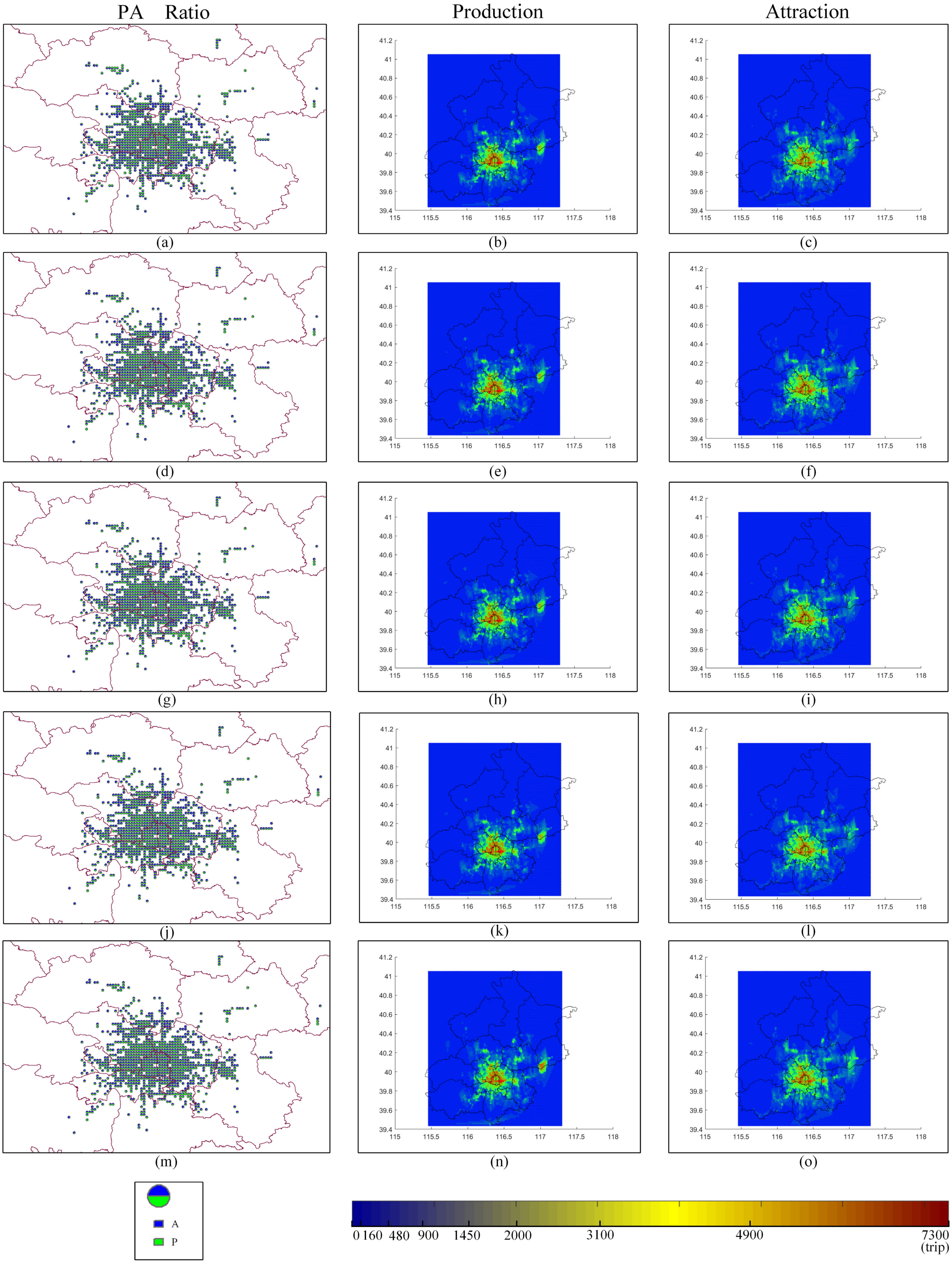
b: ECONOMICS CRITERION (E)

According to data from the official website of the Beijing Public Transport Corporation,[3] the average price of CB is 10 yuan per person per trip. Then a passenger will spend about $10 \times 2 \times 22 \times 11 = 4840$ yuan on CB a year. And according to the data from the Beijing Municipal Bureau of Statistics, the per capita housing cost was 13,347 yuan in 2017. If CR_E accounts for 45% of personal total revenue is "affordable", then PCDI should be no less than $(13347+4840)/0.45 \approx 40416$ yuan. Therefore, for Beijing, it is reasonable to set "PCDI$\geq$40416 yuan" as the E criterion.

c: ACTIVITY CRITERION (A)

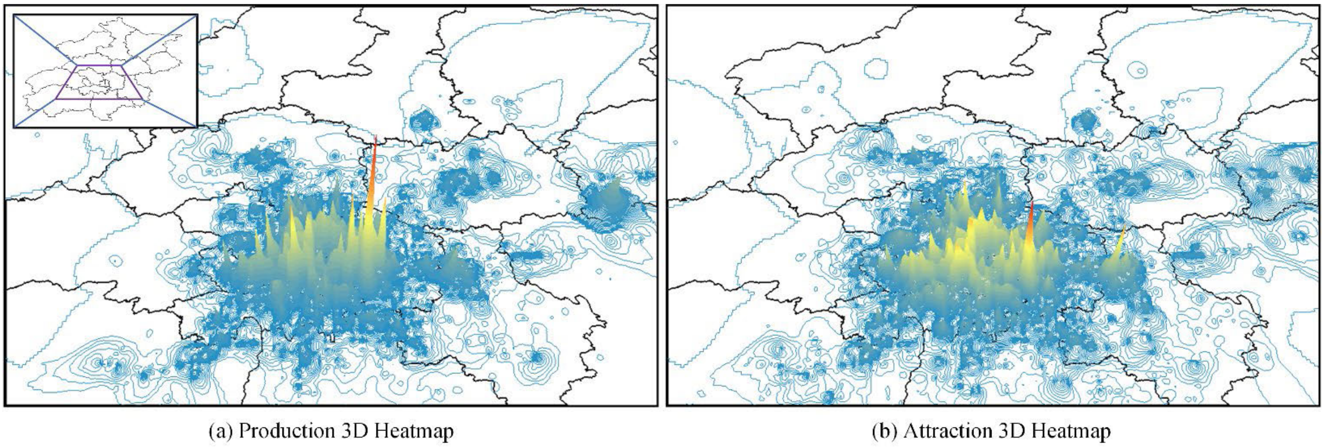
Also based on the introduction from the Beijing Public Transport Corporation, the CB's capacity is 26-30 passenger. And a route will be set when there are 20 or more subscribers.

[2] https://mp.weixin.qq.com/s/AiSrTLEZBwJk6t87wnHaOw
[3] http://www.bjbus.com/home/index.php



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (o)

FIGURE 11. Monday-Friday evening peak Production and Attraction heat maps, PA ratio map of main TAZs.


(a) Production 3D Heatmap

(b) Attraction 3D Heatmap

FIGURE 12. Evening peak 3D heat map.

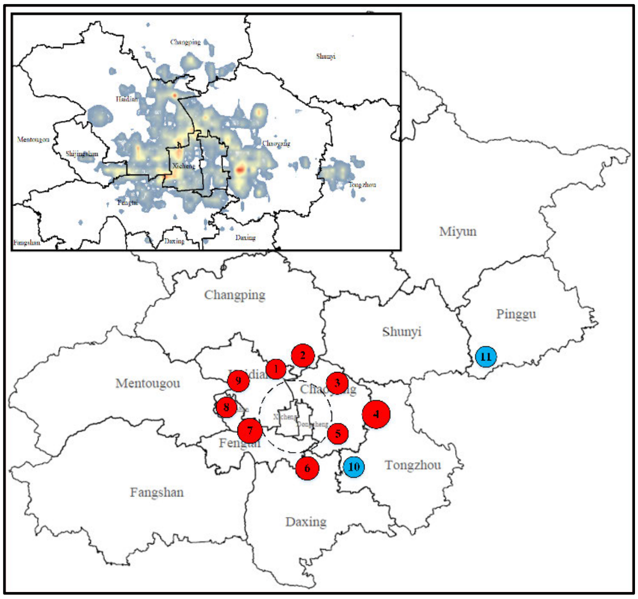


FIGURE 13. Potential demand hot areas.

According to the reference [10], about 65% of passengers are commuters in Beijing in workdays. Since there is currently no code or survey that can be used to refer for f_{cb}, the ratio of the CB route's number to the total number of bus route is used as reference. Beijing currently has 1,266 bus routes, of which 144 are CB routes. Based on all of the above, the C_v, f_{cb}, f_{co} are set as 20, 0.114, 0.65 respectively. So, CR_A should be greater than $20/0.114/0.65 \times 2 \approx 540$ passenger (The peak period lasts two hours). But it will undoubtedly narrow the area of PDA, which is not conducive to the follow-up hotspots extraction. "$CR_A \geq 500$ passenger" is set as A criterion by referring to the previous heat maps.

d: OD VOLUME CRITERION (V)

Since the passenger flow should be greater than 20 or C_v/f_{co}, and $20/0.65 \approx 30$. "$CR_V \geq 30$ passenger per hour" is a rational set for V criterion.

TABLE 10. Selection criteria for PDA in Beijing.

Name	Recommeded Value in case of Beijing
Distance (D)	CR_D ≥5 km
Economics (E)	PCDI ≥40416 yuan
Activity (A)	CR_A ≥500 passenger
OD Volume (V)	CR_V ≥30 passenger per hour

Based on the above, the quantified selection criteria for PDA are shown in the Table 10.

The heat map shows that commuters in external city zones are quite abundant, especially during the morning peak hours, so the point-to-point service is very meaningful for passengers in these zones. Considering the passenger patterns observed from the heat maps, the identified demand hotspot should be in areas far from the city center and with high heat intensity. Because the heat maps illustrate that passengers in these areas need to travel to and from the city center and these areas during peak hours, that means these passengers have long travel distances and high requirements for time reliability. Combining with the selection criteria and passengers' pattern, nine areas meeting D, E & A criterion can be temporarily selected, as shown in Figure 13 and Table 11. PCDI data come from the Beijing Municipal Bureau of Statistics.

Nine red circles represent nine selected areas. The dotted line circle in the middle of Figure 13 has a radius of 10km, and the small image in the upper left corner is the area where the activity is higher than 500. The reason why the radius of the dotted line circle being 10km is there are fewer commuting within the 2nd Ring Road, which is the passengers' pattern from previous analysis. This is equivalent to taking the area around the 2nd Ring Road as the "city center area" and expanding it with 5km.

These two settings are used to ensure that D& A criterion are met, while facilitating the selection of potential hot zones.

To further verify the rationality of the PDA selection, the "SV/ stations number" of each TAZ is defined as the station load degree of the TAZ, and bus station load degree map can also be generated by ArcGIS, as shown in Figure 14.

TABLE 11. Potential demand hot areas.

Number	Distance from city center (km)	PCDI (yuan)	Area center Activity (trips)	Location of PDA
1	17	58861	Above 3200	Southeastern part of the junction of Haidian and Changping District
2	17	45735	Above 1600	Southernmost part of Changping District, border with Haidian and Chaoyang District
3	12	64841	Above 2200	Northeastern part of Chaoyang District, border with Shunyi District
4	22	44607	Above 1600	Northwestern part of Tongzhou District, border with Chaoyang District
5	13	64841	Above 1900	South-central part of Chaoyang District
6	15	47572	Above 750	Eastern part of the junction of Daxing and Fengtai District
7	15	64656	Above 1900	Junction part of Shijingshan, Fengtai and Haidian District
8	22	66112	Above 1300	North-central part of Shijingshan District
9	20	71986	Above 1300	West-central part of Haidian District
10	22	44607	About 300	Western part of Tongzhou District, border with Daxing Distrcit
11	54	41130	About 270	Southeastern part of Pinggu District

TABLE 12. Qualified potential CB demand hot areas and their matching areas.

Number	Distance from matching area (km)	OD volume (passenger)	Qualified? (Yes / No)	Location of matching area
1	6.2	85	Yes	Northern part of the junction of Xicheng and Dongcheng District
2	10.0	346	Yes	Northern part of Dongcheng and Chaoyang District
3	7.3	62	Yes	Northern part of Dongcheng District, border with Chaoyang District
4	10.2	625	Yes	Southwest-central part of Chaoyang District
5	4.3	226	No	/
6	3.1	90	No	/
7	9.2	177	Yes	Southwestern part of Fengtai District
8	4.7	193	No	/
9	4.1	102	No	/
10	3.4	68	No	/
11	7.7	95	Yes	7.7km northeast of 11

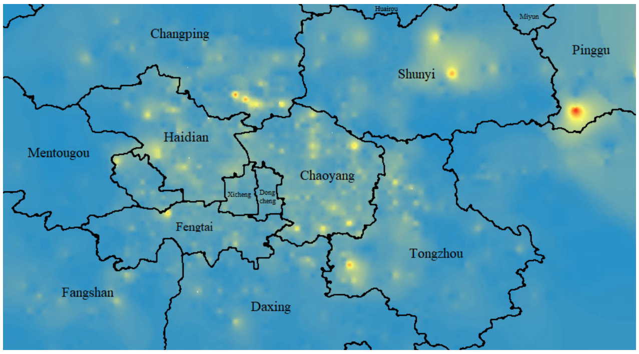


FIGURE 14. Bus station load degree map.

It can be seen from the map that the selection of the nine potential hot areas is consistent with the area where the bus station is heavily loaded, so the previous choice is reasonable. At the same time, there are three other areas that can be observed with high degree of station load. They are in the western part of Tongzhou District, the central part of Shunyi District and the southwestern part of Pinggu District. Although none of their activity degree is greater than 500, it does not mean that they are not likely to meet the V criterion, which is more important than A criterion (stated in Section II. B). And only the commuting OD map can be used to judge whether the V criterion is met. Since they all meet the D criterion, it is also reasonable to select them as PDA temporarily, except for the one in Shunyi District, whose PCDI does not meet the E criterion. They are presented by blue circles in Figure 13.

4) COMMUTING OD FOR EXPLORING POTENTIAL CB ROUTES

Far-reaching, high-traffic passenger flows have a greater potential for CB according to its demand characteristics. To judge whether those PDAs meet the V criterion, obtaining the OD between TAZs is necessary. This helps determine which hot areas have a greater potential. Moreover, it will help transit agencies make decisions to planning the CB service routes. As mentioned above, CB route setting is flexible. Therefore, the setting work only needs to determine the start and end areas of the route. And the two areas are the PDA and its matching area. Therefore, the next step is to analyze the commuting OD map to determine which PDAs are qualified and to find their matching area. The commuting OD map of the top 1000 active TAZs can be generated by the method given in Section III. B, as shown in Figure 15.

In this figure, the darker the color and the wider the width of the desired line, the greater the passenger flow it represents. So, it's easy to tell how the passenger mobile from the picture. By analyzing the OD map, whether the previous PDAs meet the V criterion can be judged. Surprisingly, some hot areas are not even able to meet the D criterion. Because the OD map not only shows the passenger volume of these potential hot areas and their matching areas (ie, the other end of the OD line), but also shows the length and direction of the OD lines. For example, the hot area 5, which has an activity of 1600 and 13km away from the city center, is very close to its matching area. They are only

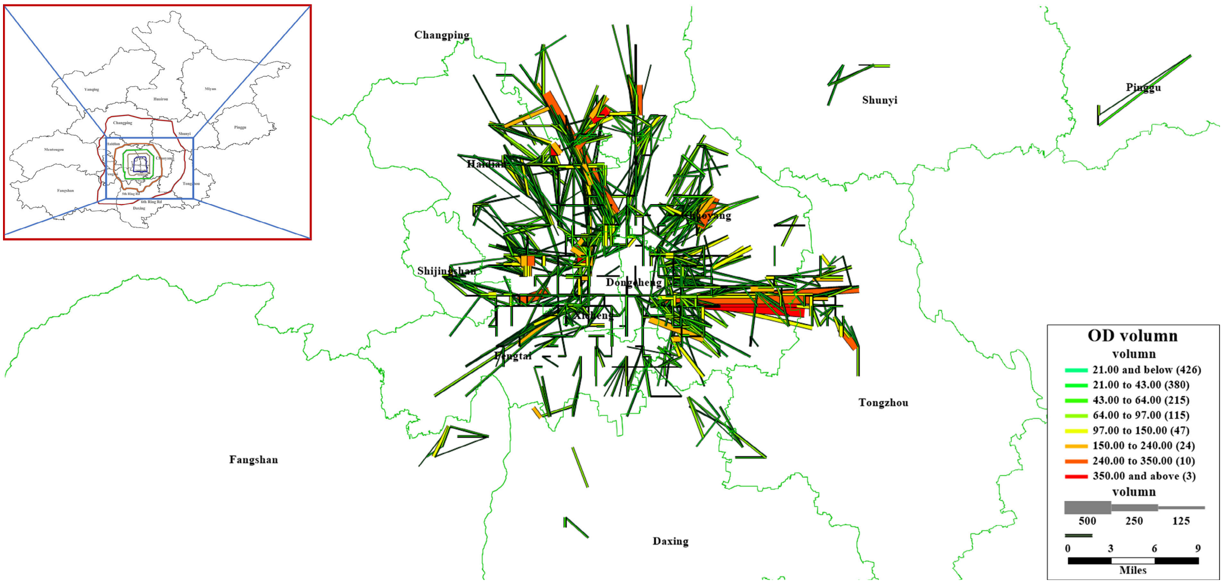


FIGURE 15. The commuting OD map.

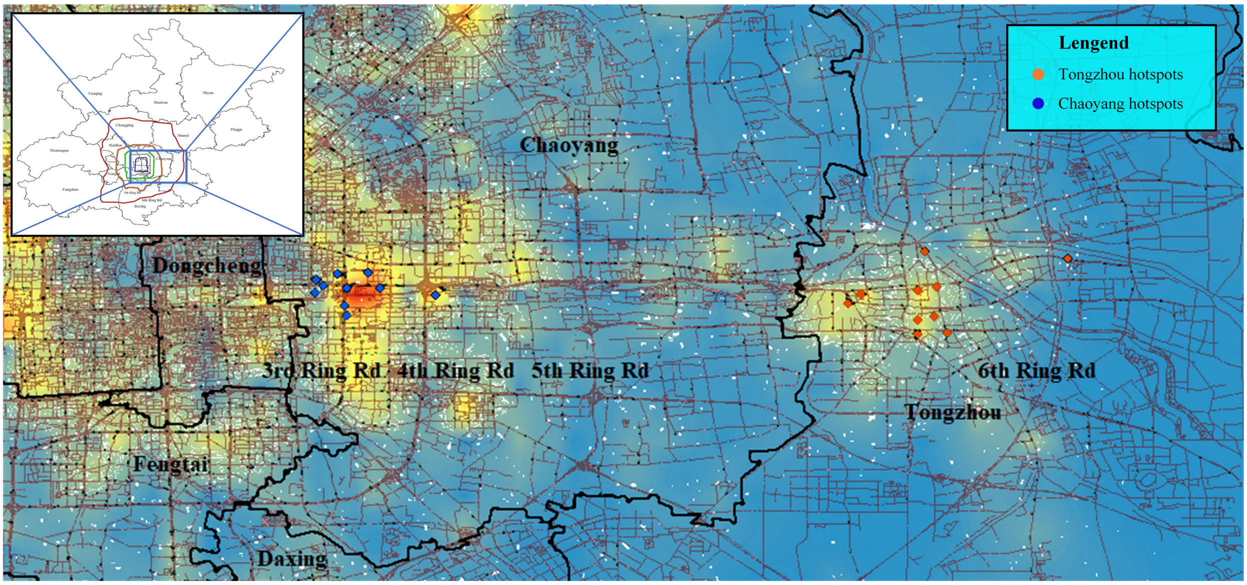


FIGURE 16. 10 hotspots in Chaoyang and Tongzhou District.

4.3km apart. Therefore, the hot area 5 does not meet the D criterion. But the hot area 11 is unexpectedly met. The PDAs that ultimately qualified is shown in the Table 12 below. And their matching areas are shown in the table correspondingly.

5) SPATIAL CLUSTER ANALYSIS RESULTS

The final step is to extract the hotspots from the PDAs by using the clustering method described in Section III. C.

The Tongzhou PDA (4 in Figure 9) and its matching area is a good example. They are more than 10.2 km apart and the passenger flow between them is $625/2 \approx 313$ passenger per hour. The Tongzhou District, where the PDA 4 belongs to, was planned as the administrative sub-center of Beijing as early as 2015,[4] meaning that a large number of residents will move to this place. And its matching area is located in the Beijing CBD, where is home of a number of advanced companies like GM and Deutsche Bank, meaning people in here being in good financial condition.

Assuming 30 people per bus (capacity is 26-30), setting ten stations will be adequate. Therefore, k should take 10, that is, ten hotspots in each area will be generated. By applying the AFW k-means, α and β are taken as 0.015 and 0.97. After using MATLAB to execute 5000 clustering treatments (or iteration), the cost value of Chaoyang and Tongzhou District are stable below 43 million and 3.3 million respectively. The clustering result is acceptable. The results are shown

[4] https://en.wikipedia.org/wiki/Tongzhou_District,_Beijing

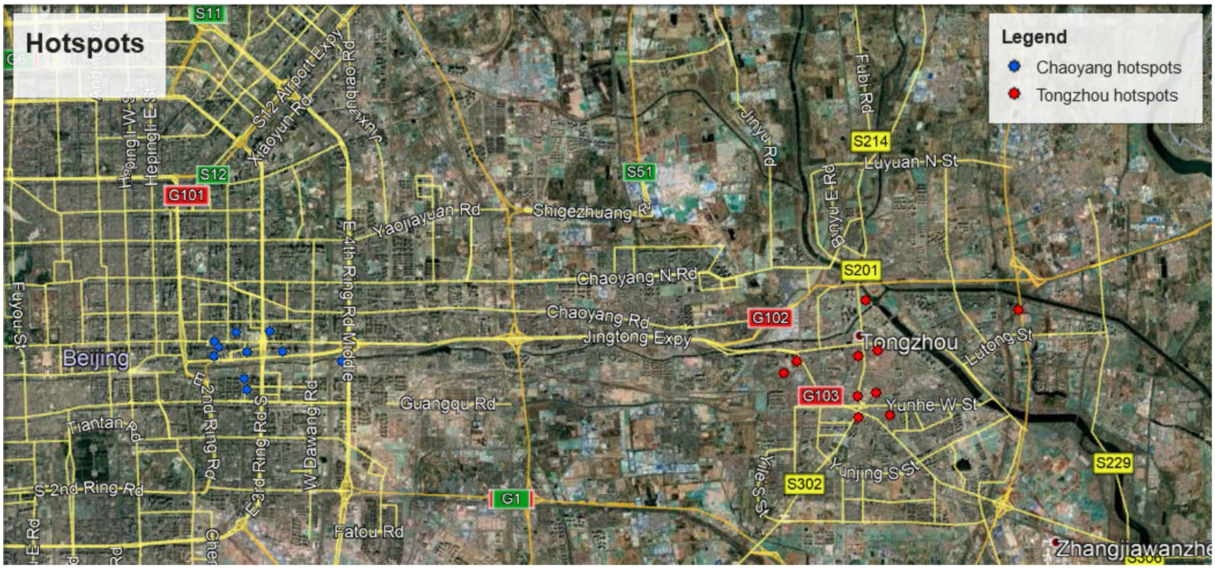


FIGURE 17. 10 hotspots in Chaoyang and Tongzhou District (satellite).

TABLE 13. 10 hotspots in Chaoyang and Tongzhou District.

Number	Chaoyang District		Tongzhou District	
	longitude	latitude	longitude	latitude
1	116.44588	39.90486	116.63172	39.89015
2	116.46818	39.90599	116.63597	39.90366
3	116.45317	39.91079	116.65842	39.91883
4	116.45671	39.90588	116.65598	39.90490
5	116.46404	39.91106	116.65576	39.89488
6	116.48751	39.90357	116.63172	39.90067
7	116.45571	39.89923	116.66178	39.89575
8	116.45656	39.89648	116.66230	39.90629
9	116.44743	39.90710	116.65603	39.88960
10	116.44603	39.90853	116.70826	39.91635

in Table 13 and Figure 16 below. And Figure 17 shows the spots in satellite imagery.

The blue and red spots in the maps are the demand spots in Chaoyang District and Tongzhou District. The small black dots are the existing stations. As can be observed from the heat map (Figure 14), the selected demand spots are close to the hot area in each PDA they belong to, and are also close to the existing site location, which facilitates follow-up site planning. As shown in the satellite imagery (Figure 15), these spots are also close to the main working areas and residential areas of each region. The above results all indicate that the selection results are acceptable.

V. CONCLUSION

To extract CB demand spots from a large amount of smart card data, the multi-step methodology, which incorporates self-designed data preprocessing algorithm, interpolation analysis, spatial clustering algorithm and GIS tools, is proposed in this paper. The contributions of this method include: (1) Massive smart card data are reasonably organized and utilized, while the calculation speed is guaranteed, rather than small amount of calculations [42], [49], which is very practical. (2) The CB demand characteristics and selection criteria of PDA are defined. It can not only help the PDA selection in this research, but also promote CB demand analysis in future researches. (3) The passenger spatial-temporal distribution patterns and mobility are clearly characterized on observing multi-day and multi-style maps, including OD map, heat maps, 3D heat maps and PA ratio maps. (4) Accurate and rational extraction from macro to micro is achieved. And a common framework is provided, which facilitates the migration of the methodology on various cities which adopt the entry-exit charging system.

In view of the good performance of proposed method in the case study of Beijing, four recommendations are proposed to the metropolitan PT planners. First, GIS should be widely used in the analysis of PT data including smart card data. Multi-style maps make it easy for PT planners to quickly narrow their focus from the entire city to a smaller region, as well as to observe passenger patterns and connections between areas. Hence it can give very positive effect on improving the digitization and information of the city's Intelligent Transportation System (ITS). Second, the bus and metro operators should enhance their contact and share each other's data to rationally allocate available PT resources. Third, in the initial stage of the CB system establishment, the criteria for PDA selection should be strictly adhered to. But the criteria can be gradually lowered in the later stage, so that more hotspots can be provided to consumers, thus attracting more potential users as much as possible. For tourist cities, during the peak season, this method can also be used to quickly establish CB sites so as to serve foreign tourists.

However, the research in this paper is still relatively basic, and there is still room for further improvement. First, the fusion of AFC data from bus and subway. In fact, the bus-subway and subway-bus transfer modes are common during commuting. The fused data can identify more commuters and their OD accurately, which facilitates follow-up research. Second, the spatial clustering method considers only one attribute, ''activity'', as the third dimension. More attributes can be added to it as well, such as POI, housing prices, bus GPS data, etc. [12], [15], [16], thus improving the accuracy of the extraction. Third, due to the limited data collected, weather, events, and holiday impacts [50] have not yet been incorporated into the methodology. The inclusion of these influence factors helps to make the methodology more general. Fourth, the discussion of the impact of the proportion will be one of the focuses in future research, because the use of mobile payment will inevitably affect the original AFC system, which is the source of data using in proposed method.

ACKNOWLEDGMENT

The authors would like to thank the editors and reviewers for their insightful comments and valuable suggestion to improve the quality of our manuscript. The authors have declared no conflicts of interest to this work. Authors would like to thank the anonymous reviewers for spending time to review our paper.

REFERENCES

[1] T. Liu and A. Ceder, "Analysis of a new public-transport-service concept: Customized bus in China," *Transp. Policy*, vol. 39, pp. 63–76, Apr. 2015.

[2] T. Liu, A. A. Ceder, R. Bologna, and B. Cabantous, "Commuting by customized bus: A comparative analysis with private car and conventional public transport in two cities," *J. Public Transp.*, vol. 19, no. 2, p. 4, 2016.

[3] J. Zhang, D. Z. Wang, and M. Meng, "Analyzing customized bus service on a multimodal travel corridor: An analytical modeling approach," *J. Transp. Eng., A, Syst.*, vol. 143, no. 11, 2017, Art. no. 4017057.

[4] L. C. Tong, L. Zhou, J. Liu, and X. Zhou, "Customized bus service design for jointly optimizing passenger-to-vehicle assignment and vehicle routing," *Transp. Res. C, Emerg. Technol.*, vol. 85, pp. 451–475, Dec. 2017.

[5] W. Huang, W. Jin, J. Huang, and B. Han, "Pricing problem of customized bus under different market strategies," *J. Guangxi Normal Univ. (Natural Sci. Ed.)*, vol. 2, p. 2, Feb. 2018.

[6] M. Bagchi and P. White, "What role for smart-card data from bus systems?" *Municipal Eng.*, vol. 157, no. 1, pp. 39–46, 2004.

[7] P. T. Blythe, "Improving public transport ticketing through smart cards," *Municipal Eng.*, vol. 157, no. 1, pp. 47–54, 2004.

[8] D. Dempsey and P. Stephen, *Privacy Issues With the Use of Smart Cards*, 2007.

[9] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 4, pp. 557–568, 2011.

[10] X. Ma, C. Liu, H. Wen, Y. Wang, and Y. Wu, "Understanding commuting patterns using transit smart card data," *J. Transp. Geogr.*, vol. 58, pp. 135–145, Jan. 2017.

[11] C. Zhong, E. Manley, S. M. Arisona, M. Batty, and G. Schmitt, "Measuring variability of mobility patterns from multiday smart-card data," *J. Comput. Sci.*, vol. 9, pp. 125–130, Jul. 2015.

[12] Q. L. Gao, Q. Q. Li, Y. Yue, Y. Zhuang, Z. P. Chen, and H. Kong, "Exploring changes in the spatial distribution of the low-to-moderate income group using transit smart card data," *Comput., Environ. Urban Syst.*, vol. 72, pp. 68–77, Nov. 2018.

[13] J. B. Ingvardson, O. A. Nielsen, S. Raveau, and B. F. Nielsen, "Passenger arrival and waiting time distributions dependent on train service frequency and station characteristics: A smart card data analysis," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 292–306, May 2018.

[14] X. Jie, Z. He, G. Wei, and B. Ran, "Optimal timetable development for community shuttle network with metro stations," *Transp. Res. C, Emerg. Technol.*, vol. 60, pp. 540–565, Nov. 2015.

[15] W. Tu, R. Cao, Y. Yue, B. Zhou, Q. Li, and Q. Li, "Spatial variations in urban public ridership derived from GPS trajectories and smart card data," *J. Transp. Geogr.*, vol. 69, pp. 45–57, May 2018.

[16] G. Qi, A. Huang, W. Guan, and L. Fan, "Analysis and prediction of regional mobility patterns of bus travellers using smart card data and points of interest data," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1197–1214, Apr. 2019.

[17] A. Delbosc and G. Currie, "Why do people fare evade? A global shift in fare evasion research," *Transp. Rev.*, vol. 39, no. 3, pp. 376–391, 2019.

[18] A. V. Reddy, J. Kuhls, and A. Lu, "Measuring and controlling subway fare evasion: Improving safety and security at New York City transit authority," *Transp. Res. Rec.*, vol. 2216, no. 1, pp. 85–99, 2011, doi: 10.3141/2216-10.

[19] B. Barabino and S. Salis, "Moving towards a more accurate level of inspection against fare evasion in proof-of-payment transit systems," *Netw. Spatial Econ.*, vol. 19, no. 4, pp. 1319–1346, 2019.

[20] M. Trépanier, S. Barj, C. Dufour, and R. Poilpré, "Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain," in *Proc. Congr. l'Association Transp. Canada*, 2004.

[21] P. Guarda, P. Galilea, L. Paget-Seekins, and J. de Dios Ortúzar, "What is behind fare evasion in urban bus systems? An econometric approach," *Transp. Res. A, Policy Pract.*, vol. 84, pp. 55–71, Feb. 2016.

[22] S. Robinson, B. Narayanan, N. Toh, and F. Pereira, "Methods for pre-processing smartcard data to improve data quality," *Transp. Res. C, Emerg. Technol.*, vol. 49, pp. 43–58, Dec. 2014.

[23] B. Barabino, M. D. Francesco, and S. Mozzoni, "An offline framework for handling automatic passenger counting raw data," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 6, pp. 2443–2456, Dec. 2014.

[24] N. Nassir, A. Khani, S. G. Lee, H. Noh, and M. Hickman, "Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system," *Transp. Res. Rec.*, vol. 2263, no. 1, pp. 140–150, 2011, doi: 10.3141/2263-16.

[25] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach," *Transp. Res. C, Emerg. Technol.*, vol. 46, pp. 179–191, Sep. 2014.

[26] A. Alsger, B. Assemi, M. Mesbah, and L. Ferreira, "Validating and improving public transport origin–destination estimation algorithm using smart card fare data" *Transp. Res. C Emerg. Technol.*, vol. 68, pp. 490–506, Jul. 2016.

[27] A. A. Nunes, T. G. Dias, and J. E. F. Cunha, "Passenger journey destination estimation from automated fare collection system data using spatial validation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 1, pp. 133–142, Jan. 2016.

[28] V. Nagy, "Theoretical method for building OD matrix from AFC data?" *Transp. Res. Procedia*, vol. 14, pp. 1802–1808, Jan. 2016.

[29] M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *J. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 1–4, 2007.

[30] T. Li, D. Sun, J. Peng, and K. Yang, "Smart card data mining of public transport destination: A literature review," *Information*, vol. 9, no. 1, p. 18, Jan. 2018.

[31] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from Santiago, Chile," *Transp. Res. C, Emerg. Technol.*, vol. 24, no. 9, pp. 9–18, 2012.

[32] M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transp. Res. C, Emerg. Technol.*, vol. 44, no. 4, pp. 70–79, 2014.

[33] S. K. Fayyaz, X. Liu, and R. J. Porter, "Genetic algorithm and regression-based model for analyzing fare payment structure and transit dwell time," *Transp. Res. Rec.*, vol. 2595, no. 1, pp. 1–10, 2016.

[34] S. K. Barai, "Data mining applications in transportation engineering," *Transp.-Vilnius*, vol. 18, no. 5, pp. 216–223, 2003.

[35] Z. He, L. Zheng, P. Chen, and W. Guan, "Mapping to cells: A simple method to extract traffic dynamics from probe vehicle data," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 32, no. 3, pp. 252–267, 2017, doi: 10.1111/mice.12251.

[36] Z. He, Y. Lv, L. Lu, and W. Guan, "Constructing spatiotemporal speed contour diagrams: Using rectangular or non-rectangular parallelogram cells?" *Transp. B, Transp. Dyn.*, vol. 7, no. 1, pp. 44–60, 2019, doi: 10.1080/21680566.2017.1320774.

[37] S. J. Lee and K. Siau, "A review of data mining techniques," *Ind. Manage. Data Syst.*, vol. 101, no. 1, pp. 41–46, 2001.

[38] M. Friendly and D. J. Denis. (2001). *Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*. [Online]. Available: http://www.datavis.ca/milestones

[39] C. E. Fritz, N. Schuurman, C. Robertson, and S. Lear, "A scoping review of spatial cluster analysis techniques for point-event data," *Geospatial Health*, vol. 7, no. 2, pp. 183–198, 2013.

[40] L. Wilkinson and M. Friendly, "The history of the cluster heat map," *Amer. Statistician*, vol. 63, no. 2, pp. 179–184, 2009.

[41] C. Yu and Z.-C. He, "Analysing the spatial-temporal characteristics of bus travel demand using the heat map," *J. Transp. Geogr.*, vol. 58, pp. 247–255, Jan. 2017.

[42] N. B. Stoll, T. Glick, and M. A. Figliozzi, "Using high-resolution bus GPS data to visualize and identify congestion hot spots in urban arterials," *Transp. Res. Rec.*, vol. 2539, no. 1, pp. 20–29, 2016.

[43] J. Wood, J. Dykes, and A. Slingsby, "Visualisation of origins, destinations and flows with OD maps," *Cartograph. J.*, vol. 47, no. 2, pp. 117–129, 2010.

[44] G. Andrienko, N. Andrienko, G. Fuchs, and J. Wood, "Revealing patterns and trends of mass mobility through spatial and temporal abstraction of origin-destination movement data," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 9, pp. 2120–2136, Sep. 2017.

[45] K. C. Clarke, "Advances in geographic information systems," *Comput., Environ. Urban Syst.*, vol. 10, nos. 3–4, pp. 175–184, 1986.

[46] D. O'Sullivan, A. Morrison, and J. Shearer, "Using desktop GIS for the investigation of accessibility by public transport: An isochrone approach," *Int. J. Geogr. Inf. Sci.*, vol. 14, no. 1, pp. 85–104, 2000.

[47] S. Mavoa, K. Witten, T. Mccreanor, and D. O'Sullivan, "GIS based destination accessibility via public transit and walking in Auckland, New Zealand," *J. Transp. Geogr.*, vol. 20, no. 1, pp. 15–22, Jan. 2012.

[48] T. Yigitcanlar, N. Sipe, R. Evans, and M. Pitot, "A GIS-based land use and public transport accessibility indexing model," *Austral. Planner*, vol. 44, no. 3, pp. 30–37, 2007.

[49] A. Agrawal and P. Nagrath, "Analysing and designing automated and dynamic bus route allocation," in *Proc. Int. Conf. Comput. Techn. Inf. Commun. Technol.*, 2016, pp. 251–256.

[50] A. Domènech and A. Gutiérrez, "A GIS-based evaluation of the effectiveness and spatial coverage of public transport networks in tourist destinations," *ISPRS Int. J. Geo-Inf.*, vol. 6, no. 3, p. 83, 2017.

[51] J. Lee and H. J. Miller, "Measuring the impacts of new public transit services on space-time accessibility: An analysis of transit system redesign and new bus rapid transit in Columbus, Ohio, USA," *Appl. Geogr.*, vol. 93, pp. 47–63, Apr. 2018.

[52] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[53] P. Zhao, K. Qin, X. Ye, Y. Wang, and Y. Chen, "A trajectory clustering approach based on decision graph and data field for detecting hotspots," *Int. J. Geograph. Inf. Sci.*, vol. 31, no. 6, pp. 1101–1127, 2017.

[54] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Anal. Prevention*, vol. 41, no. 3, pp. 359–364, 2009.

[55] H. Qi and P. Liu, "Mining taxi pick-up hotspots based on spatial clustering," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, Oct. 2018, pp. 1711–1717.

[56] P. G. Furth and A. B. Rahbee, "Optimal bus stop spacing through dynamic programming and geographic modeling," *Transp. Res. Rec.*, vol. 1731, no. 1, pp. 15–22, 2000.

[57] S. I. Chien and Z. Qin, "Optimization of bus stop locations for improving transit accessibility," *Transp. Planning Techn.*, vol. 27, no. 3, pp. 211–227, 2004.

[58] M. Nikolić and D. Teodorović, "Transit network design by bee colony optimization," *Expert Syst. Appl.*, vol. 40, no. 15, pp. 5945–5955, 2013.

[59] C. Iliopoulou, C. Milioti, E. Vlahogianni, K. Kepaptsoglou, and J. Sánchez-Medina, "The bus bunching problem: Empirical findings from spatial analytics," in *Proc. 21st Int. Conf. Intell. Transp. Syst.*, 2018, pp. 871–876.

[60] M. Enoch, S. Potter, G. Parkhurst, and M. Smith, "Why do demand responsive transport systems fail?" presented at the Transp. Res. Board 85th Annu. Meeting. [Online]. Available: http://pubsindex.trb.org/view.aspx?id=775740

[61] P. B. Goodwin, "A review of new demand elasticities with special reference to short and long run effects of price changes," *J. Transp. Econ. Policy*, vol. 26, no. 2, pp. 155–169, 1992.

[62] K. Button, *Transport Economics*. Cheltenham, U.K.: Edward Elgar Publishing, 2010.

[63] R. Cervero, "Bus rapid transit (BRT): An efficient and competitive mode of public transport," UC Berkeley, Berkeley, CA, USA, Working Paper 2013-01, 2013.

[64] L. Fu, "Planning and design of flex-route transit services," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1791, no. 1, pp. 59–66, 2002, doi: 10.3141/1791-09.

[65] H. Xiaochao, "Study on community bus development basic on trip-chain analysis," *J. Wuhan Univ. Technol. (Transp. Sci. Eng.)*, vol. 36, no. 4, pp. 675–679, 2012.

[66] *Code for Transport Planning on Urban Road*, Standard GB 50220-1995, 1995.

[67] A. F. Perrotta, "Transit fare affordability: Findings from a qualitative study," *Public Works Manage. Policy*, vol. 22, no. 3, pp. 226–252, 2017.

[68] A. Kittelson, C. R. P. Transit, and D. C. Transit, "Transit capacity and quality of service manual," Kittelson Assoc., Transp. Res. Board Nat. Acad., Washington, DC, USA, TCRP Rep. 100, 2003.

[69] M. Li and Y. Long, "The coverage ratio of bus stations and spatial pattern evaluation in Chinese major cities," *Chinese Urban Plann. Forum*, vol. 6, pp. 33–40, 2015.

[70] S. H. Li and Z. B. Man, "K-means clustering algorithm based on adaptive feature weighted," *Comput. Technol. Develop.*, vol. 23, no. 6, pp. 98–105, 2013.

[71] A. Huang, H. Michael Zhang, G. Wei, Y. Yang, and G. Zong, "Cascading failures in weighted complex networks of transit systems based on coupled map lattices," *Math. Problems Eng.*, vol. 2015, pp. 1–16, Jul. 2015.

[72] J. Wang, "Analysis of bus passenger travel characteristics based on trip chain," M.S. thesis, Dept. Transp. Eng., Beijing Jiaotong Univ., Beijing, China, 2017.

YIYI CHENG was born in Henan, China, in 1996. He received the B.S. degree in transportation from Beijing Jiaotong University, China, in 2018. He is currently pursuing the M.S. degree in transportation with Southeast University, China. He is also pursuing the degree in intelligent transportation system (ITS) with Southeast University. He has been involved in many traffic design and optimization projects. His current research interests include geography information system—transportation (GIS-T), ITS, and traffic big data application.

AILING HUANG was born in Guangxi, China, in 1977. She received the M.E. and Ph.D. degrees in systems engineering from Beijing Jiaotong University, Beijing, in 2003 and 2014, respectively.

She is currently an Associate Professor of systems engineering with the School of Traffic and Transportation, Beijing Jiaotong University. She has authored eight books and more than 30 articles. Her research interests include systems engineering, transportation planning and management, and control science and engineering.

GEQI QI was born in Inner Mongolia, China, in 1987. He received the B.S. and M.S. degrees in electrical engineering and automation from the College of Information and Electrical Engineering, China Agricultural University, Beijing, China, in 2009 and 2011, respectively, and the Ph.D. degree in civil engineering from Tsinghua University, Beijing, in 2016. He is currently a Lecturer with the School of Traffic and Transportation, Beijing Jiaotong University, Beijing. His research interests include machine learning, data mining, behavior analysis, human factors, and intelligent transport systems.

BEI ZHANG was born in Beijing, China, in 1978. She received the M.Sc. degree in transport and business management from the Newcastle University, Newcastle upon Tyne, U.K. Since 2003, she has been with the China International Engineering Consulting Corporation. She is also a Registered Consulting Engineer and a Vice Division Chief of the Urban Transit Division, Transportation Department. She is responsible for urban transit planning consultation more than 15 years.

• • •