Transportation Research Part C 86 (2018) 1–22

Contents lists available at ScienceDirect

Transportation Research Part C

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

An empirical study on travel patterns of internet based ride-sharing[☆]

Yongqi Dong[a,b], Shuofeng Wang[a], Li Li[a,*], Zuo Zhang[a,b]

[a] Department of Automation, Tsinghua University, Beijing 100084, China
[b] Tsinghua University Institute for Data Science, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Taxis
Ride-sharing
Travel pattern

ABSTRACT

The rapid growth of internet based ride-sharing brings great changes to residents' travel and city traffic. However, few studies had employed empirical data to examine the unique travel patterns of internet based ride-sharing trips. In this paper, we compare taxi trip records and internet based ride-sharing trip records provided by DiDi company. Results reveal many interesting findings that had never been reported before. From the viewpoint of service patterns, ride-sharing mainly increases supplies in hot areas and peak hours. By applying a non-negative matrix factorization method, we find that ride-sharing principally serves as an approach for commuting. So, as an effective supplement to traditional taxi service, it regulates spatial and temporal supply-demand imbalance, especially during morning and evening rush periods. From the viewpoint of individual behavior patterns, we use a clustering method to identify two kinds of internet based ride-sharing drivers. The first kind of drivers usually provides ride-sharing along daily home-work commuting. Trips served by these drivers have relatively constant origin-designation (OD) pairs. The second kind of drivers does not serve regularly and roams around the city even in working hours. Therefore, there are no constant OD pairs in their ride-sharing trips. Counterintuitively, we find that home-work commuting drivers account for only a small part of total drivers and they only serve a small number of commuting trips. In addition, internet based ride-sharing is not just traditional hitchhiking worked through mobile internet. We find that internet based ride-sharing drivers intend to make long distance trips, and they intend to detour further to pick up or drop off passengers than traditional hitchhike drivers since they are paid. All these findings are helpful for policy makers at all levels to make informed decisions about deployment of internet based ride-sharing service. This paper also verifies that big data analytics is particularly useful and powerful in the analysis of ride-sharing and taxi service patterns.

1. Introduction

Taxis act as an important transportation service in our daily lives. However, the enormous imbalance between taxis demand and taxis supply remains bothering in many cities. Let us take Beijing for example. As the capital and the second largest city in the People’s Republic of China, Beijing is one of the most populated cities in the world. Statics show that the resident population of Beijing increased from 14 million to 21 million, while the corresponding supply number of taxicabs merely changed from 65 thousand to 66 thousand. The slow development of taxis supply is far behind the ever-growing increase of taxi demand.

[☆] This article belongs to the Virtual Special Issue on Emerging Urban Mobility Services: Characterization, Modeling and Application.

[*] Corresponding author at: Room 806, Central Main Building, Tsinghua University, Beijing 100084, China.
E-mail address: li-li@tsinghua.edu.cn (L. Li).

https://doi.org/10.1016/j.trc.2017.10.022
Received 31 July 2016; Received in revised form 26 October 2017; Accepted 27 October 2017
Available online 09 November 2017

On the other hand, with the rise of residents' income level, private car ownership increases dramatically. Currently, Beijing has more than 5 million private cars. To alleviate traffic congestion and pollution, authorities had restricted residents to buy new private cars. This in return makes more residents resort to taxi services and exaggerates the imbalance between taxi demand and supply.

As a result, ride-sharing, as one typical mode of shared mobility that enabled passengers to obtain short-term access to transportation as needed rather than requiring the ownership (Shaheen et al., 2016), is proposed as a remedy for such problems. The fast growth of mobile internet, location-based service (LBS) (Artigues et al., 2012), cloud computing and other innovative techniques makes the internet based ride-sharing popular in many cities, e.g., Beijing, since the last ten years.

Ride-sharing is a means of transportation where more than one person with common origin and/or destination travel in a car to share (at least a part of) their journeys. Studies have verified that ride-sharing is beneficial to drivers, passengers, and environment, because ride-sharing can reduce travel costs, total fuel consumptions, as well as carbon emissions, and help relieve traffic congestions (Ferguson, 1997; Kelley, 2007; Morency, 2007; Caulfield, 2009; Minett and Pierce, 2010; Chan and Shaheen, 2012). Amey et al. (2011) identified and discussed the potential benefits of and obstacles to internet based real-time ridesharing. He pointed out that challenge hindering greater rideshare participation was a series of economic, behavioral, institutional, and technological obstacles. Agatz et al. (2011) developed approaches to minimize the total system-wide vehicle miles incurred by system users, and their individual travel costs, and presented a simulation study by using travel demand data from metropolitan Atlanta in 2008 to assess the merits of dynamic ride-sharing.

Hence, authorities often encourage ridesharing. Noland et al. (2006) stated that other than prohibiting driving (implementation of odd/even driving bans) and reducing vehicle speeds, promoting carpooling is one of the most effective strategies to reduce energy consumption.

We can divide ride-sharing services into several kinds, based upon the way to make the trips. According to the classification by Chan and Shaheen (2012), traditional ride-sharing includes the "acquaintance-based" ride-sharing (also called "fampools"), which is typically formed among families and friends; the "organization-based" ride-sharing, which requires participants to share their trips within formal organizations; and the "*ad hoc*" ridesharing, which is realized through casual carpooling.

Let us take hitchhiking for an example to explain the "*ad hoc*" ridesharing. Hitchhiking is gained by asking strangers for a ride via putting the thumb out as cars pass by, indicating the need for a ride. Usually hitchhikers get their ride for free, so drivers have no responsibilities to bring these passengers to their exact destinations and thus seldom detour.

However, as one type of service provided by[1] transportation network companies (TNCs), internet based ride-sharing usually operates in a different way. With the advent of LBS, mobile technologies and GPS positioning, it becomes much easier for drivers and passengers to know the need of each other (Dailey et al., 1999; Levofsky and Greenberg, 2001).

Currently most internet based ride-sharing operates as follows: the potential drivers and passengers of ride-sharing first release the planned departure times and destinations of their trips on the ride-sharing platform. The ride-sharing platform then automatically matches the drivers and passengers with the most similar itineraries and announces the sharing plan to both drivers and passengers. If the drivers and passengers all agree on the plan, they will first make a deal and then share the trip in the determined time and routes. Once a ride-sharing trip has been finished, passengers will pay the trip fare which is quite cheap (usually less than half of the fee to make the same trip by taking a taxi) through the ride-sharing platform.

The rapid growth and popularity of internet based ride-sharing brings changes and challenges to residents' travel and city traffic. This phenomenon had attracted great attentions from all sectors of society. Particularly, the factors that may influence user intensions to make ride-sharing had been addressed. For example, Deakin et al. (2010) used data collected from statistical and geographic analysis of the downtown and campus travel markets in Berkeley, California, and surveyed employees and graduate students in University of California to assess the potential for dynamic ridesharing. They found that, if parking charges are fairly high and parking supply is limited and regulated, financial incentives and carpool parking subsidies greatly increase interest in dynamic ride-sharing. Buliung et al. (2010) studied the data provided by Carpool Zone, an on-line carpool-matching tool managed by travel demand management (TMD) group at Metrolinx. Applying a logistic regression, they illustrated that the factors increasing the probability of successful ride-sharing include geographical proximity to other users, TDM policies in workplaces, the scheduling of work, and commuter role preference. Most recently, Chen et al. (2017) applied an ensemble learning approach for improving the prediction accuracy of ridesplitting choices so as to better understanding ridesplitting behavior. Taking a variety of features, that may impact ridesplitting, such as order time, trip distance, trip fare, travel time, weather, air quality, reliability of origins/destinations and so on, into account, they modeled ridesplitting behavior as a general binary classification problem through real-world city-wide on-demand ridesourcing data. Furthermore, some researchers and industrial circles have carried out attempts to perform better ride-sharing. Nourinejad and Roorda (2016) proposed an agent based model for solving a dynamic ride-sharing problem in order to make single or multiple drivers to be matched with single or multiple passengers. Li et al. (2016) defined two variants of the Share-a-Ride problem with stochastic travel times and stochastic delivery locations, and formulated them as two-stage stochastic programming model with recourse. They integrated an adaptive large neighborhood search heuristic with three sampling strategies for the scenario generation to maximize the expected profit of serving a set of passengers and parcels using a set of homogeneous vehicles. Liu and Li (2017), in dealing with the dynamic user equilibrium of the morning commute problem under ridesharing condition, proposed a

[1] Transportation network companies (also known as ridesourcing, on-demand ride service or ride-hailing) provide prearranged and on-demand transportation services for compensation through internet or mobile internet. These internet based companies use apps to connect drivers of personal vehicles with passengers. In their ridesourcing services, applications installed on smartphone are used for booking, communications between drivers and passengers, ratings (for both drivers and passengers), and payment on line.

time-varying toll combined with a flat ridesharing price to eliminate queuing delay, thus achieving system optimum, and they found that under system optimum toll, ridesharing could attract more users and enlarge its feasible area. Aïvodji et al. (2016) presented a privacy-preserving approach which combined existing privacy enhancing methods and multimodal shortest path routing algorithms to compute mutually meeting points for both drivers and riders in ridesharing, so that users themselves, not the ridesharing operators, are in control of their own location-based data. They built a prototype implementation of their proposed approach, and demonstrated that it is possible to make both the privacy and utility levels satisfactory by conducting experiments on a real transportation network. Sánchez et al. (2016) also paid considerable attention to ensure the privacy of users. Relying on the co-utility notion, they designed a fully decentralized P2P ridesharing system, which enforce trust among driver and passenger peers with a distributed reputation management protocol. They demonstrated that their protocols are mutually beneficial and self-enforcing, and further evaluated and tested their system with real mobility data according to adoption and quality metrics. As for the city traffic, ride sourcing is a double-edged sword. Recently, Schaller (2017) found app-based ride services, also called Transportation Network Companies (TNCs), such as Uber and Lyft had made up for the public transportation system, but added to congestion problems in New York City at the same time. Nie (2017) found that the growing of ridesourcing resulted in a significant loss in ridership for taxis, helped lift the capacity utilization rate of taxis especially during the off-peak period, and worsened the congestion, to a relatively mild extend, in Shenzhen, China. However, few studies had employed empirical data to retrieve the unique service and behavior patterns of internet based ride-sharing trips. In this paper, we compare taxis and internet based ride-sharing trip records provided by DiDi company (DiDi, 2016). Results reveal some interesting findings that had never been reported before.

First, from the viewpoint of service patterns, ride-sharing mainly increases supplies in hot areas and peak hours, and on the macrocosmic level, internet based ride-sharing mainly serves as an approach for commuting. Moreover, the amount of supply contributed by internet based ride-sharing is considerable especially during peak hours. Under this circumstance, on the stand of customers, we say internet based ride-sharing helps to alleviate the stress of on-demand traveling in hot places; while, on the stand of taxi drivers, we deduced that internet based ride-sharing bring competition in hot places during hot hours, implicitly driving taxis to serve in other places, and thus relieving the spatial imbalance of taxi service in a city during hot hours. So, we conclude, internet based ride-sharing, as an effective supplement to traditional taxi service, regulates spatial and temporal supply-demand imbalance.

Second, from the viewpoint of individual behavior patterns, we identify two kinds of internet based ride-sharing drivers, by using a clustering method. The first kind of drivers usually provides ride-sharing along daily home-work commuting. Trips served by these drivers have relatively constant origin-designation (OD) pairs. Thus, these drivers can be viewed as home-work commuting drivers who behave similarly as conventional ride-sharing drivers. The second group of drivers does not serve regularly and roams around the city even in working hours. Therefore, there are not constant OD pairs in their ride-sharing trips. Thus, this kind of drivers can be viewed as essentially converted taxi vehicle drivers. Counterintuitively, we find that home-work commuting drivers account for only a small part of total drivers and they only serve a small number of commuting trips.

Third, most ride-sharing drivers occasionally share their journeys for a very limited times in a month. Thus, internet based ride-sharing drivers do not directly compete with conventional taxi drivers in on-demand ride-hauling service market. Additionally and more importantly, we find that one driver is more likely to be classified into the home-work commuting group when the quantity of trips served by this specific driver is not high. These findings accord with the newly introduced policies by local authorities.

Fourth, on the whole, internet based ride-sharing drivers intend to make long distance trips. Here, we need to emphasize that some people hold the wrong viewpoint that ride-sharing is just traditional hitchhiking worked through mobile phone. We find that, since internet based ride-sharing drivers are paid, they can detour further to pick up or drop off passengers than traditional hitchhike drivers. There are possibly two reasons: (1) The cost of detouring can be compensated by the fee paid by the passengers; (2) Under the assurance of on demand car service platform, such as DiDi and Uber, the risk of failing to pick up any passengers for internet-based ride-sharing is much lower than that for conventional hitchhike.

All these findings are helpful for policymakers at all levels to make informed decisions about deployment of internet based ride-sharing service. Recently, local authorities in many Chinese cities announced that every ride-sharing driver can only serve twice or thirds per day to restrict the use of converted taxi vehicles. We think that they had also found these patterns and applied it in making the policies. Furthermore, to the best knowledge of the authors, this paper is one of the first attempts to apply big data analytics to the analysis of the most important patterns of internet based ride-sharing with comparisons to taxi, based on real-world metropolis-wide on-demand ride service data.

To better present our findings, the rest of this paper is arranged as follows. Section 2 presents the trip data used in this paper, and demonstrates the research overview. Section 3 introduces the methodologies used in the analyses of services patterns as well as unique individual behaviors. Section 4 further presents the verification of models and calibration of the parameters. Section 5 shows the results of services patterns and individual behavior analyzed in the paper. Finally, Section 6 concludes the paper.

2. Dataset, preliminary analysis and research overview

Our study is based on a set of randomly sampled and anonymized taxi and internet-based ride-sharing trip records provided by DiDi company. When it was founded in 2012, DiDi served as a taxi-hailing application, then gradually developed other types of private-car hailing services. Now, DiDi is the biggest ride-hailing service company in China and one of the largest on-demand ridesourcing service platforms in the world (Shih, 2015). China has the world's most fertile market for on-demand transportation service. It has more than 750 million potential riders and huge runway for growth, because hundreds of millions more Chinese will enter the middle class over the next decade and are potentially users of ride-hailing service. So, the empirical data from DiDi company reflects the most exciting change of ride-sharing nowadays. Currently, there are four main types of travel services provided by DiDi

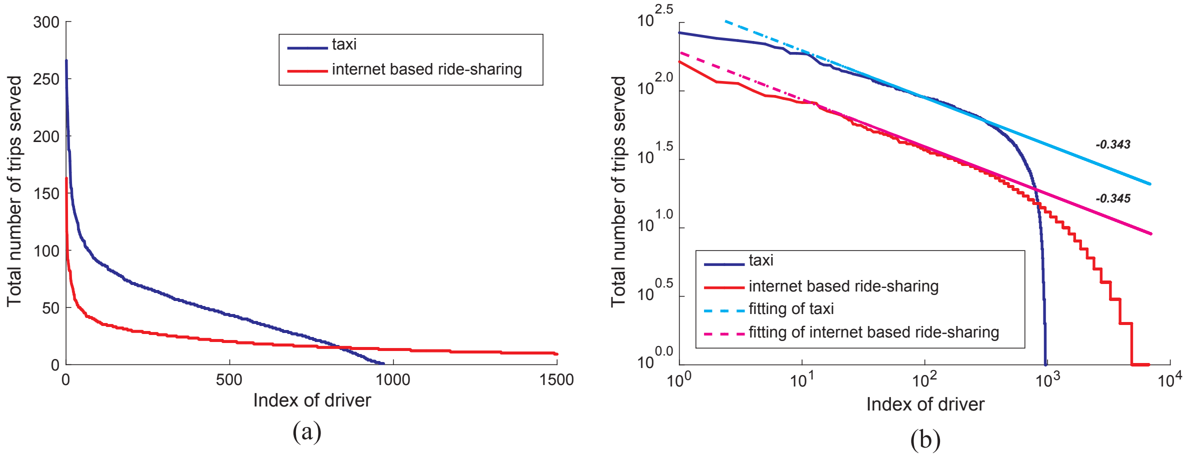


Fig. 1. The (sorted) total number of trips served by taxi drivers and internet based ride-sharing car drivers, (a) in normal coordinates and (b) in double logarithmic coordinates.

platform, e.g., Taxi Hailing, Express, Private Car, and Hitch. The internet-based ride sharing in this paper refers to the Hitch mode.

The dataset used here contains records of 48,251 taxi trips and 41,795 internet based ride-sharing trips occurred in Beijing, during December 3, 2015 to January 3, 2016 (32 days in total). 970 taxi drivers and 6471 internet based ride-sharing drivers were involved in this study. Each trip record includes the pick-up/drop-off locations (longitude and latitude) and the associated time stamp. Unfortunately, since there are only pick-up and drop-off locations provided, it is impossible to investigate problems related to trajectory which are important to comprehensive researches of taxi or ride-sharing cars. Nevertheless, in this paper, we focus on the origins and destinations and find results we are interested in.

Since human traveling activities can be quite different between workdays and weekends/holidays, we divide the trip records into two groups, namely workdays group and weekends/holidays group (which contain all the weekends in December 2015 and January 1–3, 2016 which are the New Year's holiday). There are 37,289 taxi trip records (965 drivers involved) and 33,285 ride-sharing trip records (5899 drivers involved) in the workdays group and 10,962 taxi trip records (894 drivers involved) and 8510 ride-sharing trip records (2083 drivers involved) for the weekends/holidays group.

Fig. 1 gives some basic statistics of the data and shows the (sorted) total number of trips served by taxi drivers (plotted in blue color) and by internet based ride-sharing drivers (plotted in red[2] color) during 32 days. Only the top 1500 internet based ride-sharing drivers are plotted here, since the total number of trips served by other internet based ride-sharing drivers is no more than 10. On average, most taxi drivers made significantly more trips than internet based ride-sharing drivers. The most diligent internet based ride-sharing driver only served 163 trips in 32 days. In other words, no more than 6 trips will be served per day for an internet based ride-sharing driver, according to our samples.

This finding indicates that most ride-sharing drivers occasionally share their journeys for very limited times in a month and they do not consider serving passengers as a job. This verify that, unlike other internet based ride-hailing service, internet based ride-sharing drivers do not directly compete with conventional taxi drivers in on-demand ride-hauling service market, mainly because its much lower pay rate.

Besides, we observe clear Power law (Mitzenmacher, 2005; Clauset et al., 2009) existing in the total number of trips served by taxi drivers and ride-sharing drivers. The estimated exponents for taxi and internet based ride-sharing are -0.343 and -0.345 respectively, which are roughly equal. Nevertheless, the origins of these exponents remain to be further discussed. Since this topic does not fall into the main theme of this paper, we would like to discuss it in another dedicated paper.

Here in this paper, we study (1) service patterns, in terms of (1) temporal service patterns, (2) spatial service patterns, (3) spatial-temporal patterns and (4) traveling distance patterns, and (2) individual behavior patterns, in terms of (1) divisions of commuting styles and (2) detour patterns. For service patterns, we adopt basic big data analytics and visualization methods to reveal the temporal, spatial service patterns and traveling distance patterns. Since the problems involved in those analyses are quite simple, there are not complex methodologies. However, for the analysis of spatial-temporal service patterns and for the study of individual behavior patterns, we develop a series of corresponding methodologies, which is demonstrated in detail in the following Section 3.

3. Methodology

3.1. Spatial-temporal patterns analysis

We first grid the study region of Beijing city into 200 m × 200 m square areas. To present our method more clearly, we label the

[2] For interpretation of color in Figs. 1, 2, 3, 4, 13, 14, 17 and 18, the reader is referred to the web version of this article.

squares by (i,j), as i represent the ith row and j for the jth column. Assume there are p rows and q columns in total, then $i \in [1,p] \cap \mathbb{Z}$, and $j \in [1,q] \cap \mathbb{Z}$. We use $h \in [1,24] \cap \mathbb{Z}$ to present the number of time slots during the day. Now we have a $1 \times h$ vector $\boldsymbol{M}_{i,j}$ to demonstrate the macro pattern, i.e. the number of trips, at the specific location (i,j). Previous studies had found that the macro traffic pattern can be described by some linear combinations of basis collective patterns (Peng et al., 2012). To find out basis collective patterns we need to adopt appropriate inference methods. Here we adopt the method proposed by Peng et al. (2012).

For the basis collective pattern, corresponding to the macro pattern $\boldsymbol{M}_{i,j}$, we can define a set of $1 \times h$ vectors: $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \dots, \boldsymbol{B}_\lambda$, each of which stands for one basis pattern we are seeking for.

We already know that the macro traffic pattern is a linear combination of basis collective patterns, therefore, we have

$$\boldsymbol{M}_{i,j} = \boldsymbol{S}_{i,j} \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \boldsymbol{B}_3 \\ \vdots \\ \boldsymbol{B}_\lambda \end{bmatrix} \tag{1}$$

where $\boldsymbol{S}_{i,j}$ is a row vector containing the λ coefficients for the linear combination on the right-hand side of the above equation.

All the macro patterns $\boldsymbol{M}$ in the $i \times j$ squares can be factorized in the above method, thus for abbreviation, we have

$$\boldsymbol{M} = \boldsymbol{S}\boldsymbol{B} \tag{2}$$

There are many matrix decomposition methods to deal with the above equation (Cichocki et al., 2009; Townsend and Trefethen, 2015; Udell et al., 2016), including principal components analysis (PCA), k-singular value decomposition (SVD), maximum margin matrix factorization, and nonnegative matrix factorization (NMF), etc. We know that all the entries of the two matrices, $\boldsymbol{S}$ and $\boldsymbol{B}$, should be nonnegative by their physical meaning, thus nonnegative matrix factorization (NMF) method (Lee and Seung, 1999; Lee and Seung, 2001; Lin, 2007; Cichocki et al., 2009; Peng et al., 2012) should be an appropriate solution for the decomposition in Eq. (2). In our scenario, the perspective is to decompose a matrix $\boldsymbol{M} \in \mathbb{R}_+^{pq\times h}$ into two low rank nonnegative factors $\boldsymbol{S} \in \mathbb{R}_+^{pq\times\lambda}$ and $\boldsymbol{B} \in \mathbb{R}_+^{\lambda\times h}$. Through this process, we can excavate out the basis collective patterns of the macro traffic which are demonstrated by the row vector of $\boldsymbol{B}$. As for matrix $\boldsymbol{S} \in \mathbb{R}_+^{pq\times\lambda}$, each element denotes the scale of traffic flow with respect to the corresponding pattern category, in the specific square (i,j). However, we would not focus our research on $\boldsymbol{S} \in \mathbb{R}_+^{pq\times\lambda}$ in this paper.

In the above nonnegative matrix factorization, the only problem left to be dealt with is to determine the number of basis collective patterns, λ. It is difficult to find stable results for the factorization since NFM starts with random initial conditions (Lee and Seung, 1999). Thus, experiments were carried out on taxi data with many different initial conditions in previous studies. Researchers found out that when $\lambda = 3$, the results can be stable. Furthermore, according to physical meaning and from the land-use perspective, $\lambda = 3$ is an effective and reasonable choice, since when $\lambda = 3$, the solution can perfectly describe the three main categories for the purpose of trips, i.e. commuting between home and workplace (pattern 1), business traveling between two workplaces (pattern 2), and trips from or to other places (pattern 3) (Peng et al., 2012; Wang et al., 2014). However, no previous literatures have explored the patterns for ridesourcing trips, especially for the internet based ride-sharing served through the on-demand platform. Therefore, we perform experiments on our collected data for both internet based ride-sharing and taxi trips, respectively, by exploring the value of λ from 2 to 4. The results are demonstrated in Section 5.

3.2. Analysis of individual behavior patterns for internet based ride-sharing

On individual level, internet based ride-sharing drivers show great diversities. Our interest in this paper is to verify whether they share ride mainly during their relatively constant home-work commuting.

Home-work commuting is the most important commuting form in many contemporary cities. City planners, policy makers, and public agencies have advocated jobs-housing balance policies and car-pool policies to reduce travel demand, traffic congestion and associated air/noise pollution (Buliung et al., 2010; Zhou et al., 2014; Yang et al., 2015). If relatively constant home-work commuting are observed in internet based ride-sharing trip records, we can reversely estimate which citizens provide these rides indeed.

Here we select 1000 drivers who served more than 10 trips during the monitored workdays to study their home-work commuting styles. The followings are examples of three different commuting styles.

If a driver provides ride-sharing mainly during either his/her home-to-work commuting or his/her work-to-home commuting, most of his/her trip records will fall in one cluster and the trip records fall in other clusters belong to random commuting (Murphy and Killen, 2011). If a driver provides ride-sharing mainly during both his/her home-to-work commuting and work-to-home commuting, most of his/her trip records will fall in two clusters. When the two clusters are expressed by vectors, they will have almost opposite directions. If a driver provide ride-sharing during his/her random commuting, his/her trip records may scatter in three clusters and the divergence of records in these clusters will be much larger than the divergence of records falling in the major cluster (s) of drivers that provide ride-sharing mainly during home-to-work commuting. Here the divergence of one specific cluster is defined by the distance among trip vectors in that cluster.

For example, Fig. 2 shows the trips made by internet based ride-sharing driver whose ID is 377. The distinguishable ID we allocate to each driver is numbered by the amounts of trips they served. We can see that nearly all his/her trips fall in one cluster. So, this driver belongs to the first kind of drivers who provide ride-sharing mainly during either his/her home-to-work commuting or his/her work-to-home commuting.

Fig. 3 shows the trips made by internet based ride-sharing driver labeled with ID 76. We can see that most of his/her trips fall in

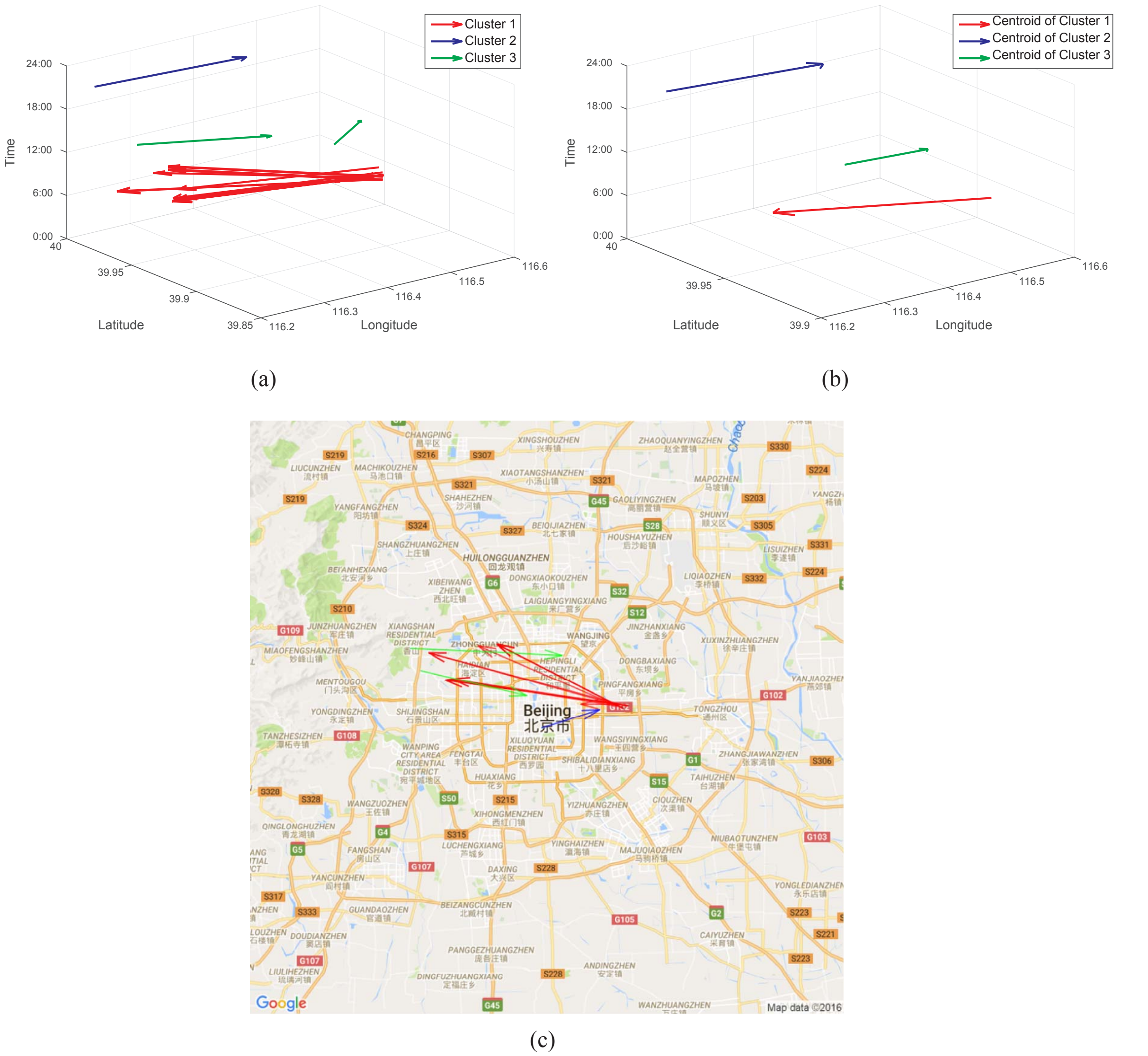


Fig. 2. An illustration of the trips made by internet based ride-sharing driver (labeled with ID 377): (a) the trips grouped in three clusters, respectively, (b) the estimated cluster centers, and (c) the trips plotted on the map of Beijing city.

two clusters whose cluster centers have opposite directions. So, this driver belongs to the second kind of drivers who provide ride-sharing mainly during both home-to-work commuting and work-to-home commuting.

Fig. 4 shows trips made by internet based ride-sharing driver labeled with ID 1. No regular patterns can be found in his/her trips. So, this driver belongs to the third kind of drivers who provide ride-sharing mainly during random commuting.

(1) Methods for divisions of drivers based on their commuting styles

To detect the possible home-work commuting style for each internet based ride-sharing driver, we adopt the K-means clustering algorithm (Jain, 2010; Duda et al., 2010; Murphy, 2012) to group all trip records of each driver into three clusters in an unsupervised manner. Similar to (Guo et al., 2012; Peng et al., 2012), the features of trips are selected as: the origin position $O(x,y)$ (in longitude and latitude degrees), the destination position $D(x,y)$ (in longitude and latitude degrees), as well as the payment time t (in hour of a day) of each trip, which then form a five-element vector $\boldsymbol{V}(x_o,y_o,x_d,y_d,t)$ (all normalized into the range [0, 1]).

K-means clustering is widely used as a method for vector clustering. It partitions some samples (in vector form) into K clusters, indicated as $\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_K$, in which each sample belongs to the cluster with the nearest mean. We usually define the mean of all samples in one cluster as the center (vector) of this cluster, which is written as $\boldsymbol{C}_i = (x_{oc}^{(i)},y_{oc}^{(i)},x_{dc}^{(i)},y_{dc}^{(i)},t_c^{(i)})$, $i = 1,2,\cdots,K$.

Here, we calculate the distance between sample vectors and the cluster center by Manhattan distance measure. Take $\boldsymbol{V}_j(x_{oj},y_{oj},x_{dj},y_{dj},t_j) \in \boldsymbol{S}_i$ for an example, the distance between $\boldsymbol{V}_j$ and the cluster center of $\boldsymbol{S}_i$, can be measured as

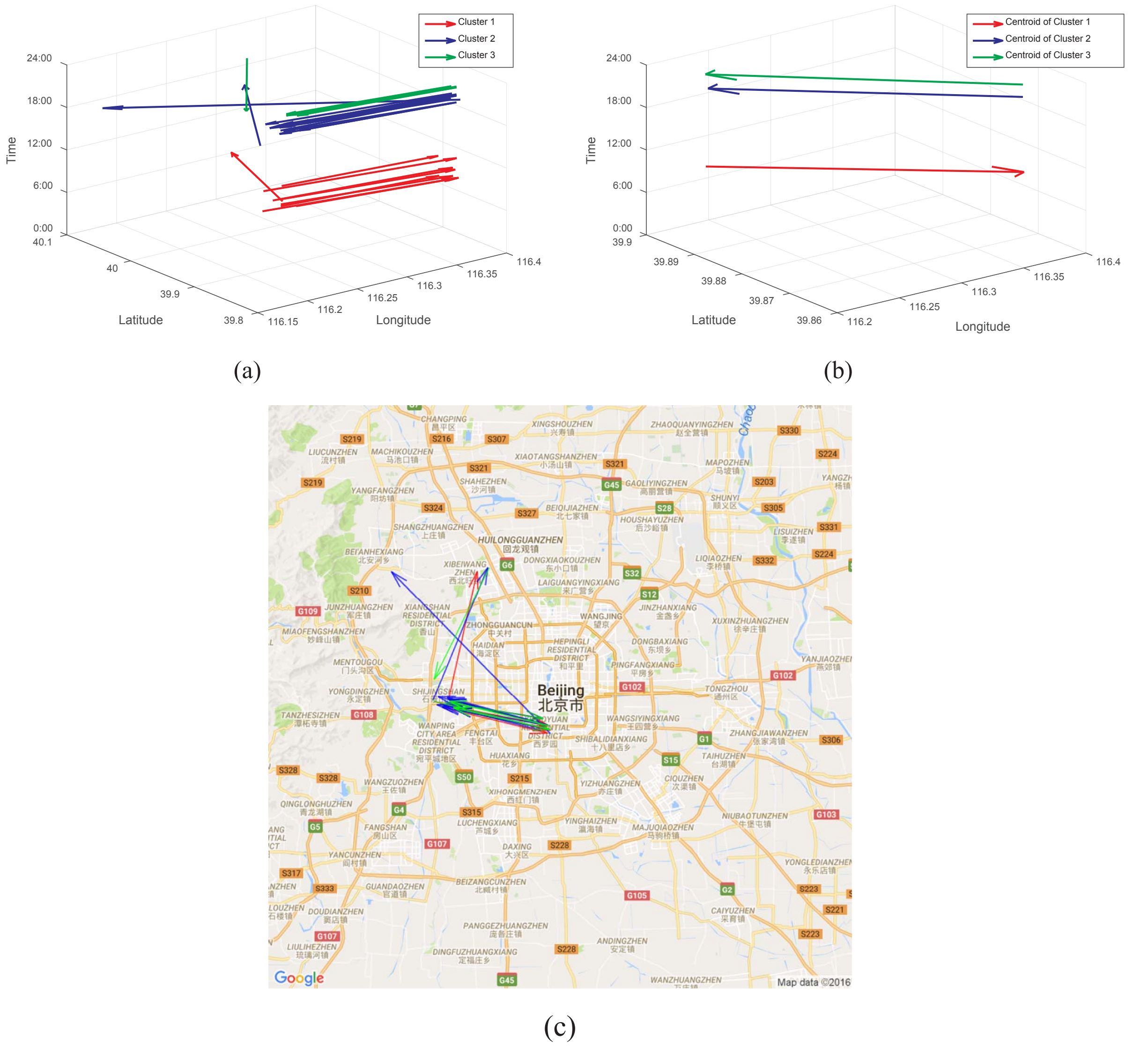
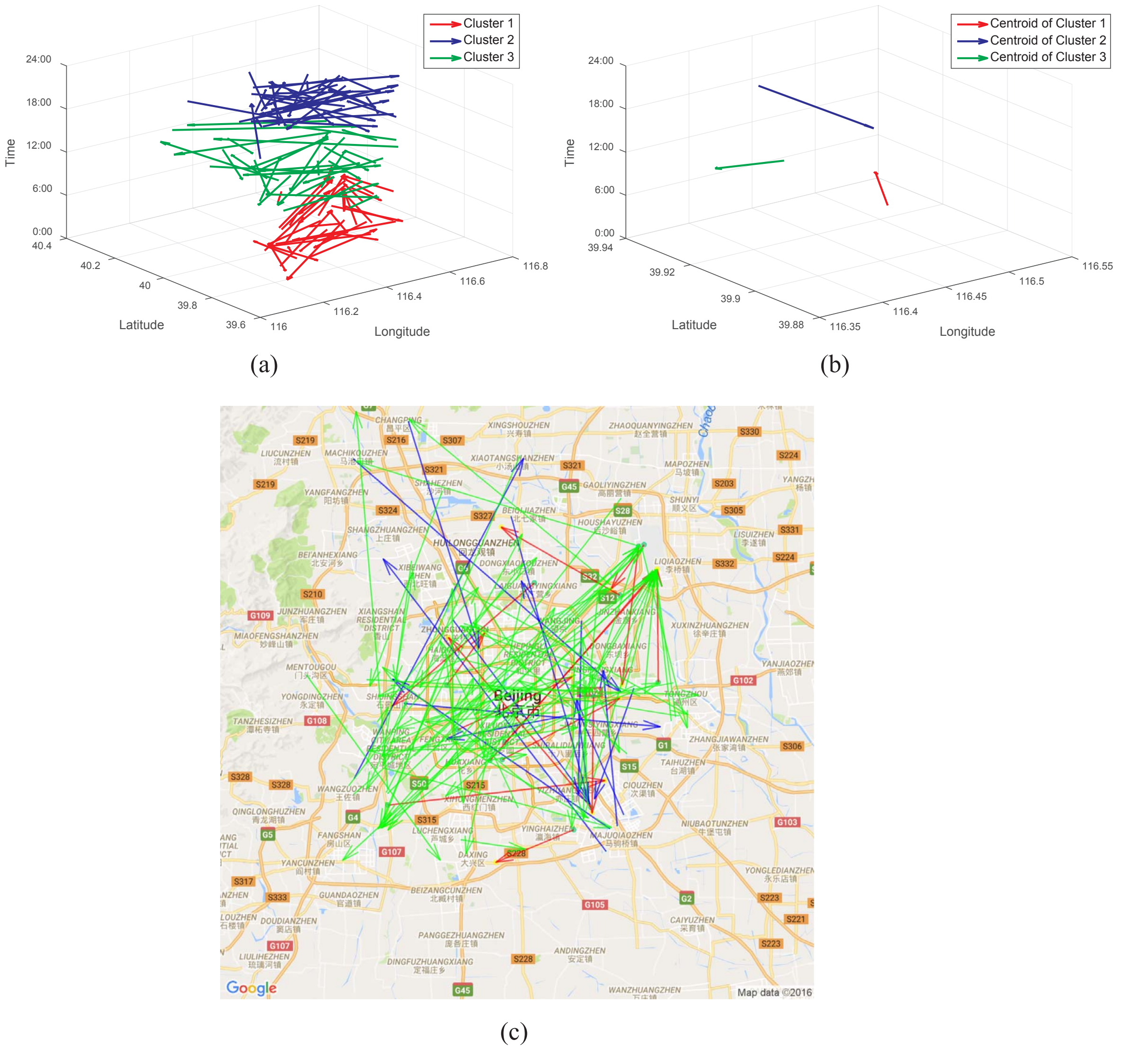


Fig. 3. An illustration of the trips made by internet based ride-sharing driver (labeled with ID 76): (a) the trips grouped in three clusters, respectively, (b) the estimated cluster centers, and (c) the trips plotted on the map of Beijing city.

$$d_j^{(i)} = |x_{oj} - x_{oc}^{(i)}| + |y_{oj} - y_{oc}^{(i)}| + |x_{dj} - x_{dc}^{(i)}| + |y_{dj} - y_{dc}^{(i)}| + |t_j - t_c^{(i)}| \tag{3}$$

To identify the percentage of home-work commuting drivers from the investigated 1000 drivers who served more than 10 trips during the monitored workdays, we carry out the classifying steps shown in Fig. 5.

Step (1) we check three cluster of the trips served by each driver to see if there exits one cluster containing more than 70% of all the trips of the particular driver. If so, we calculate the divergences within this cluster. If the divergence is below one threshold selected through sampling experiments, then this driver may belong to the first kind of drivers. To further verify which kind of group this driver belongs to, we check whether most samples in this largest cluster fall into the morning home-to-work commuting or the evening work-to-home commuting time period of one day. If so, this driver is distributed to the first kind of drivers who provide ride-sharing mainly during either his/her home-to-work commuting or his/her work-to-home commuting; and if not, to the third kind of drivers who provide ride-sharing mainly during random commuting. If not, we turn to Step (2).
Step (2) we further check whether there are two clusters containing more than 70% of all the trips of the particular driver. If so, we turn to Step (3); if not, the driver belongs to the third kind of drivers.
Step (3) we calculate the divergences of two selected clusters in Step (2), to see whether the two divergences are below the threshold we set. If so, we turn to Step (4); if not, the driver belongs to the third kind of drivers.
Step (4) we calculate the angle θ of two selected clusters in Step (2). Specifically, if we name these two clusters as



(a)

(b)

(c)

Fig. 4. An illustration of the trips made by internet based ride-sharing driver (labeled with ID 1): (a) the trips grouped in three clusters, respectively, (b) the estimated cluster centers, and (c) the trips plotted on the map of Beijing city.

$\boldsymbol{C}_a = (x_{oc}^{(a)},y_{oc}^{(a)},x_{dc}^{(a)},y_{dc}^{(a)},t_c^{(a)})$ and $\boldsymbol{C}_b = (x_{oc}^{(b)},y_{oc}^{(b)},x_{dc}^{(b)},y_{dc}^{(b)},t_c^{(b)})$, we can calculate the OD vector of the two cluster as $\overrightarrow{\boldsymbol{OD}}^{(a)} = (x_{dc}^{(a)}-x_{oc}^{(a)},y_{dc}^{(a)}-y_{oc}^{(a)})$ and $\overrightarrow{\boldsymbol{OD}}^{(b)} = (x_{dc}^{(b)}-x_{oc}^{(b)},y_{dc}^{(b)}-y_{oc}^{(b)})$. The θ is then gotten as

$$\theta = \arccos\frac{\overrightarrow{\boldsymbol{OD}}^{(a)} \bullet \overrightarrow{\boldsymbol{OD}}^{(b)}}{\|\overrightarrow{\boldsymbol{OD}}^{(a)}\|\|\overrightarrow{\boldsymbol{OD}}^{(b)}\|} \tag{4}$$

We check whether the angle θ is within [169°,180°]. If so, we turn to Step (5); if not, the driver belongs to the third kind of drivers.

Step (5) we further check whether most samples in the two selected cluster fall into the morning home-to-work commuting and the evening work-to-home commuting time period of one day, respectively. If so, this driver is distributed to the second kind of drivers who provide ride-sharing mainly during both his/her home-to-work commuting and work-to-home commuting; if not, this driver belongs to the third kind of drivers.

We can see, according to their different serving styles, the above mentioned three kinds of drivers can be further grouped into two different types. One type of drivers can be viewed as home-work commuting drivers who behave similarly as conventional ride-sharing drivers. Trips served by these drivers have relatively constant origin-designation (OD) pairs. The other type of drivers can be viewed as essentially converted taxi vehicle drivers. There are not constant OD pairs in their ride-sharing trips.

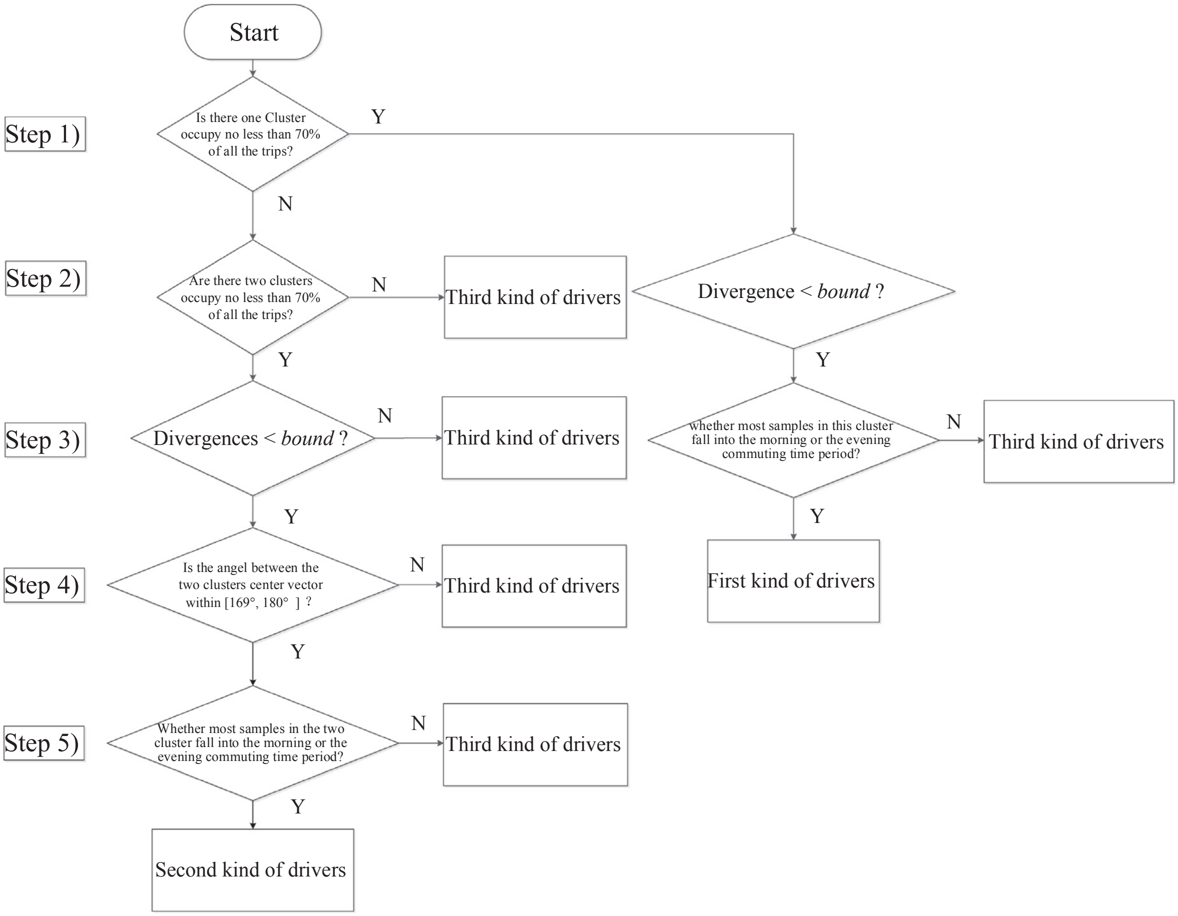

Start
Step 1)
Is there one Cluster occupy no less than 70% of all the trips?
Y
N
Step 2)
Are there two clusters occupy no less than 70% of all the trips?
N
Third kind of drivers
Divergence < bound ?
Y
Step 3)
Divergences < bound ?
N
Third kind of drivers
whether most samples in this cluster fall into the morning or the evening commuting time period?
N
Third kind of drivers
Y
Step 4)
Is the angel between the two clusters center vector within [169°, 180°] ?
N
Third kind of drivers
First kind of drivers
Y
Step 5)
Whether most samples in the two cluster fall into the morning or the evening commuting time period?
N
Third kind of drivers
Y
Second kind of drivers

Fig. 5. An illustration of classifying steps of drivers.

(2) Analyzing method of detour patterns

To study drivers' detour behaviors, we calculate the average detour distance d_{detour} of home-to-work commuting and work-to-home commuting trips served by 206 selected drivers who mainly serve home-to-work or work-to-home commuting trips and are with high probabilities to be classified into home-work commuting drivers (i.e., the first and second kinds of drivers). Here, we define d_{detour} as the average of the sum of the distances between the origin/destination of each trip to the origin/destination of the associated cluster center (in Manhattan distance measure). More precisely, for trips in one specific cluster $\boldsymbol{S}_k$ we have

$$d_{detour} = \frac{1}{|\boldsymbol{S}_k|} \sum_{V_l(x_{ol},y_{ol},x_{dl},y_{dl},t_l)\in \boldsymbol{S}_k} |x_{ol}-x_{oc}^{(k)}| + |y_{ol}-y_{oc}^{(k)}| + |x_{dl}-x_{dc}^{(k)}| + |y_{dl}-y_{dc}^{(k)}| \tag{5}$$

where $|\boldsymbol{S}_k|$ denotes the cardinality of the cluster set $\boldsymbol{S}_k$.

For the first kind of drivers, we only check the largest cluster to calculate d_{detour}; for the second kind of drivers, we check the largest two clusters to calculate d_{detour}.

Furthermore, we provide the empirical distribution of the driver-specific detour proportions to better demonstrate drivers' detour patterns intuitively (shown in Fig. 21 in Section 5). The driver-specific detour proportion is defined as the ratio between the average detour distance of each driver and the average home-work commuting distance of the same driver.

4. Model verification and calibration

In this section, we discuss the verification of our classification model in Section 3, and test the sensitivity of classification results to the parameters we select for the study of drivers' commuting styles in the analysis of individual behavior patterns.

4.1. Verification of classification of drivers based on their commuting styles

The basic idea of cross validation used here is to check the consistency between machine learning based classification results and human volunteer based classification results.

According to the classification rules introduced in Section 3.2, we design the corresponding rules for volunteers as follows.

(1) First of all, judge the trips' randomness served by specific driver to pick out drivers with messy trips who clearly belongs to random commuting category.
(2) Then, check the driver's trips to see whether there exits one majority group containing more than approximately 60% of all the trips of the particular driver. If so, check the trips in the majority group to see whether they fall in the morning rush hours or the evening rush hours. If the trips in the majority group fall in either the morning rush hours or the evening rush hours, this driver belongs into home-work commuting category. If not, go to Step (3). Otherwise, go to Step (3).
(3) Check whether there are two groups of the trips, with the approximately opposite directions, containing more than 60% of all the trips of the particular driver. If so, check whether the majority of the trips in these two groups belong to the morning home-to-work commuting and the evening work-to-home commuting, respectively. If so, this driver belongs to home-work commuting category. If not, the driver ought to belong to the random commuting category. Otherwise, the driver should be classified into the random commuting category.

We give the volunteers a detailed explanation of these rules, and made sure that they know how to classify these drivers. Then, we plot all the trips of each internet based ride-sharing driver onto a map like Figs. 2(c), 3(c), and 4(c). We let volunteers read the figure of trips and determine to which kind this driver belongs.

In cross validation, we had 19 volunteers in total. These volunteers come from a wide range of backgrounds, including undergraduate students, master students, doctoral students and teachers. Based on the voting majority, one driver would be taken into home-work commuting category, only when more than 9 volunteers did so.

Tests on 1000 drivers show that, for 92.3% of drivers, most volunteers had the same classification results with our machine learning model. This indicates that our model make satisfactory classification of drivers based on their trip data.

We also find that it is hard to give a clear classification for a small number of ambiguous drivers. However, this will not significantly influence the conclusions we draw.

4.2. Sensitivity of the classification to the parameters

As we know that the selection of parameters would affect the classification result, here, we provide the angle bound as one example illustrated by Fig. 6.

Fig. 6(a) gives the classification results, the changes of classification accuracy, when the upper side of the angle bound is fixed at 180° and the lower bound increases from 160° to 180° with a step of 1°. We can see that with the lower angle bound increasing, the classification accuracy rises at first, and then decreases when the lower bound reaches 169°, where we get the optimal classification with an accuracy rate of 92.3%, as marked by the arrow and the pink dash line.

As illustrated by Fig. 6(b), the above results can be explained as follows: with the lower angle bound increasing, the precision ratio (e.g. the ratio between the number of drivers who are classified into home-work commuting group and are really home-work commuting ones, and the total number of drivers who are classified into home-work commuting group) will rise and serve as the principal influence factor of accuracy making it increase at first; while the recall ratio (e.g. the ratio between the number of drivers

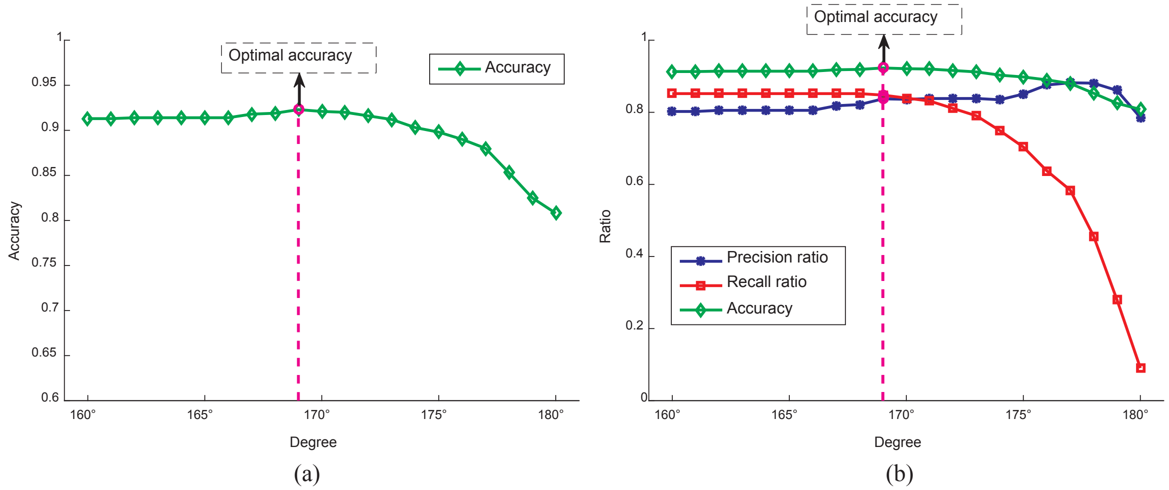
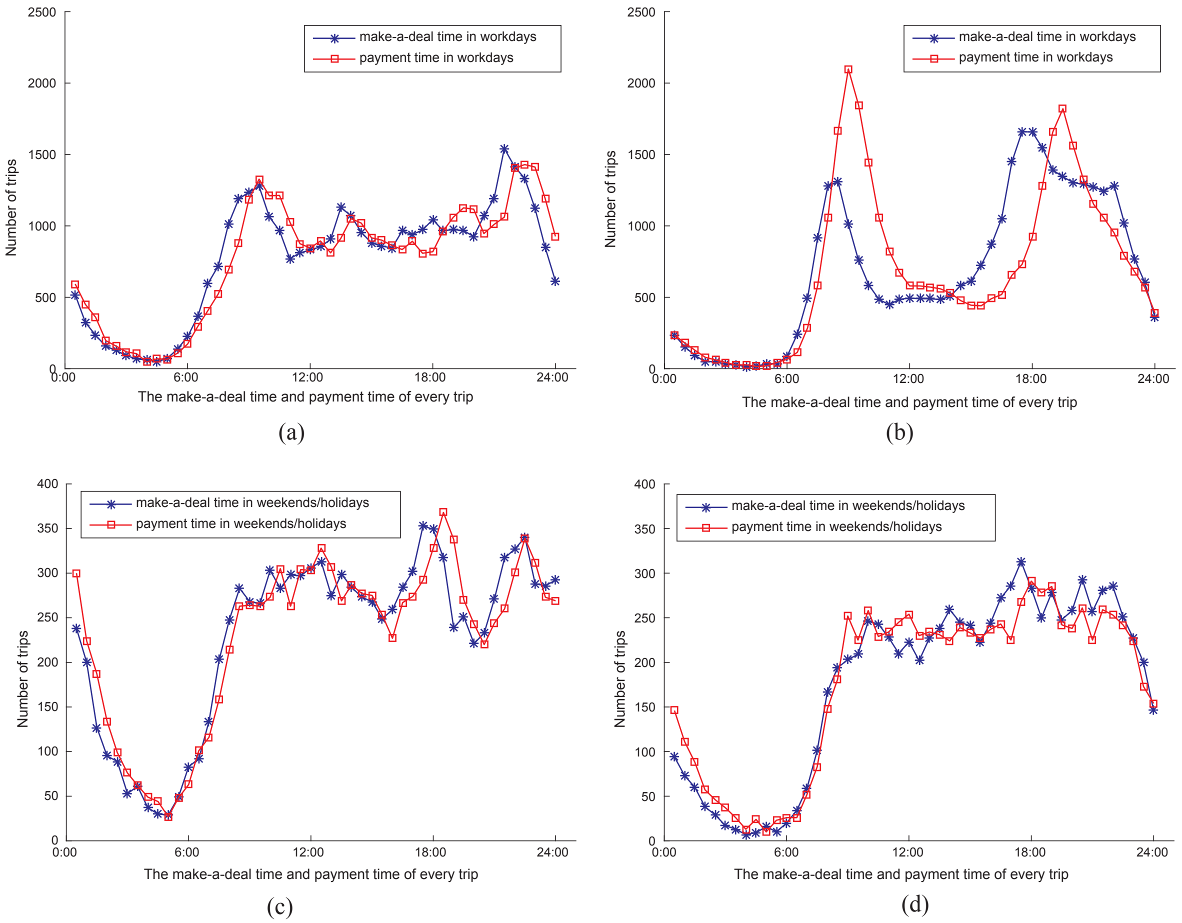


Fig. 6. Results of testing the sensitivity of the classification to the angle bound, when the upper angle bound is fixed at 180° and the lower bound increases from 160° to 180° with a step of one degree. Subfigure (a) depicts the results when only accuracy is provided. Subfigure (b) depicts the results when accuracy with precision ratio and recall ratio are all provided. The arrow and the pink dash line mark the optimal classification accuracy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) (b) (c) (d)

Fig. 7. The make-a-deal time and payment time distribution per half hour in a day during the workdays by (a) taxi and (b) internet based ride-sharing, and during the weekends/holidays by (c) taxi and the (d) internet based ride-sharing.

who are classified into home-work commuting group and are really home-work commuting ones, and the total number of drivers who are really home-work commuting ones) will decrease and serve as the principal influence factors of accuracy at a certain time making the accuracy decrease from that point.

5. Results

5.1. Service pattern analysis of taxis and internet based Ride-Sharing

(1) Temporal service patterns

Fig. 7 compares the trips (depicted by the make-a-deal times and payment times) served by 970 taxi drivers and 6471 internet based ride-sharing drivers, per half an hour during the whole day. From Fig. 7, we can draw at least three findings:

(1) In Beijing, the morning rush hours are 7:30 A.M. to 9:30 A.M., and the evening rush hours are 16:30 P.M. to 18:30 P.M, for workdays. From the drop-off times, we can see that most of internet based riding-sharing trips occurred during these two rush hours.
(2) In contrast, the trips served by taxis remain above 75% of the highest value in a workday, since 7:30 A.M. to 23:00 P.M. This shows that the taxis service in Beijing almost reaches its capacity limit in most time of a workday.
(3) During the workdays, 965 taxi drivers made 4083 trips in morning rush hours and 3482 trips in evening rush hours. Meanwhile, 5899 internet based ride-sharing drivers made 6657 trips in morning rush hours and 3585 trips in evening rush hours. In other words, 5899 internet based ride-sharing drivers made more trips than 965 taxi drivers in rush hours of a workday.

We calculate these numbers using the payment time (the time that the passenger(s) paid the drivers). Usually, the payment time is almost equal to the drop-off time for a trip.

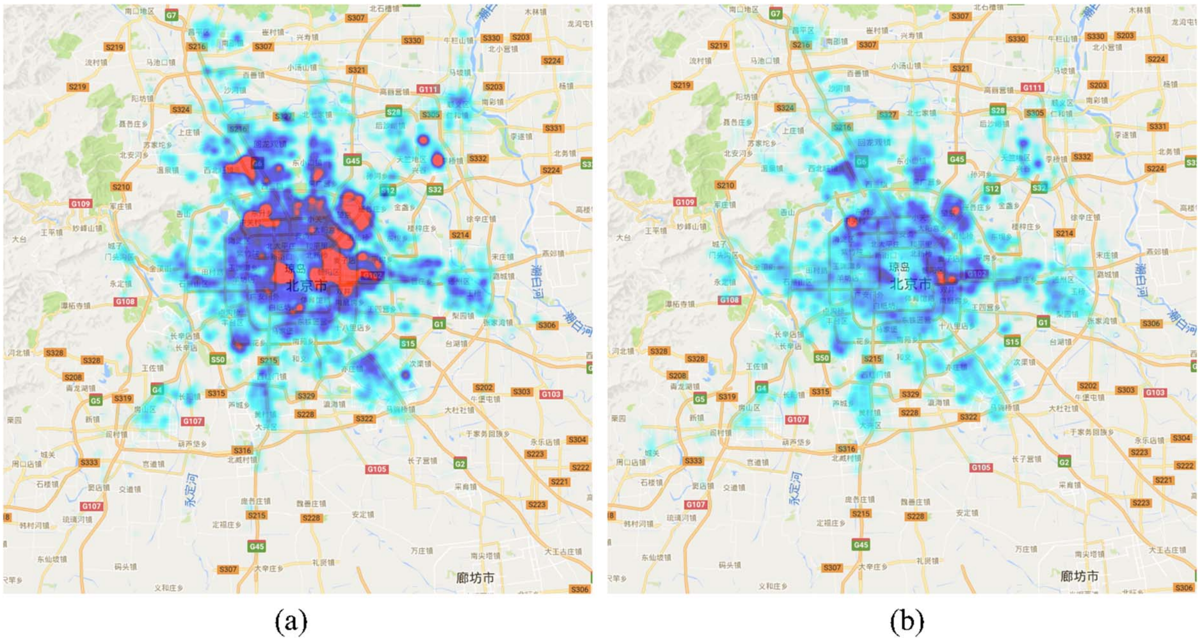


Fig. 8. Hotspot visualizations of the pick-up locations of trips served by (a) taxi and (b) internet based ride-sharing, when the absolute trip numbers are considered.

Based on the above three findings, we can conclude that the emergence of internet based ride-sharing greatly increases travel supply in rush hours and help alleviate the lack of on-demand riding service.

(2) Spatial service patterns

To highlight the scale and relative differences between taxi and internet-based ride-sharing OD, we provide the following plots. Figs. 8–11 compare the hotspots of the pick-up and drop-off locations of trips made by taxi and internet based ride-sharing, respectively. The color of each grid in the hotspot visualization indicates on average how many trips begin or end in this grid.

One should notice that, in Figs. 9 and 11, the different color shades in each region just represent the distribution of quantity of trip order in that region with respect to the overall order quantity of taxi or internet based ride-sharing. Also, it makes no sense to directly

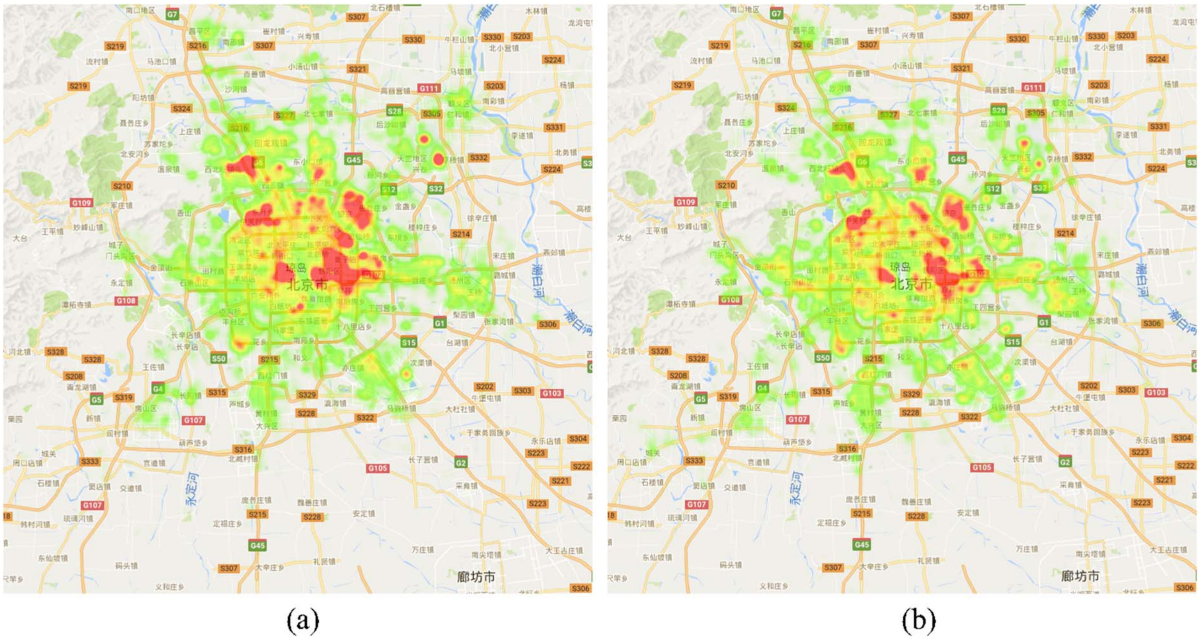


Fig. 9. Hotspot visualizations of relative distribution of pick-up places served by (a) taxi and (b) internet based ride-sharing.

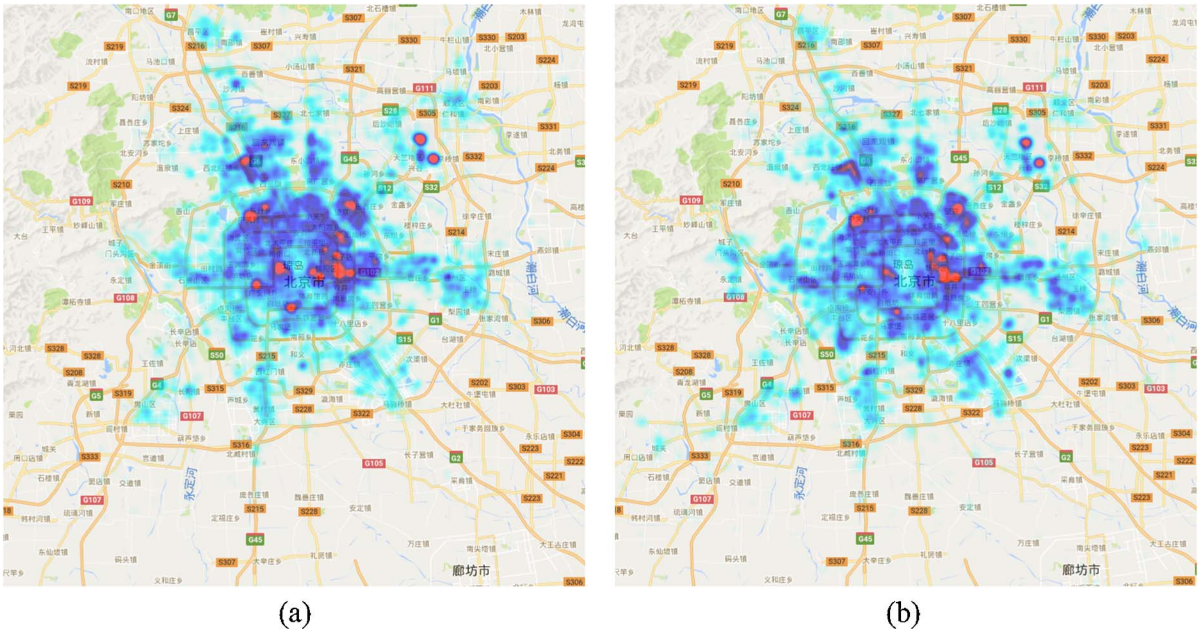


Fig. 10. Hotspot visualizations of the drop-off locations of trips served by (a) taxi and (b) internet based ride-sharing, when the absolute trip numbers are considered.

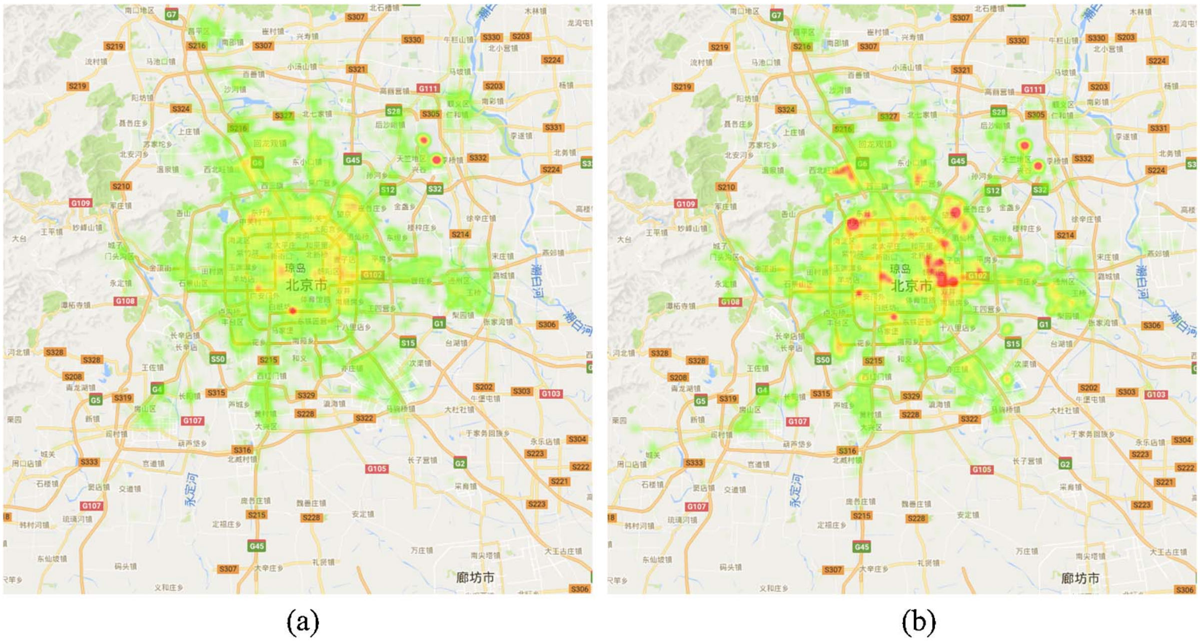


Fig. 11. Hotspot visualizations of relative distribution of drop-off places served by (a) taxi and (b) internet based ride-sharing.

compare the color deepness among these pictures.

Fig. 8(a) and (b) gives the absolute pick-up number based hotspot visualization of the pick-up locations for taxi and internet based ride-sharing, respectively. Since the number of trips for internet based ride-sharing is much smaller than that for taxi, we can only identify three most prominent hotspots in Fig. 8(b).

Accordingly, Fig. 9(a) and (b) plots the normalized pick-up number based hotspot visualization of pick-up locations for taxi and internet based ride-sharing, respectively. As shown in Fig. 9(a) and (b), the hotspot regions of internet based ride-sharing are similar to those of taxi service, if the relative level-of-service is considered. These hotspots are either workplaces, resident regions or traffic hubs.

We also provide two kinds of hotspot visualizations of the drop-off locations of trips served by taxi and internet based ride-sharing, see Figs. 10 and 11. In contrast, when the relative trip numbers are considered, we can only identify three extremely hot drop-offs (e.g. the two terminals of Beijing international airport and the Beijing South Railway Station) for taxi, see Fig. 11(a). This is mainly because the drop-off numbers served by taxi are so large in these three locations that the other locations had been suppressed in color.

As we can see, the hotspots of the pick-up/drop-off locations of trips made by internet based ride-sharing, to certain extend, roughly overlap with those made by taxi. This means that internet based ride-sharing helps to alleviate the stress of travel demand in these hot places. In addition, the highlighted hotspots of trips served by internet based ride-sharing are typical working regions or resident regions mainly because of the severe shortage of taxi service in such areas at particular time. So, internet based ride-sharing significantly makes up the need of ride service in typical hotspots and blind spots of taxi service.

The above plots also indicate that hotspot visualization comparison may not always reveal all the differences between two patterns. We should be careful about this fact in our transportation geography studies, and we provide the following analysis for more details.

(3) Spatial-temporal service patterns

Fig. 12(a) shows the hotspot plot of places in which the relative number of pick-ups is larger than that of drop-offs during morning rush hours (e.g. 7:30 A.M. to 9:30 A.M). Fig. 12(b) shows the hotspot plot of places in which the relative number of drop-offs is larger than that of pick-ups during morning rush hours (e.g. 7:30 A.M. to 9:30 A.M).

Accordingly, Fig. 12(c) shows the hotspot plot of places in which the relative number of pick-ups is larger than that of drop-offs during evening rush hours (e.g. 16:30 P.M. to 18:30 P.M), and Fig. 12(d) illustrates the hotspot plot of places in which the relative number of drop-offs is larger.

We can see that, the dominant hotspots for drop-off places during the morning rush hours, which are particular working zones, are almost the same as the dominant hotspots for pick-ups places during the evening rush hours. This indicates that internet based ride-sharing mainly serves as a commuting methods, and people who work in the working zones shown as hotspots, like Central Business Distribute and Zhongguancun, take ridesharing frequently as an important tool of commutes.

We also tested trips served by taxi in the same way, but did not find the obvious pattern as internet based ride-sharing. Limited by the length, we would not provide those plots here.

The above figures we plot demonstrate the most important spatial-temporal pattern in internet-based ride sharing in a visualization way, from which we can see an interesting feature: On the macrocosmic level, internet based ride-sharing mainly serves as an approach for commuting.

To show the above finding in a mathematical way, we apply nonnegative matrix factorization method to detect the basis patterns of the spatial-temporal demand of internet-based ride-sharing and taxi as described in the methodology section.

Fig. 13 shows the factorization results of trips served by taxi through NFM method when choosing the value of λ from 2 to 4. We can find that, in accord with previous studies, when choosing $\lambda = 3$, the result is more reasonable and stable. However, the main purpose for the different trip categories can be explained in a different way compared to description in aforementioned studies (Peng et al., 2012; Wang et al., 2014). The three main collective patterns can be explained as: commuting between home and workplace (illustrated by green line in Fig. 13(b)), workplaces to entertainment venues, or workplaces to homes (illustrated by blue line in Fig. 13(b)), and trips traveling from entertainment venues to homes or between other places (illustrated by red line in Fig. 13(b)).

Accordingly, Fig. 14 shows the factorization results of trips served by internet based ride-sharing through NFM method when choosing the value of λ from 2 to 4. Clearly, this time, it is when choosing $\lambda = 2$ that a stable and appropriate result could be received. This indicates that internet based ride-sharing performs different patterns compared to taxi. Intuitively, the two main collective patterns can be explained as: commuting from home to workplace in the morning (illustrated by blue line in Fig. 14(a)), and commuting from workplace to home in the evening (illustrated by green line in Fig. 14(a)).

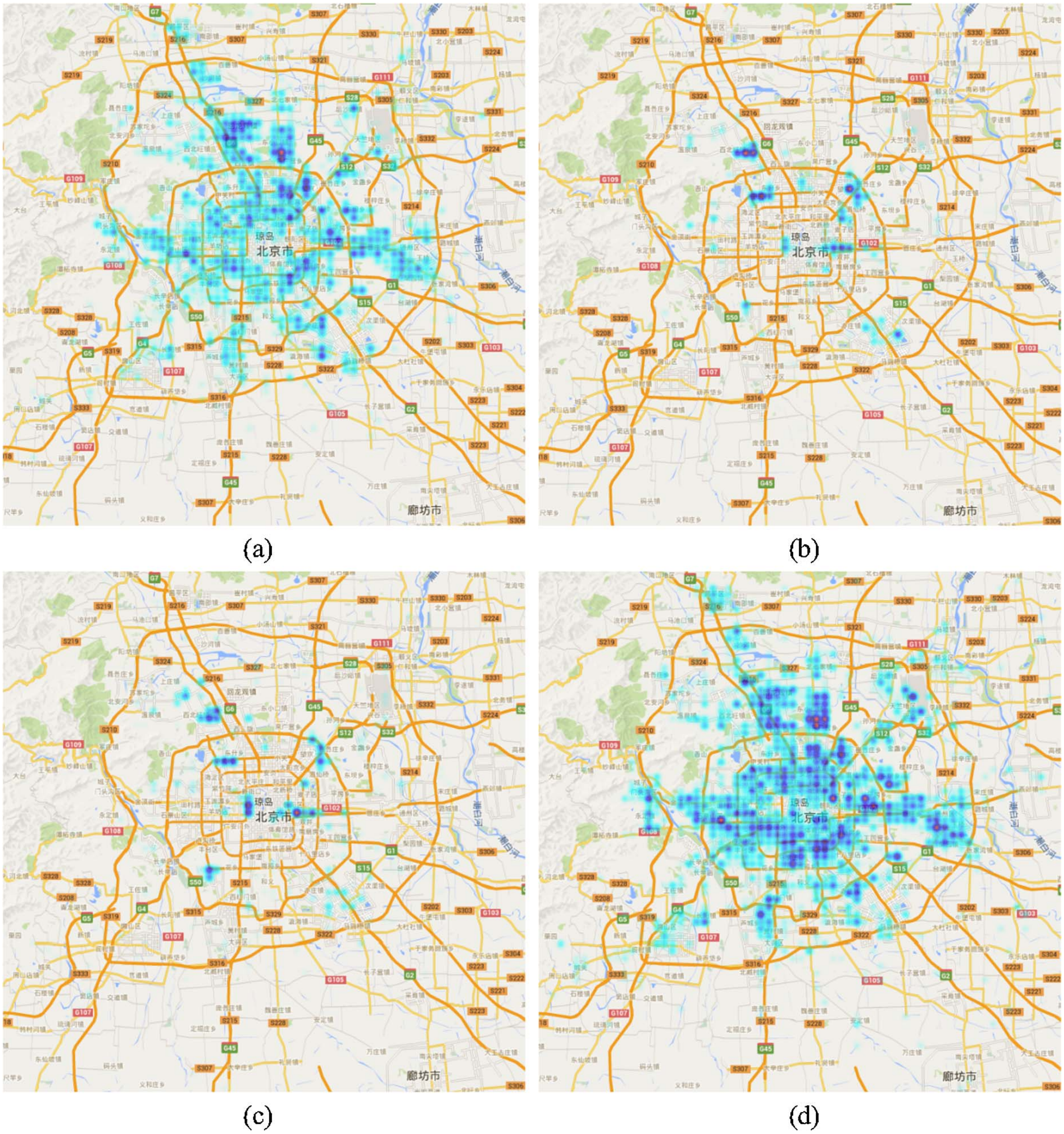
We do explore other factorization methods, including PCA (Jolliffe, 2002), *k*-SVD (Kleibergen and Paap, 2006), only to found that NFM performed the best results in our scenario. Thus, in a mathematical way, we further demonstrate the aforementioned conclusion that on the macrocosmic level, internet based ride-sharing mainly serves as an approach for commuting.

Case study of OD patterns of Central Business District of Beijing

On the basis of the above findings, we provide one case study of hotspot visualizations of the pick-up places or drop-off places of trips terminate in or start from typical workplace, Central Business District of Beijing, for the analysis of the most important patterns of OD pairs in Beijing. Results of the case study again support our conclusion that internet based ride-sharing mainly serves as an approach for commuting.

It is known that OD patterns are very important in urban transportation network. To study the most important OD patterns in Beijing, we provide the hotspot visualizations of pick-up places or drop-off places of trips terminate in or start from Central Business District of Beijing (within 2 km radius of the center of Central Business District of Beijing), shown in Figs. 15 and 16, respectively. We find that for both taxi and internet based ride-sharing, their hotspots of drop-off locations of trips start from Central Business District of Beijing are consistent to the corresponding hotspots of pick-up locations of trips terminate in Central Business District of Beijing. This indicates that passengers who take taxi or internet based ride-sharing from home to work place tend to choose the same method back from work place to home.

Significantly, we find that the hotspots for trips served by taxi are mainly distributed near Central Business District of Beijing or the terminals of Beijing international airport; while hotspots for trips served by internet based ride-sharing are generally located in



(a)

(b)

(c)

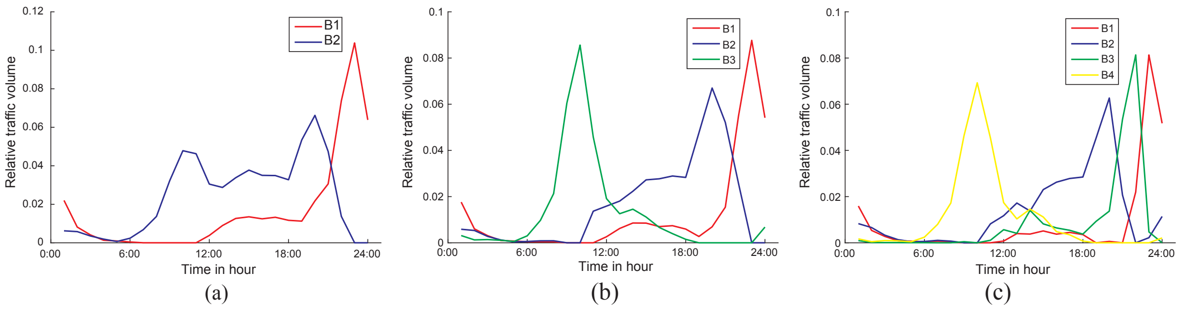
(d)

Fig. 12. Hotspot visualizations of trips served by internet based ride-sharing during morning and evening rush hours by differences between relative number of pick-up places and relative number of drop-off places. In subfigure (a) and (c) the hotspots indicate places where the relative trip numbers of pick-up is dominant to that of drop-off, while in subfigure (b) and (d) the hotspots indicate places where the relative quantity of drop-offs is dominant to that of pick-up.

resident regions, in addition to the hotspots in traffic hubs (e.g. the two terminals of Beijing international airport and the Beijing South Railway Station).

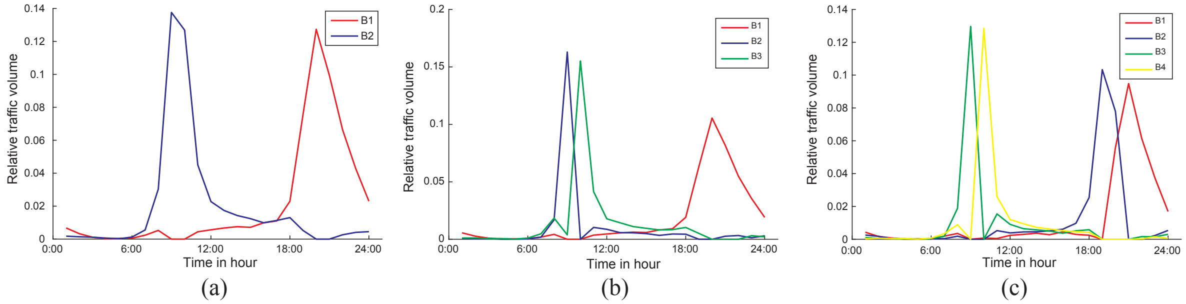
To better illustrate the temporal-spatial patterns of internet based ride-sharing and taxi, and to better understand their service patterns, we provide the hotspot plots as shown in Figs. 17 and 18, with different colors representing different time periods to make up for the insufficient of Figs. 15 and 16. Through this visualization method, we distinguish the important dynamic OD patterns during the morning rush hours and during the evening rush hours for the typical workplace.

In Figs. 17 and 18, different colors of points represent different time periods of trips, where the red points stand for origin or destination of trips during the morning rush hours, the blue points stand for that of trips during the evening rush hours, and the green points stand for that of trips during other time periods of the day. Specially, in Fig. 17, the black points illustrate the origins of every trip, while in Fig. 18, the black points illustrate the destinations of every trip. So, the black points all distribute within 2 km radius of



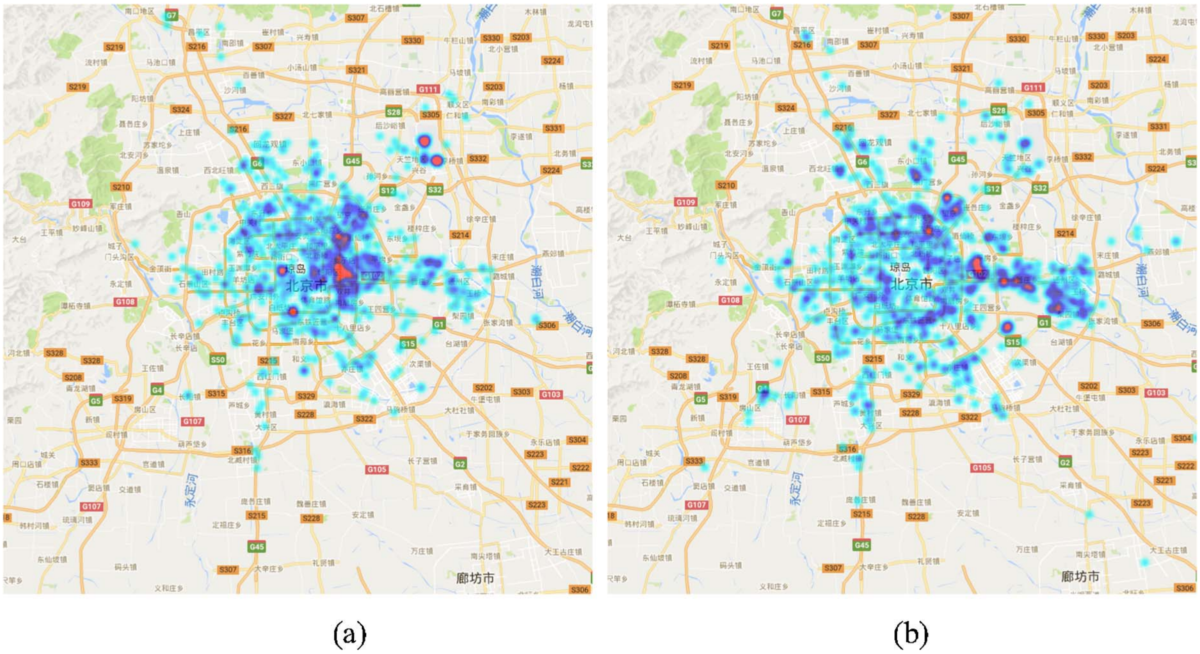
(a) (b) (c)

Fig. 13. The basis collective patterns $\boldsymbol{B}_i$ for trips served by taxi, where in subfigure (a), $\lambda = 2$, in subfigure (b), $\lambda = 3$, and in subfigure (c), $\lambda = 4$.



(a) (b) (c)

Fig. 14. The basis collective patterns $\boldsymbol{B}_i$ for trips served by internet based ride-sharing, where in subfigure (a), $\lambda = 2$, in subfigure (b), $\lambda = 3$, and in subfigure (c), $\lambda = 4$.



(a) (b)

Fig. 15. Hotspot visualizations of drop-off locations of trips start from Central Business District of Beijing, and served by (a) taxi (b) internet based ride-sharing.

Central Business District of Beijing in this case study.

From Fig. 17, we can see that trips start from this place and served by taxi mostly fall in neither morning rush hours nor evening rush hours, and lots of the drop-off places are generally located near Central Business District of Beijing. While trips served by internet based ride-sharing mainly fall in evening rush hours and the drop-off places are generally located in resident regions. For trips terminate in Central Business District of Beijing, as shown in Fig. 18, most of those served by taxi are still short-distance trips and fall in neither morning rush hours nor evening rush hours. However, the majority of trips served by internet based ride-sharing are long-distance trips located in resident regions and generally fall in morning rush hours.

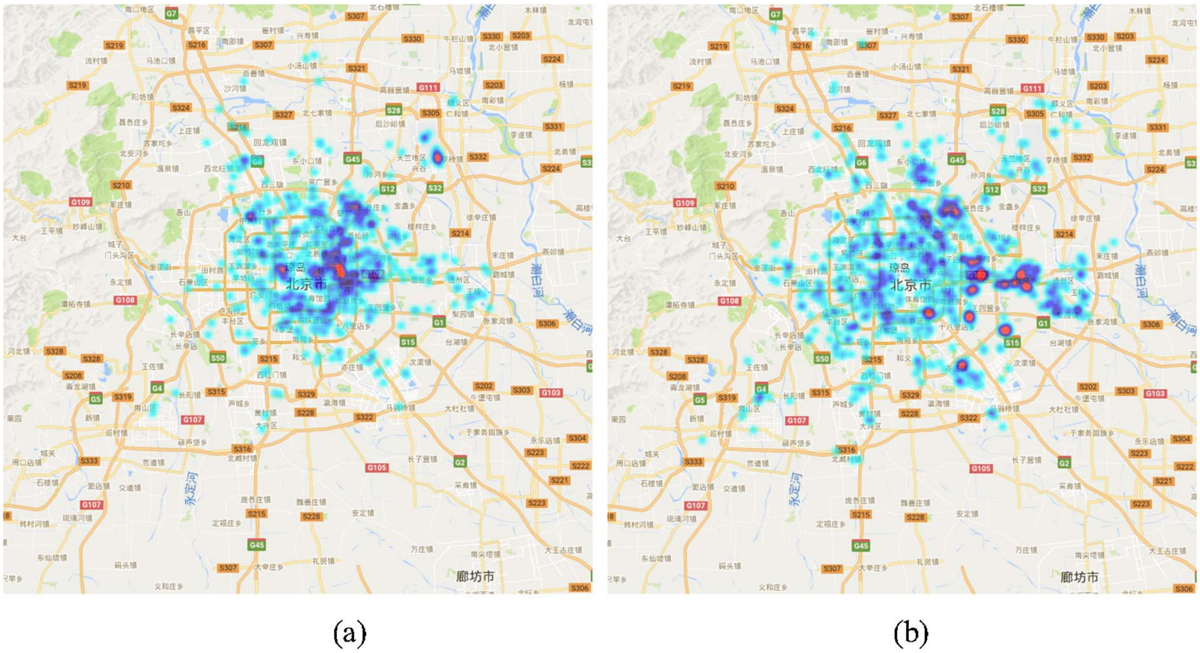


Fig. 16. Hotspot visualizations of pick-up locations of trips terminate in Central Business District of Beijing, and served by (a) taxi (b) internet based ride-sharing.

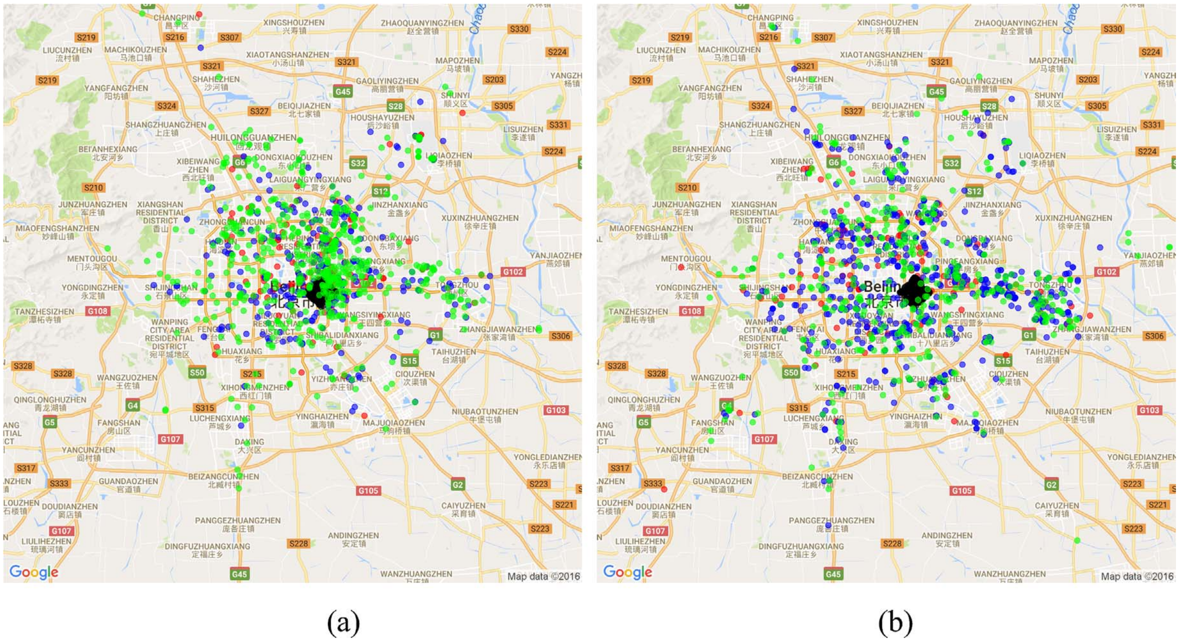


Fig. 17. Hotspot visualizations of drop-off locations of trips start from Central Business District of Beijing, and served by (a) taxi (b) internet based ride-sharing, where different colors illustrate different time periods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

These plots and findings again support our conclusion that internet based ride-sharing mainly serves as an approach for commuting.

(4) Traveling distance patterns

To demonstrate the overall traveling distance patterns, we further compare the traveling distance distribution of trips served by taxi and internet based ride-sharing in Fig. 19. Here, the traveling distance of each trip is approximately calculated as the Euclidean

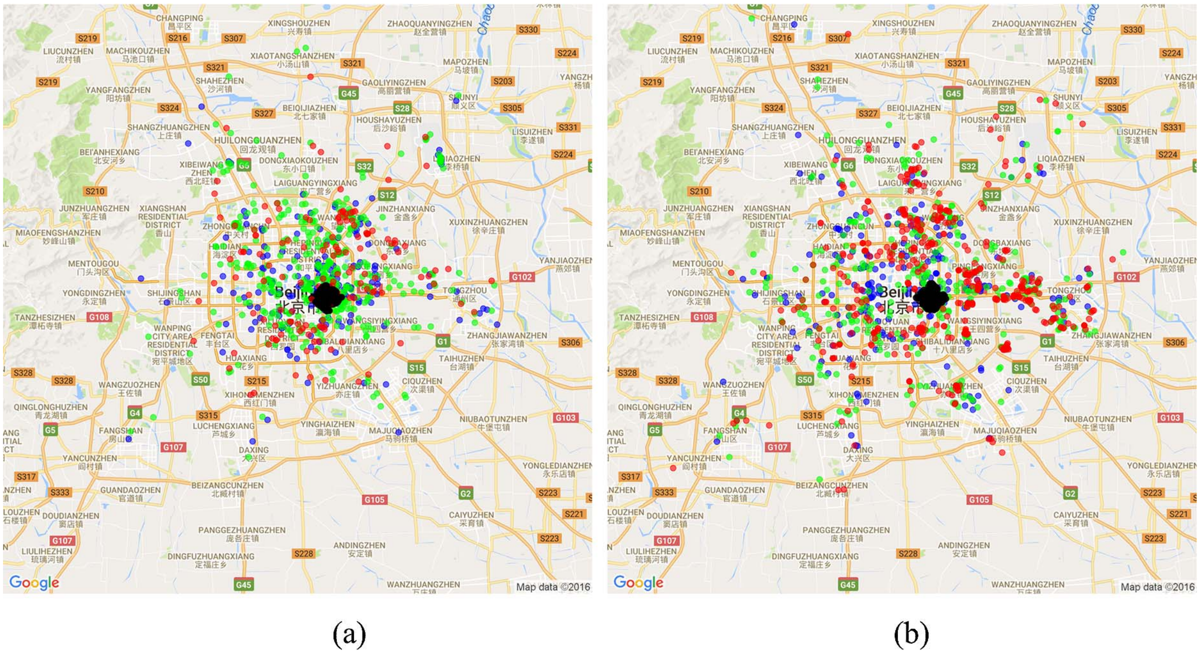


Fig. 18. Hotspot visualizations of pick-up locations of trips terminate in Central Business District of Beijing, and served by (a) taxi (b) internet based ride-sharing, where different colors illustrate different time periods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

or Manhattan distance between the original and destination points. Manhattan distance measure, also known as taxicab geometry, is widely used in urban transportation researches when detailed GPS trajectory data is unobtainable and only origin and destination (OD) are provided, especially for cities whose road network has a layout of chessboard pattern or grid pattern (Dong et al., 2016; Long and Thill, 2015). Since the layout of Beijing city has a well-regulated chessboard pattern, the estimation error of traveling distance we used here can be ignored.

We can see that the number of short-distance trips made by internet based ride-sharing is much smaller than that made by taxi. This is mainly because, in short distance trips, the cost of detouring may not be fully compensated by the fee paid by the passengers. Furthermore, drivers intend to make long distance trips in internet based ride-sharing. There are possibly two reasons: (1) Ride-sharing users pay less per kilometer traveled; (2) As we find above, from the viewpoint of the whole service patterns, internet based ride-sharing mainly serves as a commuting method, and the majority of home-work commuting trips are long-distance ones in Beijing (Pan and Ge, 2014; Long et al., 2015; Long and Thill, 2015), so internet based ride-sharing users seems to make longer distances trips compared to taxi users.

Furthermore, for the dynamic traveling distance patterns, we investigate the changes of average traveling distance with different trip times during one whole day in workdays and weekends/holidays, respectively. The results are demonstrated in Fig. 20. Here, we use payment time to represent trip complete time.

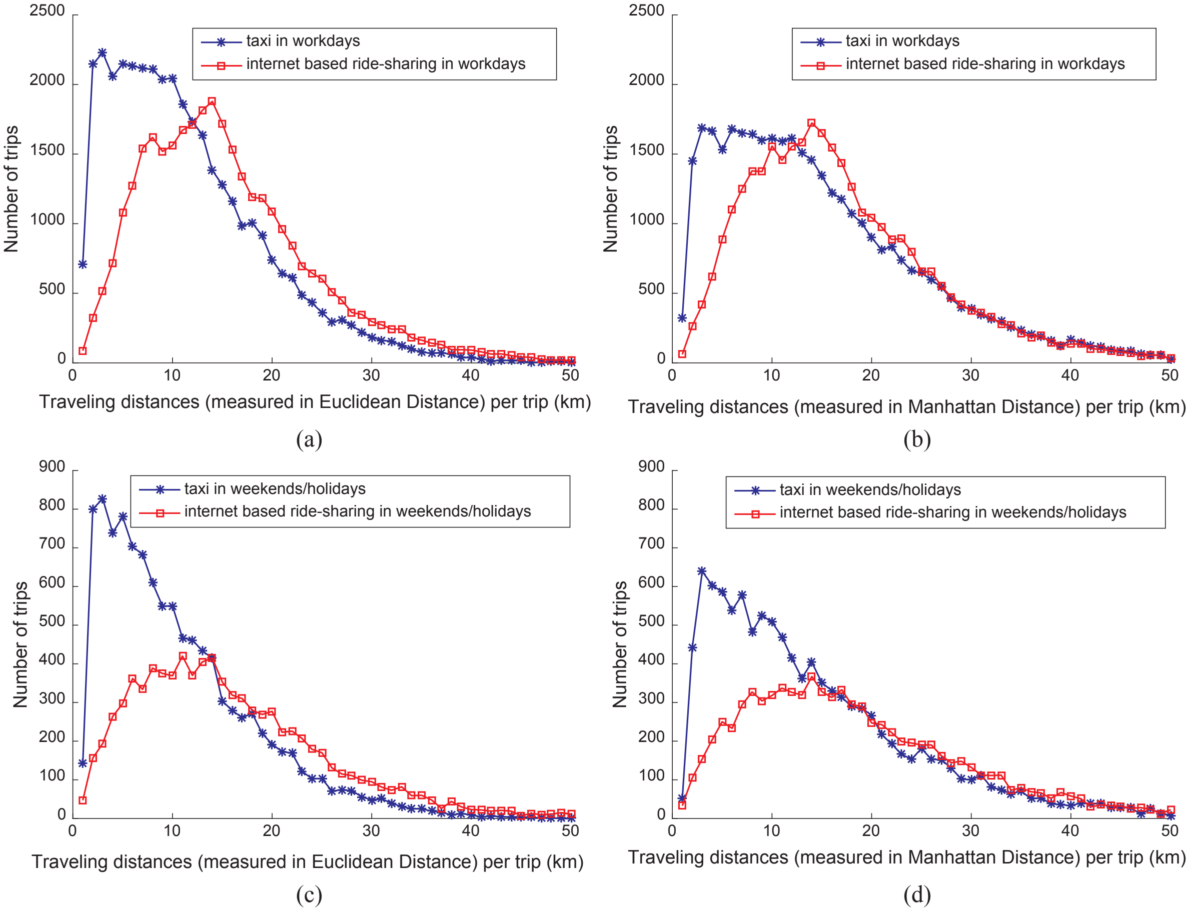
We can see the general trends for changes of average traveling distances with different trip complete times of taxi and internet based ride-sharing are similar. However the average traveling distances of internet based ride-sharing are always longer than that of taxi during the day both in workdays and in weekends/holidays. This agrees with our aforementioned findings of traveling distance patterns.

5.2. Individual behavior patterns of internet based ride-sharing

(1) Results of divisions of drivers based on their commuting styles

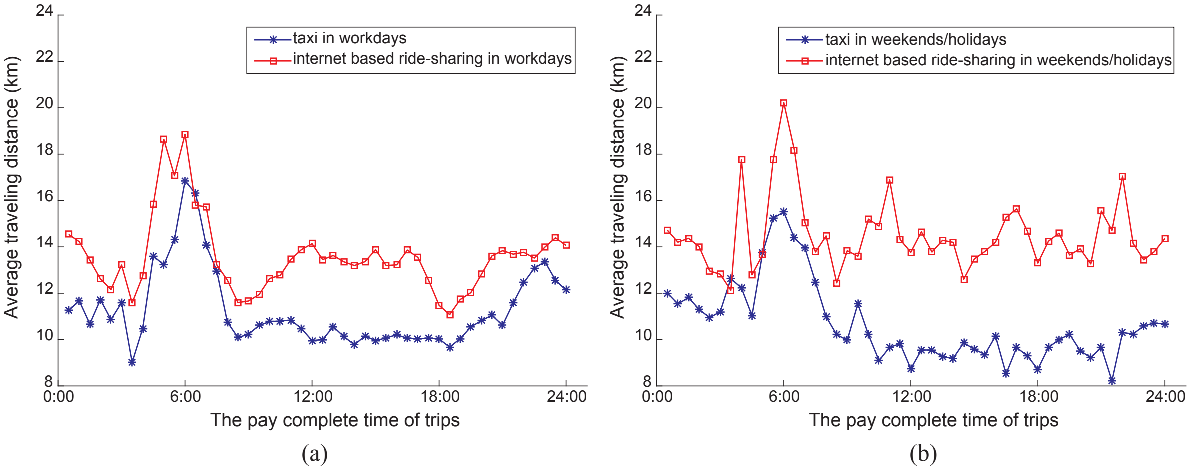
Based on the classifying rules described in Section 3, and after the verification of the classification model and the calibration of parameters described in Section 4, we find that in the selected 1000 drivers: 28 drivers belong to the first kind of drivers and they made 312 (1.77%) trips in either home-to-work or work-to-home commuting; 219 drivers belong to the second kind of drivers and they made 2937 (16.68%) trips in home-work communing.

The total number of home-work commuting drivers is 247, which only accounts for 24.7% of the selected drivers. The total number of home-work commuting trips is 3249, which only accounts for 18.45% of the whole trips that we study here. So, we find that home-work commuting drivers account for only a small part of total drivers and only serve a small number of commuting trips, that is to say, ride-sharing during home-work commuting still contributes little to those who use ride-sharing to commuting between home and work. We expect that the ride-sharing platform companies and the authorities make informed deployment of internet based



(a) (b) (c) (d)

Fig. 19. The traveling distance distribution of trips served by taxi and internet based ride-sharing, in workdays and weekends/holidays, respectively, in (a), (c) Euclidean distance measure, and (b), (d) Manhattan distance measure.



(a) (b)

Fig. 20. The changes of average traveling distances with different trip complete times, of trips served by taxi (indicated by the blue lines) and internet based ride-sharing (indicated by the red lines) in (a) workdays, and (b) weekends/holidays. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ride-sharing service to attract more home-work commuting car owners to provide ride-sharing.

Additionally, as illustrated in Table 1, the proportion of drivers to be classified into home-work commuting group, when they each served less than 30 trips, is larger than that of when these drivers each served no less than 30 trips. When the number of trips served

Table 1
The proportions of drivers to be classified into home-work commuting and random commuting groups, respectively, with different trip amounts served by specific driver.

Trip amounts served by specific driver	The proportion of drivers to be classified into home-work commuting (%)	The proportion of drivers to be classified into random commuting (%)
≥30	21	79
< 30	25	75

by one specific internet based ride-sharing driver is no less than 30, the possibility that the driver being classified into commuting groups would reduce by 4%. Thus, we find that it is more likely to be classified into the home-work commuting group when the quantity of trips served by specific driver is not high. Therefore, limiting the trip numbers served by internet based ride-sharing drivers may be an advisable solution for the authority and on-demand service company managers to restrict the use of converted taxi vehicles, thus making the internet based ride-sharing platforms serve more commuting drivers.

Recently, local authorities in many Chinese cities announced that every ride-sharing driver can only serve twice or thirds per day to restrict the use of converted taxi vehicles. We think that they had also found these patterns and apply it in making the policy.

(2) The results of detour patterns

We calculate the detour distances using methods described in *Section 3*. Results show that the average detour distance of the selected 206 internet based ride-sharing drivers' home-to-work or work-to-home trips is 5.17 km and the average detour proportion is 29.3%. More impressively, as we can see from Fig. 21, the driver-specific detour proportions are almost above 10% with the majority lying in 10–30%, some particular drivers can have a very high detour proportion of 80% or even around 100%.

It was difficult to study the traditional carpooling by data driven methods due to the inaccessibility of corresponding data. Usually, only the average detour distance can be obtained based on surveys or investigations. According to Buliung et al. (2010), traditional carpooling matching partners were usually found within one km radius of their residential locations. So, the detour distance of the surveyed carpooling trips could not be longer than two km.

Therefore, we conclude that internet based ride-sharing drivers' intent to detour further to pick up or drop off passengers than hitchhike drivers, and their driver-specific detour proportions are really high. This falsifies the wrong viewpoint held by some people that ride-sharing is just traditional hitchhiking worked through mobile phone. There are possibly two reasons accounting for the longer detour distance and high driver-specific detour proportions: (1) The cost of detouring can be compensated by the fee paid by the passengers; (2) Under the assurance of on demand car service platform, such as DiDi and Uber, the risk of failing to pick up passengers for internet-based ride-sharing is much lower than that for conventional hitchhike. We will carry out new survey to verify these hypotheses hypothesis in the near future.

Additionally, limited by the length, we are not regarding detour pattern as a key part of our paper. We will study the detour pattern part in close detail in the future when comprehensive data were obtained.

6. Conclusions

In this paper, we compare travel service patterns and driver behavior patterns of taxi and internet based ride-sharing. Many interesting features had been newly found. It is shown that internet based ride-sharing emerges as an effective supplement to

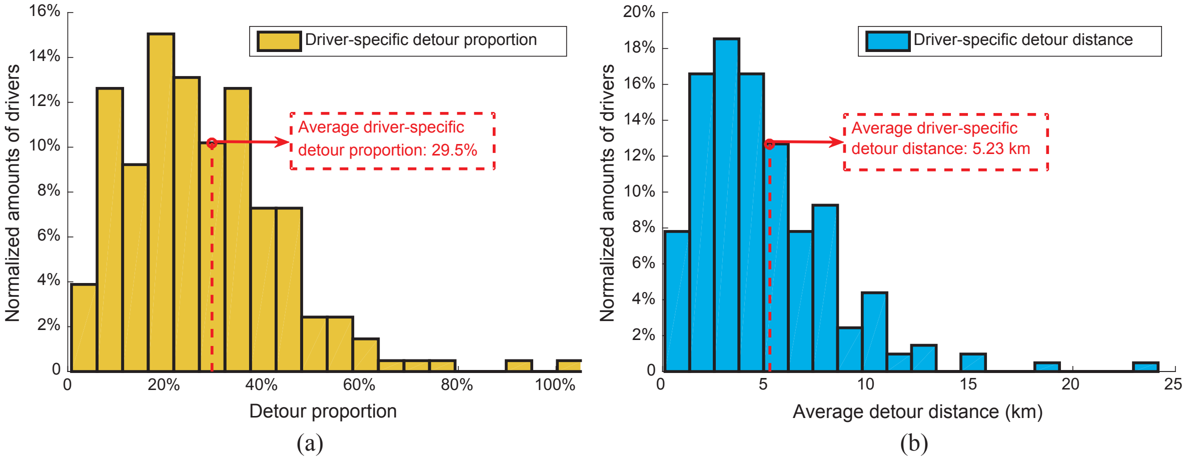


Fig. 21. The empirical distributions of (a) driver-specific detour proportion and (b) driver-specific detour distance.

traditional taxi service making up the spatial and temporal shortage of ride-hailing service, especially in home-to-work or work-to-home commuting during rush hours, and mainly serves as an approach for commuting. By excavating out the basis collective patterns of the macro traffic of internet based ride-sharing and taxi service respectively, we find that internet based ride-sharing, on the macrocosmic level, mainly serves as an approach for commuting. However, the majority of ride-sharing drivers occasionally shares journeys for a very limited times in a month. Thus, internet based ride-sharing drivers do not directly compete with conventional taxi drivers in on-demand ride-hauling service market. More importantly, in this paper, for the first time, we identify two types of internet based ride-sharing drivers based on their different commuting style, i.e., the home-work commuting drivers who behave similarly as conventional ride-sharing drivers and the essentially converted taxi vehicle drivers who do not serve regularly and roams around the city even in working hours. Counterintuitively, we find that home-work commuting drivers account for only a small part of total drivers and they only serve a limited number of commuting trips. Significantly, we find that one driver is more likely to be classified into the home-work commuting group when the quantity of trips served by this specific driver is not high, which accords with the newly introduced policies by local authorities.

It must be pointed out that the number of ride-sharing drivers is continuously increasing. The ride-matching algorithms are still under upgrading to better match drivers and passengers to reduce the detouring or waiting time. The pay rate is also varying from time to time, according the feedbacks of all the users. All these factors may change the patterns of ride-sharing. We will keep consistent attentions in this field and study the possible change in the future.

In summary, our findings indicate that shared mobility and many developing phenomena can be better understood by analyzing the associated big data. We hope that transportation network companies could open more data to researchers and public in this fast changing new era.

Acknowledgement

This work was supported in part by Research Project of Tsinghua University. Many thanks to the Institute of Policy Studies, DiDi company for providing the data in this study.

References

Agatz, N.A., Erera, A.L., Savelsbergh, M.W., Wang, X., 2011. Dynamic ride-sharing: a simulation study in metro Atlanta. Transp. Res. Part B: Method. 45 (9), 1450–1464.

Aïvodji, U.M., Gambs, S., Huguet, M.J., Killijian, M.O., 2016. Meeting points in ridesharing: a privacy-preserving approach. Transp. Res. Part C: Emerg. Technol. 72, 239–253.

Amey, A., Attanucci, J., Mishalani, R., 2011. Real-time ridesharing opportunities and challenges in using mobile phone technology to improve rideshare services. Transp. Res. Rec. J. Transp. Res. Board 2217 (2217), 103–110.

Artigues, C., Deswarte, Y., Guiochet, J., Huguet, M.-J., Killijian, M.-O., Powell, D., Roy, M., Bidan, C., Prigent, N., Anceaume, E., Gambs, S., Guette, G., Hurfin, M., Schettini, F., 2012. Amores: an architecture for mobiquitous resilient systems. In: Proceedings of the 1st European Workshop on AppRoaches to MObiquiTous Resilience, ARMOR '12. ACM, New York, NY, USA, pp. 7:1–7:6. http://doi.acm.org/10.1145/2222436.2222443.

Buliung, R.N., Soltys, K., Bui, R., Habel, C., Lanyon, R., 2010. Catching a ride on the information super-highway: toward an understanding of internet-based carpool formation and use. Transportation 37 (6), 849–873.

Caulfield, B., 2009. Estimating the environmental benefits of ride-sharing: a case study of Dublin. Transp. Res. Part D: Transp. Environ. 14 (7), 527–531.

Chan, N.D., Shaheen, S.A., 2012. Ridesharing in North America: past, present, and future. Transp. Rev. 32 (1), 93–112.

Chen, X., Zahiri, M., Zhang, S., 2017. Understanding ridesplitting behavior of on-demand ride services: an ensemble learning approach. Transp. Res. Part C: Emerg. Technol. 76, 51–70.

Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I., 2009. Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation. John Wiley & Sons.

Clauset, A., Shalizi, C.R., Newman, M.E., 2009. Power-law distributions in empirical data. SIAM Rev. 51 (4), 661–703.

Duda, R.O., Hart, P.E., Stork, D.G., 2010. Pattern Classification, second ed. Wiley.

Dailey, D.J., Loseff, D., Meyers, D., 1999. Seattle smart traveler: dynamic ridematching on the world wide web. Transp. Res. Part C: Emerg. Technol. 7 (1), 17–32.

Deakin, E., Frick, K., Shively, K., 2010. Markets for dynamic ridesharing? Case of Berkeley, California. Transp. Res. Rec. 2187, 131–137.

DiDi, 2016. < http://www.xiaojukeji.com/en/company.html > (accessed June 30, 2016).

Dong, L., Li, R., Zhang, J., Di, Z., 2016. Population-weighted efficiency in transportation networks. Sci. Rep. 6.

Ferguson, E., 1997. The rise and fall of the American carpool: 1970–1990. Transportation 24 (4), 349–376.

Guo, D., Zhu, X., Jin, H., Gao, P., Andris, C., 2012. Discovering spatial patterns in origin-destination mobility data. Trans. GIS 16 (3), 411–429.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. 31, 651–666.

Jolliffe, I., 2002. Principal Component Analysis. John Wiley & Sons.

Kelley, K., 2007. Casual carpooling enhanced. J. Publ. Transp. 10 (4), 119–130.

Kleibergen, F., Paap, R., 2006. Generalized reduced rank tests using the singular value decomposition. J. Economet. 133 (1), 97–126.

Lee, D., Seung, H., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.

Lee, D.D., Seung, H.S., 2001. Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562.

Levofsky, A., Greenberg, A., 2001. Organized dynamic ride sharing: the potential environmental benefits and the opportunity for advancing the concept. In: Transportation Research Board Annual Meeting. No. 01-0577.

Li, B., Krushinsky, D., Van Woensel, T., Reijers, H.A., 2016. The Share-a-Ride problem with stochastic travel times and stochastic delivery locations. Transp. Res. Part C: Emerg. Technol. 67, 95–108.

Lin, C., 2007. Projected gradient methods for nonnegative matrix factorization. Neural Comput. 19, 2756–2779.

Long, Y., Liu, X., Zhou, J., Chai, Y., 2015. Early Birds, Night Owls, and Tireless/Recurring Itinerants: An Exploratory Analysis of Extreme Transit Behaviors in Beijing, China. arXiv preprint arXiv:1502.02056.

Liu, Y., Li, Y., 2017. Pricing scheme design of ridesharing program in morning commute problem. Transp. Res. Part C: Emerg. Technol. 79, 156–177.

Long, Y., Thill, J.C., 2015. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. Comput. Environ. Urban Syst. 53, 19–35.

Minett, P., Pierce, J., 2010. Estimating the energy consumption impact of casual carpooling. In: TRB 89th Annual Meeting Compendium of Papers DVD, Transportation Research Board of the National Academies, Washington, DC [DVD-ROM].

Mitzenmacher, M., 2005. Editorial: the future of power law research. Internet Math. 2 (4), 525–534.

Morency, C., 2007. The ambivalence of ridesharing. Transportation 34 (2), 239–253.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.

Murphy, E., Killen, J.E., 2011. Commuting economy: an alternative approach for assessing regional commuting efficiency. Urban Stud. 48 (6), 1255–1272.

Nie, Y.M., 2017. How can the taxi industry survive the tide of ridesourcing? Evidence from Shenzhen, China. Transp. Res. Part C: Emerg. Technol. 79, 242–256.

Noland, R.B., Cowart, W.A., Fulton, L.M., 2006. Travel demand policies for saving oil during a supply emergency. Energy Policy 34 (17), 2994–3005.

Nourinejad, M., Roorda, M.J., 2016. Agent based model for dynamic ridesharing. Transp. Res. Part C: Emerg. Technol. 64, 117–132.

Pan, H., Ge, Y., 2014. Jobs-housing balance and job accessibility in Beijing. In: Transportation Research Board 93rd Annual Meeting (No. 14-5416).

Peng, C., Jin, X., Wong, K.C., Shi, M., Liò, P., 2012. Collective human mobility pattern from taxi trips in urban area. PLoS ONE 7 (4), e34487.

Sánchez, D., Martínez, S., Domingo-Ferrer, J., 2016. Co-utile P2P ridesharing via decentralization and reputation management. Transp. Res. Part C: Emerg. Technol. 73, 147–166.

Schaller, B., 2017. UNSUSTAINABLE? The Growth of App-Based Ride Services and Traffic, Travel and the Future of New York City. Available at: <http://schallerconsult.com/rideservices/unsustainable.htm>.

Shaheen, S., Cohen, A., Zohdy, I., 2016. Shared Mobility: Current Practices and Guiding Principles. U.S. Department of Transportation, Federal Highway Administration. Report No. FHWA-HOP-16-022.

Shih, G., 2015. China taxi apps DiDi Dache and Kuaidi Dache announce $6 billion tie-up, Reuters February 14, 2015. Available at: < https://www.yahoo.com/tech/china-taxi-apps-didi-dache-kuaidi-dache-announce-023235253-finance.html > .

Townsend, A., Trefethen, L.N., 2015. Continuous analogues of matrix factorizations. Proc. R. Soc. A 471 (2173), 20140585.

Udell, M., Horn, C., Zadeh, R., Boyd, S., 2016. Generalized low rank models. Found. Trends® Machine Learning 9 (1), 1–118.

Wang, J., Gao, F., Cui, P., Li, C., Xiong, Z., 2014. Discovering urban spatio-temporal structure from time-evolving traffic networks. In: Asia-Pacific Web Conference. Springer International Publishing, pp. 93–104 September.

Yang, F., Jin, P.J., Cheng, Y., Zhang, J., Ran, B., 2015. Origin-destination estimation for non-commuting trips using location-based social networking data. Int. J. Sustain. Transp. 9, 551–564.

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.