Computers, Environment and Urban Systems 53 (2015) 19–35

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing

Ying Long [a,*], Jean-Claude Thill [b]

[a] Beijing Institute of City Planning, Beijing 100045, China
[b] Department of Geography and Earth Sciences, The University of North Carolina at Charlotte, Charlotte, NC 28223, USA

ARTICLE INFO

Article history:
Available online 4 April 2015

Keywords:
Bus smart card data
Jobs–housing spatial mismatch
Commuting trip
Rule-based
Beijing

ABSTRACT

Location Based Services (LBS) provide a new perspective for spatiotemporally analyzing dynamic urban systems. Research has investigated urban dynamics using LBS. However, less attention has been paid to the analysis of urban structure (especially commuting pattern) using smart card data (SCD), which are widely available in most large cities in China, and even in the world. This paper combines bus SCD for a one-week period with a oneday household travel survey, as well as a parcel-level land use map to identify job–housing locations and commuting trip routes in Beijing. Two data forms are proposed, one for jobs–housing identification and the other for commuting trip route identification. The results of the identification are aggregated in the bus stop and traffic analysis zone (TAZ) scales, respectively. Particularly, commuting trips from three typical residential communities to six main business zones are mapped and compared to analyze commuting patterns in Beijing. The identified commuting trips are validated by comparison with those from the survey in terms of commuting time and distance, and the positive validation results prove the applicability of our approach. Our experiment, as a first step toward enriching LBS data using conventional survey and urban GIS data, can obtain solid identification results based on rules extracted from existing surveys or censuses.

1. Introduction

This paper identifies job–housing location dyads and commuting patterns in Beijing using smart card data (SCD) that store the daily trip information of bus passengers. It proposes and implements a method for deriving commuting patterns from increasingly common SCD for informing city planners and transit system managers about patterns of transit usage across space and through time as well as about mobility patterns in a large and fast growing city region. Related research on jobs–housing relationships has conventionally used data acquired through surveys or censuses. The increasing pervasiveness of location-based services (LBS) associated with the prevalence of positioning technologies has led to the creation of large-scale and high-quality space-time datasets (Jiang & Yao, 2006). This development has also created opportunities to better describe and understand urban structures[1] in multiple dimensions. These datasets have been shown to be important for analyzing urban and environmental systems such as relationships between housing and jobs (Batty, 1990). Meanwhile, a geo-tagged smart card system is an effective alternative tool for individual data acquisition necessary to analyze urban spatial structures.

Various types of fine-granularity individual data generated by LBS technologies have been extensively leveraged to analyze urban structures (Ahas & Mark, 2005; Lu & Liu, 2012). With respect to handheld Global Positioning System (GPS) devices, Newhaus (2009) used location data to record and visualize urban diaries, while Gong, Chen, Bialostozky, and Lawson (2012a) elicited travel modes of travelers in New York City. Liu, Andris, and Ratti (2010) identified taxi drivers' behavior patterns from their daily digital trajectories, and Yue et al. (2012) used these trajectories to calibrate a spatial interaction model. With respect to mobile phone systems (see Steenbruggen, Borzacchiello, Nijkamp, & Scholten, 2013 for a review), Ratti, Pulselli, Williams, and Frenchman (2006) evaluated the density and spatiotemporal characteristics of urban activities using mobile phone data in Milan, Italy, whereas Wan and Lin (2013) studied fine-scale individual activities, Yuan, Raubal, and Liu (2012) correlated mobile phone usage and city-wide travel behavior in Harbin, China and Chi, Thill, Tong, Shi, and Liu (In press) exploited network properties of mobile phone

* Corresponding author. Tel.: +86 10 88073660; fax: +86 10 68031173.
E-mail address: longying1980@gmail.com (Y. Long).

[1] The concept of "urban structure" concerns the spatial arrangement of public and private spaces in cities and the degree of connectivity and accessibility. In this paper, the concept is focused particularly on the spatial concentration of resident population and employment (Anas, Arnott, & Small, 1998).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.02.005
0198-9715/

data to reveal urban hierarchical structures at the regional scale. As for Wi-Fi, Rekimoto, Miyaki, and Ishizawa (2007) used Wi-Fi-based location detection technology to log the locations of device holders from received Wi-Fi beacon signals, a technology that works both indoors and outdoors. Torrens (2008) developed a system to detect Wi-Fi infrastructure and transmission and analyze their geographic properties, and tested this system in Salt Lake City, Utah.

Meanwhile, the discipline of time geography established by Hagerstrand (1970) also benefited from the development of LBS by retrieving more objective data. In sum, various LBS technologies have been successfully applied in urban studies. However, these technologies remain immature and most research on urban structure continues to employ data from the urban physical space or questionnaire surveys (with a few studies as exceptions, e.g. Kwan (2004)). Access to large-scale micro datasets remains a barrier to their widespread use for research, planning and management (Long & Shen, 2013).

A smart card that records full cardholder's bus trip information is an alternative form of location-acquisition technology. Smart card automated fare collection systems are increasingly deployed in public transit systems. Along with collecting revenue, such systems can capture a meaningful portion of travel patterns of cardholders, and the data are useful for monitoring and analyzing urban dynamics. Since the 1990s, the use of smart cards has become significant owing to the development of the Internet and the increased complexity of mobile communication technologies (Blythe, 2004). As of 2007, Intelligent Transportation Systems (ITS) that incorporate smart card automated fare systems either existed or were being established in over 100 Chinese cities, as well as in many other cities around the world (Zhou, Zhai, & Gao, 2007). The data generated by smart card systems track the detailed onboard transactions of each cardholder. We argue that smart card technology can deliver valuable information because it is a continuous data collection technique that provides a complete and real-time bus travel diary for all bus travelers. SCD can be used to validate traditional travel models applied to public transit. In contrast to SCD collection, conventional travel behavior surveys have the drawbacks of being expensive and infrequent. Notably, transit SCD collects data in fundamentally the same way as an AVI (automatic vehicle identification) system, which has been widely used in the United States to automatically identify vehicles. AVI is used in some states in the US for planning purposes. One such example is New York, where the E-ZPass tag is used as part of the TRANSMIT system.

Previous studies have advocated using SCD to make decisions on the planning and design of public transportation systems (see Pelletier, Trepanier, and Morency (2011) for a review). In South Korea, Joh and Hwang (2010) analyzed cardholder trip trajectories using bus SCD from ten million trips by four million individuals, and correlated these data with land use characteristics in the Seoul Metropolitan Area. Jang (2010) estimated travel time and transfer information using data on more than 100 million trips taken in Seoul on the same system. Roth, Kang, Batty, and Barthélemy (2011) used a real-time "Oyster" card database of individual traveler movements in the London subway to reveal the polycentric urban structure of London. Gong et al. (2012b) explored spatiotemporal characteristics of intra-city trips using metro SCD on 5 million trips in Shenzhen, China. Also, Sun, Axhausen, Lee, and Huang (2013) used bus SCD in Singapore to detect familiar "strangers".

There is considerable research on inferring home and job locations from individual trajectories like mobile phone call data records and location-based social networks (LBSN). For the identification of home locations, Lu, Wetter, Bharti, Tatem, and Bengtsson (2013) regarded the location of the last mobile signal of the day as the home location of a mobile user. The most frequently visited point-of-interest (POI) (Scellato, Noulas, Lambiotte, & Mascolo, 2011) or grid cell (Cheng, Caverlee, Lee, & Sui, 2011; Cho, Myers, & Leskovec, 2011) was regarded as a LBSN user's home location. It is not easy to infer home locations from LBSN with a high spatial resolution. Compared to approaches to home location identification, there are fewer studies on identifying job locations based on trajectories, with Cho et al. (2011) using LBSN and Isaacman et al. (2011) using cellular network data as notable exceptions. It should be mentioned that taxi trajectories are not well suited for identifying a passenger's home and job locations considering the passenger-sharing nature of taxis. However, less attention has been paid to using SCD to identify home and job locations as well as to analyze job–housing dyadic relationships and commuting patterns in a metropolitan region.

This paper regards job–housing dyadic relationships and commuting pattern analysis as a showcase for using SCD to urban spatial analysis. We argue that job and home locations, their dyadic relationships, and commuting trips can be identified from SCD and serve as valuable information on the modalities of use of the urban space in its residents. We propose a methodology to this effect and use Beijing as a case study to test its implementation. The identification results are validated using travel behavior survey data from Beijing. This paper is organized as follows. The retrieval of job–housing trips from conventional travel behavior surveys is discussed in Section 2, and the SCD and other related datasets used in our research are presented in Section 3. The approaches for identifying home and job locations, as well as commuting trips are elaborated in Section 4. In Section 5, the results of job–housing identification and commuting patterns are shown and analyzed in detail. Finally, we discuss our work and present concluding remarks in Sections 6 and 7, respectively.

2. Job–housing trips in conventional travel behavior surveys

Travel behavior surveys have been the primary means of data collection on urban resident travel behavior for planning and managing urban transit systems (Beijing Transportation Research Center, 2009). There is a well-established tradition in geography and urban planning to use surveys for tracking individual travel diaries (Gärling, Kwan, & Golledge, 1994; Schlich, Schönfelder, Hanson, & Axhausen, 2004). Travel behavior surveys track traveler socio-economic attributes, as well as trip origin and destination, time and duration, as well as trip purposes and travel modes. On the one hand, the traveler's home and job locations are directly recorded in the survey together with his/her socioeconomic attributes, and both locations are mostly aggregated in the traffic analysis zone (TAZ) scale. On the other hand, trips between work and home (i.e. commuting trips) can be screened using the purpose attribute. These trips are also recorded using the inter-TAZ scale rather than a finer spatial scale. Therefore, job–home location dyads and trips have already been recorded in conventional travel surveys, but mostly at the TAZ scale. Additionally, only a small portion of all households in a given city are surveyed due to time and cost constraints.

Compared with travel behavior surveys, mining the enormous volume of SCD can provide a more precise spatial resolution and a much larger sample, despite the SCD being unable to directly provide job–home location dyads and commuting trips. We will focus on using SCD to identify jobs–housing relationships. Patterns of commuting trips from typical residential communities or to typical business centers can be visualized by identifying the results at a finer scale than is available in travel surveys because residential communities or business zones are generally smaller than a TAZ. A more detailed commuting pattern is expected to reveal fresh information on jobs–housing relationships in a megacity such as

Beijing. A shortcoming of SCD however is that they are devoid of information on the cardholder's socioeconomic attributes, and the purpose of individual trips is also unknown. Conventional travel surveys can supply such additional information for use in analyzing SCD, and combining SCD with travel behavior survey data is a promising method of job–housing analysis, which will be elaborated below.

3. Data

3.1. Bus routes, bus stops, and traffic analysis zones (TAZs) of Beijing

Geographic Information Systems (GIS) layers of bus routes and stops are essential for geocoding and mapping SCD. There are 1287 bus routes[2] (Fig. 1a) in the Beijing Metropolitan Area (BMA), which totals 16,410 km^2. These bus routes have 8691 stops in total (see Fig. 1b). Note that a pair of bus platforms on opposite sides of a street is considered a single bus stop. For instance, there are two bus platforms at Tian'anmen Square, one on the south side of Chang'an Avenue and the other one on the north side. In the GIS bus stop layer, the two platforms are merged into a single bus stop feature.[3] The average distance between a bus stop and its nearest neighbor is 231 m in the city. Relying on Ji and Gao's (2010) result that the number of bus stops within an 800 m vicinity of a resident has a significant effect on their satisfaction with public transportation services in Beijing, we take the 800-m buffer zone around each stop to be its catchment zone, so that the potential service area of a bus stop is 2.0 km^2. We overlay the calculated bus catchment with the population density surface inferred from the 2010 sub-district level population census. The estimated population within the catchment zone of any of the 8691 stops is 14.8 million, or 75.5% of Beijing's 19.6 million residents.

We use Beijing TAZs to aggregate the analytical results for better visualization. In total, 1118 TAZs are defined (Fig. 1c) according to the administrative boundaries, main roads, and the planning layout in the BMA.

3.2. The one-week smart card dataset

A smart card system has been deployed in the public transit system of Beijing since April 1, 2006 (Liu, 2009). The system can automatically track cardholder's bus trip information.[4] The bus share of total trips taken in Beijing during 2008 was 28.8%, and the subway share was 8.0% (Beijing Transportation Research Center, 2009). Over 42 million smart cards (see Fig. 2) have been issued in Beijing till 2011, and over 90% of all bus trips are recorded by smart cards.[5] The smart cards used by the Beijing public transit system are anonymous and users provide no personal information when applying for them. A person can buy any number of cards in Beijing and one does not need show any photo ID. A small portion of Beijing's public transit passengers holds several cards, which may be used by relatives while visiting Beijing. It is not common for a cardholder to use several cards in a day since there is no incentive for him/her to do so. One extreme circumstance would be that one card has no cash balance left and the owner would use another card as a substitute. Considering that the bus fare in Beijing is only 0.4 CNY in 2008, such extreme case is rare in the personal experience of the authors.

The Beijing SCD used in this study were obtained from Beijing Municipal Administration & Communications Card Co., Ltd; they covered a one-week period of 2008 (April 7–13)[6] and did not include subway records. The SCD records several essential fields (see Table 1). The data comprise 77,976,010 bus trips by 8,549,072 cardholders (cardholders with only subway rides and no-bus-trips are not included), and thus each cardholder makes an average of 1.30 bus trips per day.

Two fare types exist on Beijing's bus system. The first is fixed and does not depend on the distance traveled, which is associated with short routes, while the other is a distance-fare, which is associated with longer routes (see Table 2 for a comparison). For the first type, a riding cardholder is charged a flat fee of 0.4 CNY for each single bus trip. For most trips on fixed-fare routes, the corresponding SCD record contains only the departure time, excluding the departure stop ID, arrival time and arrival stop ID. Thus, cardholders' spatiotemporal information is incomplete for this kind of route. For the latter fare type, the fare depends both on the route ID and trip distance, and the SCD contains full information. We consider the SCD of both fare types, including fixed-fare and distance-fare. However, due to the largely incomplete information on fixed-fare trips, the identification of a home or job location may not be possible. For instance, if the first trip of a cardholder on a given day is a fixed-fare trip, it is impossible to identify his/her home location since the location where they first boarded is unknown. However, in some cases, such as when a fixed-fare trip is taken as a transfer between two distance-fare trips, it is still possible to identify the home location.

The temporal and spatial dimensions of the SCD are described here. The total count of bus trips shows significant variability across days of the week (Fig. 3a). The total number of bus trips on a weekday (Tuesday) and a weekend day (Saturday) for each departure hour is displayed in Fig. 3b, and shows significant differences between the two days. Most bus trips are distributed from 6:00 to 22:00 and match peak hours in the 2005 BMA travel behavior survey. With respect to the spatial dimension, the total daily bus trip density in the inner area of the BMA exceeds that in the outer area in terms of boarding locations (Fig. 3c). Fig. 3c is prepared by using the head/tail breaks approach proposed by Jiang (2013) for data with a heavy-tailed distribution.

3.3. The Beijing travel behavior survey

Travel behavior surveys have been conducted in 1986, 2000 and 2005 in Beijing. The 2005 Beijing travel behavior survey (hereafter called the 2005 survey) is included in this paper to set rules for identifying job–housing location dyads and commuting trips. This survey covers the whole BMA, including all 18 districts, with 1118 TAZs (as shown in Fig. 1c) (Beijing Municipal Commission of Transport, 2007). The sampling size is 81,760 households/208,290 persons, with a 1.36% sampling rate. This survey adopts a travel diary form. For each trip, the survey records the departure time/location, arrival time/location, trip purpose and mode, as well as other important information such as the trip length, destination building type, and transit route number. It should be noted that the original survey data have been processed in the form of trips through the merging of consecutive trip segments reported in

[2] The distinction between fixed-fare and distance-fare routes is discussed later.

[3] We recognize that very broad streets are common in Beijing so that our stop merger rule may result in excessive loss of information. We plan to use the unique locations of stops as platforms in future research. Considering that most of the research results are aggregated at the TAZ scale, we merge those bus stops on opposite sides of a broad street in this paper, which would not influence the results of the analysis in most cases.

[4] Beijing's bus smart card is operated by Beijing Municipal Administration & Communications Card Co., Ltd. The official website is www.bjsuperpass.com. Cardholder bus trips over several weeks can be queried by users via this website by inputting a card ID (Fig. 2).

[5] See http://news.rfidworld.com.cn/2011_03/389042f2a28b3d53.html. In addition, cardholders can get discounts of 60% (for regular bus riders) or 80% (for students) off the regular fare by using smart cards in Beijing, which incentives bus riders to use smart cards.

[6] In April 2008, the policy of weekly 'no driving days' for car owners had not yet been implemented.

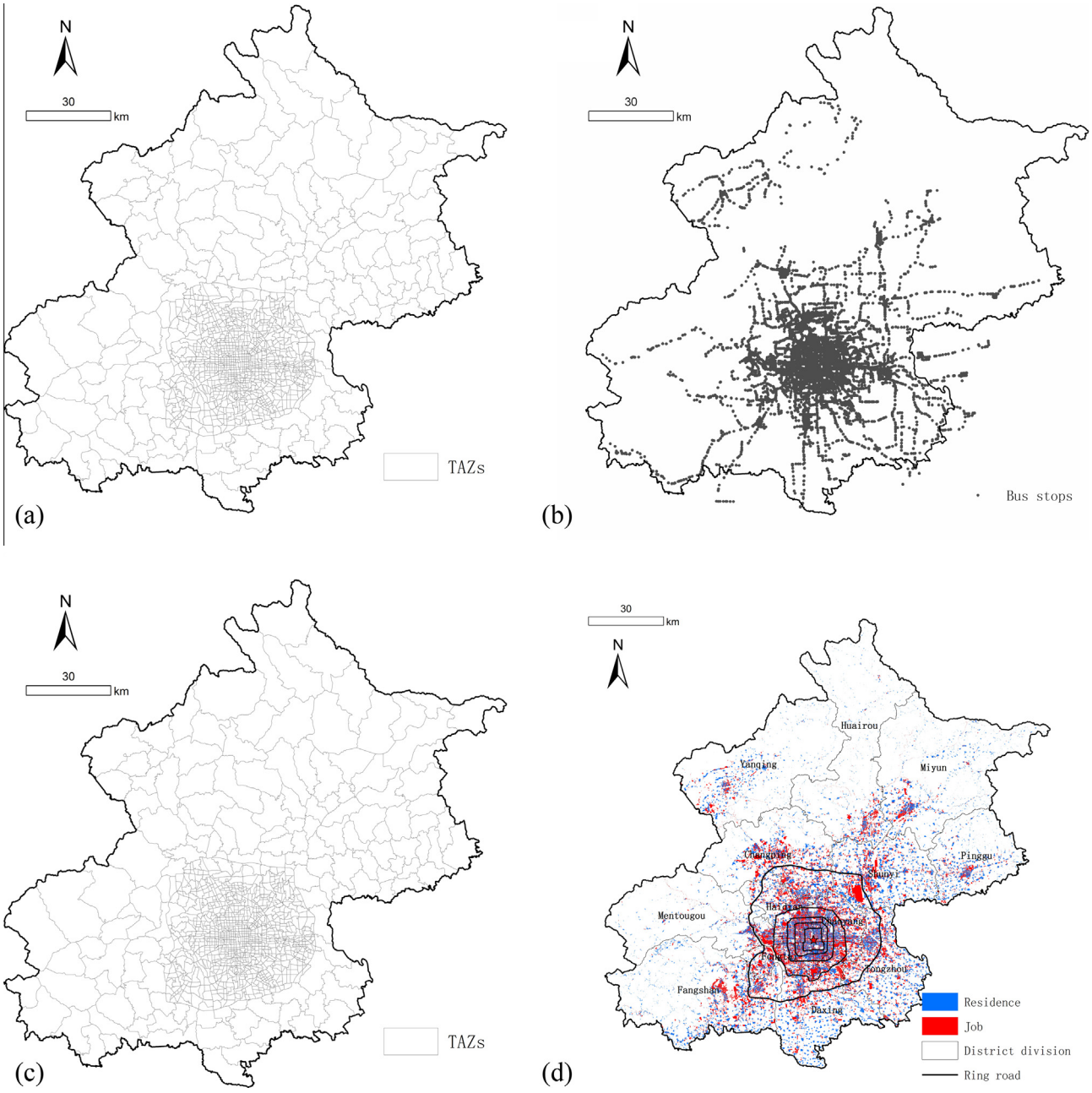

Fig. 1. Bus routes (a), bus stops (b), traffic analysis zones (TAZs) (c), and land use patterns (d) of the BMA. *Note:* Maps are from the Beijing Institute of City Planning. Some bus routes and stops are outside the BMA, as shown in (a) and (b), since some frequent traveler inside the BMA actually live outside the BMA, in neighboring towns of Hebei province. In the TAZ-level analysis conducted in this paper, trips outside the BMA recorded in the SCD are not counted due to the lack of TAZ data. The five nested circles in (d) represent the second, third, fourth, fifth, and sixth ring roads of Beijing. The star in the central area represents Tian'anmen Square.

Fig. 2. The bus smart card of Beijing ((a) front side; (b) back side).

Table 1
Data structure of the SCD.

Category	Variable	Examples of values
Card information	Card ID	"10007510038259911", "10007510150830716"
	Card type	1, 4
Route information	Route ID	602, 40, 102
	Route type	0, 1
	Driver ID	11032, 332
	Vehicle ID	111223, 89763
Trip information	Trip ID	25, 425, 9
	Departure date (YYYY-MM-DD)	2008-04-08
	Departure time (HH-MM-SS)	"06-22-30", "11-12-09"
	Departure stop	11, 5, 14
	Arrival time (HH-MM-SS)	"09-52-05", "19-07-20"
	Arrival stop	3, 14, 9

Note: 0 stands for a fixed-fare route and 1 stands for a distance-fare route for the attribute "Route Type". For the attribute "Card Type", 1–4 denote normal, student, staff and monthly pass, respectively. The attribute "Trip ID" represents the accumulated trip count on a card since its issue, including both subway and bus journeys.

Table 2
Comparisons between fixed-fare and distance-fare routes and SCD.

	Fixed-fare	Distance-fare
Route count	566	721
Total length (km)	7529.1	25812.6
Average length (km)	13.3	35.8
Trip count	50,916,739 (65.3%)	27,059,271 (34.7%)

the original travel diaries. Additionally, recognized trip purposes include: (1) work, (2) school, (3) returning home from school or work, (4) returning-home trip from other purposes, (5) shopping, (6) entertainment, (7) daily life (such as dining, medical, social visit, leisure/fitness, and pick up/delivery), (8) business, and (9) other. Trip modes include: (1) walk, (2) bicycle, (3) electric bicycle, (4) motorbike, (5) bus, (6) mini bus, (7) metro, (8) employer-provided bus, (9) private car, (10) employer-provided car, and (11) legal and illegal taxi. Among all these transportation modes, the bus accounts for 27.08% of all non-walking trips in the BMA according to this survey (3.61% for metro, 33.19% for bicycle and 25.54% for private car).

The survey also includes household and personal information. The household information includes household size, Hukou (official residency registration) status and residence location, while the personal information includes gender, age, household role, job type and location, and whether the respondent holds a driver's license or transit monthly pass. Job types include: (1) worker, (2) researcher, (3) office/public employee, (4) teacher, (5) student, (6) self-employed, (7) household attendant, (8) retiree, (9) specialized worker (such as medical staff, professional driver, and soldier/police), (10) farmer, (11) unemployed, and (12) other.

3.4. Land-use pattern of Beijing

Land-use type of each land parcel in the BMA is a critical element of our approach to uncover job–housing location dyads and commuting patterns. These data are introduced to identify home and job locations. Floor area data are available for each land parcel. We assume that the residential land-use type represents home locations, and that the commercial, public facility and industrial land-use type represents job locations. The 133,503 parcels include 29,112 residential parcels, and 57,285 parcels with job locations (labeled "job parcels" in this paper) (Beijing Institute of City Planning, 2010).[7]. The land-use pattern is used to calculate the probability of each bus stop servicing a home or job location.

[7] Land use mix at the parcel level is not considered in this study.

4. Data processing and analytical approach

4.1. Data pre-processing and data forms

As the raw data recording cardholders' bus riding information, the SCD need to be pre-processed to facilitate the job–housing analysis and evaluate bus trips spatiotemporally. First, we geocode the SCD by linking the bus stop ID in the SCD with the bus stop layer in GIS. Second, we combine the trips of each cardholder to retrieve their full bus travel diary (BTD), which records information on all the trips each cardholder takes on each day, as well as their card type. The BTD is the basic data used for further job–housing analysis. Following this, the trip count, total bus trip duration, total bus trip length, start point, and end point on each day of the week can be calculated for each cardholder using the generated BTD data. Since the BTD does not include subway rides due to restrictions on raw data availability, cardholders with non-consecutive "Trip ID" attributes are regarded as possibly engaged in a subway ride and are removed from the analysis to avoid an identification bias.[8] The jobs–housing relationships are analyzed using the pre-processed BTD data following the approach below.

We propose two data forms for representing the SCD of each cardholder on each day, trip (TRIP), and position–time–duration (PTD). In TRIP, a trip denotes one record in the SCD, which comprises a cardholder boarding and alighting, namely one bus ride. In the SCD, a trip (TRIP) is stored as its departure location (OP) and time (OT), as well as its arrival stop (DP) and time (DT), as TRIP = {OP, OT, DP, DT}. TRIP is a direct expression of the BTD.

The PTD data form, as an alternative to TRIP, is converted from the TRIP data form and can describe an activity's spatiotemporal characteristics. The generation of PTD assumes that a cardholder does not use travel modes other than the bus. For a cardholder, PTD is expressed as PTD = {P, t, D}, where P is a bus stop around which the cardholder stays to perform some activity, t is the start time of the activity at location P, and D is the temporal duration at the location P. Compared with TRIP, PTD better matches the time geography and can identify various types of urban activities. We need to convert TRIP to PTD and use an example to show how to do so. Let us assume that a cardholder leaves home (bus stop H0) at 7:00 and travels by bus to arrive at their work location (bus stop J0) at 8:00. After working for a full-day, the cardholder leaves their workplace at 17:00 and travels by bus to arrive home at 18:00. The TRIP data form for the two trips is expressed as {H0, 7:00, J0, 8:00} and {J0, 17:00, H0, 18:00}.

[8] For example, let us consider a cardholder making four trips in a week and the Trip IDs are 13, 14, 15, and 17, respectively. The trip with ID = 16 is missing from the raw SCD. We suppose the trip with ID = 16 is a subway ride. Under these conditions, all records of this cardholder are excluded from our analysis.

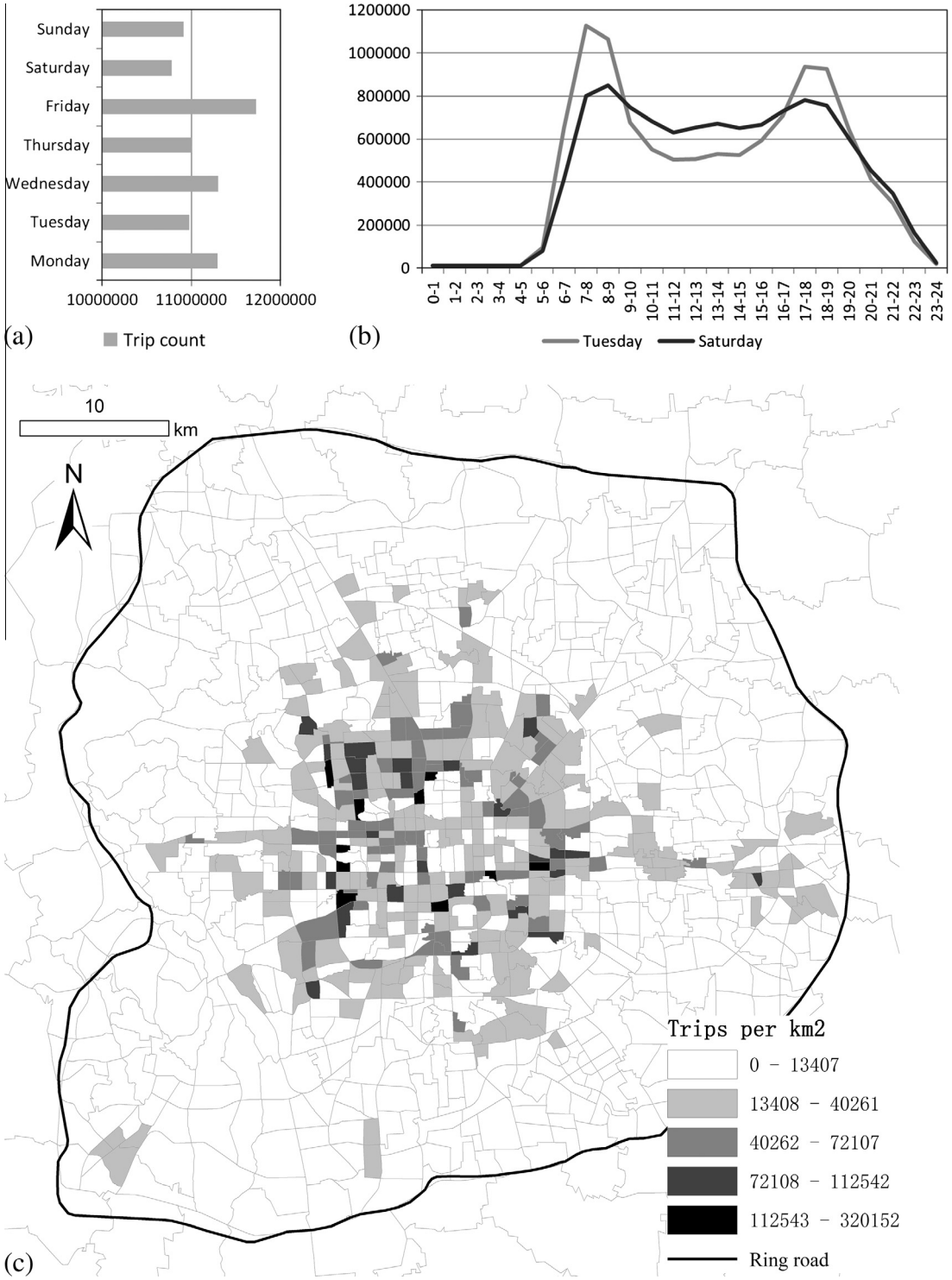

Fig. 3. Space-time characteristics of bus trips in the SCD. (a) Total count of bus trips by day; (b) Trip count by departure hour on Tuesday and Saturday; (c) Trip count densities in each TAZ within the 6th ring road of the BMA.

The converted PTD data form is then expressed as {H0, 18:00 (−1), 13 h} and {J0, 8:00, 9 h} and represents two activities, first the home activity and then the work activity. The home activity starts at 18:00 on the previous day and lasts 13 h till 7:00. The work activity starts at 8:00 and lasts for 9 h.

4.2. Identification of home and job location dyads using one-day data

We use the PTD data form to identify home and job locations for each cardholder. How to identify home and job locations using one-day data is the first step of our approach.

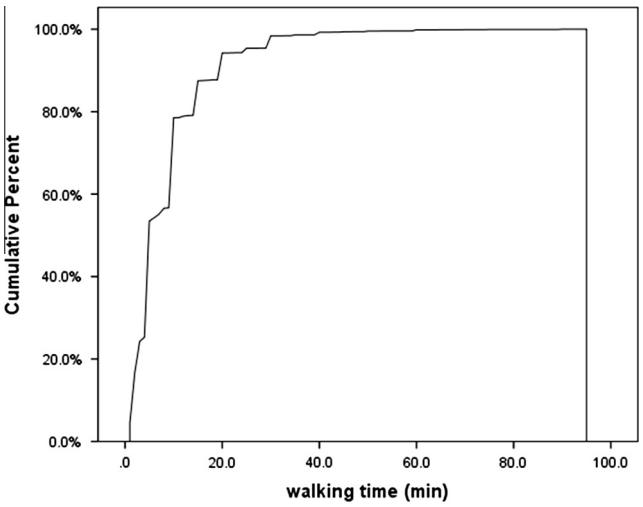

Fig. 4. Cumulative probability distribution of trip walking time in the 2005 survey.

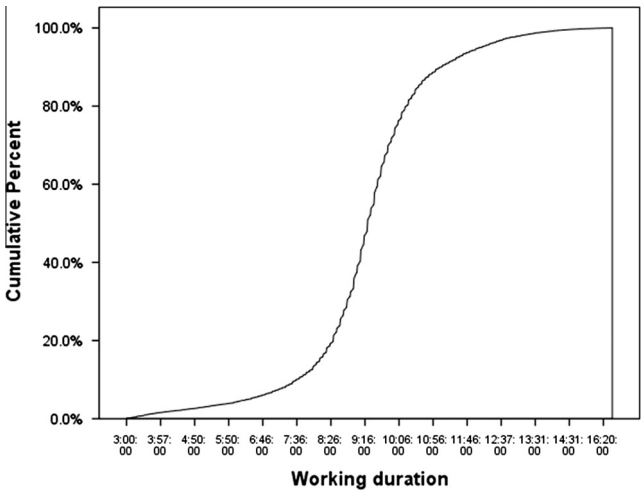

Fig. 5. Cumulative probability distribution of working time extracted from the 2005 survey.

To identify home locations, we suppose the departure bus stop of the first trip (TRIP1) to be the home location of a cardholder.[9] The home location is assumed to be within walking distance of this bus stop, which is the spatial resolution of our analytical results. In the 2005 survey, trips with the mode "walking" can be screened to calculate walking duration. The histogram of walking time illustrated in Fig. 4 shows that the average walking time is 9.0 min in the 2005 survey for trips of all purposes (walking trip segments linking other travel modes not included). Accordingly, the average walking distance is estimated to be 750 m, assuming an average walking speed of 5 km/h (Bohannon, 1997). This is the distance a bus rider is likely to walk from their home or job location to ride a bus, or to reach their destination after riding a bus. In the 2005 survey, 99.5% of first trips started from home, which supports our rule for identifying home locations. If a cardholder's first trip of the day is on a fixed-fare bus route, we cannot identify the home location for this cardholder.

To identify job locations, we need to identify work trips by bus. Full-time job locations are identified based on the interval between any two adjacent cardholder trips being long enough for a full-time job. This method assumes that the full-time job activity is the urban activity conducted for the longest time on week days. If a cardholder meets all the conditions below, the kth location P_k is regarded as their job location.

Condition 1: The card is not a student card.
Condition 2: $D_k \geqslant 360$.
Condition 3: $k <> 1$.[10]

That is, for non-student cardholders, if more than 360 min (6 h) is spent at any location other than their first location, we assume this location is their place of work. The benchmark of 6 h is based on the 2005 survey, in which the average working time is 9 h and 19 min (with standard deviation of 1 h and 41 min, see Fig. 5 for details) for a sample of 27,550 persons (210 persons went home for a rest at noon, and their data were not counted). Thus 96% of sampled persons work for over 6 h per day. Notably, the process of identifying home locations is independent from that of identifying job locations.

4.3. Identification of home and job locations using one-week data

Because the home and job locations identified for a cardholder by the approach presented in Section 4.2 may differ for each day, we propose several indicators to combine the one-day results into a single home and job location dyad that is consistent over the one-week period.

To identify a single home location from the one-week data, we apply a rule-based method to one-day results (see Fig. 6). Both the frequency and spatial distribution of locations identified using one-day data are used in this process. In addition, we introduce the new term of "cluster" to encompass locations that are within a certain distance threshold of each other. We set this threshold at 500 m, which is about twice the average distance between two adjacent bus stops ($231 * 2 = 462$ m). This threshold is also based on the findings of Zhao, Lv, and de Roo (2011) that the threshold service distance of bus stops is 500 m. If several distinct clusters are associated with a certain cardholder and each one encompasses a single bus stop, the stop that corresponds to the home location cannot be identified with confidence. Furthermore, if a single cluster has the largest number of locations (the maximum cluster), the most frequent location in this cluster is taken as the home location. Otherwise, when multiple clusters have the same largest number of locations, the most frequent location in these clusters is regarded as the bus rider's home location; this situation applies when the most frequent location counts are different in maximum clusters. Finally, when the most frequent location counts are the same in the maximum clusters, the location with highest frequency in the maximum cluster that also exhibits the greatest residential potential is regarded as the final location of the cardholder.

In the case of two locations with the same frequency, the concept of "residential potential" is introduced to impute a location (the concept of "job potential" is also used to determine the final job location).[11] The potential is calculated on the basis of the land use pattern data using Eq. (1), in which p_h^k is the housing potential of bus stop k, p_j^k is the job potential of bus stop k, and the neighborhood of bus stop k is the Voronoi polygon generated from the bus stop layer. A parcel's centroid determines the neighborhood the

[9] Notably, there could be a minor bias in identifying the home location since a cardholder may have taken a taxi to a bus stop prior to their first bus ride.

[10] According to the definition of the PTD data form, the first activity of a cardholder is the home-based activity started on the previous day. This rule guarantees that the identified job location is not the home location.

[11] The land use pattern data were used as supplementary data for inferring home and job locations of cardholders. It is not commonly used in our identification, and only used in this special condition. Therefore, the data are not a precondition for inferring home and job locations, but supplementary materials.

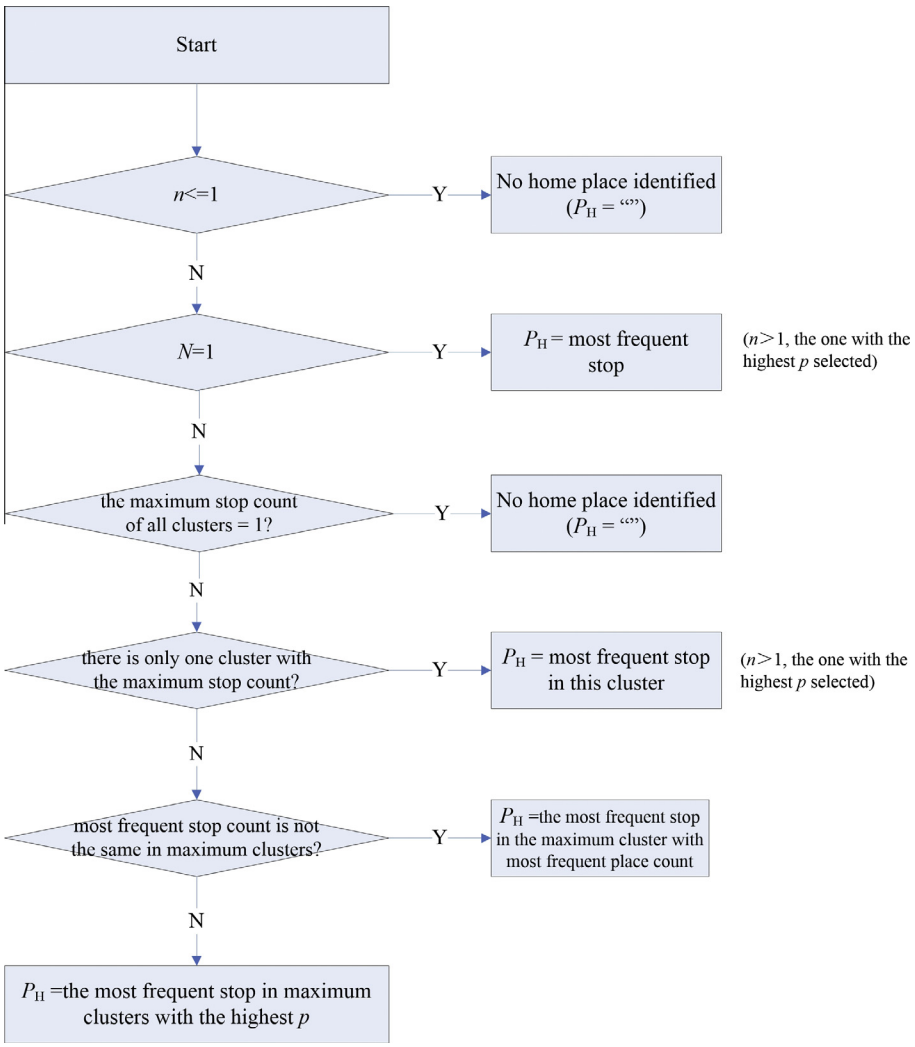

Fig. 6. Decision tree diagram for identifying home locations based on one-day results. Note that n is the count of home locations identified for a cardholder during a week, N is the count of clusters of the cardholder, and P_H is an imputed home location for this cardholder. p is the potential of a home location.

parcel belongs to. The two potential indicators are further rescaled to range from 0 to 1.

$$p_n^k = \frac{\text{the total floor area of housing parcels in the neighborhood of bus stop } k}{\text{the total floor area of all parcels in the neighborhood of bus stop } k}$$
$$p_j^k = \frac{\text{the total floor area of job parcels in the neighborhood of the stop } k}{\text{the total floor area of all parcels in the neighborhood of bus stop } k} \quad (1)$$

Similarly, the rules for imputing job locations from one week of SCD follow the same approach as the rules for home locations.

4.4. Identification of commuting trips based on identified home and job locations

We use the TRIP data form to identify the commuting trip[12] from a home location to a job location for cardholders with both an identified home location and an identified job location using one-week SCD. Commuting distance and time are used as key indicators for measuring commuting patterns. Commuting distance is measured as the Euclidean distance between the home and job location. Commuting time is taken as the time duration between the boarding time at the home location and the arrival time at the job location. The calculation requires identifying the commuting trip (corresponding to one or several bus trips/ridings in the TRIP data) from one-week trips where the cardholder meets the following three conditions: (1) The boarding bus stop of the first trip on a given day is the identified home location. (2) The job location is identified based on trips made during a given day. (3) Both the home and job locations are identified on the same day (the stops in the same cluster are considered identical in this process). For each cardholder obeying the Conditions (1)–(3) in each day, the commuting trip can be identified in the form of one or more "connected" trips in his/her TRIP records. "Connected" here is defined as, for two consecutive trips of a cardholder, the alighting time of the first trip and boarding time of the second trip are less than 30 minutes considering the local condition in Beijing, and the alighting stop and boarding stop are within 750 m considering walking distance and spatial distribution of bus stops in Beijing. In some cases, a cardholder has the same identified home and job locations in several days in a week, and commuting time from home to job location may vary across days. Then we use the average commuting time as the final commuting time for that cardholder. In addition, identified job–housing location dyads with extreme commuting times are dropped as these cases

[12] In this paper, both commuting time and commuting distance are calculated based on a one-way trip from the home location to the job location.

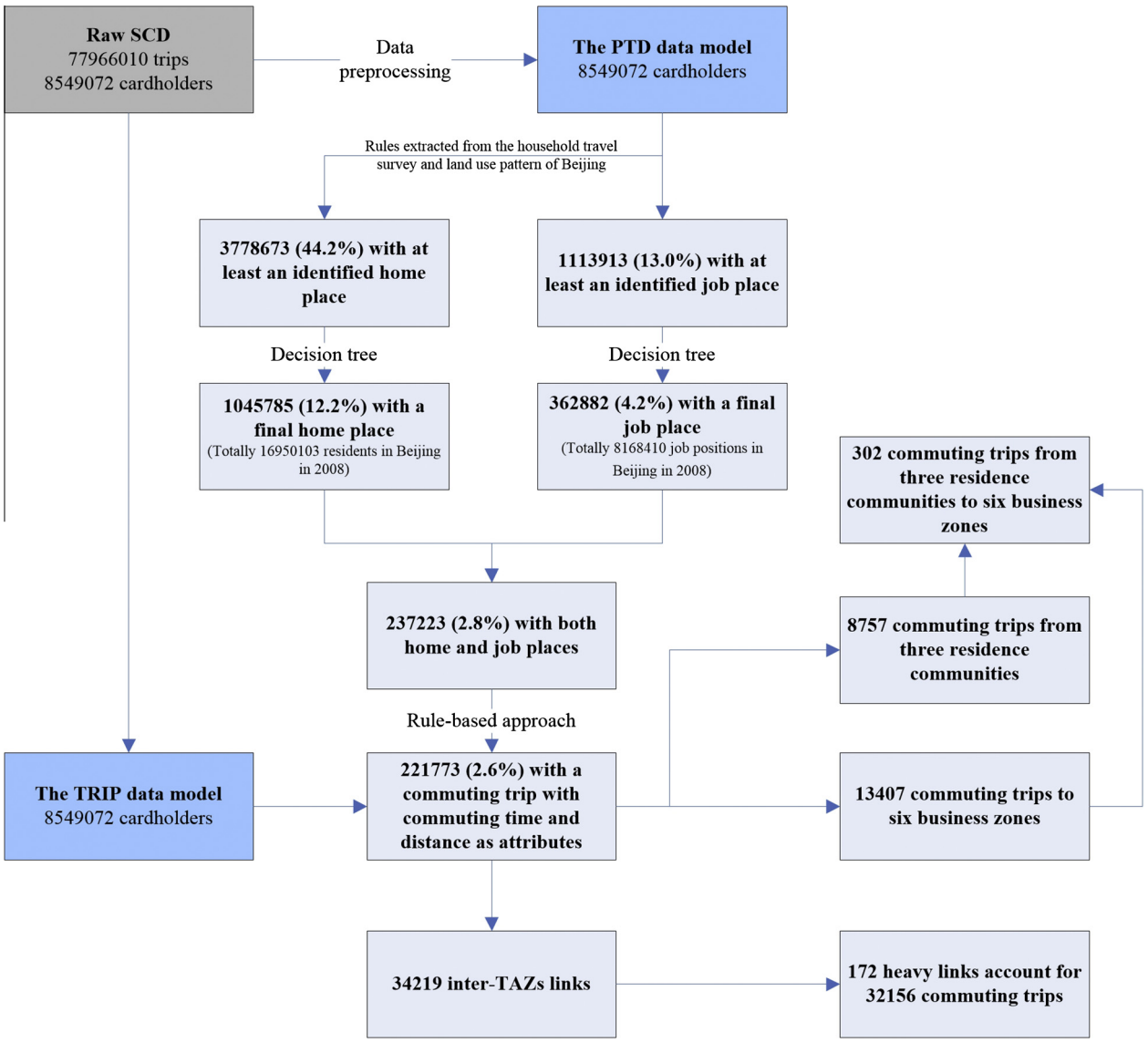

Fig. 7. Identification results for the job–housing location dyad recovery algorithm.

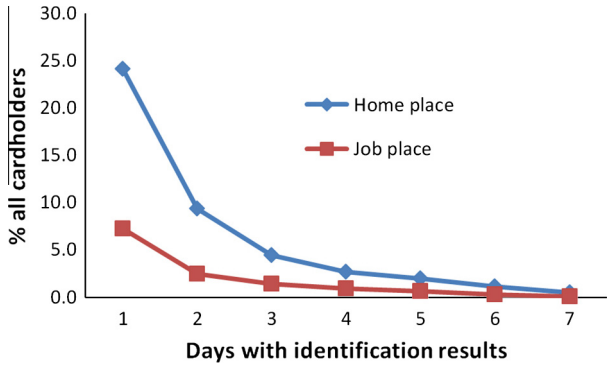

Fig. 8. Percentage of cardholders for which identification results exist by number of days.

may result from erroneous imputations. The benchmark for identifying extreme commuting time is set to 180 minutes, which covers 99% of all bus commuting trips in the 2005 survey.

5. Processing results

With SCD stored in the MS SQL Server, we developed a Python tool based on ESRI Geoprocessing to identify job and home locations and commuting trips of cardholders as well as to analyze and visualize commuting patterns using the pre-processed data. The results of data analysis are illustrated in Fig. 7, which gives a general summary of the information presented in this section.

5.1. Job–housing locations identification, aggregation of the bus stop and TAZ scales and comparison with observed data

The job–housing locations identification using one-week data is based on the identification results for each day. The distribution of successful imputations in Fig. 8 shows there are far few cardholders with more than one day with identification results, in contrast to those with only one-day results. As set by our rules in Section 4.3, it would be more robust to impute a final home or job location to cardholders for whom home or job locations are identified on two days or more.

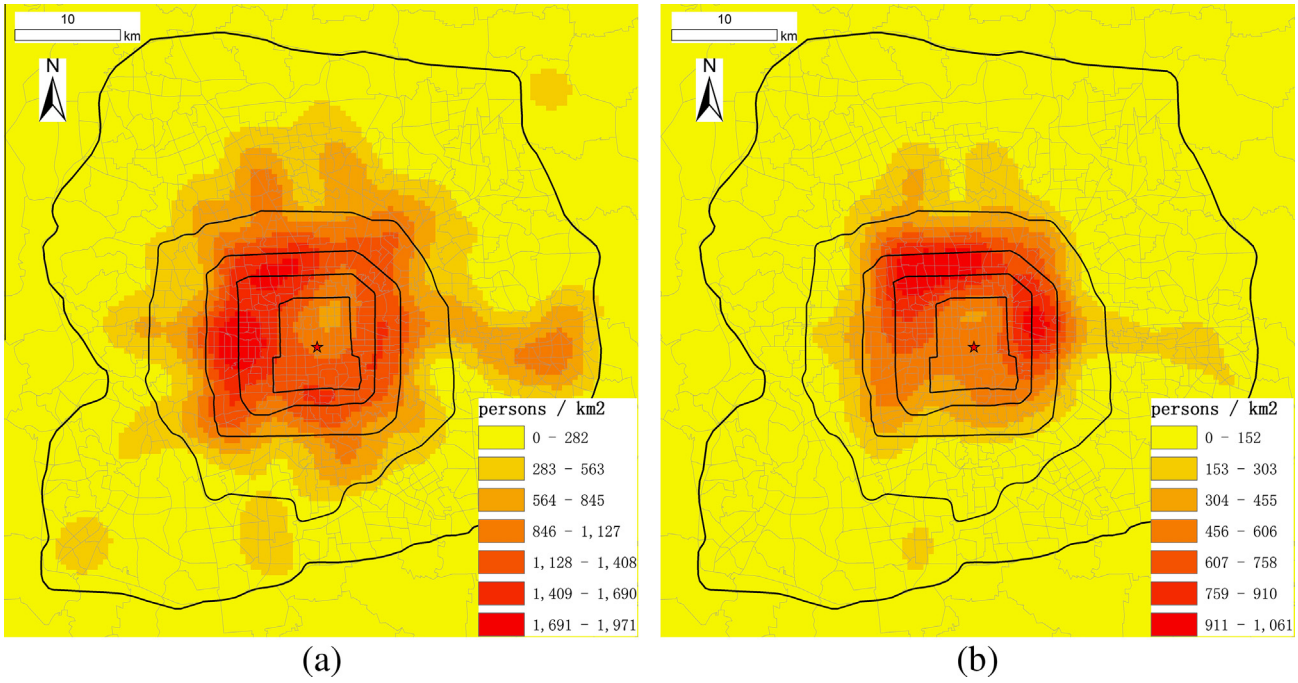

Fig. 9. Identified home (a) and job (b) kernel density maps in the central BMA. *Note:* This figure only represents data from bus smart cardholders.

Using one-week data, home locations are identified for 1,045,785 cardholders (12.2% of all 8,549,072 cardholders), and job locations are identified for 362,882 cardholders (4.2% of the total). Since the processes of identifying home and job locations are independent of each other, 237,223 cardholders (2.8% of the total) have both final home and final job locations.

The identification results are aggregated in the bus stop scale with a total of 3414 bus stops (39.3% of all 8691 bus stops) corresponding to home locations and 3329 (38.3% of the total) corresponding to job locations. The identification results are further aggregated at the TAZ scale, and 729 out of 1118 TAZs correspond to home locations. Home and job kernel density maps (Fig. 9) show that both the home and job densities in the inner area exceed those in the outer area from the perspective of bus landscapes. Both maps in Fig. 9 follow rather well the urban structure of population and employment distribution in Beijing. The imputed home density map shows a more dispersed pattern than the job density map, which is consistent with known patterns in Beijing.

The identification results are compared with the observed home and job data from 2008 at the TAZ scale. In 2008, the total number of residents in Beijing was 16,950,103, and the number of jobs was 8,168,410 (Beijing Municipal Statistics Bureau & and NBS Survey Office in Beijing, 2009). Among all TAZs, for home locations, the average identification ratio (cardholders with identified home locations in a TAZ divided by observed residents in the TAZ) was 6.2%. For job locations, the average identification ratio (cardholders with identified job locations in a TAZ divided by observed jobs in the TAZ) was 4.4%. Since we do not have access to job–housing data at the TAZ level in Beijing, we synthesize the 'observed' job–housing data for each TAZ using the resident and job numbers for each sub-district from the statistical data in the 2008 yearbook and the floor area of each parcel. Two indicators are used for comparison, the resident ratio and job ratio. The ratio maps are shown in Fig. 10a and b, whose legends are set according to the median values of resident ratio (6.5%) and job ratio (4.3%) among all 729 TAZs that have ratio values.

We find that both the resident ratio and job ratio vary significantly among TAZs (also supported by the correlation analyses reported in Fig. 10c and d), which may originate from the spatial heterogeneity of bus modal split. That is, some TAZs may have greater bus share than others because of accessibility and of the socioeconomic structure of the resident population.

5.2. Commuting trip identification and comparison with the 2005 survey and other existing researches

Using one-week data we identify the job–housing location dyads of 221,773 cardholders out of 237,223 cardholders for whom both home and job locations had been identified.[13] The identified commuting trips have an average duration of 36.0 min and standard deviation of 24.2 min. Meanwhile, the average commuting distance (Euclidean distance) is 8.2 km and the standard deviation is 7.0 km. When we use the Manhattan distance to measure commuting distance, the average commuting distance is 10.2 km and the standard deviation is 9.1 km. The average trip duration and distance of cardholders are also calculated for all TAZs by aggregating the home locations of identified commuting trips (Fig. 11a and b). For the 729 TAZs with identified commuting trips, the median commuting duration is 35.0 min, and the median commuting distance is 7.2 km. TAZs in the central area have lower commuting duration and distance than those elsewhere. The circular distribution of commuting distance reflects the mono-centric urban structure of Beijing. To measure the spatial autocorrelation of the calculated average trip time of each TAZ, we calculated the global Moran's I statistic. Moran's I is 0.024 with Z Score of 11.57, indicating less than a 1% likelihood that this clustered pattern could occur by chance. For commuting distance, Moran's I is 0.067 with Z score of 31.08. Therefore, TAZs are significantly clustered in terms of both average trip duration and length.

We compare our identification results with the 2005 survey to validate our approach (see Table 3). The 2005 survey contained

[13] Cardholders that undertake commuting trips slightly less than those with both home and job locations, which reflects the fact that home and job locations must be identified on the same day for us to identify a commuting trip for the cardholder, as elaborated in Section 4.4.

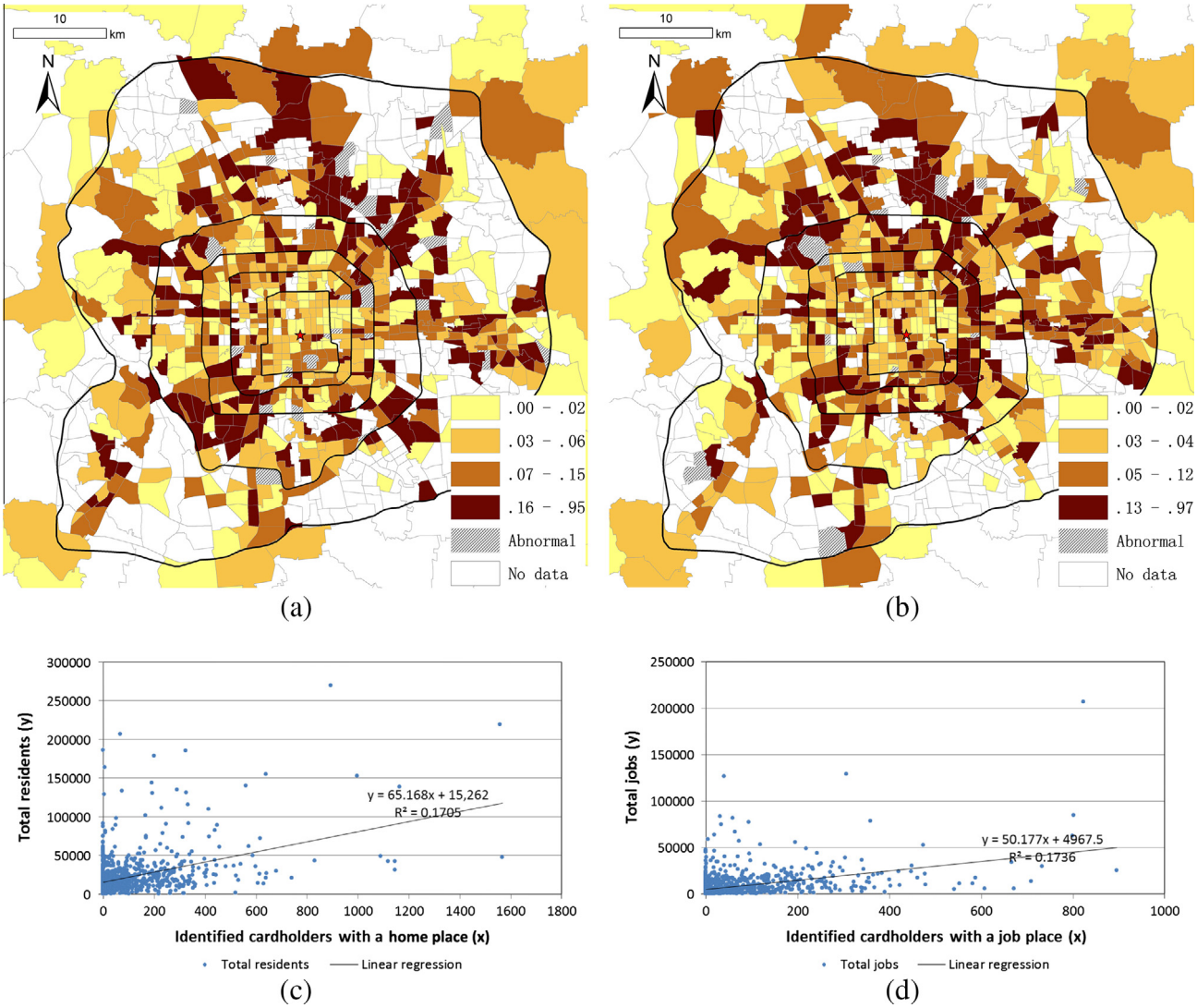

Fig. 10. Bus rider ratio in terms of home (a) and job (b) locations, correlation of identified residents and total population (c), and correlation of identified jobs and total jobs (d) at the TAZ scale. Note that "abnormal" means a ratio over 1.

6651 persons who commute by bus, or about 3% of the number identified on the basis of the SCD. The 2005 survey included three levels of validation, as follows:

(1) The average commuting duration was 40.5 min (Std.D = 23.1), and the average commuting distance (both going to work and returning home) was 8.4 km (Std.D = 8.3 km) in the 2005 survey. The *t*-test between the 2005 survey and our results reveals a significant difference ($t = 15.7$, $df = 7095$, two-tailed $p = 0.000 < 0.05$). This may result from the sample size and strategy of the survey, as well as from potential bias of our rule-based identification algorithm since a sizeable proportion of trip could not be identified.

(2) We further compare the cumulative distribution function (CDF) of commuting trips in the 2005 survey with our results (Fig. 12). The two CDFs of commuting duration generally overlap, although the CDF is not smooth for the survey because commuting duration was recorded by commuter's memory and discretized into various categories. For commuting distance the CDFs overlap almost perfectly.

(3) Since the commuting trip count in the 2005 survey is not large enough to conduct the comparison at the TAZ scale, the comparison is instead conducted in the district scale. The BMA includes 18 districts, as shown in Fig. 1d, including eight in the central area, five in the near suburbs and five in the remote suburbs. While some deviations exist in certain districts like Fangshan, Huairou, Pinggu, Chaoyang and Changping, the comparison in Table 4 shows that our results coincide well overall with those of the survey, especially in the central area where more trips are generated. This is supported by Wilcoxon signed-rank tests for both commuting duration ($p = 0.286$) and distance ($p = 0.267$) of 18 districts; with these results we cannot reject the Null Hypothesis that the two groups are not different. Deviation between our results and those of the survey (e.g. the commuting duration ratio of 1.54 for the Huairou district) may lie in the limited sample size in some districts in the 2005 survey. To summarize, the multiple validations tasks demonstrate the applicability of our proposed approach for identifying home and job locations as well as commuting trips.

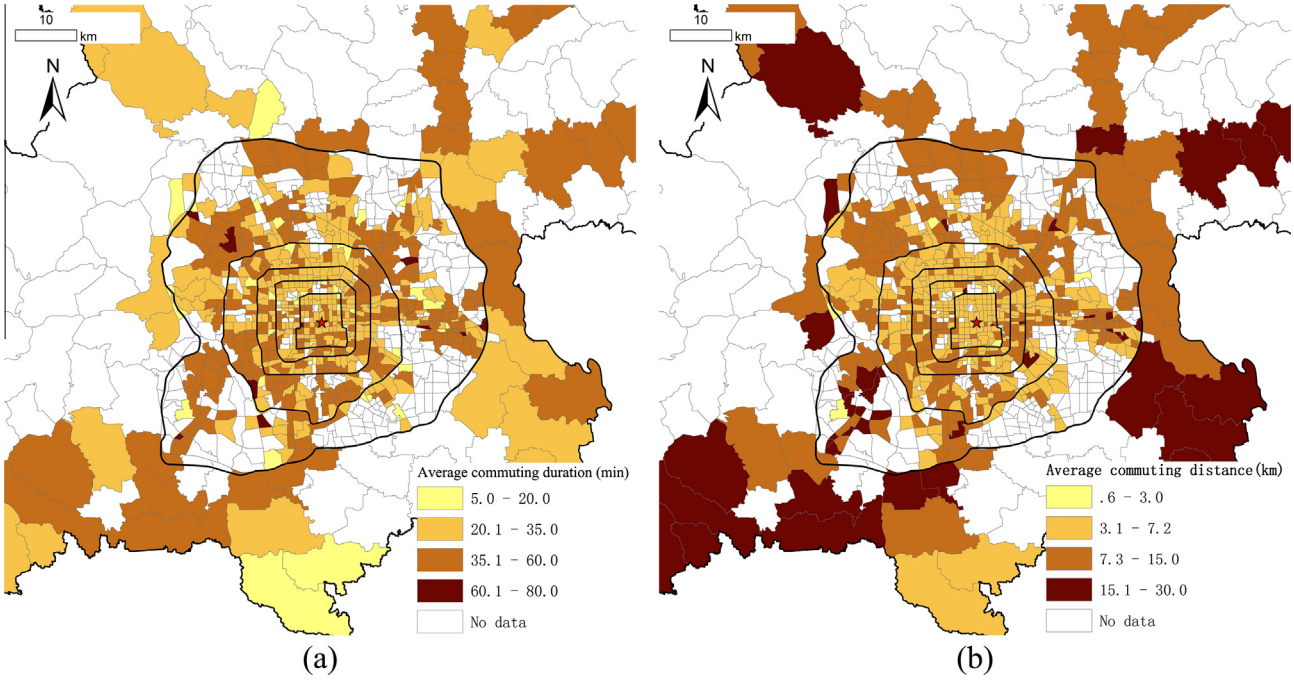

Fig. 11. Average commuting duration (a) and distance (b) by TAZ in Beijing.

Table 3
Commuting duration and distance data for Beijing based on existing studies.

Name of Study	Travel modes and year of research	Sample size	Average commuting duration (min)	Average commuting distance (km)
Present study	Bus, 2008	221,773	36.0 (24.2) 48.6 including walking	8.2 (7.0)
2005 Survey	Bus, 2005	6651	40.5 (23.1)	8.4 (8.3)
Liu & Wang, 2011	Bus, 2007	307	46.3 (N/A)	N/A
Wang & Chai, 2009	Bus, 2001	227	55.1 (30.4)	N/A
Zhao et al, 2011	Bus and metro, 2001	220	52.4 (26.6)	N/A

Note that the numbers in brackets are the standard deviation of the average commuting duration and distance. Bus samples in studies other than the present one are extracted from the survey of all travel modes.

Finally, it should be recognized that our identified results do not include connecting trip segments like walking. According to the 2005 survey, the average walking time to board a bus is 6.0 min (N = 11783 and Std.D = 4.7) and average walking time after alighting a bus to the destination is 6.6 min (N = 11781 and Std.D = 4.7). Considering this information, the average commuting duration estimated from the SCD should be adjusted to 48.6 min (6.0 + 36.0 + 6.6). The average adjusted bus commuting duration would then be 20% greater than the duration reported in the 2005 survey. Given the time lapse between the 2005 survey and the time stamp on the SCD used in this study and given that the average commuting duration by bus increased to 64.6 min (N = 12067 and Std.D = 45.0) in 2010 according to the 2010 survey (the counterpart of 2005 survey), the SCD estimates of bus commuting duration can be regarded to be in line with the longitudinal trend. This trend points to a significant increase during the 2005–2010 period. Therefore the bus/metro commuting condition has worsened in this period, which could be ascribed to the explosion of private car ownership in this timeframe and to the resulting traffic congestion. We also compare our results with other research, as shown in Table 3. Given the small sample sizes used in these studies and the time lag between studies, our estimates exhibit broad consistency with this existing body of evidence on commuting travel.

5.3. Visualization of commuting trips for the whole region and for selected zones

The mapping of identified commuting trips is an effective method of understanding commuting patterns in Beijing. Each commuting trip is visualized as a line that links the departure (home) and arrival (job) bus stops, and has associated commuting duration, commuting distance and card ID as GIS attributes. To identify the dominant commuting patterns in the BMA, commuting trips are further aggregated into the TAZ scale to produce trip counts between different pairs of TAZs (termed trip links in this paper). The inter-TAZ commuting pattern contains 34,219 links.

We use the head/tail division rule proposed by Jiang (2013) to classify the links into six levels based on their trip counts. In levels 4 to 6 with heavy commuting traffic, 175 links (0.5% of all links) account for 32,156 commuting trips (14.8% of all trips). Fig. 13 illustrates the dominant commuting patterns in the BMA and the heavy links can support route design of customized shuttle buses. These heavy links are mainly located within the 6th Ring Road and only a few links connect the city core with the suburban districts, which suggests the need for express public transportation services that would cross this ring road.

Beijing traffic congestion and its roots in overly large residential communities and overly agglomerated business zones are widely

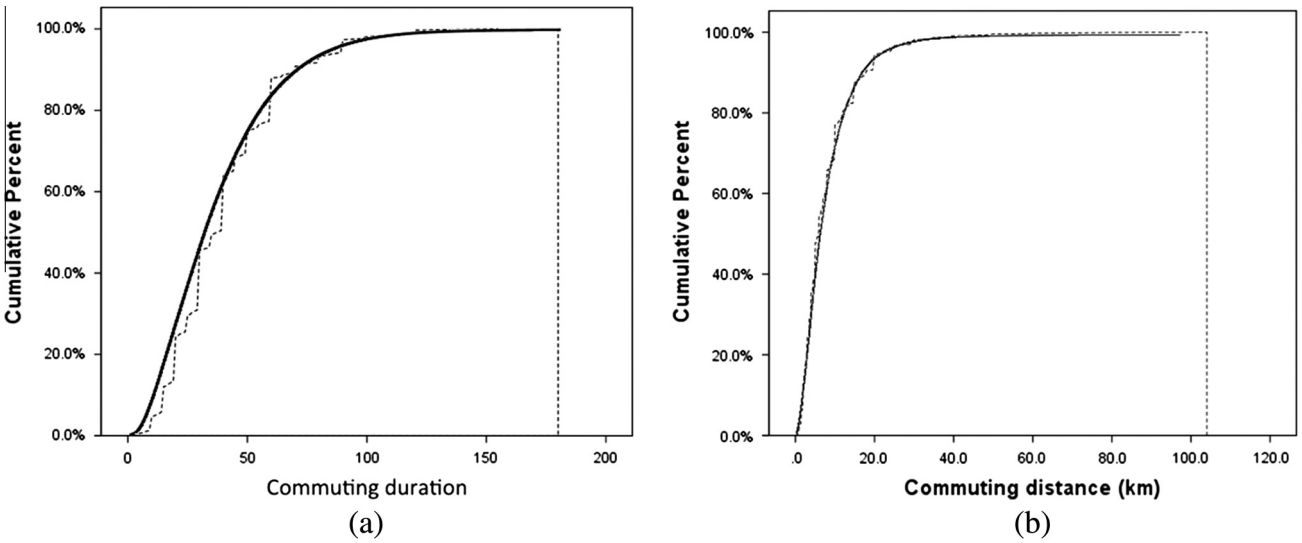

Fig. 12. Comparison of CDFs of commuting duration (a) and distance (b) between commuting trips in the 2005 survey (dashed lines) and results identified by our algorithm (solid lines).

Table 4
District-level comparison for commuting duration (t) and distance (d) of commuting trips between our identification results and the 2005 survey.

District	Our results			The 2005 survey			Our results/survey results	
	Count	t (min)	d (km)	Count	t (min)	d (km)	t ratio	d ratio
Central area								
Dongcheng	4179	35.1	6.5	317	37.7	5.8	0.93	1.12
Xicheng	9145	33.7	7.1	467	35.2	6.3	0.96	1.13
Chongwen	3762	39.8	7.6	276	37.6	5.8	1.06	1.31
Xuanwu	4377	36.6	8.2	432	40.3	6.9	0.91	1.19
Chaoyang	66,918	37.2	7.5	2031	42.7	8.7	0.87	0.87
Haidian	48,888	35.7	7.3	1277	39.8	8.0	0.90	0.92
Fengtai	32,170	38.6	9.0	678	46.6	9.9	0.83	0.91
Shijingshan	4561	34.3	7.6	313	30.3	6.2	1.13	1.21
Near suburbs								
Changping	13,035	36.5	8.8	202	47.4	11.1	0.77	0.79
Tongzhou	10,400	38.4	10.1	181	40.9	12.8	0.94	0.79
Daxing	9455	38.9	9.1	94	40.1	10.1	0.97	0.91
Fangshan	3057	47.4	15.7	157	31.7	11.5	1.49	1.37
Mentougou	1196	31.1	9.9	113	36.7	9.1	0.85	1.08
Remote suburbs								
Huairou	299	44.3	12.5	8	28.8	11.6	1.54	1.08
Miyun	149	43.7	13.1	7	34.6	16.1	1.26	0.82
Pinggu	730	43.8	15.7	8	42.5	23.8	1.03	0.66
Shunyi	5497	34.3	10.0	80	39.5	14.1	0.87	0.71
Yanqing	254	36.8	12.1	10	56.0	41.9	0.66	0.29

discussed in Chinese media. To shed more light on this situation, we now analyze commuting flows associated with specific areas of the BMA. We extract commuting trips originating from home locations in three major residential communities, Huilongguan, Tiantongyuan and Tongzhou (see Fig. 14a). The former two zones are the biggest communities in Northern Beijing and were built in the 1990s, while the Tongzhou area in Eastern Beijing contains several newly-built residential developments. Similarly, we extract commuting trips associated with job locations in six dominant business zones (Fig. 14b and c), the Central Business District, Shangdi (an IT park), Yizhuang (the biggest industrial zone), Tianzhu (the airport base), Shijingshan (a business zone in western Beijing) and Jinrongjie (a financial, banking and insurance district).

Commuting trips originating from each community or ending in each zone are further aggregated in terms of commuting duration and distance (Table 5). Residents of TTY tend to commute much shorter distances than those of TZH, and few residents of either group work on the south side of Beijing. Some residents of TZH work in new cities on the outskirts of Beijing, and a few work in western Beijing. Regarding commuting trips to business zones, the CBD attracts workers from locations that are more spread out geographically, thus resulting in the highest average commuting duration of all six business zones. Workers in BDA have the lowest commuting duration and distance, which may result from its status as a local job center. Surprisingly, only 302 commuting trips (0.14% of the total number identified) are from the three major residential communities to the six major business zones in Beijing (see also Fig. 13). In fact, most commuters between these zones travel by car or ride the subway rather than travel by bus, due to long distance between them and good access to subway stations in most of these zones. Interestingly, customized commuting shuttle buses (both morning and evening services) linking HLG and

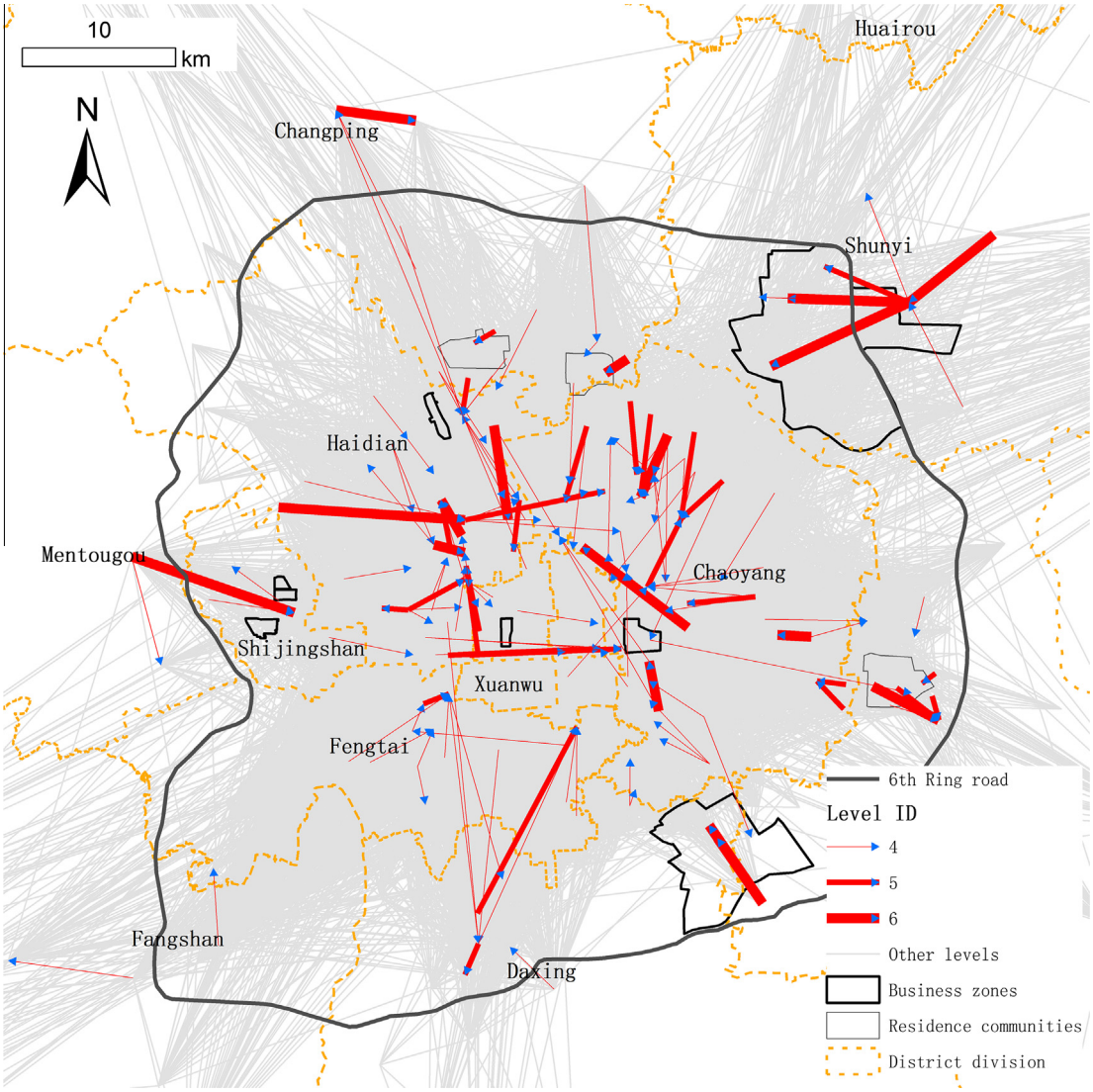

Fig. 13. Trip links at the TAZ scale illustrating dominant commuting patterns. *Note:* Arrows denote the commuting direction from home location to job location.

TTY with JRJ (http://news.sina.com.cn/c/2011-04-26/013522356172.shtml), and linking TZH with the CBD (http://news.dichan.sina.com.cn/bj/2011/06/14/333255.html) have since been inaugurated. These were the first customized shuttle bus services aimed at alleviating traffic congestion in Beijing. Government plans call to operate more shuttle bus routes linking the major residential communities and major business zones mentioned in this paper.

6. Discussion

6.1. Our contributions

This paper makes three main contributions. First, we investigate urban dynamics based on ubiquitous LBS data using rules generated from conventional travel behavior surveys and GIS layers, which is a promising approach for analyzing LBS data (Batty, 2012). Comparison with the 2005 survey has validated this approach on three levels, showing a significant and sound methodology. These points to the benefits of combining the spatiotemporal dimensions of rich LBS data with the social dimensions of conventional surveys to enhance our understanding of urban dynamics, particularly in fast changing environment. Additionally, the identification results obtained from the SCD can provide useful information during the multi-year gaps that separate surveys due to their demanding nature in terms of human and financial resources. Second, the processing of a whole week of SCD tracking individual cardholder bus trips is applied to analyze commuting patterns in Beijing, thus making our identification results more solid than those based on one-day data only. We also proposed a decision tree framework for identifying job–housing location dyads of cardholders over a weeklong period using the longitudinal information and spatial distribution of the one-day results. Third, we retrieved explicit spatial commuting patterns for Beijing based on more accurate information than conventional questionnaires or travel behavior surveys.

To our knowledge, our commuting pattern analysis of Beijing involves a larger sample size and more precise spatial and temporal information than any previous studies, although it is limited to bus riders. We hope our study provides more solid information on deriving policy implications for urban and transportation planning. To sum up, our test use of SCD is promising for analyzing urban dynamics, especially commuting patterns, and offers a new approach for the study and monitoring of commuting issues in a mega region in addition to conventional travel surveys. Our

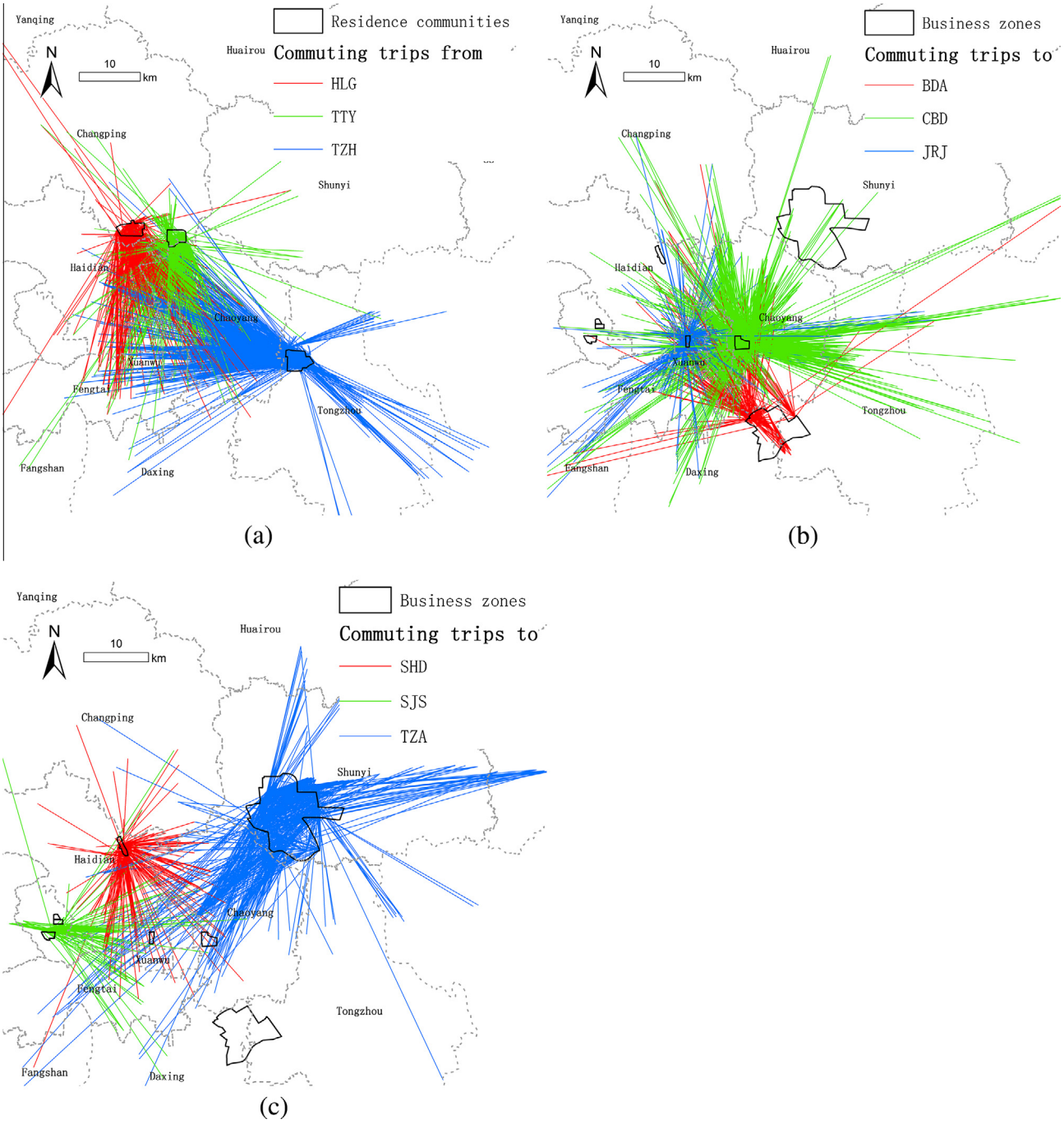

Fig. 14. Commuting trips (a) from three main residential communities; (b and c) to six major business zones. *Note:* HLG = Huilongguan community, TTY = Tiantongyuan community, TZH = Tongzhou community, CBD = Central Business District, SHD = Shangdi Industrial Park, BDA = Beijing Development Area at Yizhuang, TZA = Tianzhu Airport Park, JRJ = Jinrongjie (financial district), and SJS = Shijingshan Park.

contention is that SCD is a complement to travel behavior surveys generally. For instance, future surveys could focus more on socio-economic attributes of frequent bus travelers as well as their travel modes, and leave the specific travel pattern to SCD via documenting their smart card IDs. When the focus is only on travel pattern of bus passengers, SCD can advantageously substitute for surveys, considering that most bus passengers use smart cards in Beijing.

With the booming availability of information technology, large volumes of individual commuting data are increasingly ubiquitous, thus making individual travel diaries available for further data mining and decision support. This study represents a typical application of this sort of data. The SCD we obtained for a one-week period in Beijing totals 21 GB. Data pre-processing and data form building in the SQL Server took 8 h, while the identification of job and home locations and commuting trips (job–housing location dyads) using a custom Python tool was faster (72 and 113 minutes, respectively). This research was conducted using a workstation with a CPU of 3.0 GHz $*$ 2 and memory of 4 GB.

6.2. Limitations of the smart card data

While we appreciate the rich information provided by the smart card data used in this paper, using SCD to investigate urban systems suffers several limitations. First, SCD is limited to bus riders and trips by other modes are excluded. Future studies should attempt to combine SCD with data sources on other modes for a

Table 5
Commuting duration and distance for various residential communities and business zones in Beijing.

Zone name	Commuting duration (min)	Commuting distance (km)	% of all identified commuting trips
Trips from residential communities			3.9
TZH	45.1	10.0	1.4
HLG	39.4	7.0	1.0
TTY	36.2	6.1	1.5
Trips to business zones			6.0
CBD	41.4	9.4	2.7
SHD	40.4	6.7	0.3
JRJ	34.9	7.1	0.5
TZA	31.6	10.0	1.3
SJS	28.4	6.9	0.3
BDA	26.6	6.4	0.8

more comprehensive depiction of mobility in the city environment. Given the data infrastructure that is in place in Beijing, we could combine bus/metro trips recorded in more recent SCD with taxi trajectories released by Microsoft Research Asia (http://research.microsoft.com/en-us/projects/urbancomputing/), and scale other modes of trips in surveys to the observed totals released by official departments. Second, the validation of SCD-based estimations could be strengthened by analyzing socioeconomic attributes of bus travelers in the travel surveys and surveying local bus passengers with smart cards as well. Third, the spatiotemporal information of SCD generated by fixed-fare bus routes is incomplete. Most short fixed-fare routes are distributed in the central city area, while longer distance-fare routes are distributed across the central and outskirt areas. Thus part of the spatial and temporal information of bus rides in the central city area is lost, which means identification results, and hence policy implications, are more accurate and complete in the outskirt areas. We hope Yuan, Wang, Zhang, Xie, and Sun (2014)'s method could contribute to reconstruct partial fixed-fare records. Third, bus trips paid for by cash and cases of card sharing are not counted, although in Beijing they comprise only a small ratio of total trips. Fourth, the anonymity of the smart cards in this study prevents the inclusion of any socio-demographic information, thus making it hard to conduct behavioral study at the cardholder level. The above limitations of SCD may be addressed in future studies involving more comprehensive SCD from cards that store more information.

6.3. Future work

In the near term, our work plan will seek to extend this work in several directions. First, smart cards are also widely used in the metro system in Beijing, and the data format is the same as the SCD used in this paper. The subway's share of total journeys relative to all modes of household transportation in Beijing has recently increased, from 8.0% in 2008 to 10.0% in 2009, stimulated by the rapid construction of subway lines. SCD for bus and subway lines will allow us to get more complete passenger travel data and identify more realistic and comprehensive commuting patterns. Second, PTD can be used to study other urban activities (e.g. shopping, hospital and recreation) besides the home and work activities considered in this paper. The 2005 survey and trip purpose information can be leveraged to retrieve rules for identifying various urban activities through an approach similar to that used in this paper.

7. Concluding remarks

This paper demonstrated the effectiveness of using SCD for urban job–housing relationships analysis, including evaluating spatiotemporal dynamics of bus commuting system, identifying job–housing locations and commuting trips, and analyzing commuting patterns in terms of duration and distance.

First, we proposed two data forms, the original TRIP and location-time-duration (PTD), for processing and mining the raw SCD. The PTD data form is for identifying home and job locations, and the TRIP data form is for identifying commuting trips.

Second, we proposed an algorithm for identifying home and job locations using one-day data based on rules extracted from the 2005 survey and land use patterns in Beijing. We then used a decision tree to combine the one-day results to retrieve one-week results, thus providing more accurate identification results. With this approach, home locations were identified for 1,045,785 cardholders, job locations for 362,882 cardholders, and both home and job locations for 237,223 cardholders.

Third, commuting trips were identified and further mapped based on identified home and job locations. In total there were 221,773 cardholders with identified commuting trips. Commuting duration and distance were aggregated at the TAZ scale to present the overall commuting patterns in Beijing. We analyzed commuting trips from a set of three residential communities and six business zones to illustrate the 'tidal traffic' phenomenon in Beijing, an analysis that represents the first explicitly spatial test of commuting patterns in Beijing. We also aggregated commuting trips on the TAZ scale and generated links between pairs of TAZs by recording the commuting trip count. Dominant links were identified to demonstrate the mainstream commuting patterns. Both forms of analysis can obtain information useful to urban and transportation planners, and decision makers.

Fourth, we validated our identified commuting trips on three levels (the average and standard deviation values, cumulative distribution function and spatial distribution in the district scale) by comparing them with those in the 2005 survey in terms of commuting duration and distance. The validation results prove the applicability of our approach.

The findings of this study demonstrate the feasibility of spatiotemporal analysis of urban structure using smart card data as an alternative to conventional travel behavior surveys. This paper also tests novel methods of identifying interesting information from massive geo-tag datasets using rules retrieved from conventional questionnaires or surveys (e.g. the Beijing travel behavior survey) and urban GIS datasets (e.g. land use pattern, bus stops, and TAZs). Future research can further develop and highlight such novel methods for using ubiquitous geo-tagged data.

Acknowledgments

We would like to acknowledge the financial support of the National Natural Science Foundation of China (No. 51408039). Dr. Zhenjiang Shen, Kanazawa University, is also recognized for his valuable comments on an early draft of this paper. Last but not least, we are grateful to three anonymous reviewers for their in-depth comments which improve the quality of the manuscript significantly.

References

Ahas, R., & Mark, U. (2005). Location based services: New challenges for planning and public administration? *Futures, 37*, 547–561.
Anas, A., Arnott, R., & Small, K. A. (1998). Urban spatial structure. *Journal of Economic Literature*, 1426–1464.
Batty, M. (1990). Invisible cities. *Environment and Planning B: Planning and Design, 17*, 127–130.
Batty, M. (2012). Editorial. *Environment and Planning B: Planning and Design, 39*, 191–193.
Beijing Institute of City Planning (2010). *Beijing Job Position Allocation for 2008*. Internal Working Report (in Chinese).
Beijing Municipal Commission of Transport, Beijing Transportation Research Center (2007). *The 3rd Transportation Comprehensive Survey Report of Beijing*. Internal Working Report (in Chinese).

Beijing Municipal Statistics Bureau, and NBS Survey Office in Beijing (2009). *Beijing Statistical Yearbook 2009*. Beijing: China Statistics Press.

Beijing Transportation Research Center (2009). *Beijing Transportation Annual Report 2009*. Unpublished Official Report (in Chinese). Retrieved from <http://210.75.218.99/InfoCenter/Func/OpenNews.asp?NewsID=INN20090706001> (accessed on 30.10.10).

Blythe, P. (2004). Improving public transport ticketing through smart cards. *Proceedings of the Institute of Civil Engineers, Municipal Engineers, 157*, 47–54.

Bohannon, R. W. (1997). Comfortable and maximum walking speed of adults aged 20–79 years: reference values and determinants. *Age and Ageing, 26*(1), 15–19.

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. *ICWSM, 2011*, 81–88.

Chi, G., Thill, J.C., Tong, D., Shi, L., & Liu, Y. (in Press). *Uncovering regional characteristics from mobile phone data: A network science approach*. Papers in Regional Science.

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1082–1090), ACM.

Gärling, T., Kwan, M. P., & Golledge, R. G. (1994). Computational-process modelling of household activity scheduling. *Transportation Research Part B: Methodological, 28*(5), 355–364.

Gong, H., Chen, C., Bialostozky, E., & Lawson, C. T. (2012a). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems, 36*, 131–139.

Gong, Y., Liu, Y., Lin, Y., Yang, J., Duan, Z., & Li, G. (2012b). Exploring spatiotemporal characteristics of intra-urban trips using metro smartcard records. *Proceedings of Geoinformatics, Hong Kong*.

Hagerstrand, T. (1970). What about people in regional science? *Papers and Proceedings of the Regional Science Association, 24*, 7–21.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., et al. . *Identifying important places in people's lives from cellular network data. Pervasive Computing*. Berlin Heidelberg: Springer (pp. 133–151). Berlin Heidelberg: Springer.

Jang, W. (2010). Travel time and transfer analysis using transit smart card data. *Transportation Research Record, 2144*, 142–149.

Ji, J., & Gao, X. (2010). Analysis of people's satisfaction with public transportation in Beijing. *Habitat International, 34*(4), 464–470.

Jiang, B. (2013). Head/tail breaks: A new classification scheme for data with a heavy-tailed distribution. *The Professional Geographer, 65*(3), 482–494.

Jiang, B., & Yao, X. (2006). Location-based services and GIS in perspective. *Computers, Environment and Urban Systems, 30*(6), 712–725.

Joh, C-H., & Hwang, C. (2010). A time-geographic analysis of trip trajectories and land use characteristics in Seoul metropolitan area by using multidimensional sequence alignment and spatial analysis. *AAG 2010 Annual Meeting, Washington, DC*.

Kwan, M. P. (2004). GIS methods in time-geographic research: Geocomputation and geovisualization of human activity patterns. *Geografiska Annaler: Series B, Human Geography, 86*(4), 267–280.

Liu, X. (2009). *Public transit smart cards will be extensively adopted in Beijing on April 1, 2006*. Retrieved from http://www.enet.com.cn/article/2005/1229/A20051229488080.shtml> (accessed on 10.03.14) (in Chinese).

Liu, L., Andris, C., & Ratti, C. (2010). Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems, 34*(6), 541–548.

Liu, Z., & Wang, M. (2011). Job accessibility and its impacts on commuting time of urban residents in Beijing: From a spatial mismatch perspective. *Acta Geographica Sinica, 66*(4), 457–467 (in Chinese and with the abstract in English).

Long, Y., & Shen, Z. (2013). Disaggregating heterogeneous agent attributes and location from aggregated data, small-scale surveys and empirical researches. *Computers, Environment and Urban Systems, 42*, 14–25.

Lu, Y., & Liu, Y. (2012). Pervasive location acquisition technologies: Opportunities and challenges for geospatial studies. *Computers, Environment and Urban Systems, 36*, 105–108.

Lu, X., Wetter, E., Bharti, N., Tatem, A. J., & Bengtsson, L. (2013). Approaching the limit of predictability in human mobility. *Scientific Reports, 3*, 2923.

Newhaus, F. (2009). Urban diary – A tracking project. *UCL Working Paper Series*, 151.

Pelletier, M.-P., Trepanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C, 19*, 557–568.

Ratti, C., Pulselli, R. M., Williams, S., & Frenchman, D. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design, 33*(5), 727–748.

Rekimoto, J., Miyaki, T., & Ishizawa, T. (2007). LifeTag: WiFi-based continuous location logging for life pattern analysis. In *Third International Symposium on Location- and Context-Awareness (LoCA 2007)* (pp. 35–49).

Roth, C., Kang, S. M., Batty, M., & Barthélemy, M (2011). Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE, 6*(1), e15923.

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. *ICWSM, 11*, 329–336.

Schlich, R., Schönfelder, S., Hanson, S., & Axhausen, K. W. (2004). Structures of leisure travel: Temporal and spatial variability. *Transport Reviews, 24*(2), 219–237.

Steenbruggen, J., Borzacchiello, M. T., Nijkamp, P., & Scholten, H. (2013). Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: A review of applications and opportunities. *GeoJournal, 78*, 223–243.

Sun, L., Axhausen, K. W., Lee, D. H., & Huang, X. (2013). Understanding metropolitan patterns of daily encounters. *Proceedings of the National Academy of Sciences, 110*(34), 13774–13779.

Torrens, P. M. (2008). Wi-Fi geographies. *Annals of the Association of American Geographers, 98*(1), 59–84.

Wan, N., & Lin, G. (2013). Life-space characterization from cellular phone collected GPS data. *Computers, Environment and Urban Systems, 39*, 63–70.

Wang, D., & Chai, Y. (2009). The jobs–housing relationship and commuting in Beijing, China: The legacy of Danwei. *Journal of Transport Geography, 17*, 30–38.

Yuan, N. J., Wang, Y., Zhang, F., Xie, X., & Sun, G. (2014). Reconstructing individual mobility from smart card transactions: A space alignment approach. In *Proceedings of IEEE 13th International Conference on Data Mining (ICDM)* (pp. 877–886).

Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior: A case study of Harbin, China. *Computers, Environment and Urban Systems, 36*, 118–130.

Yue, Y., Wang, H., Hu, B., Li, Q., Li, Y., & Yeh, A. G. O. (2012). Exploratory calibration of a spatial interaction model using taxi GPS trajectories. *Computers, Environment and Urban Systems, 36*, 140–153.

Zhao, P., Lv, B., & de Roo, G. (2011). Impact of the jobs–housing balance on urban commuting in Beijing in the transformation era. *Journal of Transport Geography, 19*, 59–69.

Zhou, T., Zhai, C., & Gao, Z. (2007). Approaching bus OD matrices based on data reduced from bus IC cards. *Urban Transport of China, 5*(3), 48–52 (in Chinese with the abstract in English).