Journal of Transport Geography 58 (2017) 135–145

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Understanding commuting patterns using transit smart card data

Xiaolei Ma [a,b,*,1], Congcong Liu [b], Huimin Wen [c,d,**,1], Yunpeng Wang [b], Yao-Jan Wu [e]

[a] Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), Beihang University, Beijing 100191, China
[b] School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure, Systems, and Safety Control, Beihang University, Beijing 100191, China
[c] Beijing Transportation Research Center, Beijing 100073, China
[d] Beijing Key Laboratory of Energy Conservation and Emission Reduction Detection and Evaluation in Transport, Beijing 100073, China
[e] Dept. of Civil Engineering and Engineering Mechanics, The University of Arizona, 1209 E 2nd St. Room 324F, Tucson, AZ 85721, United States

ARTICLE INFO

Article history:
Received 30 April 2016
Accepted 1 December 2016
Available online xxxx

Keywords:
Commuting
Human mobility
Public transportation
Travel behavior
Transit smart card data

ABSTRACT

Commuting reflects the long-term travel behavior of people and significantly impacts urban traffic congestion and emission. Recent advances in data availability provide new opportunities to understand commuting patterns efficiently and effectively. This study develops a series of data mining methods to identify the spatiotemporal commuting patterns of Beijing public transit riders. Using one-month transit smart card data, we measure spatiotemporal regularity of individual commuters, including residence, workplace, and departure time. This data could be used to identify transit commuters by leveraging spatial clustering and multi-criteria decision analysis approaches. A disaggregated-level survey is performed to demonstrate the effectiveness of the proposed methods with a commuter identification accuracy that reaches as high as 94.1%. By visualizing the spatial distribution of the homes and workplaces of transit commuters, we determine a clear disparity between commuters and noncommuters and confirm the existence of job–house imbalance in Beijing. The findings provide useful insights for policymakers to shape a more balanced job–housing relationship by adjusting the monocentric urban structure of Beijing. In addition, the extracted individual-level commuting patterns can be used as valuable information for public transit network design and optimization. These strategies are expected to reduce car dependency, shorten excess commute, and alleviate traffic congestion.

1. Introduction

As the primary component of personal daily travel, commuting considerably influences urban traffic conditions (Zhou et al., 2014). The job–housing spatial imbalance forces people to endure long commuting time, and thus, results in excessive travel time and fuel consumption (Van Acker and Witlox, 2011; Charron, 2007). To mitigate these adverse effects, transportation planners and operators strive to reduce travel demand and improve commuting efficiency through policies or technological countermeasures (Lovelace et al., 2014; Li et al., 2013a). Among these measures, prioritizing public transportation systems is considered as one of the most effective strategies because it can significantly reduce car dependency, mitigate traffic congestion, and alleviate air pollution (Ma and Wang, 2014; Li et al., 2013b; Zhao, 2013). Understanding transit commuting patterns offers valuable insights into the spatial and temporal relationship between transit commuters' residences and workplaces (Zhao, 2013; Louf and Barthelemy, 2014). These insights will highlight the critical need for properly designing public transport networks to establish job–housing balance (Zhou et al., 2014). Unlike traditional zonal or regional travel commuting behavioral studies, individual-level commuting patterns provide more detailed and useful information to refine existing travel demand forecasting models with high-resolution human mobility (Yang et al., 2014; Ren et al., 2014). In addition, targeting individual transit commuters using fare reduction for ridership attraction is necessary to improve public transportation systems usage rates (Ma et al., 2013). However, extracting the commuting behavior of an individual transit rider is not a straightforward task. Conventional public transport behavioral studies rely on household travel surveys or diaries to obtain personal profile, socioeconomic and demographic information, and travel patterns (Louail et al., 2014; Schneider et al., 2013; Jiang et al., 2012). This process is costly, time-consuming, and frequently results in a low sampling rate and a small population size. When ask to participate in multi-day travel surveys, people are usually reluctant to respond because of survey fatigue, which will further decrease data availability and accuracy (Mahrsi et al., 2014).

The recent developments of emerging data sources and statistical methods have created opportunities to analyze and determine transit commuting behavior at an individual level over long-term periods. Transit smart card data from automatic fare collection systems are

* Correspondence to: X. Ma, Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), Beihang University, Beijing 100191, China.
** Correspondence to: H. Wen, Beijing Transportation Research Center, Beijing 100073, China.
E-mail addresses: xiaolei@buaa.edu.cn (X. Ma), vincent_liu@buaa.edu.cn (C. Liu), wenhm@bjtrc.org.cn (H. Wen), ypwang@buaa.edu.cn (Y. Wang), yaojan@email.arizona.edu (Y.-J. Wu).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jtrangeo.2016.12.001
0966-6923/

widely adopted by transit authorities to manage revenue as well as to gather abundant passenger boarding and alighting information in a disaggregated manner (Kusakabe and Asakura, 2014). Compared with traditional data collection methods, smart card data can record the day-to-day variability in the travel patterns of an individual transit rider and have the potential to identify transit commuters and detect their spatial and temporal regularity through a continuous long-term observation period (Schneider et al., 2013). A large number of behavioral studies using smart card data have gained more and more popularities. Morency et al. (2007) applied data mining techniques to examine the spatial and temporal variability of transit use based on smart card data. Ma et al. (2013) developed a Density-Based Scanning Algorithm with Noise (DBSCAN) to categorize different groups of transit riders with varying travel patterns. Kieu et al. (2015a, b) improved the original DBSCAN algorithm proposed by Ma et al. (2013), and significantly reduced the algorithm complexity with the same clustering performance. Langlois et al. (2016) utilized four-week transit smart transaction data to identify transit rider heterogeneity, and generated 11 clusters with distinct activity sequences and demographic attributes. Ali et al. (2015) developed a large-scale activity based public transport simulation platform based on MATSim, and used the smart card as an input. They demonstrated the feasibility of applying smart card data into microsimulation travel demand models. As documented by Pelletier et al. (2011), new analytical methods and disaggregate approaches based on smart card data are renovating traditional travel behavior research.

In the current literature on transit behavioral studies on commuting travel patterns, the definition of a transit commuter is oversimplified. The majority of these studies only considered the repeatability of temporal activities (e.g., transit riders traveling for at least a number of days are determined to be transit commuters) (Long et al., 2012; Zhou et al., 2014). Several researchers simultaneously incorporated both the spatial and temporal regularities of recurring travels into transit passenger segmentation studies. Ortega-Tong (2013) analyzed the spatial and temporal travel patterns of London transit riders with Oyster Cards, and grouped them into 8 clusters based on different sociodemographic characteristics and activity patterns. The transit riders traveling 4 days a week or more were categorized as regular users. Kieu et al. (2015b) proposed a transit passenger segmentation method based on smart card data. The method identified the spatial travel pattern of each transit rider using a two-step DBSCAN algorithm, and a k-means algorithm was the applied to distinguish frequent and infrequent transit users based on the number of travel days and journeys made. Ma et al. (2013) developed a density-based clustering algorithm to mine each transit rider's spatial and temporal travel pattern in Beijing, and then proposed a K-means++ algorithm and Rough Set based approach to measure travel regularity. More than 40% transit riders were identified as frequent passengers. Kung et al. (2014) used mobile phone data to understand home-work commuting behaviors at three countries (Portugal, Ivory Coast, and Saudi Arabia) and one city (Boston). Individual's home and work locations can be identified by a spatial and temporal filtering approach. They found that the home-work commuting time distribution is independent of commute distance. However, the above studies are not specifically designed for transit commuter identification. Even if each transit rider's spatial and temporal travel pattern can be mined in some literatures, how the identified spatiotemporal travel pattern tie to one's commuting behavior is still not clear. Fortunately, the rapid development of data mining and statistical techniques has facilitated finding underlying and previously hidden information through large-scale data processing (Jiang et al., 2012), and thus can be applied to transit commuting pattern mining using smart card data. Both the spatial and temporal features of the commuting trips of transit riders should be considered in a quantitative and synthetic manner.

This study seeks to answer the following important questions: Which transit riders can be classified as transit commuters? Can we infer the residence and workplace of an individual transit commuter, as well as his/her commuting departure time, based on his/her long-term smart card transaction records? To answer these questions, we first extract the spatial and temporal features of an individual transit rider that can represent the regularity of commuting patterns. These features are calculated based on continuous trip-chaining behavior from one-month of smart card data in Beijing. These features fully incorporate the heterogeneity of individual route choices and the uncertainty of departure times using a modified DBSCAN clustering algorithm. A spatial clustering algorithm, called iterative self-organizing data analysis technique (ISODATA), is then utilized to categorize group transit riders into three clusters based on their spatiotemporal features. The three clusters are automatically determined and can reflect the intensity of transit commuting travel. Consequently, the residence and workplace of each transit commuter can be derived by summarizing his/her most frequently visited locations. For each cluster, we also develop a transit commuting score based on the technique for order of preference by similarity to ideal solution (TOPSIS) method. This method can score each individual transit rider based on his/her commuting patterns via a multi-criteria decision analytical framework. The score threshold for distinguishing commuters from noncommuters routinely remains at approximately 51.7 even when different groups of transit riders are randomly selected. Instead of clustering the total population of transit riders, we can identify transit commuters only by calculating the individual commuting score. This approach will significantly reduce computational power. Finally, the proposed methods are validated via a survey conducted in Beijing and visualized in map platform to demonstrate their effectiveness.

2. Methodology

2.1. Trip generation

The public transportation systems in Beijing consist of over 1000 bus routes and 18 subway lines in 2015 (Fig. 1), and these figures result in over 28,000 bus and subway stops. The percentage of subway and bus trips among all motorized trips reached 60.1% at the end of 2014 because of the expansion of subway lines and the governmental subsidies for public transit (Beijing Transportation Research Center, 2015). Over 90% of transit riders utilize smart cards to pay fares because a huge discount rate is received by smart card holders (i.e., 75% fare reduction for students and 50% fare reduction for regular passengers) (Ma et al., 2012). Since January 2015, all buses and subway lines have adopted distance-based fare strategies in which both passenger tap-in and tap-out data (e.g., route ID, transaction times, and boarding and alighting stops) for an individual transit rider are recorded. Beijing Automatic Fare Collection (AFC) system includes urban transit, rural transit and subway systems. Buses running in the rural area record passengers' boarding times/stops and alighting time/stops. Similarly, subway AFC system also contains passengers' full trip information. However, buses running in the urban area only store individual's boarding stop and alighting time/stop. The percentage of records with missing boarding times is approximately 40% of total smart card transactions. For the urban AFC system, the bus boarding time of each transit rider can be calculated by using the transaction time (i.e. alighting time) minus the average in-vehicle travel time. According to the 2015 Beijing Transport Annual Report, the average in-vehicle travel time during morning peak hours is 59.6 min, while this time increases to 65.3 min during evening peak hours (Beijing Transportation Research Center, 2015). We collected one-month smart card data (June 2015). A total of 18 million active transit riders were identified, thus generating a total of over 364 million smart card transaction records.

We generate individual trips on a daily basis. A trip is composed of a sequence of activities for a particular purpose (Primerano et al., 2008). In the context of public transport, a transit rider may transfer from one bus route to a subway line and continue to take another bus route to his/her destination. In the above example, the trip is associated with three smart card transactions (bus, subway and bus). Time

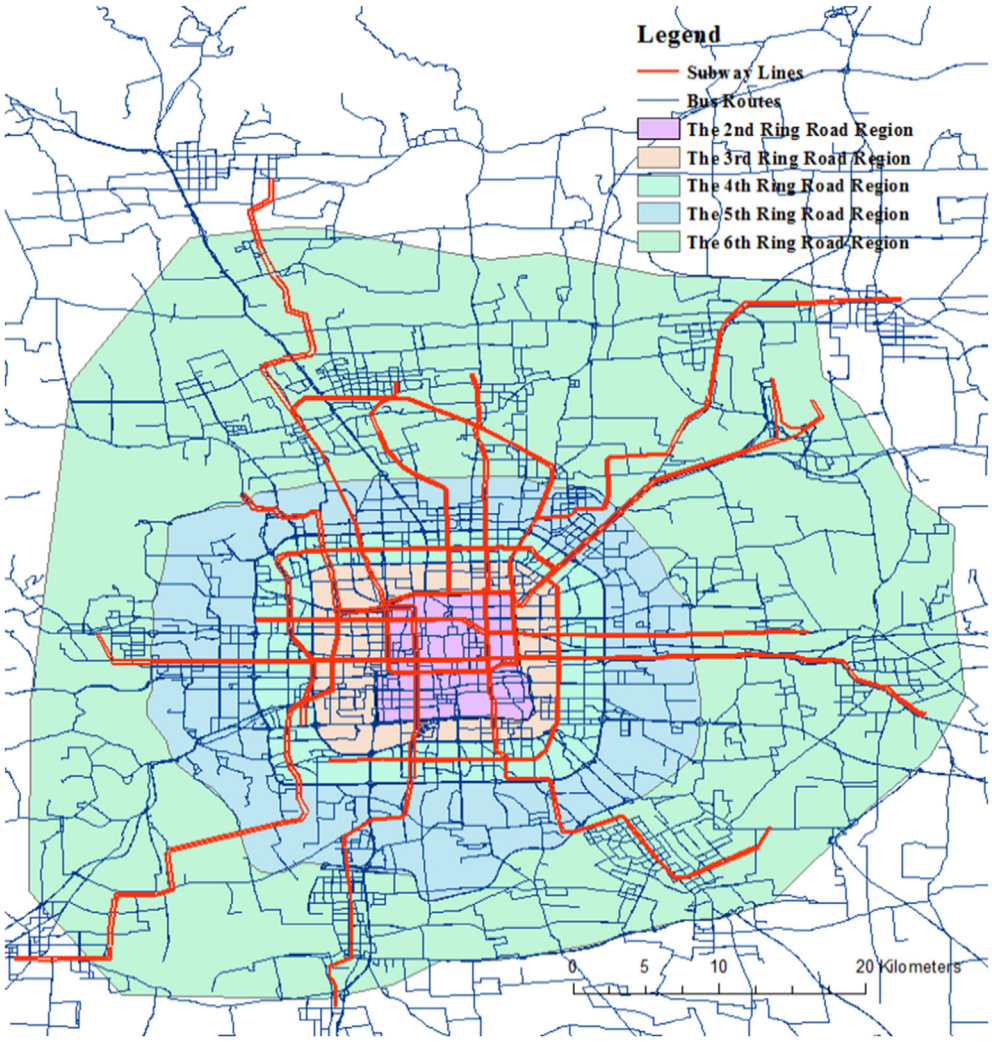

Fig. 1. Spatial distribution of Beijing bus routes and subway lines. The map displays only the bus routes and subway lines within the Sixth Ring Road region of Beijing, which is located at the outer fringes of Beijing and covers most of its administrative districts.

thresholds are adopted to link these transactions. Three possible transfer activities can be identified: subway to bus, bus to bus, and bus to subway. A comprehensive field survey conducted in Beijing suggested that the time thresholds for these three transfer activities should be set to 104, 112, and 20 min, respectively (Wang, 2014). Due to missing boarding times for buses, in-vehicle travel time, passenger's walking speed and waiting time are jointly taken into account to estimate a maximum transfer time for each transfer mode. The maximum transfer times for the three transfer activities (i.e. subway to bus, bus to bus, and bus to subway) are 140 min, 140 min and 30 min, respectively. Then, the transaction time differences that fall within the maximum transfer times are extracted and ordered, and the 95th percentile transaction time difference for each transfer activity is selected as the time threshold to form up a complete trip. If the transaction time difference for two consecutive smart card records for an individual passenger exceeds any of the thresholds, then a trip is separated. All trips can be generated based on the aforementioned criteria.

In this study, a commuter is defined as a regular transit rider who performs periodically recurring travels between home and other non-residential locations (primarily refers to the repeated trip between home and work/school) (Kung et al., 2014). We assume that only the first trip (i.e., home-to-work trip) and last trip (i.e., work-to-home trip) of an individual transit rider for each day contribute to their commuting behavior. For individual's first trip and last trip, we merge the multiple trips with short dwell times (i.e. 2 h) as a single trip. This treatment can eliminate the error caused by lacking trip purpose information to some extent: If passengers take buses/subway for meal or picking up children at school, the destination of the first trip and the origin of the last trip are not workplaces. This fact is confirmed by the findings of Zhou et al. (2014) because the majority of transit riders start their trips from their homes to their workplaces and returns to their homes at the end of a day. The rationale of choosing both the first trip and last trip of each transit rider is that one's home-to-work trip may be distinct from his/her work-to-home trip given that there could be multiple alternative bus/subway routes to commute.

2.2. Commuting feature extraction and travel pattern recognition

The regularity of commuting should be spatially and temporally measured. The repeatability of temporal patterns can be quantified by the similarity of departure time and the number of traveling days, whereas the repeatability of spatial patterns can be represented by the frequency of the most visited stops and the number of recurring travels on similar bus routes or subway lines. The underlying rationale for selecting the four commuting features is that transit commuters are likely to take buses or subways regularly with relative fixed stops at similar times within a long time span.

For the repeatability of temporal patterns, we divided the 24 h in a day into 30-minute intervals, and the transaction times of the first trip and the last trip can be converted into integer values from 0 to 47 (0

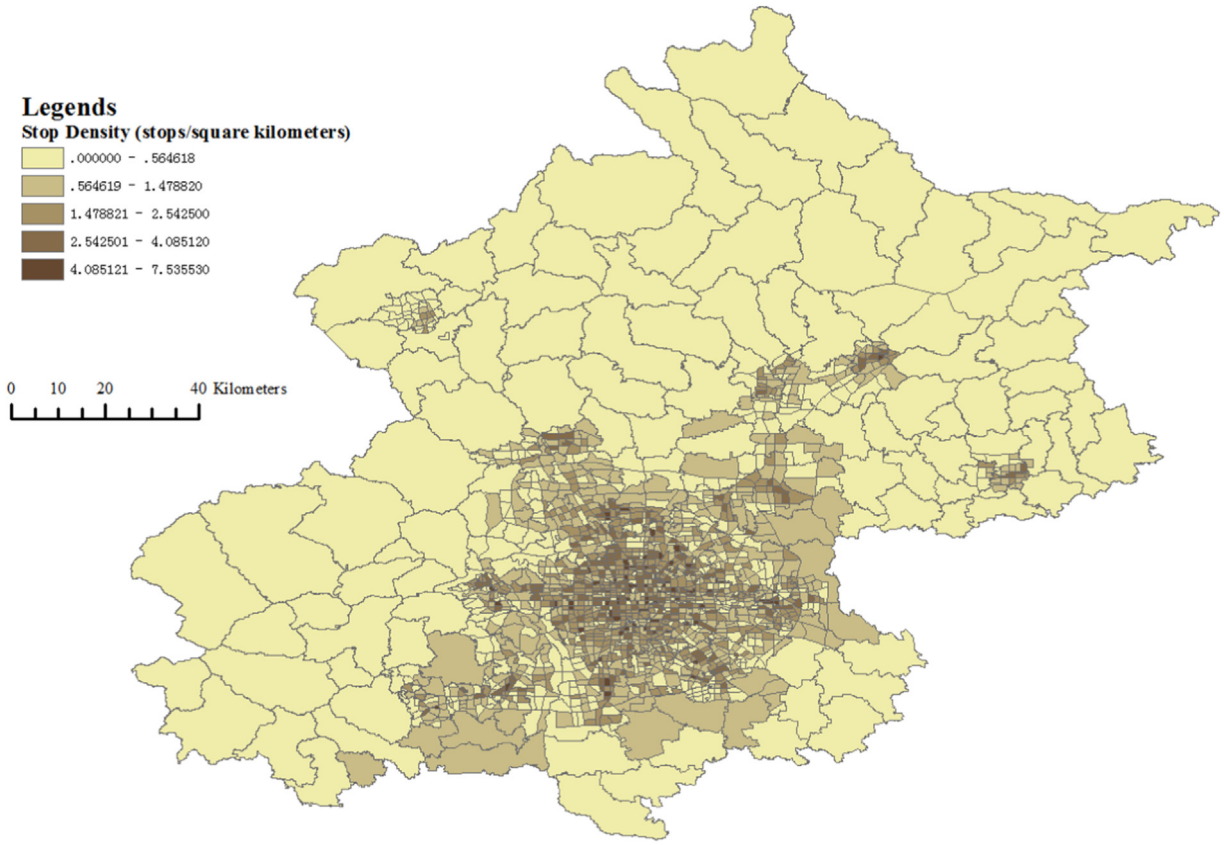

Fig. 2. Density map of clustered bus and subway stops.

indicates that a trip starts between 00:00 AM and 00:30 AM, whereas 47 denotes that the trip starts between 11:30 PM and 00:00 AM). Integer values are defined as departure time indices. For each transit rider, we record the set of departure time indices of the first trip and the last trip on a daily basis over a one-month period and count the number of occurrences of each index. The departure time index with the largest number of occurrences is considered the most frequent departure time index. For the home-to-work trip, the most frequent departure time index is defined as T_h. For the work-to-home trip, the most frequent departure time index is defined as T_w. The number of occurrences of T_h and T_w are represented as N_{T_h} and N_{T_w}, respectively. The similarity of departure time N_{time} can be calculated using Eq. (1):

$$N_{time} = N_{T_h} + N_{T_w} \tag{1}$$

Thus, the value of N_{time} represents the regularity of departure times for each transit commuter.

In addition, we can count the number of days when each transit rider travels as N_{day}. A large N_{day} indicates that the transit rider has likely become a regular transit commuter.

For the repeatability of spatial patterns, we quantify route-level and stop-level similarities. However, route-level similarity can only describe partial commuting regularity because several distinct route choices may exist between the home and workplace of an individual, which leads to different sequences of routes connecting the same bus or subway stops. To address this issue, we propose an improved DBSCAN algorithm to cluster the spatially adjacent stops into several groups and renumber these groups as new stop IDs. The traditional DBSCAN algorithm can categorize these points with a number of nearby neighbors but leave other points with only a few neighbors as outliers. When the original DBSCAN algorithm is applied into bus or subway stop partitioning, one primary disadvantage should be overcome. That is, depending on the density of points in each cluster, intra-cluster distances may be high, which causes the cluster boundary, including multiple distinct stops that should belong to multiple clusters, to form only one cluster. This occurrence may lead to the wrong identification of commuting stops. We improve the original algorithm by allowing the reclustering of these abnormal groups and by splitting each large cluster into several small clusters (see Appendix A).

A total of 28,871 bus/subway stops are clustered into 6544 stop groups. The distribution of these stop groups can be visualized in Fig. 2. The grouped stops are aggregated based on traffic analysis zones. Most clustered stop groups are densely scattered in the Central Business District (CBD) area of Beijing but sparsely located in suburban areas.

Table 1
Statistical data of commuting patterns of transit riders. S_h is the most frequent stop of home. S_w is the most frequent route sequence of home. R_h is the most frequent stop of workplace. R_w is the most frequent route sequence of workplace. T_h is the most frequent departure time of home. T_w is the most frequent departure time of workplace. N_{day} is the number of traveling days. N_{route} is the number of similar route sequences. N_{stop} is the number of similar stops. N_{time} is the number of similar departure times.

Card ID	S_h	S_w	R_h	R_w	T_h	T_w	N_{day}	N_{route}	N_{stop}	N_{time}
43207	20152	365	00986	00958	19	29	5	2	3	3
41610	217	102	6–2	10–6	14	31	2	2	4	2
32558	1533	5503	00359	00359	15	35	24	21	33	11
32664	265	723	57300–60366	60366–00741	16	36	15	10	19	9
86147	10485	20295	00012	00012	19	25	3	5	4	2

We can compute stop-level similarity by following the calculation procedure for departure time similarity. The first trip and last trip of each transit rider are assumed to be home-to-work trip and work-to-home trip, respectively. We record the set of origin stop IDs of home-to-work trip and the destination stop IDs of work-to-home trip and then count the number of occurrences of each stop ID. The stop ID with the largest number of occurrences is considered the most frequent stop of residence S_h. Similarly, we can record the set of destination stop IDs of home-to-work trip and origin stop IDs of work-to-home trip and then count the number of occurrences of each stop ID. The stop ID with

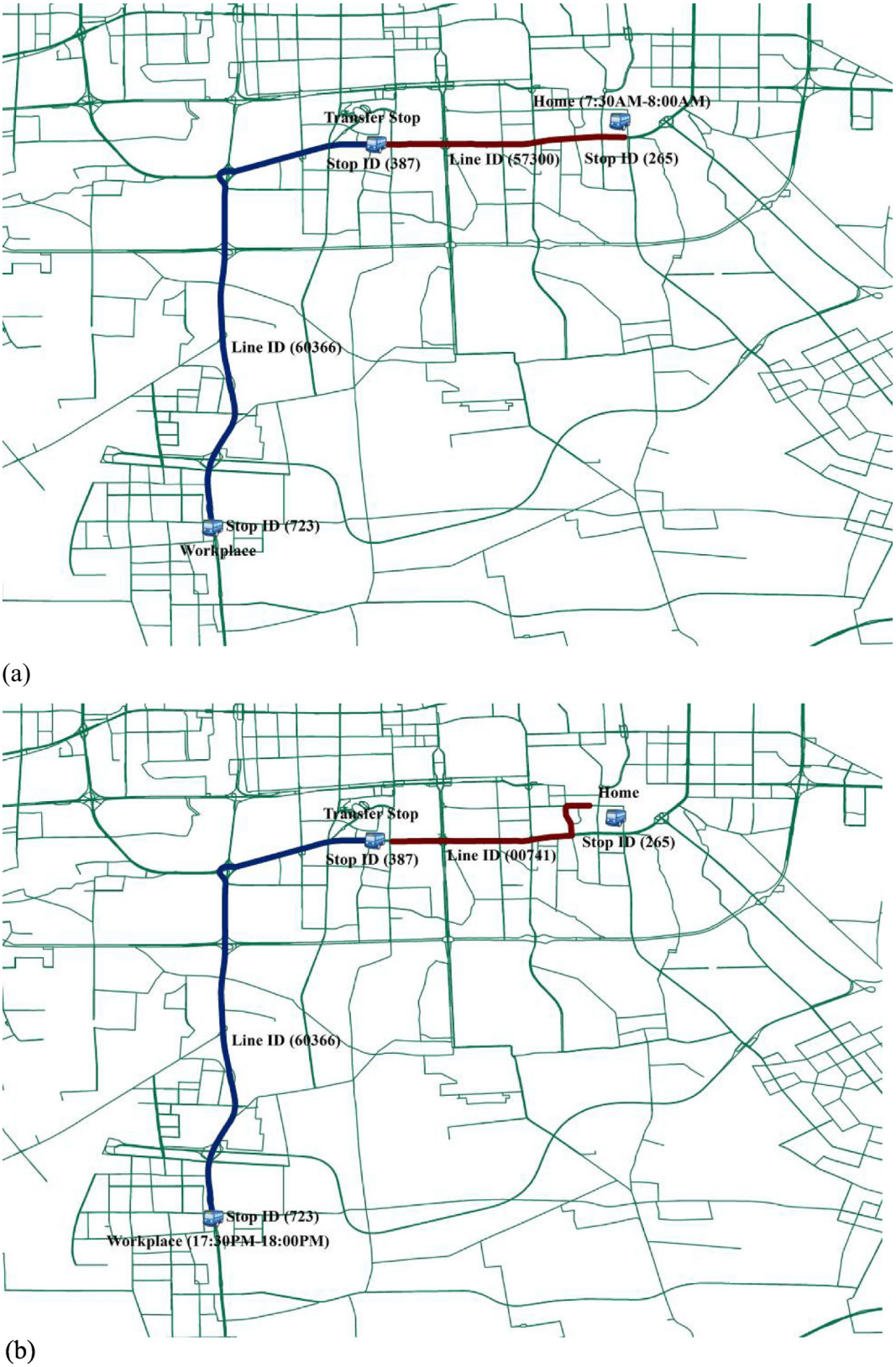

Fig. 3. Daily commuting trips of transit rider with smart card ID 32664: (a) home-to-work trip and (b) work-to-home trip.

Table 2
Clustering results of the three sample clusters from the randomly selected 500,000 transit riders.

Sample 1	N_{day}	N_{stop}	N_{route}	N_{time}	Number of transit riders	Commuting score (CS)
Absolute commuters	23.28	28.16	55.39	21.68	50,228	CS ≥ 71.30
Average commuters	18.31	13.85	21.81	11.93	79,301	
71.30 > CS ≥ 51.67						
Noncommuters	4.33	2.80	4.31	2.49	370,471	CS ≥ 51.67
Sample 2	N_{day}	N_{stop}	N_{route}	N_{time}	Number of transit riders	Commuting score (CS)
Absolute commuters	23.27	28.18	55.40	21.68	51,177	CS ≥ 71.27
Average commuters	18.23	13.77	21.66	11.87	78,970	
71.27 > CS ≥ 51.70						
Noncommuters	4.32	2.80	4.31	2.50	369,853	CS ≥ 51.70
Sample 3	N_{day}	N_{stop}	N_{route}	N_{time}	Number of transit riders	Commuting score (CS)
Absolute commuters	23.31	28.18	55.52	21.73	50,307	CS ≥ 71.33
Average commuters	18.31	13.85	21.86	11.93	78,922	
71.33 > CS ≥ 51.69						
Noncommuters	4.35	2.81	4.32	2.50	370,771	CS ≥ 51.69
Sample 4	N_{day}	N_{stop}	N_{route}	N_{time}	Number of transit riders	Commuting score (CS)
Absolute commuters	23.31	28.17	55.33	21.70	50,523	CS ≥ 71.34
Average commuters	18.32	13.88	21.87	11.98	78,836	
71.34 > CS ≥ 51.72						
Noncommuters	4.33	2.80	4.31	2.49	370,641	CS ≥ 51.72
Sample 5	N_{day}	N_{stop}	N_{route}	N_{time}	Number of transit riders	Commuting score (CS)
Absolute commuters	23.29	28.22	55.45	21.72	50,416	CS ≥ 71.36
Average commuters	18.37	13.89	21.88	11.97	79,252	
71.36 > CS ≥ 51.73						
Noncommuters	4.34	2.80	4.32	2.50	370,332	CS≥51.73

the largest number of occurrences is considered the most frequent stop of workplace S_w. The number of occurrences of S_h and S_w are represented by N_{S_h} and N_{S_w}, respectively. The similarity of stop N_{stop} can be calculated using Eq. (2):

$$N_{stop} = N_{S_h} + N_{S_w} \tag{2}$$

Each transit rider may take several different bus routes or subway lines to complete a trip. The route sequence associated with each trip is defined as $R: r_1 \rightarrow r_2 \rightarrow \cdots \rightarrow r_n$, where r_n represents either a bus or subway route connecting clustered stops, and n indicates the number of transfers during each trip. We record the set of route sequences of the first trip and last trip on a daily basis over a one-month period and count the number of occurrences of each route sequence. The route sequence with the largest number of occurrences is considered the most frequent route sequence. For home-to-work trip, the most frequent route sequence is defined as R_h. For work-to-home trip, the most frequent route sequence is defined as R_w. The number of occurrences of R_h and R_w are represented as N_{R_h} and N_{R_w}, respectively. However, R_h and R_w can only represent the most frequent route sequences taken by transit riders. Several alternative route sequences connect the residences and workplaces of riders, and are defined as all possible route sequences between home and work places excluding the most frequent route sequence. These alternative route sequences also contribute to route-level similarity calculation. This is the issue that Ma et al. (2013)'s algorithm cannot properly address. We find alternative route sequences from S_h to S_w and summarize the total number of occurrences of these route sequence as $N_{R_h'}$. Similarly, we find alternative route sequences from S_w to S_h and summarize the total number of occurrences of these route sequence as $N_{R_w'}$.

The similarity of route N_{route} can be calculated using Eq. (3):

$$N_{route} = N_{R_h} + N_{R_w} + N'_{R_h} + N'_{R_w} \tag{3}$$

N_{route} does not only represent the spatial regularity of these frequently visited routes but also includes the occasional routes that share the common origins and destinations for each transit rider, which significantly improves the accuracy of transit commuter identification.

In summary, N_{time} and N_{day} measure the temporal repeatability of commuting behavior, whereas N_{stop} and N_{route} quantify the spatial repeatability of commuting behavior. The most frequent travel patterns for several transit riders over a one-month period are listed in Table 1 as examples.

Table 3
Key questions in Beijing public transit commuting behavior survey.

Question content	Question type
What is your smart card ID?	Open-ended question
What is your primary purpose to take bus or subway? (e.g. commuting, personal business, shopping, recreational, social activities, and others)	Close-ended question
How many days do you take bus or subway per week?	Close-ended question
What is the stop that is adjacent to your home for your daily travel?	Open-ended question
What is the stop that is adjacent to your work place for your daily travel?	Open-ended question
What is your frequent departure time of your first trip?	Close-ended question
What is your frequent departure time of your last trip?	Close-ended question

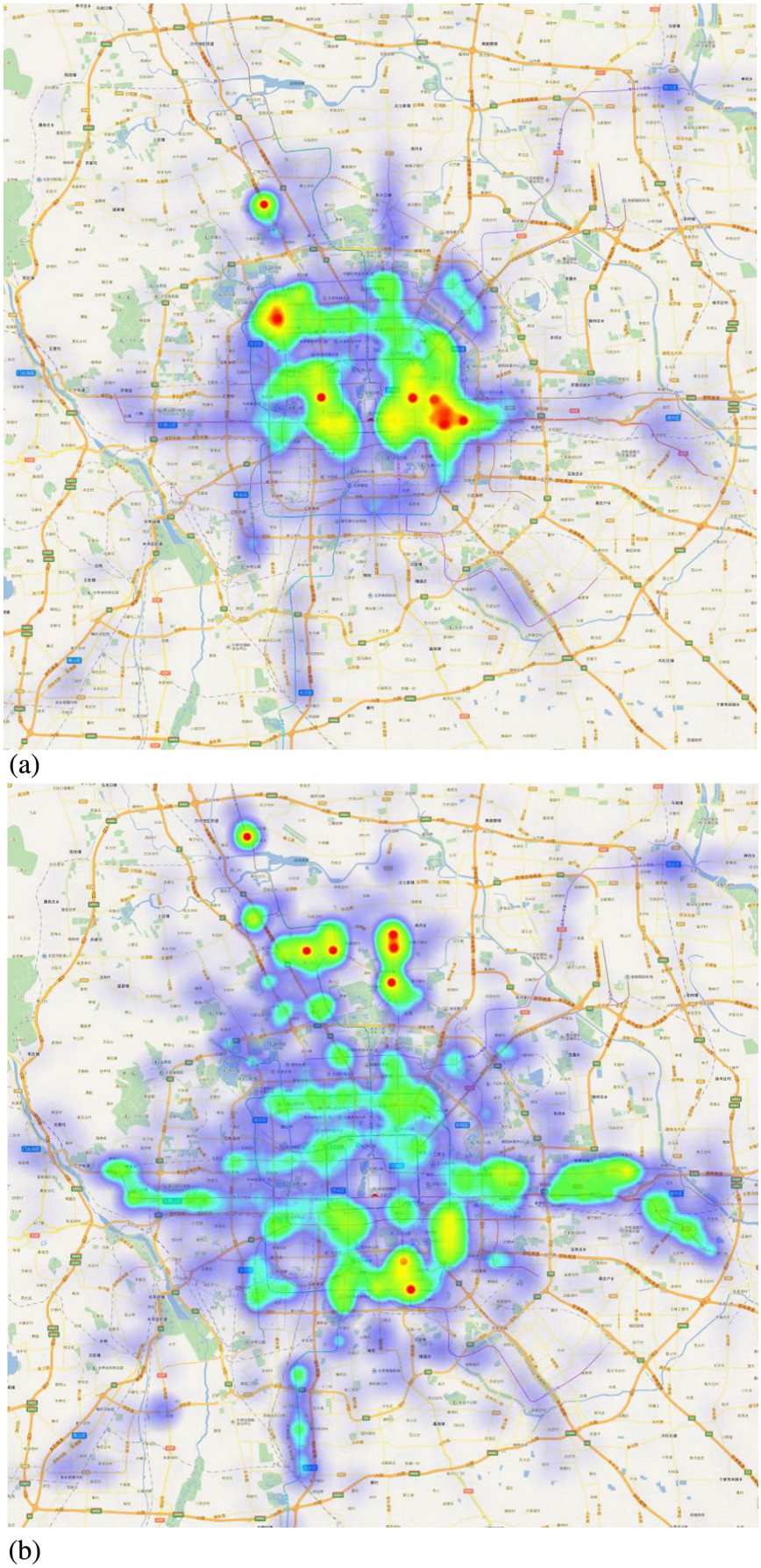

Fig. 4. Spatial distribution of Beijing transit commuters in June 2015: (a) workplace and (b) residence. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Table 1 profiles the spatiotemporal travel pattern of each transit rider over a one-month period. Each smart card ID has been hashed to protect the privacy of the riders. Using the transit rider with card ID 32664 as an example, this passenger frequently takes bus route 57300 from stop 265 between 7:30 AM–8:00 AM, transfers to route 60366, and alights at stop 723. In the afternoon, this passenger frequently leaves from stop 265 by taking route 60366 and transfer to route 00265. His/her most frequent alighting stop is stop 723. On the average, this passenger travels at least 15 days by bus or subway with the regularity of routes, stops, and departure times at 10, 19, and 9, respectively. The commuting trips of this transit rider can be further visualized in Fig. 3.

2.3. Commuter identification

We identify transit commuters using the 4D inputs (N_{day}, N_{route}, N_{stop}, N_{time}) based on the mined spatiotemporal regularities of transit riders. TOPSIS (Hwang and Yoon, 1981) is used to rate the commuting intensity of each transit rider (see Appendix A). The TOPSIS method can measure the distance among target alternatives, positive idea solutions, and negative idea solutions, and consequently, yield a similarity ratio for multi-criteria decision analysis. Each transit rider is assigned a normalized value as a commuting score between zero and one to consider the multi-attribute of commuting behavior. A score that approaches one indicates that the associated transit rider is likely to be a commuter. Instead of applying an arbitrary threshold (e.g., transit riders with scores exceeding 0.6 are classified as commuters) to distinguish transit commuters, we propose to use ISODATA to cluster transit riders automatically based on their spatiotemporal regularities and to determine the cutoff score for quantifying commuting intensity. The ISODATA method is an unsupervised learning method that can split and merge clusters iteratively and optimize the number of clusters (Ball and Hall, 1965). However, one of the disadvantages of ISODATA is the high computational cost for large data sets (Memarsadeghi et al., 2007). The amount of smart card data is considerable; that is, clustering 18 million transit riders is unrealistic. We randomly select 500,000 transit riders for 5 times and perform ISODATA for each subsample of transit riders based on their spatiotemporal regularities (N_{day}, N_{route}, N_{stop}, N_{time}). The random selection is implemented in Microsoft SQL Server using the newid() function, which generates a globally unique identifier (GUID) for each transit rider (Microsoft Cooperation, 2016). The procedure randomly samples from the entire population, and thus will not yield biased results. Three clusters can be consistently received, and the transit riders within these clusters can be categorized as absolute commuters, average commuters, and noncommuters. The commuting scores of transit riders within each cluster can be calculated and ordered based on the TOPSIS method. The minimum and maximum commuting scores are respectively selected as the lower bound and upper bound to distinguish absolute commuters, average commuters and noncommuters. Among each group, the average commuting score is stable, as presented in Table 2. Both commuting score and spatiotemporal travel pattern statistics are averaged within each cluster. We can define transit commuters as transit riders with commuting scores higher than 51.7 based on the clustering result.

2.4. Validation and visualization

To validate the effectiveness of the proposed transit commuting pattern identification methods, we conducted a detailed and anonymous survey about the travel behavior of smart card holders in Beijing via social media (i.e. WeChat) in August 2015. The population of the survey sample is diverse and includes students, government officials, and company employees, among others. Each respondent is required to input his/her smart card ID. We apply the proposed methods to assess whether the respondent is a transit commuter and infer his/her spatiotemporal travel patterns. In addition, several questions are designed to determine the travel purposes, home and workplace locations, departure times, and transit usage frequency of each respondent (see Table 3). To validate whether a particular transit rider belong to a commuter, the travel records of the transit rider over a one-month period can be extracted from the smart card database based on the inputted smart card ID. Then, the four-dimensional commuting features (N_{day}, N_{stop}, N_{route}, N_{time}) can be calculated as inputs to generate the commuting score for this transit rider. This score will be compared with the derived thresholds in Table 2 to judge whether this transit rider is a commuter and identify the corresponding commuting patterns. The inferred spatiotemporal behavioral information (e.g. travel purpose, home and workplace, departure time, number of traveling days, the most frequently taken routes, etc.) will be further matched with the survey result filled by the same transit rider. A total of 118 copies of effective questionnaires were received. Among the respondents, 63 consider themselves commuters, whereas the remaining 55 are noncommuters. The proposed method can successfully classify 56 out of the 63 respondents as commuters and categorize the remaining 55 respondents as noncommuters. The detection accuracy is 94.1%. For the group of commuters, 40 respondents leave the approximate location information of their residences and workplaces. The stops adjacent to the homes and workplaces of 37 respondents are correctly inferred with an accuracy of 90.0%.

We further compute the commuting scores of all transit riders in June 2015 and apply the clustering threshold (i.e., the commuting score is equal to 51.7) to distinguish both commuters and noncommuters. A total of 18,137,393 transit smart card holders are identified with a total of 364,846,374 transactions in June 2016. A total of 1,831,799 and 2,865,604 transit riders are identified as absolute commuters and average commuters, respectively. This number indicates that the number of transit commuters is 4,697,403, whereas the remaining 13,439,990 transit riders are noncommuters. We can define the percentage of daily commuters as the number of daily transit commuters divided by the total number of daily transit riders. The percentage of daily commuting transactions can be calculated as the number of home-to-work and work-to-home transactions for commuters divided by the total number of home-to-work and work-to-home transactions for all transit riders on a daily basis. We can then compute the percentages of transit commuters and commuting transactions during weekdays and weekends. The percentage of daily transit commuters in June 2015 is 61.70%. This percentage increases to 65.14% during weekdays but decreases to 50.90% during weekends. The percentage of daily commuting transactions is 53.70%, which increases to 59.50% during weekdays and decreases to 37.60% during weekends.

In Figs. 4 and 5, we further visualize the commuting spatial patterns of transit commuters in a map-based platform according to the locations of residences and workplaces.

Both Fig. 4(a) and (b) are heat maps in which the region with a dark color (from light blue to dark red) implies high transit commuter population density. In terms of workplaces, most transit commuters work in the core areas of Beijing CBD, ZhongGuanCun (ZGC), Beijing Financial Street (BFS), and ZhongGuanCun Software Park (ZSP). Beijing CBD and BFS are the key international financial and business centers in China, where a large number of Fortune 500 enterprises and governmental agencies are located. The majority of office workers choose public transit (bus and subway) to commute considering the limited parking resources and expensive parking price. ZGC and ZSP are the technology hubs in Beijing. These areas are known as “China's Silicon Valley,” where several universities and high-tech companies are located. These places attract a myriad of nonpermanent residents and students who do not own private cars and commute by bus or subway. Unlike the spatial distribution of workplaces, residential areas are located away from the central areas of Beijing. Typical places include HuiLongGuan (HLG), FangZhuang, TianTongYuan (TTY), and ShaHe. These areas are all large suburban residential neighborhoods where most residents work in downtown Beijing. For example, over 400,000 residents live in TTY and commute via subway or bus at two adjacent public transport

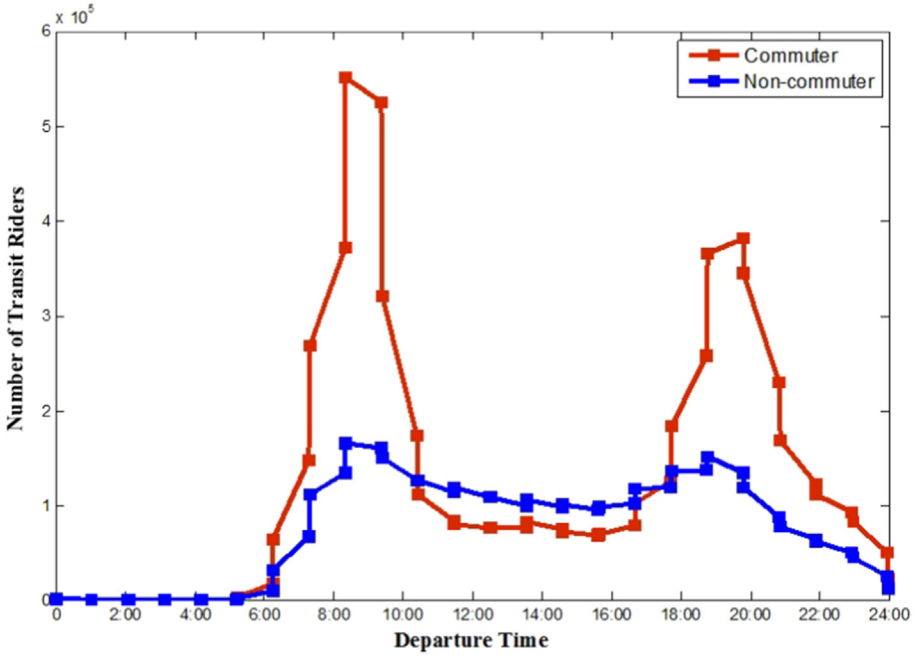

Fig. 5. Distribution of departure times of transit commuters and noncommuters.

hubs on a daily basis. By integrating this information into Fig. 1, we further summarize the percentages of the residences and workplaces of transit commuters that fall within the Sixth Ring Road region in Table 4.

A strong job–housing imbalance can be observed in Table 4. Over 80% of transit riders live outside the 3rd Ring Road region of Beijing, whereas approximately 34% of workplaces are located within the central urban area within the 3rd Ring Road region. This observation may be attributed to the fact that a number of transit commuters working in the central urban area cannot afford the high house rent of residences near their workplaces, and thus, have to live at the fringes of Beijing City (Zhao, 2013).

3. Discussion

We further analyze the commuting behavior between transit commuters and noncommuters in terms of their departure times, travel distances, and number of traveling days (Figs. 5 to 7).

The majority of transit commuters depart from their homes around morning peak hours (7:00 AM–9:00 AM) and return during evening peak hours (5:00 PM–7:00 PM), whereas noncommuters take the bus or subway at any time of a day without a clear shape of the double-peak temporal distribution. In terms of traveling days in the one-month period, the most frequent number of traveling days for transit commuters is 21 days. This number is approximately equal to the total number of weekdays, excluding weekends, during a typical month. By contrast, 90% of noncommuters travel below 10 days, which indicates that most noncommuters are more likely to be sporadic travelers with low transit usages. This fact is consolidated by the travel distance distribution of commuters during weekdays. The average commuting distance for transit commuters is 10.99 km. When the distance between residences and workplaces is far, people are unlikely to choose public transit to commute. For noncommuters, the number of transit riders declines rapidly as travel distance increases. Thus, the average travel distance of noncommuters is 9.15 km. Most noncommuters prefer short-distance travels below 5 km.

4. Conclusion

In summary, this study proposes a series of data mining methods to identify transit commuting patterns based on smart card data. The proposed framework can identify transit commuters by mining spatiotemporal travel regularities over continuous long-term observation, as well as extract individual-level residence and workplace. This approach will significantly alleviate the burden of manual data collection in longitudinal surveys and travel diaries. The transit commuting score is defined to measure the commuting intensity of each transit rider. We find that the threshold for distinguishing commuters from noncommuters remains consistently at 51.7 by leveraging the spatial clustering algorithm and the TOPSIS approach. The spatial and temporal disparity of the two groups is further presented by comparing their departure times, travel distances, numbers of traveling days, and home/workplace distributions. We demonstrate the effectiveness of the proposed methods through a disaggregated-level survey with a transit commuter identification accuracy that reaches as high as 94.1%.

We confirm the existence of job–house imbalance in Beijing, which is partly attributed to the monocentric urban structure and the centralized public transport network that result in the emergence of "excess commuting" in the Beijing public transportation system. Job

Table 4
Percentages of residences and workplaces of transit commuters that fall within the ring road regions of Beijing.

Region	Number of workplaces	Percentage of workplaces	Number of residences	Percentage of residences
Outside the 6th Ring Road	391,572	8.43%	576,471	12.42%
Between the 5th and 6th Ring Roads	821,770	17.68%	1,638,422	35.31%
Between the 4th and 5th Ring Roads	725,858	15.62%	783,818	16.90%
Between the 3rd and 4th Ring Roads	1,146,259	24.66%	846,079	18.24%
Between the 2nd and 3rd Ring Roads	918,226	19.76%	513,459	11.07%
Within the 2nd Ring Road	644,010	13.86%	281,424	6.07%

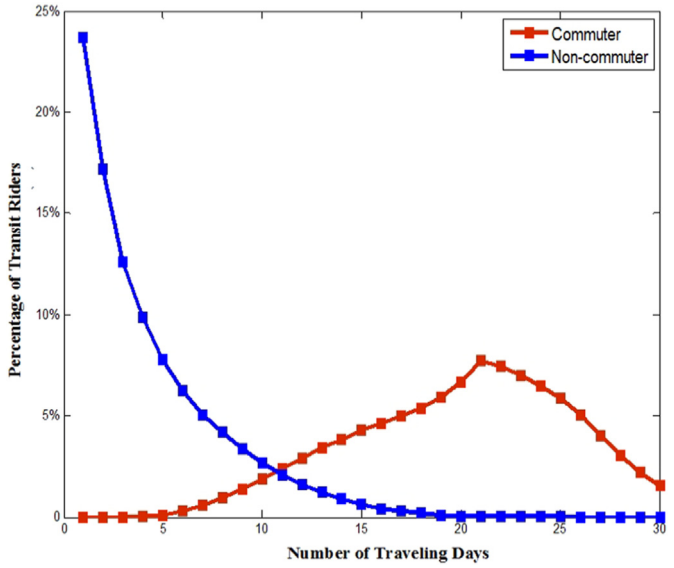

Fig. 6. Distribution of the numbers of traveling days of transit commuters and noncommuters.

opportunities are densely located in the central urban area with high housing prices and limited parking resources. Thus, transit commuters are forced to live in suburban residential regions that will require long-distance travel. The findings provide useful insights for policymakers to shape a more balanced job and housing relationship. In addition, the extracted individual-level commuting patterns can be used as invaluable information for public transit network design and optimization. For example, implementing a corridor-level bus route between HLG and CBD will significantly shorten travel time for transit commuters. These operational and planning strategies from the proposed commuting pattern mining approach are expected to reduce car dependency and alleviate traffic congestion.

Future research can be expanded to investigate the household commuting behavior using smart card. Under this situation, both parents need to work and their children also need to go to school by public transport. The proposed transit commuting pattern mining and commuter identification approaches will be correspondingly modified. In addition, how to quantitatively measure the spatial distribution of job-housing imbalance by leveraging geostatistical models is another interesting and useful research direction.

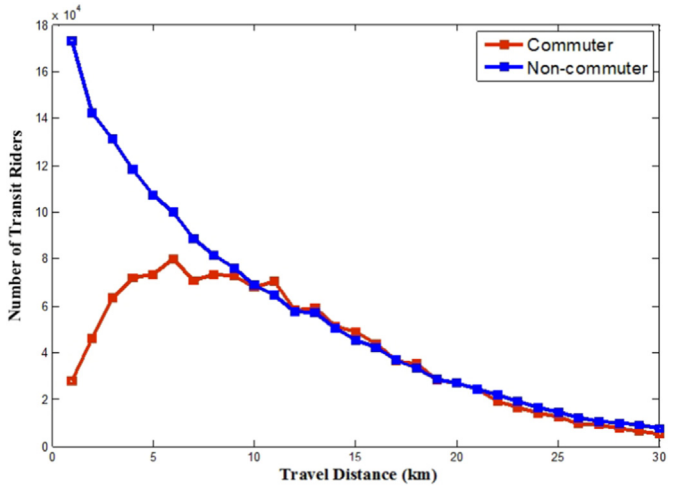

Fig. 7. Distribution of the travel distances during weekdays of transit commuters and noncommuters.

Appendix A

Bus/subway stop clustering

The DBSCAN algorithm is applied to group bus/subway stops that are adjacent to each other and to renumber each cluster as new IDs. This solution can address the issue of heterogeneous route choices. Transit riders can choose different routes with distinct boarding or alighting stops in commuting. In the traditional DBSCAN algorithm, the shape of a cluster is arbitrary. Thus, an excessively large radius of a narrowly shaped cluster is formed to include a series of stops that should be separated by multiple clusters. We improve the original DBSCAN algorithm by allowing the oversized clusters to recluster themselves until certain criteria are satisfied. Several useful parameters are defined as follows.

Minpts: the minimum number of stops that are included in a cluster. We set *Minpts* as one.

ε distance: it defines the density-reachable range. If at least *Minpts* stops are within ε distance of a certain stop, then that stop is a core point and the surrounding stops are directly reachable points from the core point. We set ε distance to 300 m.

$D_{\max}$ distance: we measure the maximum distance between the stops of each cluster. If the maximum distance of a cluster exceeds $D_{\max}$ distance, then reclustering is required with a new ε distance. We set $D_{\max}$ distance to 1000 m.

Δ distance: Δ distance is the decremental distance to adjust ε distance for reclustering. The new ε distance is computed by using ε distance minus Δ distance on each iteration. We set Δ distance to 25 m.

$\varepsilon_{\min}$ distance: The minimum ε distance of each cluster, which is set as 120 m.

The improved DBSCAN algorithm is documented as follows.

Step 1: randomly select one stop S that is not visited and search stops within the ε distance of stop S.

Step 2: if sufficient neighbor stops are found (i.e., the number of stops exceeds *Minpts*), then a cluster is formed and stop S is marked as visited. Otherwise, stop S is labeled as noise.

Step 3: select each neighbor stop of stop S and continue to search its ε neighborhood by repeating *Steps 1* and *2* until all stops within the cluster are visited.

Step 4: continue to process the remaining unvisited stops from *Steps 1* to *2* until all the stops are flagged as visited.

Step 5: examine the existing clusters and calculate the maximum distance between the stops of each cluster.

Step 6: if the maximum distance of a cluster exceeds $D_{\max}$ distance, then set the new ε distance as ε distance minus Δ distance and perform reclustering within this cluster.

Step 7: repeat *Steps 5* and *6* until any of the following criteria is satisfied: the maximum distance of each cluster is less than $D_{\max}$ distance and ε distance is less than $\varepsilon_{\min}$ distance.

Step 8: calculate the average latitude and longitude of all the stops with each cluster and assign a new stop ID to each cluster.

Calculation of transit commuting score

We use the TOPSIS method to compute the transit commuting score of each transit rider based on spatiotemporal regularities. The detailed calculation procedure is listed as follows.

Step 1: define the spatiotemporal regularities of each transit rider as $X=[N_{day},N_{route},N_{stop},N_{time}]$, and x_{ij} is each element of X, where $i=1,2,\cdots,m;j=1,2,3,4$; and m is the total number of transit riders. The normalized value r_{ij} of each spatiotemporal regularity can be calculated as $r_{ij}=x_{ij}/\sum_{i=1}^{m}x_{ij}$.

Step 2: calculate the entropy weight value w_j of each spatiotemporal regularity as $w_j = -k \cdot \sum_{i=1}^{m}(r_{ij} \lg r_{ij})$, where $k = 1/\ln m$. The normalized weighted matrix value v_{ij} is computed as $v_{ij} = r_{ij} \cdot w_j$.

Step 3: determine the positive and negative ideal solutions and calculate the Euclidean distance to each of these solutions for each transit rider:

Positive ideal solution: $A^+ = \{v_1^+, v_2^+, \cdots, v_j^+\} = \{\max v_{ij}\}$,

Negative ideal solution: $A^- = \{v_1^-, v_2^-, \cdots, v_j^-\} = \{\min v_{ij}\}$,

Distance to the positive ideal solution: $L_i^+ = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^+)^2}$,

Distance to the positive ideal solution: $L_i^- = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^-)^2}$.

Step 4: calculate the relative closeness to the ideal solution: $C_i = \frac{L_i^-}{L_i^+ + L_i^-}$.

Step 5: convert C_i into the transit commuting score: $score(i) = 50 \cdot \log 10(100 \cdot C_i) + 50$.

References

Ali, A., Kim, J., Lee, S., 2015. Travel behavior analysis using smart card data. KSCE J. Civ. Eng. 20 (4), 1532–1539.

Ball, G.H., Hall, D.J., 1965. Isodata: A Novel Method of Data Analysis and Pattern Classification. Stanford Research Institute, Menlo Park, United States.

Beijing Transportation Research Center, 2015. Annual Report of Beijing's Transportation Development (in Chinese). (Available at: http://www.bjtrc.org.cn/JGJS.aspx?id=5.2&Menu=GZCG).

Charron, M., 2007. From excess commuting to commuting possibilities: more extension to the concept of excess commuting. Environ. Plan. A 39, 1238–1254.

Hwang, C.L., Yoon, K., 1981. Multiple attribute decision making: methods and applications. Lect. Notes Econ. Math. Syst. 59–191 (New York).

Jiang, S., Ferreira, J., González, M.C., 2012. Clustering daily patterns of human activities in the city. Data Min. Knowl. Disc. 25, 478–510.

Kieu, L., Bhaskar, A., Chung, E., 2015a. A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data. Transp. Res. C 58 (Part B), 193–207.

Kieu, L., Bhaskar, A., Chung, E., 2015b. Passenger segmentation using smart card data. IEEE Trans. Intell. Transp. Syst. 16 (3), 1537–1548.

Kung, K., Greco, K., Sobolevsky, S., Ratti, C., 2014. Exploring universal patterns in human home-work commuting from mobile phone data. PLoS One 9 (6), e96180.

Kusakabe, T., Asakura, Y., 2014. Behavioural data mining of transit smart card data: a data fusion approach. Transport. Res. C 46, 179–191.

Langlois, G.G., Koutsopoulos, H.N., Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transp. Res. C 64, 1–16.

Li, Z., Wang, W., Yang, C., Ragland, D.R., 2013a. Bicycle commuting market analysis using attitudinal market segmentation approach. Transport. Res. A Policy Pract. 47, 56–68.

Li, Z., Wang, W., Yang, C., Jiang, G., 2013b. Exploring the causal relationship between bicycle choice and trip chain pattern. Transp. Policy 29, 170–177.

Long, Y., Zhang, Y., Cui, C., 2012. Analysing jobs-housing relationship and commuting patterns of Beijing using bus smart card data (in Chinese). Acta Geograph. Sin. 67, 1339–1352.

Louail, T., et al., 2014. Uncovering the spatial structure of mobility networks. Nat. Commun. 6.

Louf, R., Barthelemy, M., 2014. How congestion shapes cities: from mobility patterns to scaling. Sci. Rep. 4.

Lovelace, R., Ballas, D., Watson, M., 2014. A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels. J. Transp. Geogr. 34, 282–296.

Ma, X., Wang, Y., 2014. Development of a data-driven platform for transit performance measures using smart card data and GPS data. J. Transp. Eng. 140, 04014063.

Ma, X., Wang, Y., Chen, F., Liu, F., 2012. Transit smart card data mining for passenger origin information extraction. J. Zhejiang Univ. Sci. C 13, 750–760.

Ma, X., Wu, Y., Wang, Y., Chen, F., Liu, F., 2013. Mining smart card data for transit riders' travel patterns. Transport. Res. C 36, 1–12.

Mahrsi, M.K.E., Côme, E., Baro, J., Oukhellou, L., 2014. Understanding passenger patterns in public transit through smart card and socioeconomic data. UrbComp'14 (New York, NY, USA).

Memarsadeghi, N., Mount, D.M., Netanyahu, N.S., Le Moigne, J., 2007. A fast implementation of the ISODATA clustering algorithm. Int. J. Comput. Geom. Appl. 17 (1), 71–103.

Microsoft Cooperation, 2016. Transact-SQL Reference (Database Engine). (Available online: https://msdn.microsoft.com/en-us/library/bb510741.aspx (accessed on 5 October, 2016)).

Morency, C., Trépanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transp. Policy 14 (3), 193–203.

Ortega-Tong, M.A., 2013. Classification of London's Public Transport Users Using Smart Card Data. (Master thesis). MIT.

Pelletier, M.P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. C 19 (4), 557–568.

Primerano, F., Taylor, M.A.P., Pitaksringkarn, L., Tisato, P., 2008. Defining and understanding trip chaining behavior. Transportation 35, 55–72.

Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M.C., Toroczkai, Z., 2014. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nat. Commun. 5.

Schneider, C.M., Rudloff, C., Bauer, D., González, M.C., 2013. Daily travel behavior: lessons from a week-long survey for the extraction of human mobility motifs related information. UrbComp'13 (Chicago, Illinois, USA).

Van Acker, V., Witlox, F., 2011. Commuting trips within tours: how is commuting related to land use? Transportation 38, 465–486.

Wang, Y., 2014. Research on Methods of Extracting Commuting Trip Characteristic Based on Public Transportation Multi-source Data (in Chinese). (Master Thesis). Beijing University of Technology.

Yang, Y., Herrera, C., Eagle, N., González, M.C., 2014. Limits of predictability in commuting flows in the absence of data for calibration. Sci. Rep. 4.

Zhao, P., 2013. The impact of the built environment on individual workers' commuting behavior in Beijing. Int. J. Sustain. Transp. 7, 389–415.

Zhou, J., Murphy, E., Long, Y., 2014. Commuting efficiency in the Beijing metropolitan area: an exploration combining smart card and travel survey data. J. Transp. Geogr. 41, 175–183.