Travel Behaviour and Society 10 (2018) 42–52

Contents lists available at ScienceDirect

Travel Behaviour and Society

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

Revealing group travel behavior patterns with public transit smart card data

Yongping Zhang[a,b,c], Karel Martens[c,d], Ying Long[e,*]

[a] Centre for Advanced Spatial Analysis, University College London, W1T 4TJ, 90 Tottenham Court Road, London, UK
[b] School of Planning and Geography, Cardiff University, CF10 3WA Cardiff, Wales, UK
[c] Nijmegen School of Management, Radboud University, Thomas van Aquinostraat 5, 6525 GD Nijmegen, The Netherlands
[d] Faculty of Architecture and Town Planning, Technion - Israel Institute of Technology, Amado Building, Technion City, Haifa 32000, Israel
[e] School of Architecture, Tsinghua University, Beijing 100084, PR China

ARTICLE INFO

Keywords:
Group travel behavior
Smart card data
Proxemics
Identification
Beijing

ABSTRACT

Most analyses of travel patterns are based on the assumption of isolated individuals and ignore interpersonal relationships between travelers. In this paper, we develop a straightforward method to identify group travel behavior (GTB), defined as two or more persons intentionally traveling together from a single origin to a single destination, with public transit smart card data based on proxemics theory. We apply our method to Beijing to reveal the patterns of GTB, using all records generated by the subway system during a one-week period in 2010. Our data and method do not allow a reliable estimate of GTB share in overall travel, but do enable a description of the characteristics and the spatiotemporal pattern of GTB. The results reveal that the group size and GTB frequency follow a long tail distribution: far more people travel in small groups than in large groups and far more group travelers can be observed carrying out only one group trip than travelers making multiple group trips. Group trips tend to occur in weekends, in afternoons, and during public holidays. Furthermore, stations and lines serving leisure destinations show the highest GTB scores. We conclude that the GTB pattern is distinctly different from the pattern of individual travel in terms of both time and space, and is essentially influenced by urban land uses surrounding subway stations.

1. Introduction

Travel behavior is a well-developed research area with an extensive body of literature describing, explaining and predicting travel behaviors in various contexts (Handy, 1996; Golob, 2003; Ewing and Cervero, 2010). While traveling with other persons has been studied in the past, the typical starting point of most travel behavior studies is that persons travel on their own. The consequence is that there is limited understanding of what we call group travel behavior (GTB), which we define as two or more persons intentionally traveling together between a single origin and a single destination.[1] The aim of this paper is to develop a method to identify GTB with public transit smart card data, and to present some first empirical results about the patterns of GTB on the subway system in Beijing, China.

The paper is organized as follows. Following this introduction, we first present a brief review of studies that have addressed GTB (Section 2). We then present our smart card data based method (Section 3). Section 4 presents the study area and data set. In Section 5, we present the results for Beijing using smart card data of the subway system during a one-week period in 2010. We end with a brief conclusion and discussion about the potential applications of the proposed method across a range of contexts.

2. Literature review

While travel behavior research typically focuses on individual travel behavior, a number of strands of literature can be distinguished that directly or indirectly explore group travel behavior.

Group walking behavior may be the most well understood type of GTB. In line with most travel behavior research, early studies into walking behavior have treated pedestrians as isolated individuals, each having a desired speed and direction of motion (Moussaïd et al., 2010). More recently, GTB among pedestrians has received substantial attention (Moussaïd et al., 2010; Polzer, 2011; Vizzari et al., 2013; Zanlungo et al., 2014; Bruneau et al., 2015). Among these studies, identification of pedestrian groups is usually done manually using data collected by video recordings (Moussaïd et al., 2010; Polzer, 2011), but other methods have also been adopted, like interviews (Reuter et al., 2014), 3D laser range sensors (Zanlungo et al., 2014), and accelerometer sensors (Katevas et al., 2015). Besides identification and spatial

* Corresponding author at: School of Architecture, Tsinghua University, Beijing 100084, PR China.
E-mail addresses: yongping.zhang.15@ucl.ac.uk (Y. Zhang), ylong@mail.tsinghua.edu.cn (Y. Long).

[1] Note that this GTB definition does not include persons who travel together for part of their trip in this paper.

https://doi.org/10.1016/j.tbs.2017.10.001
Received 10 November 2016; Received in revised form 14 September 2017; Accepted 18 October 2017
Available online 11 November 2017

formation analysis of pedestrian groups, some research focused on group-considered crowd simulation using approaches like social force modeling (Moussaïd et al., 2010; Xu and Duh, 2010), cellular automata (Sarmady et al., 2009) and agent-based modeling (Manenti et al., 2012; Vizzari et al., 2013, 2015). While these studies help us understand pedestrian behavior from a group perspective, group walking behavior has so far only been analyzed at a micro scale or in a relatively small area, like a commercial street, a shopping mall, or a metro station. To the best of our knowledge, this type of analysis has not been conducted at a macro or a city scale. In addition, most methods for data collection are relatively labor intensive, which implies that only a limited pedestrian data set can be analyzed. As a result, these approaches have not been able to provide an understanding of the characteristics of group walking behavior versus individual walking behavior at a larger spatial scale, such as a neighborhood, a city center, or an entire town or city.

Analysis of household travel behavior, somewhat related to GTB, emphasizes the household as the basic analysis unit, rather than the individual as is common in travel behavior research. Drawing on notions derived from time geography, various approaches have been developed to analyze an individual's travel behavior while accounting for the interaction and interdependency between household members. This focus on the household is typical for activity-based travel models, which have developed since at least the early 1990s (Axhausen and Gärling, 1992; Ettema and Timmermans, 1997; Timmermans and Zhang, 2009). For instance, one of the main functions of UrbanSim (Waddell, 2002) is to simulate household mobility. Buliung and Kanaroglou (2006) proposed a system designed to support exploration of household level activity and travel behavior. Chatman (2008) investigated the relationship between development density and household travel behavior. While these and other studies do address the interrelationship between individuals' travel behavior, and activity-based models could theoretically also account for GTB, studies along these lines hardly ever aim to reveal GTB as part of overall travel patterns. Exceptions include studies such as conducted by Kang and Scott (2008), who identified joint episodes in persons' activity and travel diaries using restrictive and flexible criteria, respectively. Restrictive criteria require that joint episodes have the same start/end time and same activity type/travel mode, while flexible criteria for joint travel allow for a 10-min difference in the start/end time. This study does provide some understanding of GTB pattern, but is limited in terms of the population covered and the relative coarse way for identifying joint travel patterns.

Carpooling is a specific form of GTB in which persons who either differ in terms of their origin or destination travel together in a car for at least part of the trip. Carpooling has been well studied, covering issues like the rise and fall of carpooling in the US (Ferguson, 1997), the emergence of the carpooling club model (Correia and Viegas, 2011), and carpooling patterns in different countries (Wang, 2011; Ciari and Zurich, 2012). Paraphrasing group walking behavior, carpooling could be seen as group driving behavior, and thus as a distinct form of GTB. Yet, most studies into carpooling have sought to explain the decision to carpool or not, and seldom to compare the carpooling pattern with the spatial and temporal pattern of drivers traveling alone in their vehicles. It is precisely this comparison which we take up in this paper.

3. Methodology

3.1. Theoretical background

Proxemics is the study of human use of space and the effects that population density has on behavior, communication, and social interaction (Hall, 1959, 1966). Hall (1966) identified four interpersonal distances (or zones) within public space: intimate, personal, social and public distances. Generally, intimate distance (0–0.46 m) is reserved for close interpersonal interactions, and kept by two or more people having a strong bond, like family members and close friends; personal distance (0.46–1.22 m) is kept by casual friends or people with close social contacts, like friendly acquaintances and co-workers; social distance (1.22–3.66 m) is maintained by people who are somewhat acquainted but do not really know each another and who come together for a common purpose, like friends of friends and casual acquaintances; and public distance (3.66–7.62 m) is used by people whose only association is being in the same place at the same time (Thompson, 2013). In public situations, individuals usually prefer to keep close to familiar persons. If strangers come too close, uncomfortable feelings, like stress, can be caused. As a result, individuals might engage in compensatory behavior, such as avoiding eye contacting or moving away. Proxemics suggests that persons traveling in groups will tend to maintain a small distance between each other during large parts of a trip.

Referring to the theory of proxemics, we define group distance as the distance that is typical for communication between persons with emotional ties, i.e., between members of 'group'. Group distance thus encompasses intimate and personal distances. Similar definitions of group distance can be found in the literature. For example, Manenti et al. (2012) use the term proxemic distance to refer to the preferred distance pedestrians maintain with other group members. When interpersonal distance of group members exceeds their group distance, they will move closer to each other, making sure their maximum distance is below the group distance again. Thus, it is possible to distinguish between groups and non-groups based on particular values of interpersonal distances. In what follows, we will build on this understanding to identify persons engaging in GTB from among all users of Beijing's metro system.

3.2. A smart card data based method

Against the theoretical background presented in the previous section, in what follows we propose a straightforward method to identify GTB by utilizing public transit smart card data.

Smart card data, generated by automatic fare collection systems, provide detailed onboard and outboard transactions of each cardholder and thus give a (near) complete listing of all public transit trips in an area. Clearly, the availability of smart card data provides enormous opportunities for public transport research (see Pelletier et al., 2011 for a broad review). Much of the existing literature has sought to propose various methods to investigate travel behavior using smart card data (Morency et al., 2007; Chu and Chapleau, 2010; Ma et al., 2013; Zhou et al., 2014; Kusakabe and Asakura, 2014; Langlois et al., 2016; Tao et al., 2016; Kerkman et al., 2015). However, most of these smart card data-related travel behavior analyses do not make an explicit distinction between individual travel behavior and GTB. One exception is the study by Sun et al. (2013), who identified familiar strangers, understood as individuals who are recognized because of regular encounters in the (semi-) public sphere (i.e., public transport vehicles), but with whom one does not interact. To some extent, we can say ties exist among familiar strangers, but they are not what we have defined as group travelers.

Generally speaking, smart card data contain the basic attributes of public transit trips. Depending on the exact smart card system that is used in a particular country or city, this may include data on entrance and exit time, entrance and exit stations or stops, the ID of train, subway or bus line, card ID, etc. Furthermore, both the proxemics theory briefly discusses above and previous psychological studies (Cheyne and Efran, 1972; Polzer, 2011) suggest that group travelers have a preference to tap their cards shortly after one another, while strangers usually try to avoid tapping cards between members of a group. Based on this, we develop our smart card data-based identification method for the case in which travelers tap their smart cards when entering and exiting the transit system and each transit line has separate entrance and exit points (at least in terms of smart card technology).

The basic idea of our identification method is as follows. We consider the time between two persons tapping their smart cards to enter or

exit a transit system as an indicator of interpersonal distance. We term this time interval 'interpersonal time distance'. A short interpersonal time distance indicates that two persons are more likely to have a close bond and are thus traveling together. A large time difference suggests that persons are not related and are thus traveling alone and are engaging in what we call individual travel behavior. In other words, if two persons tap their cards shortly after each other to enter the transit system at the same entrance point and if the same two persons tap their cards shortly after each other to exit the transit system at the same exit point, they are very likely to be group travelers and in our method we identify them as such. We use the term 'group time distance' to refer to the interpersonal time distance that is typically maintained between members of a group. Clearly, we cannot know based on smart card data whether persons are actually traveling together or merely tap their smart cards by coincidence shortly after each other, within the predefined group time distance. However, a relatively small interpersonal time difference reflects that persons are more likely to be related. We will return to this issue in our application to Beijing.

It is more complicated to identify travel groups with three or more persons, for theoretically we have to check whether any combination of two among them are group members, and then find out whether all of them are in the same group. This would, however, result in a very inefficient identification process. Taking the characteristics of tapping smart cards into account, we have developed a simplified process, by taking three persons as an example: if A and B are group travelers, and B and C are group travelers, then A, B and C are group travelers. Although it may fail to identify GTB when the spatial order of group travelers entering the transit system is different from that exiting the transit system, it can ensure the efficiency of our identification method and the identified group travelers are really traveling in a group of three or more persons if they meet the condition.

More specifically, the smart card data-based method consists of two steps allowing a detailed identification process:

1. Identify group co-traveler(s) for each traveler *i*.
   - 1.1. Sort travelers according to their entrance and exit time to improve the calculation efficiency, then execute the following sub-steps starting from the first traveler *i*;
   - 1.2. Check whether traveler *j*'s entrance point is the same as *i*'s;
   - 1.3. If yes, check whether interpersonal time distance between *i* and *j* at the entrance point is within the pre-defined group time distance;
   - 1.4. If yes, check whether *j*'s exit point is the same as *i*'s;
   - 1.5. If yes, check whether interpersonal time distance between *i* and *j* at the exit point is within the group time distance;
   - 1.6. If yes, traveler *j* is *i*'s group co-traveler.
2. Check whether traveler *i*, for whom group co-traveler(s) have been identified in *step 1*, has another group co-traveler(s) who has (have) not been identified in *step 1*.
   - 2.1. Sort group travelers according to their entrance and exit time, then start the following sub-steps from the first group traveler *i*;
   - 2.2. Check whether traveler *j* is *i*'s group co-traveler;
   - 2.3. If yes, check whether traveler *j* has another group co-traveler *k*;
   - 2.4. If yes, check whether *k* has already been identified as *i*'s group co-traveler(s) in step 1;
   - 2.5. If no, add traveler *k* to *i*'s group of co-travelers;
   - 2.6. If traveler *k* is added as *i*'s group co-traveler, proceed to check whether *k* has another group co-traveler *m*, who has not been identified as *i*'s group co-traveler before. For this purpose, follow the process described in sub-steps 2.2–2.5.
   - 2.7. Repeat step 2.6 until no new co-traveler can be found for traveler *i*.

In Fig. 1, there are four travelers A, B, C, and D. They travel from the same departure station to the same arrival station. The predefined group time distance is *dist*. According to our smart card data identification method, we identify that A, B, and C are group travelers, while D is an individual traveler.

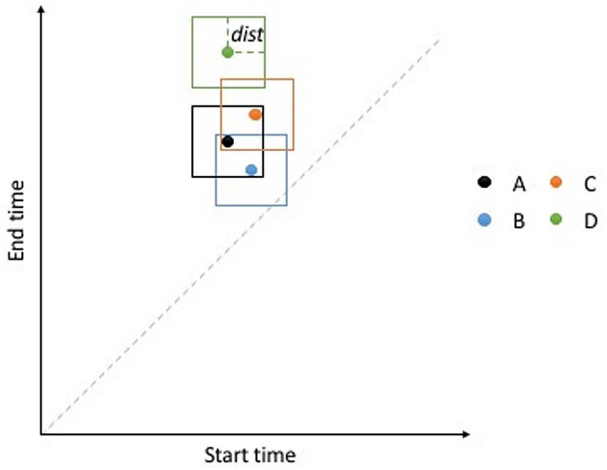

Fig. 1. An example of the smart card data identification method.

Based on these rules, it is possible to identify groups of two or more persons from an entire data set and thus to reveal the pattern of GTB and compare it with the (well-known) patterns of individual travel behavior. It should be noted that the proposed method can be seen as a general formulation of the GTB identification method using smart card data. Based on this general description, it is possible to specify more specific versions, depending on data characteristics and research requirements. For example, we can consider several trips together and only select those travelers who check in and out at least *two* times in a week within the predefined group time distance as engaging in GTB, or apply different time distances at the entrance and exit points to identify GTB (as is necessary for the Beijing case, as we will discuss below). In addition, it should also be noted that Step 2 is a preliminary algorithm to estimate the size of each travel group. Some clustering algorithms may be helpful to identify groups consisting of three or more persons.

4. Study area and data

4.1. Study area

The Beijing metropolitan region (Fig. 2A) covers an area of 16,410 km^2. It has experienced rapid growth in terms of population and GDP since the Reform and Opening Policy of 1978, established by the Chinese central government. Beijing had over 20 million residents in 2010 and is becoming one of most populous cities in the world. In 2010, the shares of bus, subway, car, and other modes in Beijing were 29%, 10%, 34%, and 27%, respectively (Beijing Transportation Research Centre, 2011). The subway system in that year consisted of 9 lines. The smart card data set we obtained contains 147 stations (Fig. 2B). It should be mentioned that stations at the intersection of two or more subway lines are designated as multiple stations. For example, there are two stations, named Xidan (14) and Xidan (69), at the Xidan intersection of Line 1 and 4. Of these 147 stations, 6 stations had only 0 or 1 trip record. All of these stations are intersection stations. They were already constructed but were not opened to the public in the week for which we have obtained smart card data. The station with one trip record may be ascribed to a trip made by a staff member. Considering this, these stations were excluded from the analysis, reducing the total number of stations included in the study to 141.

4.2. Data

In this paper, we identify GTB using smart card data of trips on the Beijing subway. The subway smartcard records cover a one-week period

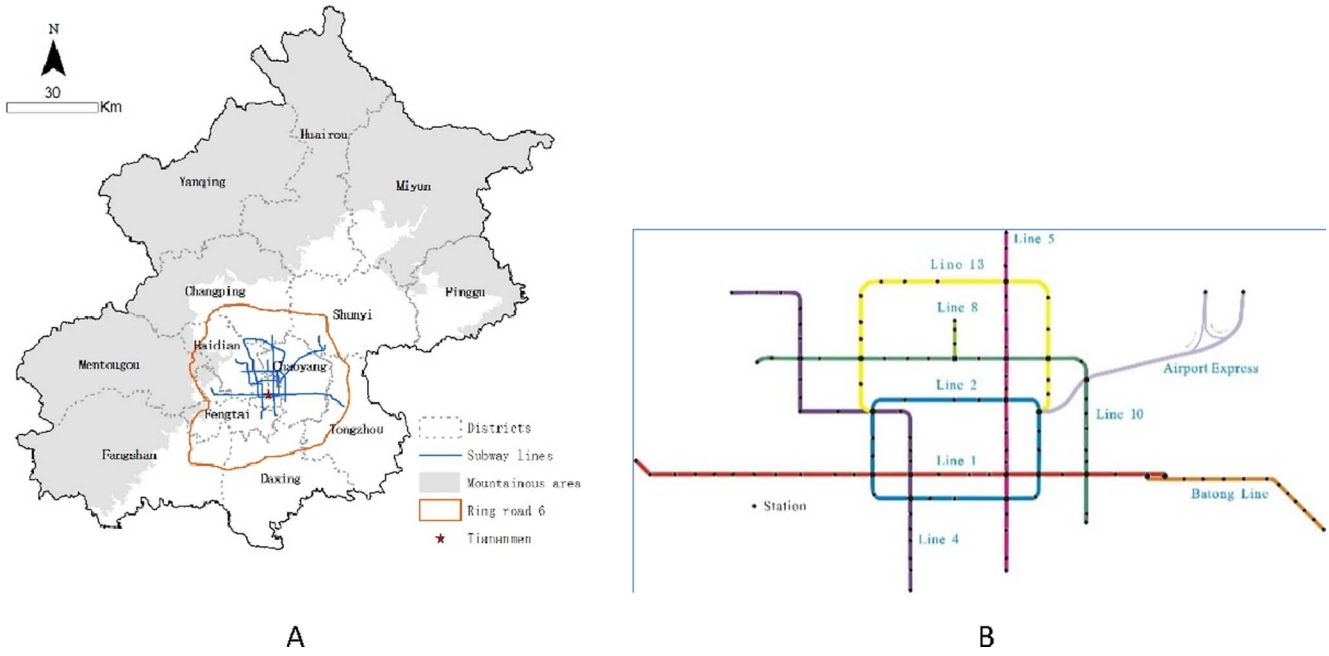

Fig. 2. Beijing metropolitan area (A) and subway map (B), in 2010. Note: the subway pattern in A was mapped according to the real situation; the subway pattern in B gives an abstract representation of the network and was mapped according to the map produced by Beijing Subway Operation Company.

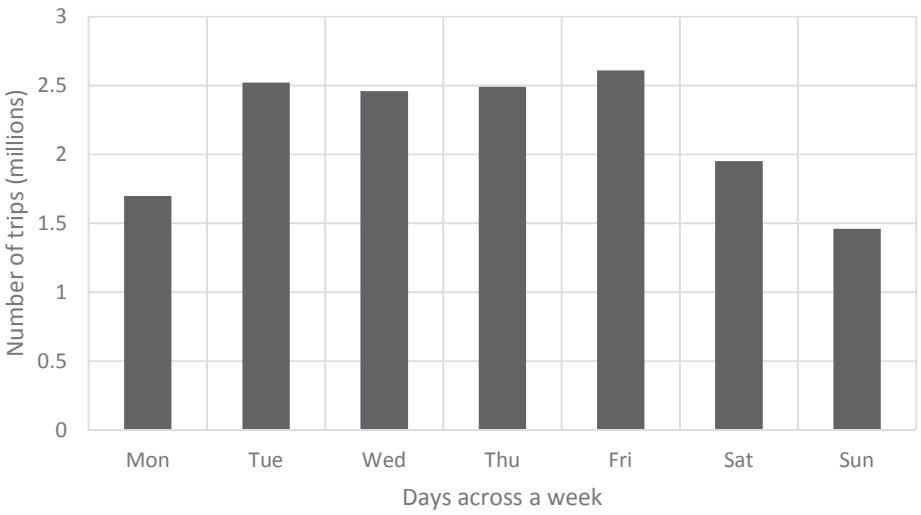

Fig. 3. Number of trips over the study period of one week. The X-axis represents the day of the week (from Monday 5 April 2010 to Sunday 11 April 2010).

in 2010 (5–11 April, Monday to Sunday) and comprise 15,204,632 trip records. This week comprises the last day of the Qingming Festival, which is a three-day public holiday in China (5 April), weekdays (6–9 April) and a weekend (10–11 April). Qingming is a time for people to go outside and enjoy the greenery of spring, but it is mostly known for its connection with Chinese ancestral veneration. Fig. 3 shows the number of trips over the study period of one week. Results show the trip number is lower on Qingming and weekends than on weekdays. For weekdays, the trip number is quite stable, with a relatively smaller amount on Wednesday and a relatively larger amount on Friday.

In Beijing, most persons use smartcards to pay their fares when traveling by subway. In 2010, one subway ride costs 2 Chinese Yuan (approximately 0.3 US dollar), regardless of the trip distance and time. The only exception is the price for the Airport Express line, which is 15 Yuan (about 2.2 US dollar). Note that there was no group ticket available for the public transit system of Beijing in 2010 (as is the case now). When cardholders use their smart cards to pay for public transit services, card readers installed in the station automatically record information. A sample of the available smart card data is shown in Table 1. All the attributes needed for GTB identification can be found in subway smart card data. However, the original entrance time (*Entrance time 1*) of each trip only has information on the hour and minute the person tapped the card, and not on the second at which he/she did so. A detailed discussion about its influence for the identification of GTB will

Table 1
A sample of smart card data.

Trip ID	Card ID	Entrance time 1	Entrance time 2	Exit time	Entrance station	Exit station
1	00001008	07:09:00	07:09:57	07:41:59	50	140
2	00001008	17:12:00	17:12:40	17:42:07	140	128
3	00001066	13:27:00	13:27:11	14:15:17	145	69
4	00001066	16:33:00	16:33:47	17:20:21	69	145
5	00001095	15:18:00	15:18:25	15:41:06	139	146

Note: The values of entrance and exit time were changed into seconds during analysis. For example, a time 18:23:12 (hour:minute:second) equals 66192 s (18*3600 + 23*60 + 12).

be provided in what follows.

To identify GTB among a large amount of travelers, it is necessary to adopt a uniform group time distance for all travelers at both entrance and exit points. If the group time distance is very short (e.g., 3 s), less group travelers can be identified, but we can be more certain that persons identified in this way are indeed group travelers. If the group time distance is very large (e.g., 3 min), more group travelers will be identified, but we can be nearly sure that not all persons so identified indeed engage in GTB. To make sure we primarily identify persons who engage in GTB, we prefer to adopt a relatively small group time distance (e.g., no more than 10 s).

However, given our data set, adopting a small time distance is not possible at the entrance point, as the smart card records does not contain a second value. To address this problem, the first approach (termed the 'M approach' for 'minute') applies different time distances at the entrance and exit points to identify GTB (e.g., 0 min for the entrance and 3 s for the exit point). This solution increases the risk of identifying travelers as engaging in GTB who do not do so, but it is still possible to identify a near-GTB pattern. The second approach (termed the 'R approach' for 'random') consists of adding a random second value (0–59) to the entrance time of each trip. Clearly, in comparison to the first strategy, this latter approach may fail to identify some GTB travelers by randomly assigning second values to two or more travelers that exceed the predefined group time distance. Likewise, it may also 'generate' GTB travelers by ascribing second values within the predefined time distance to persons who actually did not check in shortly after each other at the entrance point, although the chances that this will occur are smaller.

Given our goal to identify the *patterns* of GTB and the drawback of data, we will apply both approaches in this paper. We do so for two reasons. First, it will enable us to illustrate how the basic version of our identification method can be applied. Second, it allows us to compare the GTB patterns as identified across both approaches. If the results are comparable, we can be more certain that we have been able to identify GTB patterns. Note that neither approach allows us to obtain reliable estimates of the *volume* of GTB as a share of total travel.

In what follows, we thus use two identification approaches. We have added a random second value to all observed entrance times, using the random function of ArcGIS. An example of the resulting random values is shown as *Entrance time 2* in Table 1.

5. Beijing study

5.1. Sensitivity analysis

To find out the influence of the adoption of different group time distances, ten time intervals (1–10 s) are tested to identify GTB using the M and R approaches discussed in Section 4.2. For the M approach, we set the group time distance for entrance at 0 min, which means that only the travelers for whom the entrance time is the same can be identified as group travelers, and then investigate the influence of group time distance (1–10 s) at the exit point on the share of GTB amongst total travel. For the R approach, we adopt the same group time distance at the entrance and exit points and analyze the impact of the same range of group time distance (i.e., 1–10 s). To simplify the process, only the smart card data on Sunday 11 April 2010 are used for the sensitivity analysis. The trip number on this day is 1464,720, the lowest number among the one-week smart card data. Note that we identify group trips rather than group travelers, as every trip has a unique trip ID (e.g., a group of two persons traveling back and forth to the same destination and tapping in and out within the pre-set group time distance will be counted as four GTB trips). The identification process was done using Python.

As may be expected, the M approach generates a much higher share of GTB than the R approach (Fig. 4). For the M approach, the share of GTB trips increases from ~17% for 1 s time distance to ~29% for a 10 s time distance. For the R approach, the numbers are ~1% respectively ~11%. To the best of our knowledge, the existing literature does not provide a suitable reference point to assess the validity of either of these approaches. The high numbers for the M approach and the large differences with the R approach do seem to suggest, however, that the former method is likely to overestimate the share of GTB. It is more difficult to assess whether the R approach results in an over- or underestimation of GTB in our case study.

Based on local observation and expert advice from planning professionals in Beijing, we estimate that a group time distance between 2 and 5 s would be most suitable for identifying GTB in the Beijing case. In what follows, we have therefore adopted a time distance of 3 s for both the artificially generated entrance times and the empirically observed exit times. In other words, in the M approach, we use a group time distance of 0 min at the entrance and 3 s at the exit point. In the R approach, we use a group time distance of 3 s at both the entrance and exit stations. It should be mentioned that larger time intervals may be necessary in other cases, depending for instance on the smart card technology and the cultural norms regarding proximity between persons. At the same time, with every rise in time distance, the chances rapidly increase that persons will be identified as engaging in GTB who merely accidently tap their cards within the predefined time distance, certainly on public transport links that are traveled by many passengers.

5.2. GTB on the Beijing subway system

5.2.1. Overall results

Among all 15204,632 trips, 3164,931 (20.8%) and 431,548 (2.8%) are identified as group trips using the M and R approaches, respectively. Table 2 shows that most groups consist of only two travelers (79.9% and 95.2% using the M and R approaches, respectively). The group size follows a long tail distribution: far more people travel in small groups than in large groups. These meet existing psychological studies well that people like to perform GTB in small groups, and even large groups tend to split themselves into small ones (mainly dyads and triads) (Costa, 2010). The correlation coefficient between the results using the two approaches is 0.993 (significance level < 0.01), indicating two approaches capture a similar distribution.

Each group traveler may perform a different number of group trips during the week. Table 3 shows the percentages (probabilities) of group travelers by number of group trips taken (correlation coefficient is 0.985, significance level < 0.01). Most group travelers only took one group trip (77.1% and 89.9% using the M and R approaches, respectively), while for both approaches, only 0.002% of group travelers took more than six group trips during this week. The GTB frequency thus also follows a long tail distribution: most group travelers make a small number of group trips.

As noted before, neither the M nor the R approach allows us to obtain reliable estimates of the shares of GTB. To focus on revealing the general patterns of GTB, in the following investigation of spatio-temporal GTB patterns, we set the average GTB share across the week at 100 and use it as a bench score, and then transfer all GTB shares at different time or stations as a relative GTB score. For example, when using the M approach, we equate the average GTB share of 20.8% with 100; the GTB share on Monday, which is 30.4%, will subsequently be transferred to a relative GTB score of 146.2. For the R approach, we equate the average GTB share of 2.8% with 100; the GTB share on Monday of 4.2% is accordingly transferred to a relative GTB score of 148.7.

5.2.2. Temporal patterns on different days

Fig. 5 shows that the variation in GTB across the days of the week is very similar between the two approaches (correlation coefficient of 0.999, significance level < 0.01). The highest GTB score occurs on Monday (146.2 and 148.7 for the M and R approaches, respectively),

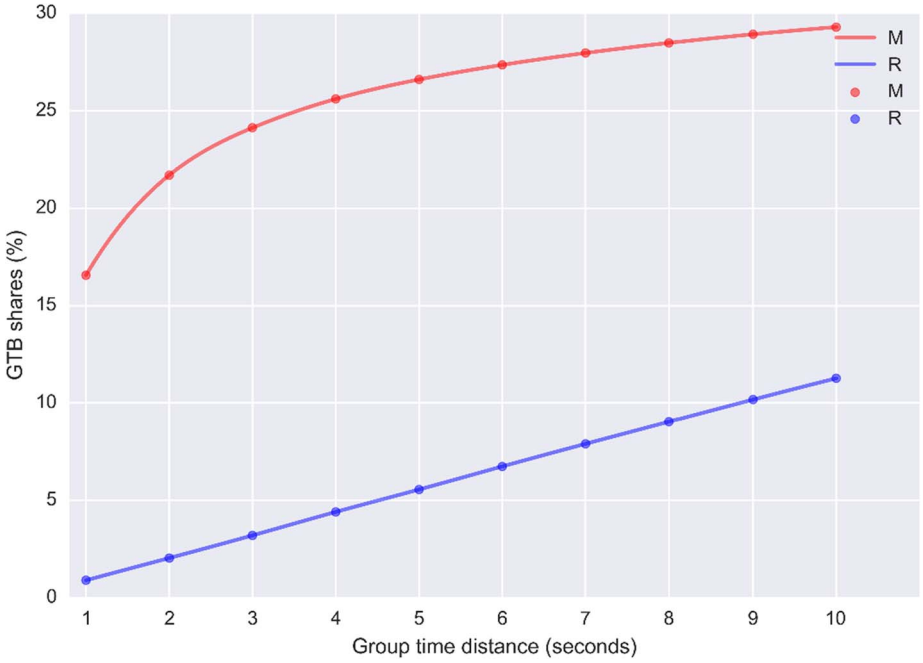

Fig. 4. Sensitivity analysis results for the M and R approaches.

Table 2
Percentages (probabilities) of group trips by group size.

Group size	2	3	4	5	6	7	8	> 8	In total
Percentages using the M approach	79.900	13.400	4.000	1.300	0.600	0.300	0.200	0.300	100.000
Percentages using the R approach	95.200	4.300	0.400	0.100	0.020	0.010	0.006	0.000	100.000

Table 3
Percentages (probabilities) of group travelers by number of group trips taken.

Number of group trips taken	1	2	3	4	5	6	> 6	In total
Percentages using the M approach	77.106	20.905	1.743	0.214	0.026	0.004	0.002	100.000
Percentages using the R approach	89.902	8.820	1.069	0.167	0.032	0.008	0.002	100.000

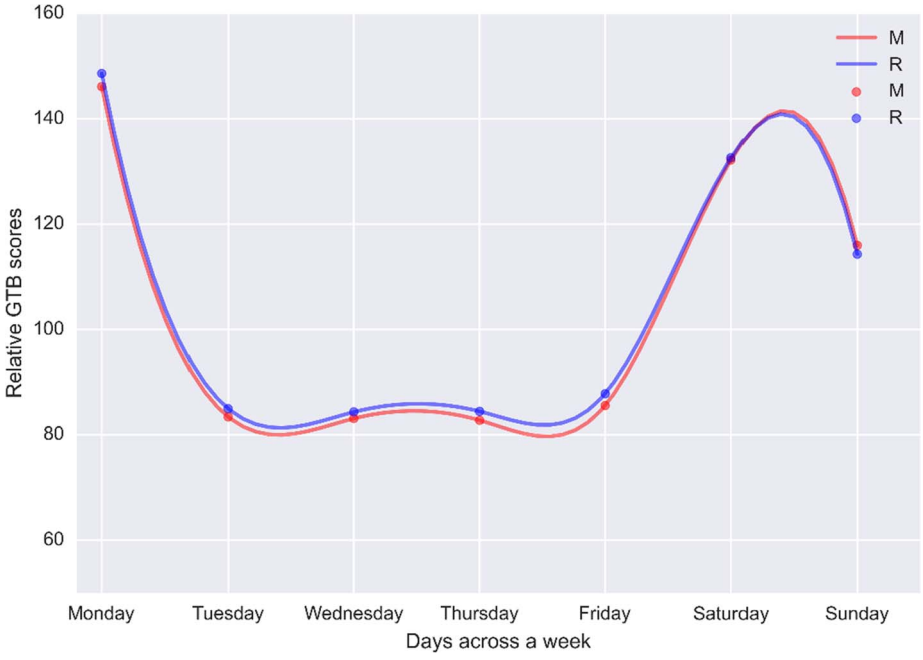

Fig. 5. Temporal patterns of GTB at the day level.

Table 4
Results of *t*-tests for the different patterns of GTB and ITB.

Pattern	M approach		R approach		*N* (the number of observations)
	t-value	Significance	*t*-value	Significance	
Across week-days	−13.37	< 0.001	−6.89	< 0.001	7
Across hours of the day	−24.41	< 0.001	−536.87	< 0.001	18[a]
Across stations	−65.45	< 0.001	−1367.90	< 0.001	141

[a] Because the numbers at hour 0:00–5:00 (sharp) and at hour 23:00–24:00 (sharp) are

the last day of the Qingming Festival, while the lowest score is observed on Thursday for the M approach (82.8) and on Wednesday for the R approach (84.4). During the Qingming Festival, many people, especially those working or studying in Beijing but originally from other cities, choose to make an excursion together with friends, while many local citizens choose to go to cemeteries with family members to memorize ancestors. This results in a relatively high GTB scores. During weekdays, when travel for leisure purposes is relatively limited, the GTB scores are much smaller than on Qingming. During weekends, more people engage in leisure-related activities together with friends or family members, resulting in a higher GTB score than on weekdays, although still below the level observed on Qingming. Interestingly, the GTB score is higher on Saturday than on Sunday, perhaps because on Saturday more people travel towards destinations well served by the metro system (e.g., shopping or other leisure centers), while on Sunday people engage more in family visits, with more GTB reverting to travel by car. Clearly, this observation warrants additional investigation.

We have conducted a *t*-test in order to determine whether the temporal pattern of GTB across weekdays is statistically different from the pattern of individual travel behavior (ITB). Here, ITB is defined as encompassing all trips performed by persons individually.[2] The closer the *t*-value is to 0, the more likely it is that there is no significant difference between GTB and ITB patterns. The significance level indicates the probability of the deviation of the *t*-value from zero, i.e. it indicates the probability that the patterns of GTB and ITB are indeed significantly different from each other. Results show that when using both the M and R approaches, the temporal patterns of GTB are significantly different from those of ITB (See the first row in Table 4).

5.2.3. Temporal patterns across a day

Fig. 6 shows the temporal GTB patterns using trip departure time over a 24-h period across the week. We only present scores for the averaged scores across all days for the period between 5:00 and 23:00, which covers ~ 100% of all trips conducted on each day. Results show that the patterns identified using the two approaches are again very similar, although the correlation coefficient of 0.850 (significance level < 0.01) is lower than that at the day level (0.999). When identifying GTB using the M approach, the GTB scores are above the average between 9:00 and 22:00 (i.e., above the GTB score average of 100). In addition, we observe much lower scores during the morning peaking hours (at hours 8:00 and 9:00), relatively lower scores during the afternoon peaking hours (about at hours 17:00, 18:00, and 19:00). When identifying GTB using the R approach, we have similar findings. Several reasons can explain this temporal pattern. First, persons may tend to engage more in group activities in the afternoon, when they have more discretionary time available. It may also be that persons traveling in groups tend to avoid peak hours. Finally, the lower score of GTB in especially the morning peak hours may be the result of the dominance of home-to-work travel during these hours, which is typically an individual activity (as implied by many studies, e.g., (Kang and Scott, 2008; Whyte, 1980)). The results of the *t*-test show that for both the M and R approaches, the temporal patterns of GTB across a day are significantly different from those of ITB (for the M approach, the *t*-value is −24.41, significance level < 0.001; for the R approach, the *t*-value is −536.87, significance level < 0.001, see the second row in Table 4).

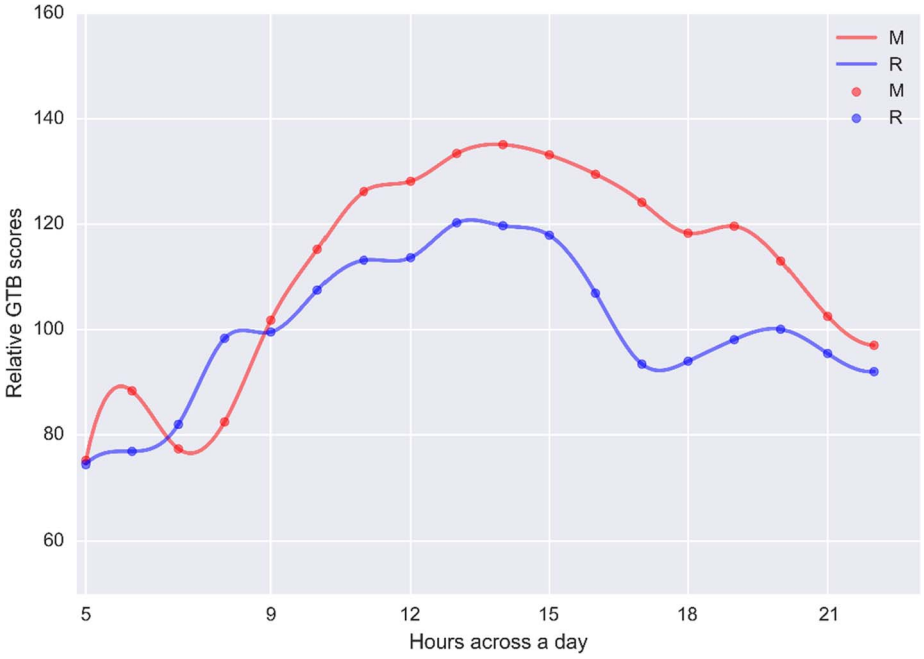

Fig. 6. Temporal patterns of GTB at the hour level.

[2] We assume in this paper that a trip is either conducted individually or with at least one other person and thus either belongs to ITB or GTB (i.e., we assume no trip is partially

(footnote continued)
GTB and partially ITB). This implies that the trip number and relative scores of ITB can be easily calculated based on the corresponding trip number and relative scores of GTB.

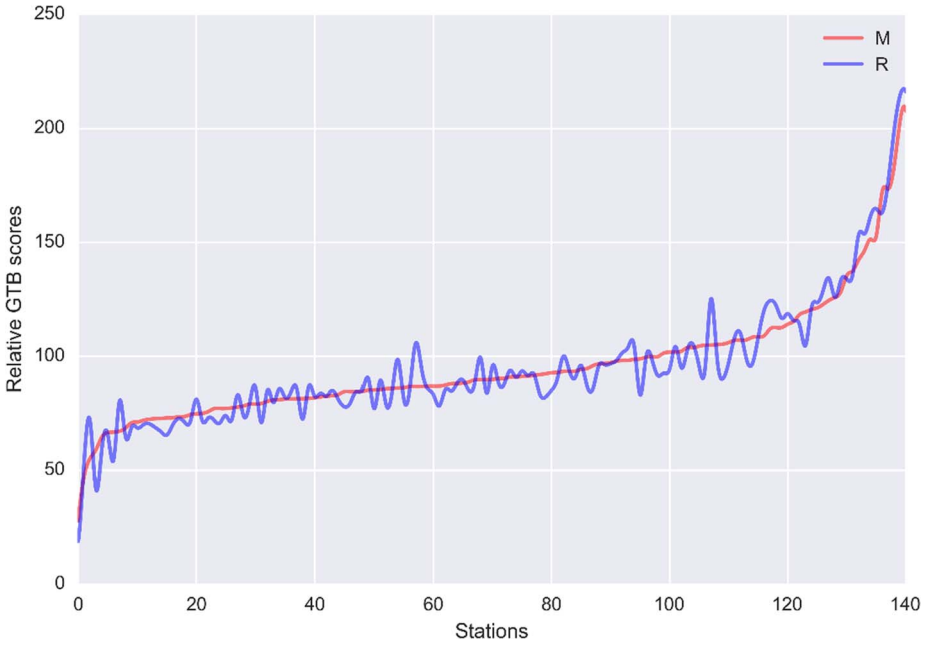

Fig. 7. Spatial patterns of GTB at the station level using the M and R approaches. The station is ordered according to their GTB scores identified using the M approach.

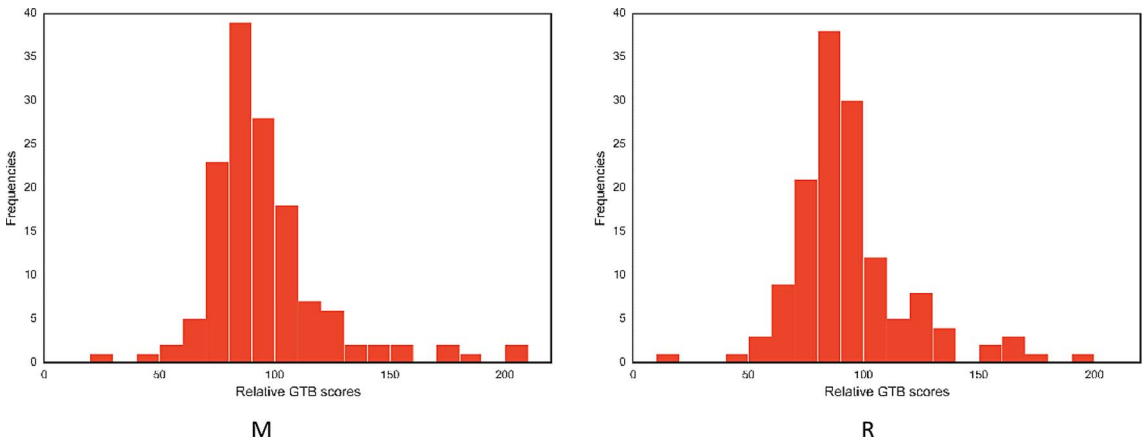

Fig. 8. Histograms of the GTB scores of stations using the M and R approaches.

5.2.4. Spatial patterns of GTB at the station level

Fig. 7 shows the spatial GTB patterns identified at the station level are very similar for the two approaches (correlation coefficient is 0.972, significance level < 0.01). When it comes to the histograms of the GTB scores of stations using the M and R approaches. They also show a similar distribution of frequencies (Fig. 8). The majority of the stations has a GTB score below the average (~70%), while a very small fraction has a GTB score 50% above the average (~5% of the stations). The results of the *t*-test show again significant differences for both the M and R approaches. In both cases, the GTB share at the station level is significantly different from the ITB share (for the M approach, the *t*-value is −65.45, significance level < 0.001; for the R approach, the *t*-value is −1367.90, significance level < 0.001, see the third row in Table 4).

In order to investigate the spatial GTB patterns more in depth, Fig. 9 shows the ten stations with highest and lowest GTB scores identified using the M and R approaches. Both approaches result in the identification of the same ten stations with highest GTB scores, although the ranking of the stations shows minor differences. The situation is different for the ten stations with lowest GTB scores. In this case, only half of the stations appear in the lists generated by each approach. Generally speaking, all ten stations with highest GTB scores are close to leisure-related or public facilities, like famous attractions (e.g., Beigongmen, adjacent to the royal garden; Beijing Zoo) or shopping centers (e.g., Wangfujing). Two areas stand out in particular for their high GTB scores: the Olympic Area, in which the high GTB score stations are associated with Line 8, and Old Beijing City, in which the high GTB score stations are associated with Line 1. The pattern of the stations with lowest GTB scores shows a different relationship with the land uses surrounding the stations. Generally, most of them do not serve leisure-related or public facilities. They may be a transport hub or close to transport hubs (e.g., Terminal 2 and 3), be close to residential communities (e.g., Hepingli Beijie), or be far away from the city center (e.g., Songjiazhuang). The spatial distribution of the stations with low GTB percentages is more dispersed, but two areas around Line 10 and the Airport Express line can still be identified as areas with particularly low scores of group travelers.

6. Conclusion and discussion

Group travel behavior (GTB) is defined as two or more persons intentionally traveling together from a single origin to a single destination. In this paper, we proposed a method to identify GTB using public transit smart card data based on proxemics theory. We applied our method to Beijing using all records generated by the subway system during a one-week period in 2010. Our data and method do not allow a reliable estimate of GTB share in overall travel, but do enable a

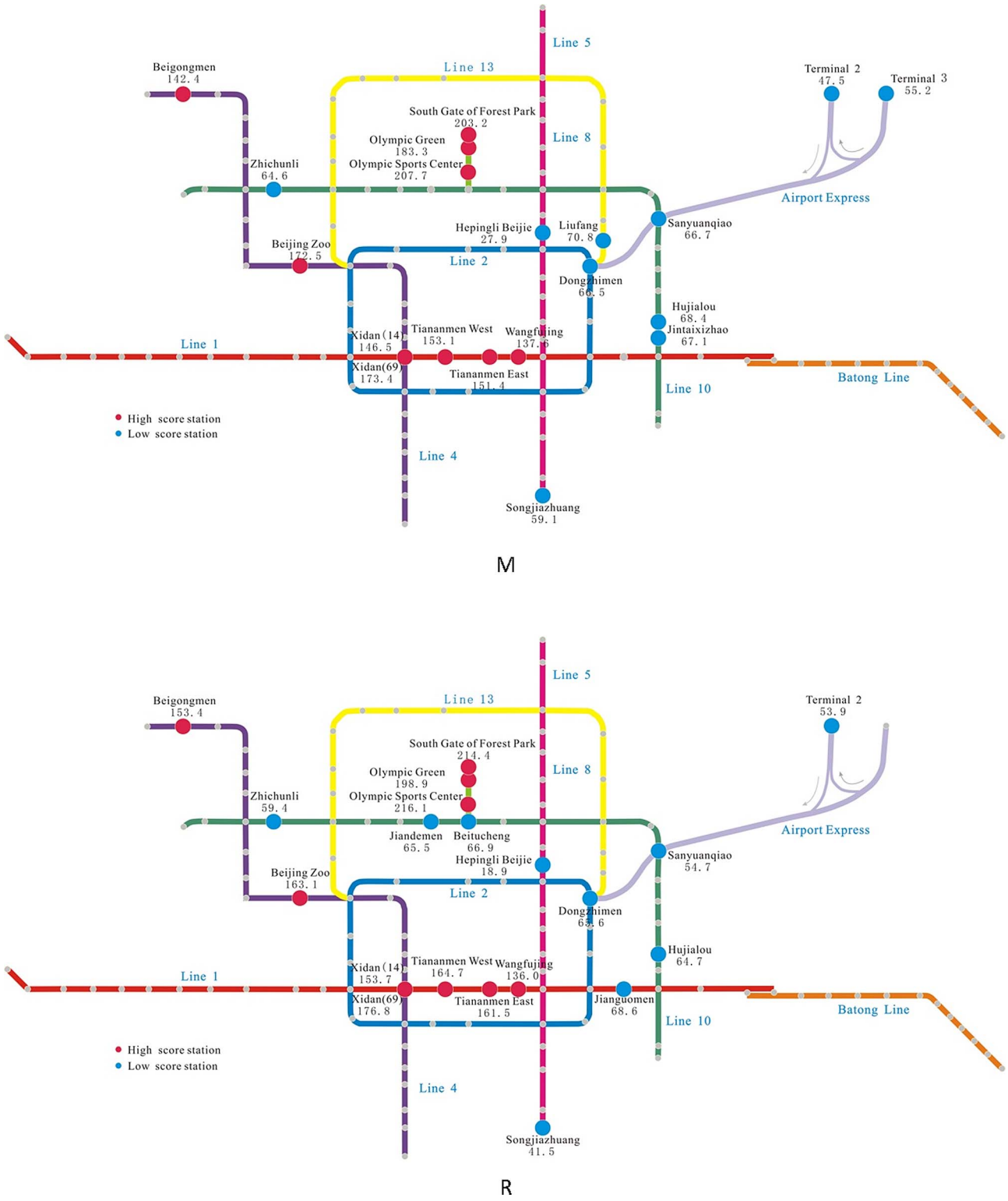

Fig. 9. Ten stations with highest and lowest GTB scores using the M and R approaches.

description of the characteristics and the spatiotemporal pattern of GTB. The results show that the group size and GTB frequency follow a long tail distribution: far more people travel in small groups than in large groups and far more group travelers can be observed carrying out only one group trip than travelers making multiple group trips in a one-week period. Group trips tend to occur in weekends, in afternoons, and during public holidays. Furthermore, stations and lines serving leisure destinations show the highest GTB scores. We conclude that the GTB pattern is distinctly different from the pattern of individual travel in terms of both time and place, and is essentially influenced by urban land uses surrounding subway stations. We have similar findings about the patterns of GTB when using the M and R approaches. This gives us the confidence that our method can successfully reveal the general spatiotemporal patterns of GTB, and that it can also be applied in other cities and countries, provided adequate smart card data are available.

At the same time, our study also has some limitations. First, neither

the M nor the R approach can fully solve the lack of detailed data on the entrance time of travelers. The former approach is likely to overestimate the GTB share, while underestimation is expected in the latter. Yet, our analyses do seem to suggest that both approaches are suitable to identify the GTB pattern. Even if more detailed data on entrance and exit time are available, the proposed method will always mistakenly identify some individuals as group travelers while failing to observe others as engaging in GTB, irrespective of the selected group time distance. For instance, passengers boarding and alighting at 'quiet' stations or bus stops may well do so within a short time limit, although they do not travel together. The other way around, group travelers may not always succeed to check-in or check-out within the predefined time distance if vehicles or stations are very busy or if they travel with luggage. In addition, since persons doing GTB do not always start or end at the same public transport stops, it would be impossible to capture them as group travelers using our smart card data-based method. Adopting a relatively small value of group time distance can ensure the accuracy of GTB identification, but cannot totally solve the problems mentioned here. Finally, it has been pointed out by others that passively generated smart card data usually lack ground truth to be validated against, yet can provide valuable information about a range of phenomena (Langlois et al., 2016). When it comes to our analysis, it is very difficult to validate the results of our method, as it requires the collection of a large sample of group travelers at the city scale with a fine temporal granularity, which is a near impossibility. In spite of these limitations, we are confident that the proposed method does shed a first light on the pattern of group travel behavior at the metropolitan scale.

Our result may also be relevant for policy purposes. A better understanding of GTB may especially inform public transport ticketing policy. Public transit group tickets, offered in parallel to regular tickets, are widely used around the world (e.g., London, Paris, Munich, Seoul, and New York). The aim of such ticketing policies is typically to promote public transport use among persons engaging in GTB. However, due to a lack of understanding of GTB patterns among transport professionals, group ticket policies in most cities are designed mainly based on an 'educated guess'. Our GTB analysis has the potential to provide information that can support the design of an effective public transit group ticket policy. Furthermore, the identification method may also enable the measurement of the success of a particular group ticketing policy. Drawing on our identification method, we can also imagine the introduction of an automatic group checking and pricing system. The basic logic of this system is in line with our identification method: passengers can enjoy a group ticket discount with (an)other passenger (s), if they tap their cards continuously and in the same order when entering and exiting the public transit system at the same entrance and exit points, respectively. The adoption of such a group ticket policy would imply that smart cards cannot only substitute the purchase of individual tickets, but can also provide a replacement for the (sometimes tedious) process of purchasing a group ticket.

The relevance of the identification method we propose in this paper reaches beyond the study of GTB on subway or bus systems. For example, the method can also be applied to study GTB at the inter-urban level, for instance GTB by train or high-speed rail, if adequate data are available. Furthermore, the method could be extended to study other types of group behavior, whenever the starting and/or ending status are recorded by smart card systems or other comparable 'big data' sources, like group eating behavior using student eating-card data, group shopping behavior using payment transaction records in shops, group chatting behavior using What's App data, and even group gaming behavior using the data of online computer games. Ideally, an identification method covering various types of group behavior may be proposed, although it may require an adaptation in the method presented here.

Acknowledgements

This research was funded by a scholarship from China Scholarship Council (CSC NO. 201508060122). Karel Martens acknowledges the support of The Leona Chanin Development Chair.

References

Axhausen, K.W., Gärling, T., 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transp. Rev. 12, 323–341.
Beijing Transportation Research Centre, 2011. Annual report of Beijing's transportation development (In Chinese). Beijing.
Bruneau, J., Olivier, A.-H., Pettre, J., 2015. Going through, going around: a study on individual avoidance of groups. Vis. Comput. Graph. IEEE Trans. On 21, 520–528.
Buliung, R.N., Kanaroglou, P.S., 2006. A GIS toolkit for exploring geographies of household activity/travel behavior. J. Transp. Geogr. 14, 35–51.
Chatman, D.G., 2008. Deconstructing development density: quality, quantity and price effects on household non-work travel. Transp. Res. Part Policy Pract. 42, 1008–1030.
Cheyne, J.A., Efran, M.G., 1972. The effect of spatial and interpersonal variables on the invasion of group controlled territories. Sociometry 477–489.
Chu, K., Chapleau, R., 2010. Augmenting transit trip characterization and travel behavior comprehension: Multiday location-stamped smart card transactions. Transp. Res. Rec. J Transp. Res. Board 29–40.
Ciari, F., Zurich, I., 2012. Why do people carpool: Results from a Swiss survey. In: 12th Swiss Transport Research Conference, Ascona.
Correia, G., Viegas, J.M., 2011. Carpooling and carpool clubs: clarifying concepts and assessing value enhancement possibilities through a Stated Preference web survey in Lisbon. Portugal. Transp. Res. Part Policy Pract. 45, 81–90.
Costa, M., 2010. Interpersonal Distances in Group Walking. J. Nonverbal Behav. 34, 15–26. http://dx.doi.org/10.1007/s10919-009-0077-y.
Ettema, D., Timmermans, H., 1997. Activity-based approaches to travel analysis:[selected papers presented at the workshop, May 1995].
Ewing, R., Cervero, R., 2010. Travel and the built environment: a meta-analysis. J. Am. Plann. Assoc. 76, 265–294.
Ferguson, E., 1997. The rise and fall of the American carpool: 1970–1990. Transportation 24, 349–376.
Golob, T.F., 2003. Structural equation modeling for travel behavior research. Transp. Res. Part B Method 37, 1–25.
Hall, E.T., 1966. The Hidden Dimension. Anchor Books, New York.
Hall, E.T., 1959. The Silent Language. Anchor Books, New York.
Handy, S., 1996. Methodologies for exploring the link between urban form and travel behavior. Transp. Res. Part Transp. Environ. 1, 151–165.
Kang, H., Scott, D.M., 2008. An integrated spatio-temporal GIS toolkit for exploring intrahousehold interactions. Transportation 35, 253–268.
Katevas, K., Haddadi, H., Tokarchuk, L., Clegg, R.G., 2015. Walking in Sync: two is company, three's a crowd. ACM Press 25–29.
Kerkman, K., Martens, K., Meurs, H., 2015. Factors influencing stop-level transit ridership in Arnhem-Nijmegen City Region, Netherlands. Transp. Res. Rec. J Transp. Res. Board 23–32.
Kusakabe, T., Asakura, Y., 2014. Behavioural data mining of transit smart card data: A data fusion approach. Transp. Res. Part C Emerg. Technol. 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. Part C Emerg. Technol. 64, 1–16.
Ma, X., Wu, Y.-J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transp. Res. Part C Emerg. Technol. 36, 1–12.
Manenti, L., Manzoni, S., Vizzari, G., Ohtsuka, K., Shimura, K., 2012. An agent-based proxemic model for pedestrian and group dynamics: motivations and first experiments. In: Multi-Agent-Based Simulation XII. Springer, pp. 74–89.
Morency, C., Trépanier, M., Agard, B., 2007. Measuring transit use variability with smart-card data. Transp. Policy 14, 193–203.
Moussaïd, M., Perozo, N., Garnier, S., Helbing, D., Theraulaz, G., 2010. The walking behaviour of pedestrian social groups and its impact on crowd dynamics.
Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. Transp. Res. Part C Emerg. Technol. 19, 557–568.
Polzer, U., 2011. Nonverbal behavior in public space as a function of density and group size, Universität Wien, Wien.
Reuter, V., Bergner, B.S., Köster, G., Seitz, M., Treml, F., Hartmann, D., 2014. On modeling groups in crowds: empirical evidence and simulation results including large groups. In: Pedestrian and Evacuation Dynamics 2012. Springer, pp. 835–845.
Sarmady, S., Haron, F., Talib, A.Z.H., 2009. Modeling groups of pedestrians in least effort crowd movements using cellular automata. In: Modelling & Simulation, 2009. AMS'09. Third Asia International Conference on. IEEE, pp. 520–525.
Sun, L., Axhausen, K.W., Lee, D.-H., Huang, X., 2013. Understanding metropolitan patterns of daily encounters. Proc. Natl. Acad. Sci. 110, 13774–13779.
Tao, S., Corcoran, J., Hickman, M., Stimson, R., 2016. The influence of weather on local geographical patterns of bus usage. J. Transp. Geogr. 54, 66–80. http://dx.doi.org/10.1016/j.jtrangeo.2016.05.009.
Thompson, S., 2013. The Applications of Proxemics and Territoriality in Designing Efficient Layouts for Interior Design Studios and a Prototype Design Studio. California State University, Northridge.
Timmermans, H.J., Zhang, J., 2009. Modeling household activity travel behavior: examples of state of the art modeling approaches and research agenda. Transp. Res. Part

B Methodol. 43, 187–190.
Vizzari, G., Manenti, L., Crociani, L., 2013. Adaptive pedestrian behaviour for the preservation of group cohesion. Complex Adapt. Syst. Model. 1, 1–29.
Vizzari, G., Manenti, L., Ohtsuka, K., Shimura, K., 2015. An agent-based pedestrian and group dynamics model applied to experimental and real-world scenarios. J. Intell. Transp. Syst. 19, 32–45.
Waddell, P., 2002. UrbanSim: modeling urban development for land use, transportation, and environmental planning. J. Am. Plann. Assoc. 68, 297–314.
Wang, R., 2011. Shaping carpool policies under rapid motorization: the case of Chinese cities. Transp. Policy 18, 631–635.
Whyte, W., 1980. The Social Life of Small Urban Spaces. The Conservation Foundation, Washington, DC.
Xu, S., Duh, H.B.-L., 2010. A simulation of bonding effects and their impacts on pedestrian dynamics. Intell. Transp. Syst. IEEE Trans. On 11, 153–161.
Zanlungo, F., Brščić, D., Kanda, T., 2014. Pedestrian group behaviour analysis under different density conditions. Transp. Res. Procedia 2, 149–158. http://dx.doi.org/10.1016/j.trpro.2014.09.020.
Zhou, J., Murphy, E., Long, Y., 2014. Commuting efficiency in the Beijing metropolitan area: an exploration combining smartcard and travel survey data. J. Transp. Geogr. 41, 175–183.