Transportation Research Part C 96 (2018) 251–269

Contents lists available at ScienceDirect

Transportation Research Part C

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

# Modeling real-time human mobility based on mobile phone and transportation data fusion

Zhiren Huang[a], Ximan Ling[a], Pu Wang[a,*], Fan Zhang[b], Yingping Mao[c], Tao Lin[c], Fei-Yue Wang[d]

[a] School of Traffic and Transportation Engineering, Central South University, Changsha, Hunan 410000, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China
[c] Shenzhen Urban Transport Planning Center, Shenzhen, Guangdong 518021, China
[d] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

*Keywords:*
Human mobility
Travel demand estimation
Big data
Data fusion

ABSTRACT

Even though a variety of human mobility models have been recently developed, models that can capture real-time human mobility of urban populations in a sustainable and economical manner are still lacking. Here, we propose a novel human mobility model that combines the advantages of mobile phone signaling data (i.e., comprehensive penetration in a population) and urban transportation data (i.e., continuous collection and high accuracy). Using the proposed human mobility model, travel demands during each 1-h time window were estimated for the city of Shenzhen, China. Significantly, the estimated travel demands not only preserved the distribution of travel demands, but also captured real-time bursts of mobility fluxes during large crowding events. Finally, based on the proposed human mobility model, a predictive model is deployed to predict crowd gatherings that usually cause severe traffic jams.

## 1. Introduction

Data-driven human mobility modeling has experienced rapid developments in recent years. Since the first data-driven human mobility study was conducted using the displacement records of 5 million U.S. dollar bills (Brockmann et al., 2006), several pioneering works have been produced. Given that the trajectory of a dollar bill was not generated by the same individual, call details records (CDR) data, which provide trajectories of millions of mobile phone users over a long period (i.e., several months to several years), have been widely used in human mobility studies. González et al. (2008) employed the CDR data of 100,000 mobile phone users to explore the universal laws of individual human mobility. Schneider et al. (2013) analyzed the CDR data of five countries to discover the motifs of human mobility. Lu et al. (2012) analyzed the human mobility patterns after an earthquake in Haiti. Song et al. (2010a) discovered that individual human mobility has a surprisingly high predictability, which is independent of an individual's age, gender, and traveling distance. Ahas et al. (2010a) extracted and analyzed the home and work anchor points of mobile phone users. Järv et al. (2014, 2017) analyzed individual monthly spatial travel behavior and proposed a generic conceptual framework to enhance the spatial accuracy of CDR data. Human mobility models based on principal component analysis (PCA) (Eagle et al., 2009) or universal human mobility laws (González et al., 2008), the explore-return model (Song et al., 2010b; Pappalardo et al., 2015), radiation models (Simini et al., 2012; Ren et al., 2014) and diverse scale model (Yan et al., 2017) have been proposed and received widespread attention from researchers in both the scientific and engineering fields (Semanjski and Gautama, 2016; Chen et al., 2016).

* Corresponding author.
*E-mail address:* wangpu@csu.edu.cn (P. Wang).

https://doi.org/10.1016/j.trc.2018.09.016
Received 13 December 2017; Received in revised form 14 September 2018; Accepted 16 September 2018
Available online 03 October 2018
0968-090X/ © 2018 Elsevier Ltd. All rights reserved.

In recent years, CDR data were also used in modeling individual activity patterns (Cheng et al., 2013; Duan et al., 2017), dividing traffic zones (Dong et al., 2015) and identifying crowd flow transportation hub (Zhong et al., 2017).

The low collection cost and high penetration rate of CDR data make them a good candidate for estimating travel demands. This has greatly changed the situation in which the estimation of origin–destination (OD) matrices were mostly relying on high-cost and hard-to-update travel surveys (Nitsche et al., 2014). Zhang et al. (2010) applied a statistical approach to extract daily travel demands from CDR data. Wang et al. (2012, 2014) combined CDR data, census data, and vehicle usage ratio data to estimate travel demands in two US urban areas. Alexander et al. (2015) extracted OD matrices through inferring the purpose of user trip. Toole et al. (2015) combined CDR data and route information collected from a crowdsourcing platform to estimate travel demand. To more accurately estimate OD matrices, some studies combined mobile phone data with traffic count data (Iqbal et al., 2014), traffic sensor data (Wu et al., 2015), location-based service data (SA et al., 2015) or sighting data (Wang and Chen, 2018).

One disadvantage of CDR data is that the location of a mobile phone user is recorded only when the user calls or sends a message, which means that a user's whereabouts are unknown when the user did not use his or her phone (Wang et al., 2012). The sparse and irregular records of CDR data prevent their application in real-time travel demand estimations. Different from CDR data, user location is recorded regularly (every hour) in mobile phone signaling data, independent of whether a mobile phone user is using his/her phone or not. Therefore, the mobility information extracted from mobile phone signaling data is usually more credible. Several pioneering works have been produced based on mobile phone signaling data covering a small group of users. González et al. (2008) employed mobile phone signaling data that recorded hourly location of 1000 mobile phone users to validate the human mobility laws obtained from the much larger CDR dataset. Ahas et al. (2010b) studied the daily rhythms of the movements of suburban commuters based on the positioning data of 277 mobile phone users. Mobile phone signaling data covering large population are characterized by a short period of data collection. To our knowledge, the durations of existing mobile phone signaling data are all less than three weeks (González et al., 2008; Ahas et al. 2010b; Duan et al., 2017; Zhong et al., 2017; Ni et al., 2018). In practice, real-time mobility information cannot be obtained from mobile phone signaling data.

Urban transportation data usually record accurate time and location information of individual trips. For example, vehicle global-positioning-system (GPS) data record the coordinate of a vehicle (taxi, bus, or private car) every few seconds; smartcard data record the time and the station when a passenger enters or exits a subway station, or the time and the bus number when a passenger boards a bus. Urban transportation data have been also widely used in human mobility studies. Giannotti et al. (2011) used GPS trajectories of 15,000 automobile users to investigate the recurrent patterns of human mobility. Tang et al. (2016) studied urban human mobility using large-scale taxi GPS data. Leng et al. (2016) investigated the mobility patterns of taxi passengers. Ma et al. (2018) designed a genetic algorithm to optimize the taxi carpooling path. Briand et al. (2016) proposed a generative mixture model to partition smartcard users into different clusters. After that Briand et al. (2017) used the clustering method to analyze the long-period smartcard data in Gatineau and Ottawa, Canada. The authors found that the usage of public transport has relative stability. He et al. (2016) used smartcard data to analyze the mobility of 5 million subway passengers in Beijing, China. Ma et al. (2013) and Zhao et al. (2017) analyzed mobility patterns of bus passengers in Beijing and Shenzhen, China. Despite that urban transportation data are characterized with real-time, and high-accuracy features, data for automobile users, walkers, and destination information of bus passengers are not available for most cities.

In this study, we developed a novel human mobility model based on one-day mobile phone signaling data and three-month subway smartcard data and taxi GPS data of Shenzhen, a major city in China (Fig. 1a). The estimated mobility flows preserve the distribution of urban travel demands and capture real-time bursts of mobility fluxes during large crowding events. The paper is organized as follows. In Section 2, we introduce the mobile phone signaling data and urban transportation data used in the present study. In Section 3, we develop the human mobility model based on the fusion of multi-source human mobility data. In Section 4, we validate the human mobility model from two different perspectives, and, in Section 5, we propose an application of the proposed human mobility model by developing a crowd gathering predictive model. Section 6 discusses our contribution to the existing literature and elaborates the benefits of combing multi-source human mobility data. Section 7 concludes with a summary of methods and findings and discusses the limitation of the present work and possible future directions.

## 2. Data

High-penetration and large-scale mobile phone signaling data, subway smartcard data, and taxi GPS data of Shenzhen were used in the present study. Before introducing the data properties and data processing approaches, we clarify the assumptions used in our analysis and modeling:

Assumption 1: the distribution of mobility flows in Shenzhen is represented by the distribution of mobility flows of the mobile phone users in our data.
Assumption 2: the distribution of smartcard passenger trips is used to represent the distribution of all subway trips, which is a common way to use smartcard data when the data for ticket passengers are not available (Ma et al., 2013; Zhao et al., 2017).
Assumption 3: the number of taxi trips is converted to the number of taxi passenger trips using the estimated average number of passengers per trip, which is a common way to deal with this problem (Yuan et al., 2013).

### 2.1. Mobile phone signaling data

The employed mobile phone signaling data were collected in an ordinary weekday of 2012 using the second generation (2G)

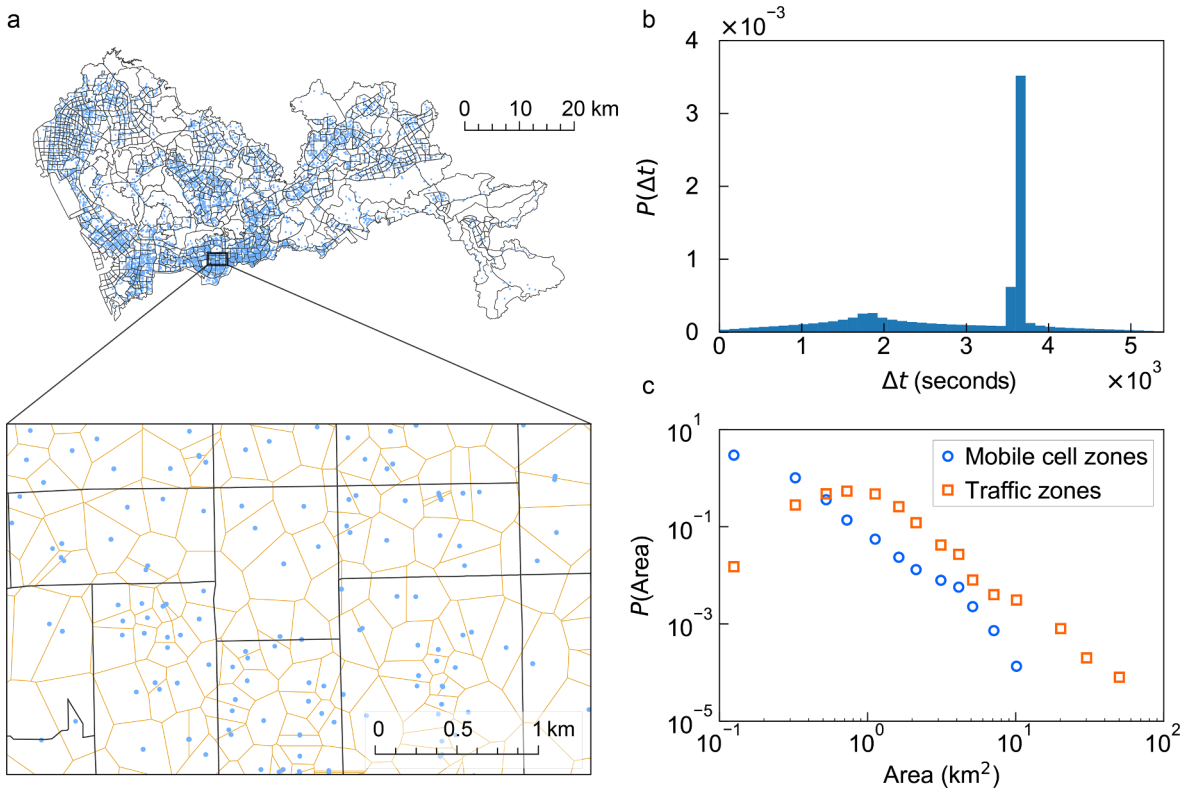


**Fig. 1.** (a) Spatial distribution of mobile phone towers, mobile cells and traffic zones. (b) Time interval between consecutive records of a mobile phone user. (c) The areas of traffic zones (red squares). The average, minimum and maximum areas of traffic zones are 2.305 $km^2$, 0.159 $km^2$ and 68.649 $km^2$ respectively. The areas of mobile cells (blue circles). The average, minimum and maximum areas of mobile cells are 0.387 $km^2$, 0.001 $km^2$ and 17.52 $km^2$ respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cellular network of one large mobile operator in China. The market share of the operator was $m = 66.7\%$ in 2012. In the mobile phone signaling data used, user location is recorded regularly in time, independent of whether the user is using his/her phone or not. There are small variations in recording frequencies (Fig. 1b). The average time and the median time between two consecutive records of a user are 2996 s and 3601 s respectively, the standard deviation of the time intervals is 1132 s.

In the mobile phone signaling data used, the location of a user is represented by the coordinates of the mobile phone tower assigned to provide the service. There are 5952 mobile phone towers in the 2G cellular network, which belongs to the operator providing the data. The cell of each mobile phone tower was estimated using Voronoi polygons (Fig. 1a and c). To estimate mobility flows of mobile phone users, traffic zone is used as the statistical unit. The geographic information systems (GIS) data of traffic zones were provided by the Shenzhen Transportation Authority. In the data, the geographic area of Shenzhen was partitioned into 996 traffic zones based on the road network and land functionalities (Fig. 1a and c).

It was reported that 82.6% of Chinese people use mobile phones in 2012 (China Daily, 2013). Shenzhen is one most developed city in one most developed province (Guangdong) of China. A consensus is that the higher the GDP per capita the higher the penetration rate of mobile phones. Hence, we can use the penetration rate 82.6% (for the whole country) as the lower bound of the penetration rate for Shenzhen. Multiplying the penetration rate of mobile phones by the market share of the operator ($m = 66.7\%$) gives that the used mobile phone data at least covered 55.1% of the Shenzhen population.

The employed mobile phone signaling data were cleaned and processed as follows:

Step 1: We removed duplicate records with the same user ID, timestamp and coordinate, and only one copy of the duplicate records was used. In this process, 328,373,094 records were selected from the original 543,024,013 records.

Step 2: In the selected 328,373,094 data records, 16,251,644 unique user IDs were identified. Given that electric meters (for generating electricity bills) also use the subscriber identity module (SIM) card provided by the mobile phone operator, we only used the data of users with more than one tower location recorded. After filtering out single-location user IDs, 14,233,961 unique IDs and 302,128,043 records remained.

Step 3: Following the method proposed in (Schneider et al., 2013), if a user was continuously found to jump back and forth between two neighboring mobile phone towers for more than three times, we regard the jumps as cell hopping. For example,

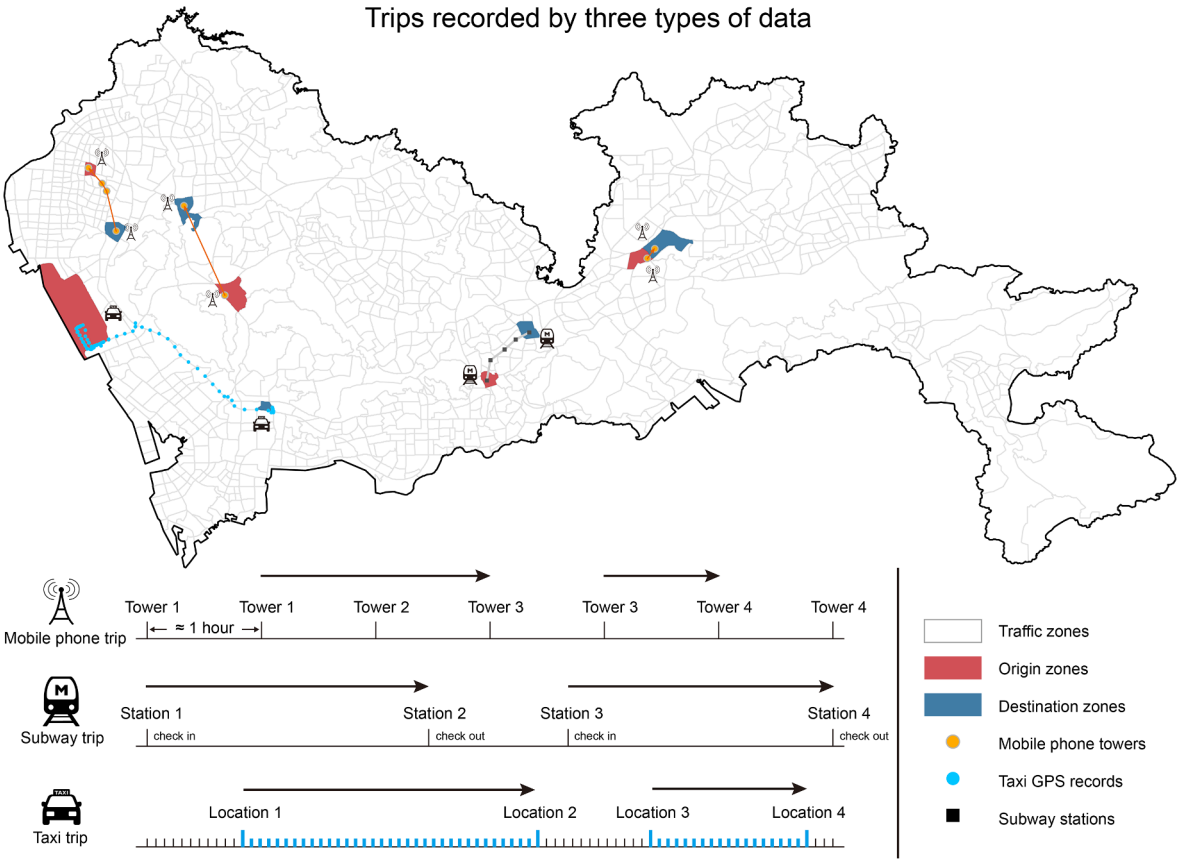


**Fig. 2.** Trips extracted from the mobile phone signaling data, the subway smartcard data, and the taxi GPS data. A mobile phone user trip can be constituted by more than one user movement.

when finding three consecutive transitions between two neighboring towers A and B (e.g., A-B-A-B-A-B), the user location observed during these jumps are all assigned to the origin of the first jump (e.g., A-A-A-A-A-A). Using this method, 1,842,377 (0.61%) cell hopping records were found from the overall 302,128,043 records.

Step 4: Given that there are much more towers than traffic zones, if a user is recorded at a mobile phone tower, we simply map the user to the traffic zone where the tower is resided in (Fig. 1a). One day was split into 24-time windows. For each hourly time window $t$, we pinpointed the traffic zone that each mobile phone user resided in. We follow a common way to extract trips from mobile phone data (Schneider et al., 2013; Alexander et al., 2015; Toole et al., 2015). If the user was found in two different traffic zones $z_1$ and $z_2$ in two consecutive records, we defined that the user has a movement $z_1 \rightarrow z_2$. Totally 36,433,440 user movements are identified using this way. Next, we checked whether the user kept moving until the user was found in the same zone $z_n$ for two consecutive records. We defined that the user had a trip from the origination zone $z_1$ to the destination zone $z_n$ (Fig. 2). Roughly 61% of the trips were constituted with a single movement, 18% of the trips were constituted with two consecutive movements, and 21% of the trips were constituted with more than two consecutive movements. The durations of 61.8% of the trips are under 1 h and 10 min. Because some single movements are slightly longer than 1 h. In this step, 25,906,745 trips are identified.

Step 5: According to assumption 1, the number of trips between two traffic zones $i, j$ during time window $t$ are scaled up by a factor $\gamma = 1/m = 1.5$ to quantify the mobility flows contributed by mobile phone users from other operators. The generated user mobility flows $T_M(i, j, t)$ were used as the ground truth mobility flows to calibrate the mobility flows of subway and taxi passengers.

Schoolchildren and elders who do not use mobile phones are underrepresented in the present approach, because their locations are not recorded in mobile phone data. The lack of mobility information for schoolchildren and elders is a general problem in transportation research. From traditional travel surveys to emerging GPS trajectories of volunteers and mobile phone data, the mobility information for this group of people is lacking (Stopher and Greaves, 2007). We conservatively choose to not consider the mobility of this small group of residents in the present study rather than performing upscaling calibrations without necessary supporting data. In addition, trips within the same mobile network cell and some very short trips may not be captured by the mobile phone data.

### 2.2. Urban transportation data

Two sets of widely available urban transportation data, subway smartcard data, and taxi GPS data were used to obtain real-time information of human mobility. Here, the term "real-time" means that the two sets of urban transportation data can be viewed as nearly real-time data using current infrastructure. According to JT/T 905.1-2014, the GPS coordinates of taxies are typically reported to the transportation data center with a delay of only 1 min. According to GB/T 20907-2007, subway smartcard data are sent to the subway data center every 15 min. Subway smartcard data or taxi GPS data are missing for a few hours or the whole day for 23 days during the three-month observation period. To avoid introducing bias into the estimation of mobility flows, only data collected in the remaining 69 days were used in this study.

In the subway smartcard data, the time, card ID, subway line ID, and subway station ID were recorded each time a passenger entered or exited a subway station (Fig. 2). During the three-month data collection period from October 1, 2014 to December 31, 2014, the subway smartcard data recorded a total of 225.9 million boarding records for 3.2 million subway passengers. The number of subway passenger trips, $N_{\text{sub}}(i, j, t)$, generated during each 1-hour time window $t$ for the $i$th origination zone and the $j$th destination zone were determined. In the subway smartcard data, the average number of daily trip records was 1.62 million. Given that the number of subway trips per day was approximately 2.84 million (Shenzhen Transportation Committee, 2015), the subway smartcard data recorded roughly 57% of all the subway trips. According to assumption 2, an up-scaling factor $\beta_{\text{sub}} = 1.76$ was obtained, and the mobility flows of subway passengers were estimated as: $T_{\text{sub}}(i, j, t) = \beta_{\text{sub}} \times N_{\text{sub}}(i, j, t)$

The taxi GPS data were also collected during the three-month data collection period (October 1, 2014 to December 31, 2014). The GPS receiver installed in each taxi in Shenzhen recorded the coordinates of the taxi roughly every 20 s. For each taxi, the service states are recorded. The origin, destination, and route of a taxi trip can be obtained (Fig. 2). The taxi GPS data were cleaned and processed as follows:

Step 1: Only GPS coordinate records located within Shenzhen were used. In this step, 8,712,104 (0.2%) records are removed from the original 4,415,684,424 coordinate records produced by 15,275 taxies.

Step 2: The total distance of a taxi trajectory $D$ is calculated using the Haversine function:

$$D = \sum_{i=1}^{n-1} \text{Haversine}(v_i \cdot \text{long}, v_i \cdot \text{lat}, v_{i+1} \cdot \text{long}, v_{i+1} \cdot \text{lat}) \tag{1}$$

where $v_i$ is a GPS coordinate of trajectory $D$, and $n$ is the number of GPS coordinates in trajectory $D$. The Haversine function calculates the shortest distance between two coordinates on the surface of the earth. The library "pyproj" (https://jswhit.github.io/pyproj/) is used in the calculation. Next, we calculated the average speed $\bar{v}$ of each taxi during a day based on the time and coordinate information of the taxi. We excluded the taxies with average speed $\bar{v} > 120$ km/h or $\bar{v} < 3$ km/h. After this step, 13,731 taxies are used in following analysis, and roughly 5% coordinate records are filtered out.

Step 3: Taxi trips lasted less than 60 s (2,622,712 trips in total) and taxi trips lasted more than 3 h (404,988 trips in total) are not used. Finally, 43,471,766 taxi trips are used to generate mobility flows of taxi passengers.

The number of taxi trips, $N_{\text{taxi}}(i, j, t)$, generated during each 1-hour time window $t$ for the $i$th origination zone and the $j$th destination zone were determined. In the taxi GPS data, we estimated 0.43 million trip records per day. Given that taxies transported an average of 1.2 million passengers per day (Shenzhen Transportation Committee, 2015), the average number of passengers per taxi trip was estimated to be $\beta_{\text{taxi}} = 2.774$. According to assumption 3, the mobility flows of taxi passengers were estimated as: $T_{\text{taxi}}(i, j, t) = \beta_{\text{taxi}} \times N_{\text{taxi}}(i, j, t)$.

Mobility flows of taxi and subway passengers are estimated for each pair of traffic zones during each 1-hour time window:

$$T(i, j, t) = T_{\text{sub}}(i, j, t) + T_{\text{taxi}}(i, j, t). \tag{2}$$

Mobility flows $T(i, j, t)$ are used as sensors of real-time human mobility.

## 3. Model

This study aims to uncover the relation between average mobility flow $\langle T(i, j, t)\rangle$ of subway and taxi passengers and the user mobility flow $T_M(i, j, t)$, and develop a human mobility model that can capture the real-time human mobility of the city.

### 3.1. Sampling rates of transportation data for different OD pairs

We defined a scaling factor $\beta(i, j, t) = \langle T(i, j, t)\rangle / T_M(i, j, t)$ for each pair of origin zone and destination zone during each 1-hour time window from 05:00 to 24:00. The motivation to calculate the scaling factor $\beta$ is to account for mobility flows contributed by other transportation modes. Clearly, scaling factor $\beta$ varies for different origin–destination (OD) pairs, because transportation mode shares vary for different OD pairs. For example, scaling factor $\beta(i, j, t)$ is usually large in downtown area where subway and taxi services are plentiful, while $\beta(i, j, t)$ is usually small between suburban traffic zones due to the lack of subway and taxi service. Taking time window 09:00–10:00 as an example, we analyzed the mobility flow patterns for OD pairs with two ranges of scaling

**Table 1**
Percentages of different ranges of scaling factor $\beta(i, j, t)$ (averaged over all time windows).

| $\beta(i, j, t)$ | 0 | (0, 0.05) | [0.05,1] | > 1 | N/A |
|---|---|---|---|---|---|
| OD pair percentage (%) | 53.88 | 2.22 | 19.67 | 8.31 | 15.92 |
| $T$ percentage (%) | 0 | 4.31 | 41.54 | 38.54 | 15.61 |
| $T_M$ percentage (%) | 26.38 | 41.17 | 29.36 | 3.09 | 0 |

factors $\beta(i, j, t) < \delta$ and $\beta(i, j, t) \geq \delta$. Here, the threshold $\delta = 0.05$ was selected based on the analysis in Section IV.

### 3.1.1. Phase I: For OD pairs with scaling factor $\beta(i, j, t) < 0.05$

Data of automobile users, bus passengers, walkers, and other modes of transportation were not available and used in this study. This means that in most cases, the average passenger mobility flow $\langle T(i, j, t)\rangle$ was smaller than the user mobility flow $T_M(i, j, t)$. Indeed, for roughly 54% of the detected OD pairs, no taxi or subway passenger trips were observed during the three-month data collection ($\beta(i, j, t) = 0$ as shown in Table 1). These OD pairs contain roughly 26% of all trips obtained from the mobile phone signaling data. In Fig. 3a, each OD pair with $0 < \beta(i, j, t) < 0.05$ is represented by a link connecting the origin zone and destination zone, whereas the mobility flow volume $T_M(i, j, t)$ between the two zones is illustrated by the width of the link. This group of links are usually distributed between two suburban zones or traffic zones that are close to each other. The small $\beta$ between suburban zones can be explained by the lack of subway and taxi service, whereas the small $\beta$ between neighboring zones can be attributed to trips using the walk mode. Fig. 3a highlights the low penetration of transportation data between some areas of the city, which also reveals the regions in which public transportation services are possibly weak and need improvement.

### 3.1.2. Phase II: For OD pairs with scaling factor $\beta(i, j, t) \geq 0.05$

For OD pairs with scaling factors $0.05 \leq \beta(i, j, t) \leq 1$, mobility flows of subway and taxi passengers represent a moderate portion of overall mobility flows (Fig. 3b). Surprisingly, the average passenger mobility flows $< T(i, j, t) >$ sometimes are even larger than the user mobility flows $T_M(i, j, t)$, and these large $\beta(i, j, t)$ are usually found between two far-away traffic zones with a subway connection (Fig. 3c). An explanation is as follows. Passengers entering a subway station may not have originated from the zone in which the station is located since subway stations can attract travelers from neighboring zones, especially in suburban areas. Passengers may first use other transportation modes before using the subway. In a similar way, passengers exiting a subway station may not stay in the zone in which the station is located; they may arrive at their final destination using other modes of transportation. Fig. 3c also reveals that mobility flows may be overestimated for some specific pairs of zones when only using transportation data.

Fig. 3d shows the trip distances of the three groups of OD pairs with $0 < \beta(i, j, t) < 0.05$, $0.05 \leq \beta(i, j, t) \leq 1$, and $\beta(i, j, t) > 1$ respectively. Trip distances of OD pairs with $\beta(i, j, t) < 0.05$ are mostly smaller than 3 km, whereas OD pairs with $\beta(i, j, t) > 1$ are dominated by long-distance trips.

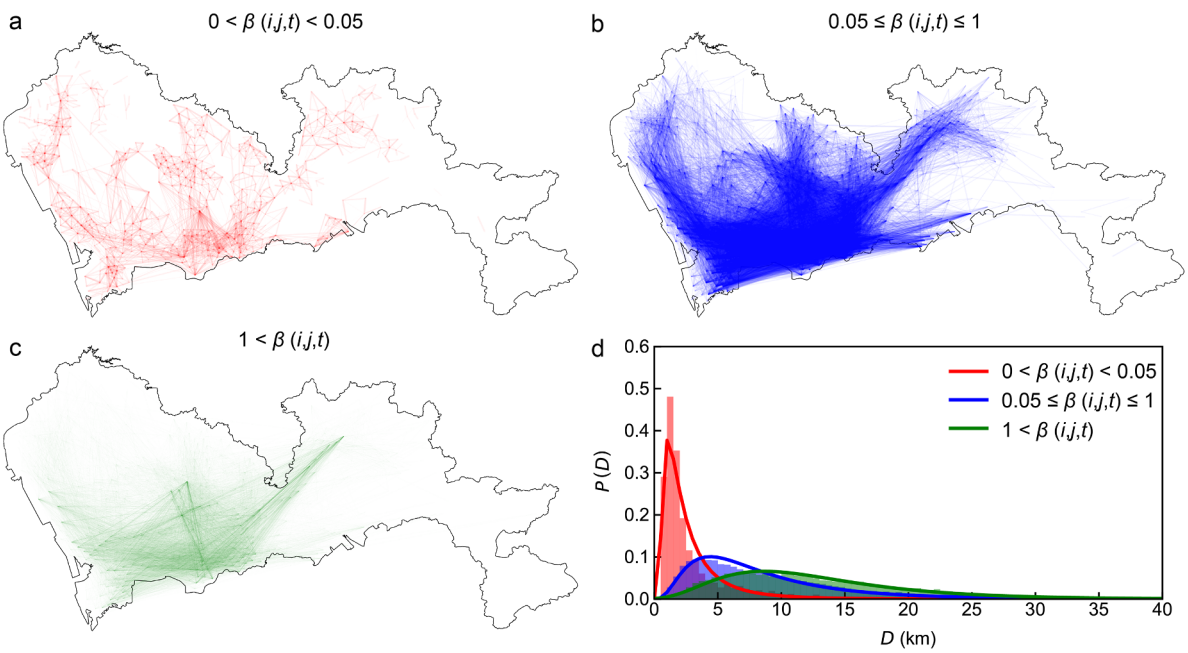


**Fig. 3.** Spatial distributions and trip distances of OD pairs with scaling factors $0 < \beta(i, j, t) < 0.05$, $0.05 \leq \beta(i, j, t) \leq 1$, and $\beta(i, j, t) > 1$.

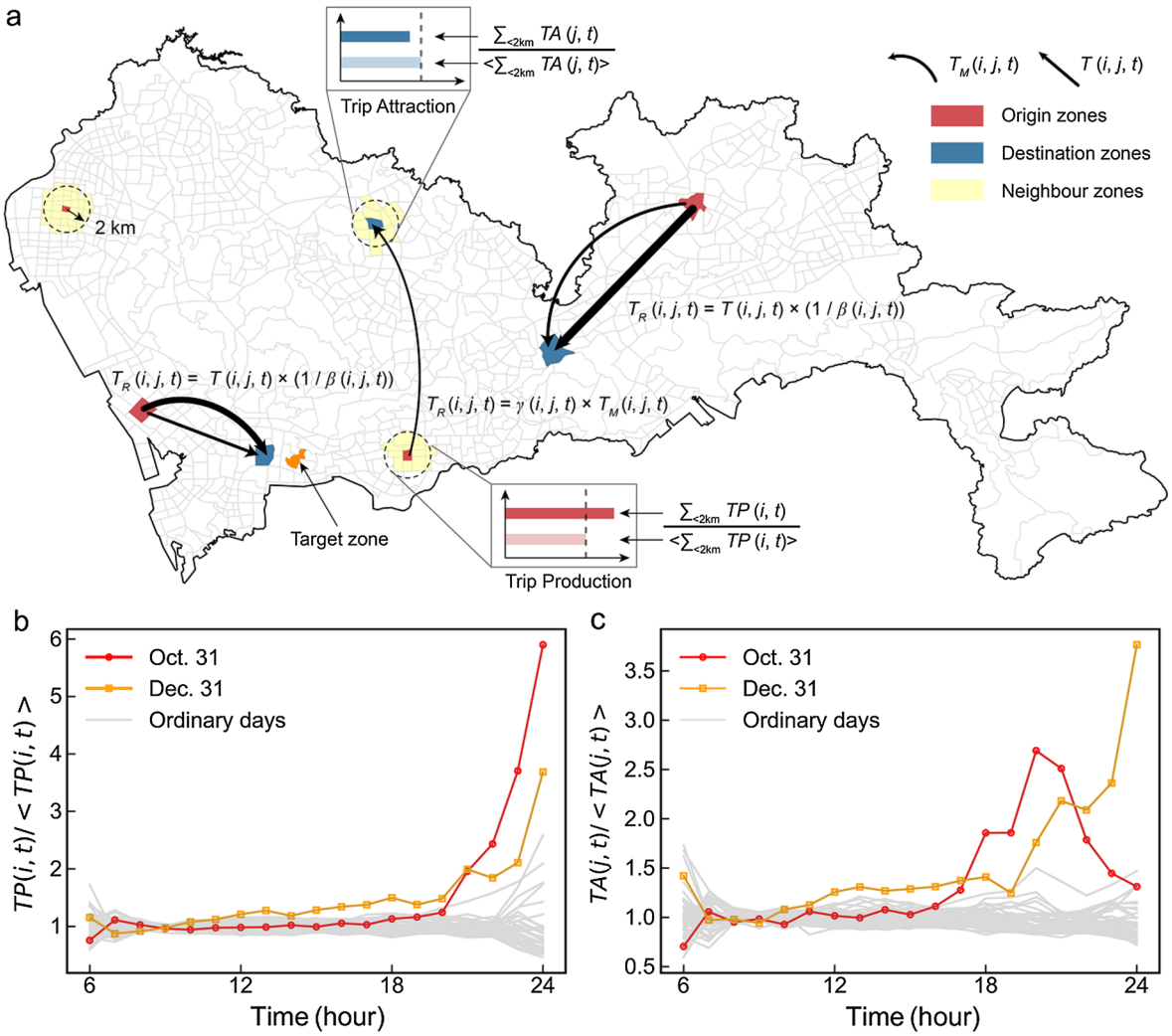


**Fig. 4.** (a) Estimating real-time mobility flows between OD pairs with different ranges of scaling factors $\beta$. The direct lines represent passenger mobility flow $T(i, j, t)$, and the arcs represent the user mobility flow $T_M(i, j, t)$. (b) The activity of hourly trip production. (c) The activity of hourly trip attraction. Much higher levels of mobility were observed during the two recreational events held on October 31 and December 31, 2014.

### *3.2. Estimating mobility flows with multi-source human mobility data*

Given the different penetration rates of transportation data between OD pairs, the human mobility model is generated for two situations according to the scaling factor $\beta$.

#### *3.2.1. Phase I: For OD pairs with scaling factor $\beta(i, j, t) < 0.05$*

For these OD pairs, the penetration rates of subway smartcard data and taxi GPS data are so low that passenger mobility flows cannot be simply up-scaled using the reciprocal of scaling factor $1/\beta$. Mobility flow $T_M(i, j, t)$ obtained from mobile phone signaling data is more credible. Here, we develop a mobility activity index $\gamma(i, j, t)$ to account for the dynamics of mobility activity in an origin traffic zone $i$ and a destination traffic zone $j$. The mobility activity index $\gamma(i, j, t)$ is estimated using the subway and taxi trip productions $\sum_{<2\text{km}} TP(i, t)$ in traffic zone $i$ and $i$'s neighboring traffic zones (whose centers are within 2 km of the center of traffic zone $i$) and the subway and taxi trip attractions $\sum_{<2\text{km}} TA(j, t)$ in traffic zone $j$ and $j$'s neighboring zones (Fig. 4):

$$\gamma(i, j, t) = \frac{1}{2} \times \left( \frac{\sum_{<2km} TP(i, t)}{\langle \sum_{<2km} TP(i, t) \rangle} + \frac{\sum_{<2km} TA(j, t)}{\langle \sum_{<2km} TA(j, t) \rangle} \right), \tag{3}$$

where $\langle \sum_{<2km} TP(i, t) \rangle$ and $\langle \sum_{<2km} TA(j, t) \rangle$ are the average total trip production and the average total trip attraction during time window $t$ for all weekdays of the three-month observation. Trip productions and trip attractions of neighboring zones are used to avoid the sparseness of urban transportation data in some areas and to obtain enough samplings to avoid sampling biases.

The mobility activity index $\gamma(i, j, t)$ quantifies the time-varying mobility level of different regions, which can encapsulate the influence of weather, seasons, holidays, and special events. As shown in Fig. 4, the mobility activity level in an ordinary day and that in a day with mass events are dramatically different. Even for different ordinary days, mobility activities have certain fluctuations. For the OD pairs with scaling factor $\beta(i, j, t) < 0.05$, mobility flow $T_R(i, j, t)$ from origin zone $i$ to destination zone $j$ is estimated by adjusting $T_M(i, j, t)$ using the mobility activity index $\gamma(i, j, t)$:

$$T_R(i, j, t) = \gamma(i, j, t) \times T_M(i, j, t). \tag{4}$$

In Eq. (4), the distribution of mobility flows is obtained from mobile phone signaling data to guarantee the confidence of travel demand distributions, while at the same time, the real-time level of mobility activity is also considered using the index $\gamma(i, j, t)$.

#### 3.2.2. Phase II: For OD pairs with scaling factor $\beta(i, j, t) \geq 0.05$

When enough penetrations of transportation data are obtained, the scaling factor $\beta(i, j, t)$ is used to calibrate the observed real-time passenger mobility flow $T(i, j, t)$:

$$T_R(i, j, t) = T(i, j, t) \times (1/\beta(i, j, t)) \tag{5}$$

Different from the static user mobility flows $T_M(i, j, t)$, $T_R(i, j, t)$ has time-variant features that can capture real-time mobility flows under anomalous situations. For example, a large crowding event occurring at traffic zone $j$ will cause an abrupt increase of mobility flows from other zones to zone $j$. This anomalous crowd gathering pattern can be captured by passenger mobility flows $T(i, j, t)$, and, hereafter, by the estimated mobility flow $T_R(i, j, t)$.

In an extreme case that no mobile phone trip was recorded between an origin traffic zone $i$ and a destination traffic zone $j$ during time window $t$, a default value of $\beta(i, j, t) = 1$ was used in Eq. (5). For this group of OD pairs, only urban transportation data were used to estimate the mobility flow $T_R(i, j, t)$. As shown in Table 1, trips on these group of OD pairs constitute 15.6% of all passenger trips collected, hence 84.4% of the subway smartcard data and taxi GPS data were combined with the mobile signaling data by the proposed human mobility model.

## 4. Validation

The model validation is performed indirectly, specifically by checking whether mobility flow distribution is well reproduced and anomalous conditions of mobility flows are well captured. This compromise is made under the background that real-time human mobility data are currently not available. Actually, such indirect way of result validations is useful and widely used in transportation research (Zhou and Mahmassani, 2006; Zhang et al., 2010; Wang et al., 2013; Hawelka et al., 2014).

### 4.1. Distribution of travel demands is well preserved

The Jensen-Shannon divergence algorithm (Lin, 1991) is usually used for measuring the similarity between two probability distributions. The Jensen-Shannon divergence JSD$(P\|Q)$ is based on Kullback-Leibler (KL) divergence, and is defined as follows:

$$\mathrm{JSD}(P\|Q) = \frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M),\ M = \frac{1}{2}(P + Q), \tag{6}$$

where KL divergence,

$$D_{\mathrm{KL}}(P\|M) = \sum_x P(x)\log(P(x)/M(x)), \tag{7}$$

is a measure of the non-symmetric difference between two probability distributions, and $P(x)$ and $Q(x)$ are the two probability distributions to be compared. We first used $P(x)$ to represent the probability distribution of normalized passenger mobility flows $T(i, j, t)/\sum_{i,j} T(i, j, t)$ in a randomly selected weekday, and $Q(x)$ to represent the probability distribution of normalized user mobility flows $T_M(i, j, t)/\sum_{i,j} T_M(i, j, t)$. The JSD values between $P(x)$ and $Q(x)$ are distributed around 0.6, indicating that $T(i, j, t)$ shows a considerable difference from $T_M(i, j, t)$ (Fig. 5a).

According to Eqs. (3)–(5), the threshold $\delta$ determines the respective conditions for using the two data-fusing methods. The selection of threshold $\delta$ needs to guarantee that the estimated mobility flows well preserve the distribution of travel demands, while at the same time capturing as much real-time human mobility information as possible. We used $P(x)$ to represent the probability distribution of normalized estimated mobility flows $T_R(i, j, t)/\sum_{i,j} T_R(i, j, t)$ of a randomly selected weekday, and $Q(x)$ to represent the probability distribution of $T_M(i, j, t)/\sum_{i,j} T_M(i, j, t)$. As the threshold $\delta$ increases, the JSD values between $P(x)$ and $Q(x)$ first drop quickly and then saturate to a stationary value JSD $\sim 0.04$ after $\delta > 0.05$ (Fig. 5b). Here, threshold $\delta = 0.05$ is selected to determine the conditions for using two data-fusing methods. Selecting this threshold can use as much real-time transportation information, while at the same time obtaining near-minimum JSD values. The low JSD value also implies great consistency between $T_R(i, j, t)$ and $T_M(i, j, t)$.

To validate our model, half of the mobile phone data is used to train the model, and the other half of the data is used to test the modeling results. We split the one-day mobile phone signaling data into two datasets (Fig. 6a), **the training dataset** includes the time windows starting with an even hour and ending with the next odd hour (i.e., 6:00–7:00, 22:00–23:00), whereas **the test dataset** includes the time windows stating with an odd hour and ending with the next even hour (i.e., 7:00–8:00, 23:00–24:00). The

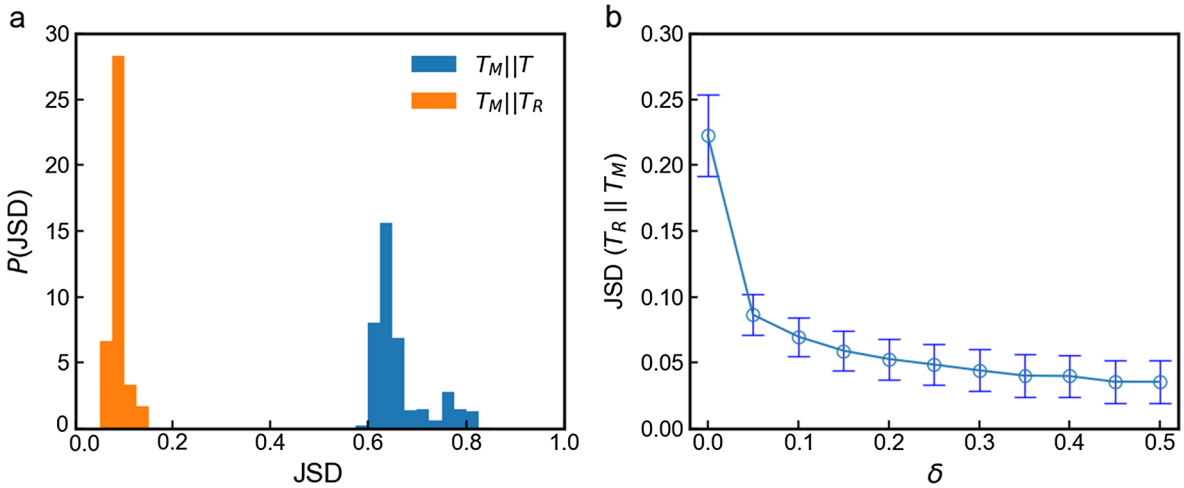


**Fig. 5.** (a) Distribution of JSD values between passenger mobility flows $T$ and user mobility flows $T_M$ (blue bars), and distribution of JSD values between estimated mobility flow $T_R$ and $T_M$ (orange bars). (b) JSD values between $T_R$ and $T_M$ under different selections of $\delta$. Error bars represent the standard deviations.

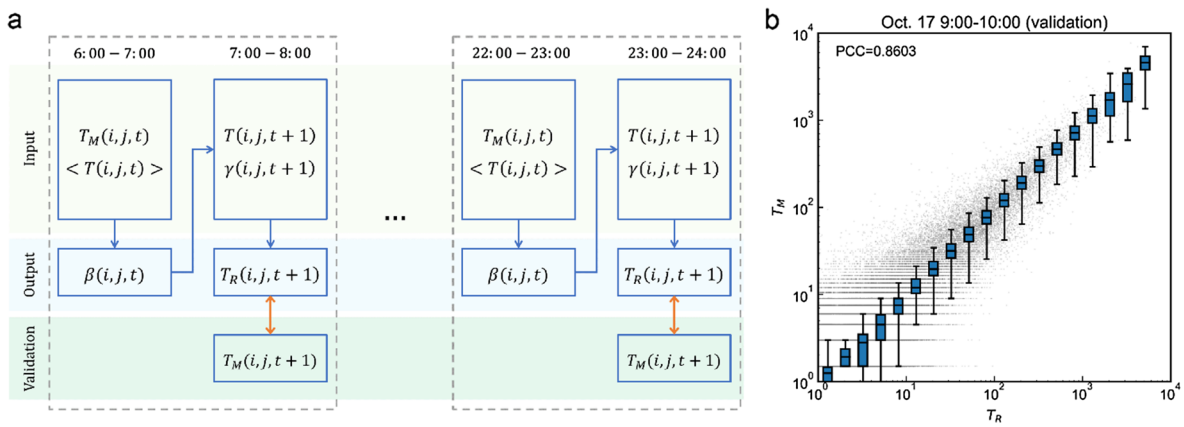


**Fig. 6.** (a) Illustration of the validation method. (b) Pearson Correlation Coefficient (PCC) between $T_R(i, j, t+1)$ and $T_M(i, j, t+1)$. Error bars indicate the range of ± 1.5 interquartile. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

validation process is as follows:

Step 1: The average weekday passenger mobility flow $\langle T(i, j, t)\rangle$ and the user mobility flow $T_M(i, j, t)$ during time window $t$ (a time window in the training dataset) are used to calculate the scaling factor $\beta(i, j, t)$ for each pair of OD. We use $\beta(i, j, t)$ of time window $t$ to approximate the scaling factor $\beta(i, j, t+1)$ of time window $t+1$.
Step 2: Mobility activity index $\gamma(i, j, t+1)$ and passenger mobility flow $T(i, j, t+1)$ are estimated based on the transportation data of time window $t+1$. Using the proposed model (Eq. (4) and Eq. (5)) and the scaling factor $\beta(i, j, t+1) = \beta(i, j, t)$, we estimate the real-time mobility flow $T_R(i, j, t+1)$ in time window $t+1$ (a time window in the test dataset).

The estimated mobility flow $T_R(i, j, t+1)$ does not use any information obtained from the mobile phone signaling data during time window $t+1$. A good positive correlation is found between $T_R(i, j, t+1)$ and $T_M(i, j, t+1)$, indicating that the proposed model can well preserve the distribution of travel demands (Fig. 6b). The average Pearson Correlation Coefficient (PCC) is 0.896 for all time windows in the test dataset.

### 4.2. Real-time burst of mobility flows is captured

Based on the estimated mobility flows $T_R(i, j, t)$, we first analyzed the dynamical population in each traffic zone. Using mobile phone signaling data, we estimate the night-time population in each traffic zone. The most frequently observed location for each mobile phone user from 22:00 to 06:00 is defined as the user's *home* location. We calculate the total number of mobile phone users,

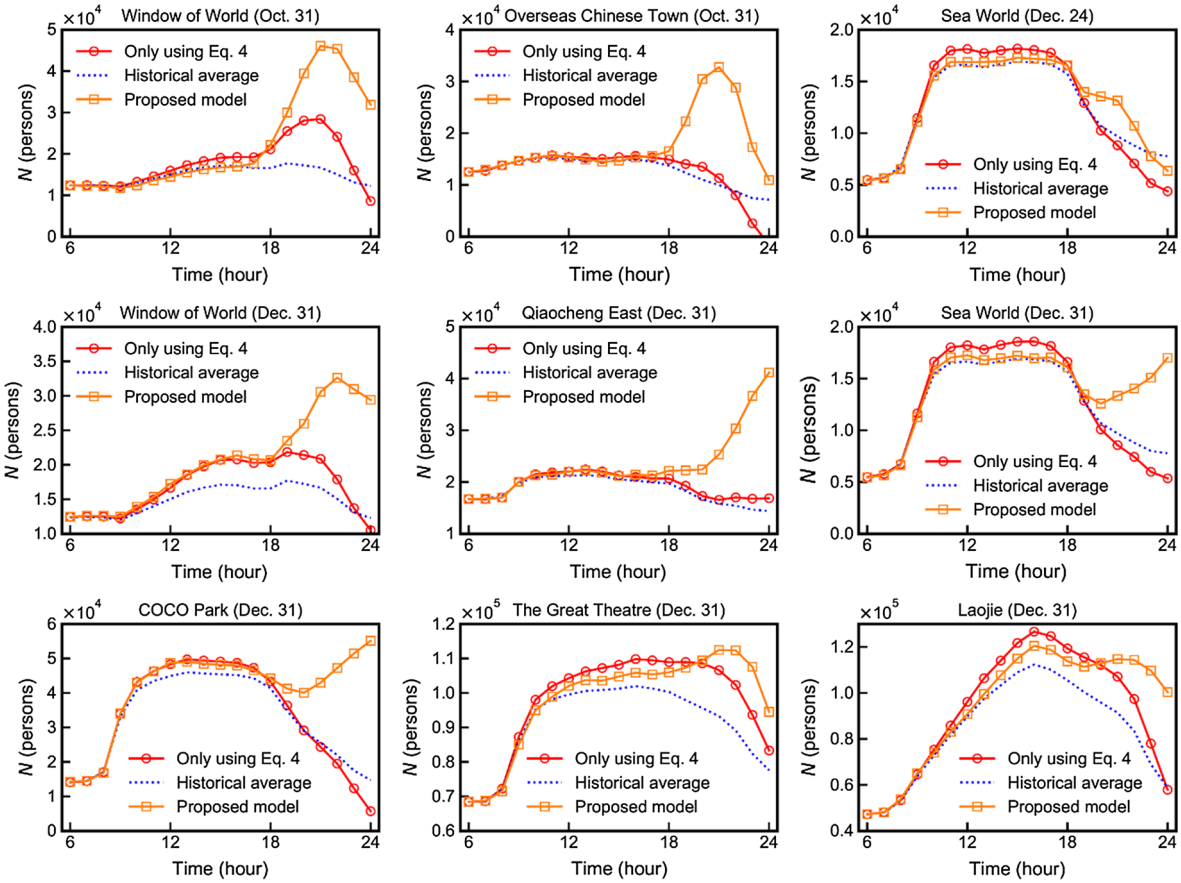


**Fig. 7.** Hourly populations of traffic zones in which crowd events occurred during the weekdays of the three-month data collection period (orange solid curves), hourly populations of these traffic zones during ordinary days (blue dashed curves), and hourly populations of these traffic zones estimated only using Eq. (4) (red solid curves). The location and date of each crowding event are given at the top of each sub-figure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$N_p(z)$, with their home location in each traffic zone $z$. Next, in the early morning when urban transportation systems return to activity (05:00), the estimated mobility flows $T_R(i, j, t)$ are used to infer the time-varying population in each traffic zone. The change in the number of individuals, $\Delta N(z, t)$, in traffic zone $z$ during each time window $t$ is estimated as follows:

$$\Delta N(z, t) = -\sum_{j\neq z}^{n} T_R(z, j, t) + \sum_{i\neq z}^{n} T_R(i, z, t) \tag{8}$$

where $\sum_{j\neq z}^{n} T_R(z, j, t)$ and $\sum_{i\neq z}^{n} T_R(i, z, t)$ represent the number of individuals leaving $z$ and the number of individuals arriving at $z$ during $t$, $n$ is the number of traffic zones. Consequently, the number of people, $N(z, t)$, in each traffic zone $z$ during each $t$, can be inferred as follows:

$$N(z, t) = N_p(z) + \sum_{t'=5am}^{t} \Delta N(z, t'), \tag{9}$$

where $t$ begins at 05:00 and ends at 24:00 every day, and $N(z, t)$ is set to $N_p(z)$ for each time window from 24:00 to 05:00.

We obtained the information on large crowding events occurring in Shenzhen from social media websites (https://goo.gl/Fa88Bn or https://1drv.ms/u/s!AnW3k7kgsODxg0KiN6aS7ovoeBET). We confirmed that a total of nine large crowd events occurred in seven traffic zones on the weekdays of the data collection period (Huang et al., 2018). The hourly population $N(z, t)$ in each of these seven traffic zones shows anomalous growth during the same time periods that crowd events occurred, implying that the estimated mobility flows $T_R(i, j, t)$ captured the real-time mobility fluxes of travelers (Fig. 7).

We find that only using Eq. (4) will result in problems. As shown in Fig. 7, crowding populations are underestimated when only using Eq. (4) to estimate $T_R$. For some crowding events, crowd gathering patterns are not observed at all. This is because the mobility activity index $\gamma(i, j, t) = \frac{1}{2} \times (\sum_{<2km} TP(i, t)/\langle \sum_{<2km} TP(i, t)\rangle + \sum_{<2km} TA(j, t)/\langle \sum_{<2km} TA(i, t)\rangle)$ underestimates the growth of mobility flows from some major sources of crowd population to the crowding zone, due to that $\gamma(i, j, t)$ incorporates trip production and trip attraction information of many neighboring zones of $i$ and $j$. Therefore, some major crowding flows are not well captured. Eq. (4) is used to estimate $T_R$ for OD pairs with low value of $\beta$. However, the penetrations of transportation data are usually high for OD pairs with the destinations at the crowding zones (usually located at the popular areas of a city). Using Eq. (4) is not suitable for

such conditions. The necessity to use two different models (Eqs. (4) and (5)) according to the transportation data sampling rates $\beta$ highlights the importance of the proposed model and the proposed JSD method for selecting suitable data-fusing methods.

## 5. Application

The information on dynamical population distribution is highly important for traffic control and transportation management. Since the end of the last century, population mapping has gained fast-growing interest and attention (Clarke, 1979; Deichmann, 1996). Representative works include population mapping using a survey census (Tobler et al., 1997), mobile phone data (Deville et al., 2014), and remote sensing based on satellite images (Dobson et al., 2000). Compared with sensing the anomalous increase of the population in some areas, predicting the anomaly in advance is more useful. An early and reliable warning can offer transportation agencies sufficient response time to prevent severe traffic jams or dangerous crowd situations (Huang et al., 2018).

As an application of the present human mobility model, we proposed a model for predicting the distribution of urban populations. Estimated zone population data were split into two parts to develop and validate the predictive model. The first part, covering October and November 2014, was used as the training dataset. The second part, covering December 2014, was used as the testing dataset. Training datasets are denoted $M_i = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$, $i = 1, 2, \cdots, n$, where $\boldsymbol{x}_m \in R^d$ comprise the input features and $y_m \in R$ comprise the output results. Here, the sample size $m$ equals 33 (the number of weekdays without missing data), the number of zones $n$ equals 996, and data dimensions $d = 2 \times n$ represents the number of input features used in the predictive model.

When predicting the population of a traffic zone $i$ during time window $t$, traffic zone $i$ is denoted as the *target zone* (Fig. 4), and time window $t$ is denoted as the *target time window*. Here, population $N(x, t)$ of every traffic zone $x$ during time windows $t = t_{\text{target}}-1$ and $t = t_{\text{target}}-2$ are used as the candidate input features (i.e., $d = 2 \times n$). This is equivalent to predicting the population in the target zone 1-hour in advance. Since too many features may cause over-fitting, an important step is to select an appropriate number of the most correlated features as the inputs of the predictive model. The recursive feature elimination (RFE) algorithm (Guyon and Elisseeff, 2003) is used to find the most important features. The fundamental idea of the RFE algorithm is to iteratively generate linear regression models and select the optimal result in each iteration and sequentially eliminate uncorrelated features. Root-mean-square error (RMSE) and mean absolute percentage error (MAPE) are used to evaluate the prediction accuracy (10):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}, \qquad \text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y_i} - y_i}{\widehat{y_i}}\right| \tag{10}$$

where $y_i$ is the predicted value, $\widehat{y_i}$ represents the actual value, and $n$ is the number of traffic zones.

**RFE Algorithm.**

---

**Input:** Feature matrix $X$ consists of $n \times 2$ features, population values in the target zone $Y$, the number of features eliminated in each loop $s$ (default $s = 5$), the number of selected top features$p$.
**Output:** The selected feature set $R$ and linear model $M$.
Initialize the selected feature set $R$ with all $n \times 2$ features.
1. **While** $|R| > p$ **do.**
2. Estimate the coefficient vector $\vec{w}$ of selected features in $R$ with linear model $M$

$$\begin{bmatrix} 1 & & x_{1p} \\ \vdots & x_{11} \cdots & \vdots \\ \vdots & \ddots \quad \ddots & \vdots \\ \vdots & x_{m1} \cdots & \vdots \\ 1 & & x_{mp} \end{bmatrix} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_p \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_m \end{bmatrix} \tag{11}$$

3. Calculate the coefficient vector $\vec{w}$ by using ordinary least squares (OLS), as $\vec{\omega} = (X^T X)^{-1} X^T y$.
4. Sort $R$ by the descending order according to the absolute value of coefficient $|\vec{\omega}|$.
5. $R_s$ consists of the worst $s$ features, $R = R - R_s$
6. **end**
7. Build linear model $M$ on the selected feature set $R$.
8. Return $R$ and $M$.

---

The top $p$ features of target traffic zone $X_i^{1\times p}$ are used to construct the linear regression function $y = \omega_0 + \omega_1 x_1 + \cdots + \omega_p x_p$, and, consequently, the population in the target traffic zone is predicted. As shown in Fig. 8a, the RMSEs of the prediction results reach a minimum when the ten most correlated features are selected to build the model, implying that the model was over-fitted when more than ten features were selected. The multiple linear regression algorithm is used to obtain the functional relations between the population feature matrix $X_i^{m\times p}$ and the population vector $y_i^{m\times 1}$ based on the training dataset. In formula (11), parameter $m$ represents the $m$th sampling date, $\omega_p$ quantifies the weight of the relation between $y$ and $x_p$, and the coefficients were estimated using ordinary least squares (OLS), as $\omega = (X^T X)^{-1} X^T y$.

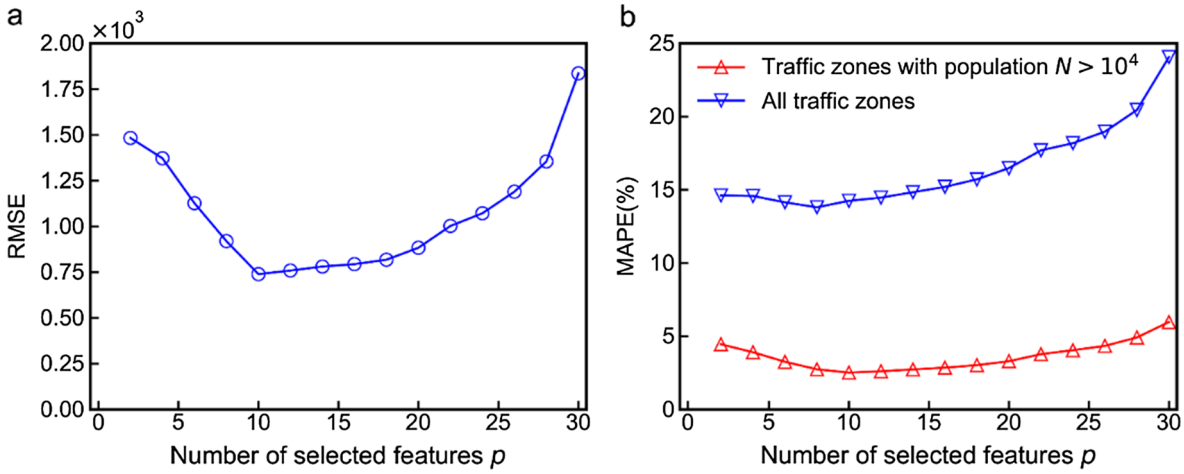


**Fig. 8.** Number of selected features vs. prediction accuracy. (a) Root-mean-square errors (RMSEs) of the predictive models using a different number of input features. (b) Mean absolute percentage errors (MAPEs) of the predictive models using a different number of input features. The MAPE value for traffic zones with population $N > 10,000$ (top 45.39%).

The deployed predictive model can accurately predict the population of a traffic zone 1 h in advance, and the average mean absolute percentage errors (MAPE) of the prediction results is 14.25%. For traffic zones with a relatively large population ($N > 10,000$), the average MAPE decreases to 2.52% (Fig. 8b). In addition, the predictive model shows the ability to predict anomalous crowd gathering situations (Fig. 9). There are six mass events occurred on December 31, 2014, many people joined the New Year's celebrations. This caused the population of many traffic zones deviated from their ordinary conditions around midnight. The regression model applied in our prediction model is linear regression, which is sensitive to outliers. Hence, the prediction for the last or the second last time window are not ideal. In future studies, a variety of new predictive methods (e.g. Lv et al., 2015; Li et al., 2017; Zhang et al., 2017; Ling et al., 2018) can be employed to improve the prediction accuracy.

## 6. Discussion

With the emergence of various big transportation data, a number of works have combined multi-source data to study human space–time activity patterns. Wu et al. (2015) combined cellular network data, vehicle count data, and travel survey to develop a convex optimization framework for real-time route flow estimation. Zhang et al. (2014) proposed a real-time OD estimation method through the superposition of historical travel demand information in CDR data. Smartcard data, and bus and taxi GPS data were used to control the superposition process. Munizaga and Palma (2012) combined subway smartcard data, bus smartcard data and GPS data of buses to infer the alighting time and station of a bus passenger and built public transport OD matrices. Gómez et al. (2015)

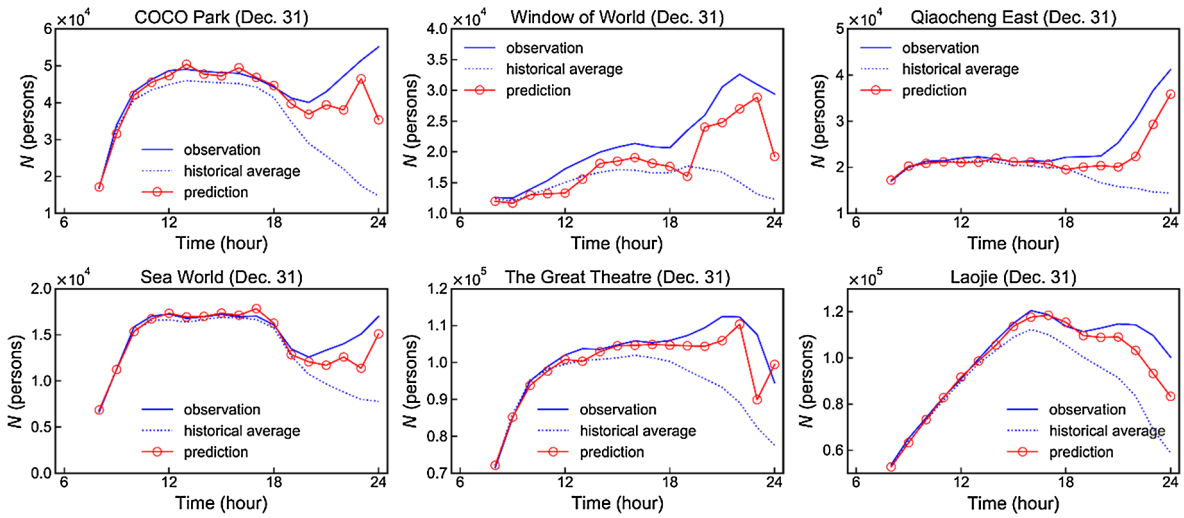


**Fig. 9.** Prediction of populations in six traffic zones with mass events and crowd gathering patterns on December 31, 2014. Prediction results show good consistency with the empirical data.

combined loop detector data and floating car data (FCD) to estimate real-time OD matrices in small networks. The minimum penetration rate of FCD to ensure a given accuracy level on OD estimation was determined. Nantes et al. (2016) combined loop detector data with partial observations from Bluetooth and GPS devices and proposed a real-time traffic flow prediction model for arterial corridors. Pan et al. (2013) and Liu and Qu (2016) combined social media data with GPS data to verify anomalous traffic condition in roads and assist route guidance. Wang et al. (2017) proposed a coupled matrix and tensor factorization model to fuse traffic information from GPS probe data and social media data. Weather data and POI data were also included to study their impact on traffic condition. Palubinskas et al. (2008) combined optical remote sensing image data with the road geographic information and vehicle property information to identify the location of vehicles, and estimate vehicle speed, vehicle density and traffic congestion state on roads.

Compared with existing works in the field of multi-source mobility data fusion, our work deals with a whole big city with more than 20 million population. This is different from existing works that usually focus on small regions or specific roads. Our work provided the new insights on what are the representativeness of different mobility data on the overall mobility flows of a big city and how can multi-source human mobility data be combined to reduce the weakness of each data; Our work also provides the new discoveries on the regions where public transportation services are possibly weak and need improvement, and the pairs of traffic zones where mobility flows may be overestimated when only using urban transportation data.

Another new contribution of our work is that we proposed an approach to determine the specific data fusion methods under different situations of transportation data sampling rates on overall mobility flows. We believe that the idea and the proposed approach for determining the threshold for selecting respective data fusion methods is new and may be also useful in a wide range of future studies which are not limited to mobility modeling and travel demand estimation.

Here, we further discussed more insights about the limitation of single data source and the benefits of combining diverse data sources in the proposed model.

(1) Only using the mobile phone signaling data:

Mobile phone signaling data are characterized with the good features such as high penetration, and regular recording frequency. However, mobile phone signaling data cannot be obtained in a real-time manner. To better use the mobile phone signaling data, it is combined with the transportation data that can be continuously collected in real-time. The integration of transportation data enables the historical, static, short-duration mobile phone signaling data to be used in real-time mobility flow estimation.

(2) Only using the subway smartcard data:

In this study, the geographic area of Shenzhen is divided into 966 traffic zones, while only 118 traffic zones contain at least one subway station. We find mobility flows existing between 410,703 pairs of origin and destination zones when analyzing the three datasets. However, there are only 13,222 OD pairs with subway passenger flows, which only represents 3.2% of all OD pairs with mobility flows. This problem can be greatly mitigated by combining the taxi GPS data that covered 219,552 OD pairs (representing 53% of all OD pairs with mobility flows). The integration of taxi GPS data greatly enhances the spatial coverage of real-time mobility sensing.

(3) Only using the taxi GPS data:

The mobility flows obtained from taxi GPS data have a large spatial coverage. However, taxi trips only represent a small share ($\sim$6%) of urban transportation (Shenzhen Transportation Committee, 2015). Here, we only use the taxi GPS data to estimate the hourly zone population during each crowding event. As shown in Fig. 10, the low-volume taxi trips had limited effect on crowd gathering processes. The integration of subway smartcard data greatly enhances the sampling rate of transportation data on overall mobility flows, and consequently enhances the quality of real-time mobility sensing.

(4) Only using the subway smartcard data and the taxi GPS data:

Scaling factor $\beta(i, j, t) = \langle T(i, j, t)\rangle / T_M(i, j, t)$ is defined to account for mobility flows contributed by other transportation modes. We measured the JSD values (similarities) between mobility flows of subway passengers $T_{\text{sub}}$, taxi passengers $T_{\text{taxi}}$, subway and taxi passengers $T$, estimated mobility flows $T_R$ and the user mobility flows $T_M$ for each time window. Only $T_R$ show prominent similarity with $T_M$ ($\langle \text{JSD}(T_R || T_M)\rangle = 0.099$) (Fig. 11). The JSD values between $T_{\text{sub}}$ and $T_M$, $T_{\text{taxi}}$ and $T_M$, and $T$ and $T_M$ are all greater than 0.5, indicating the inconsistency of mobility flows $T_{\text{sub}}$, $T_{\text{taxi}}$, $T$ with $T_M$ (Fig. 11). The integration of mobile phone signaling data calibrates the deviations of passenger mobility flows from the overall mobility flows.

## 7. Conclusions

In summary, we proposed a human mobility model that combines mobile phone signaling data and urban transportation data. The model provides a way to predict real time urban travel demand at high temporal and spatial resolution. It incorporates the comprehensive mobility features provided in mobile phone signaling data with the real time mobility features provided in subway smartcard data and taxi data. Under the general background that no data are available to evaluate the estimated real-time mobility

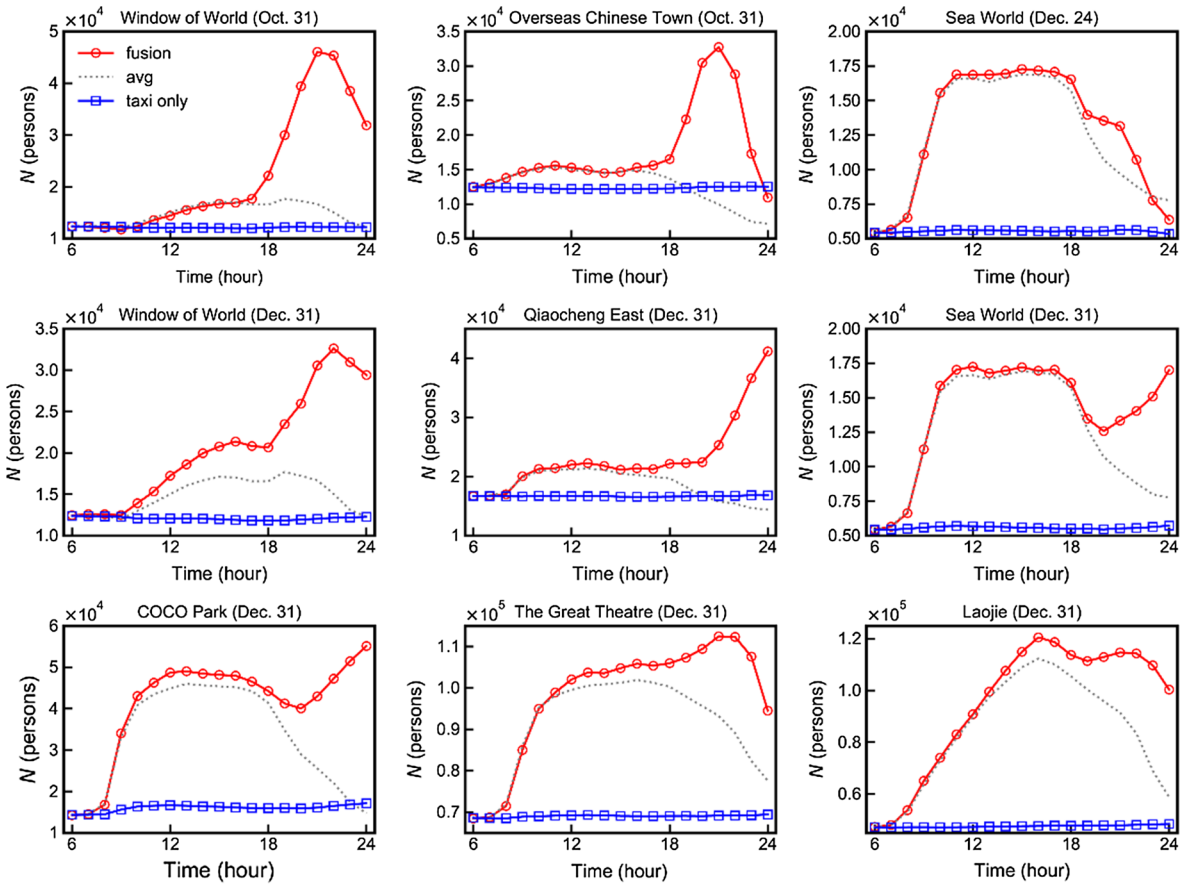


**Fig. 10.** The dashed line shows the average zone population estimated by combining three datasets. The red circles represent the estimated zone population during a crowding event. The blue squares represent the zone population estimated using the taxi GPS data only.

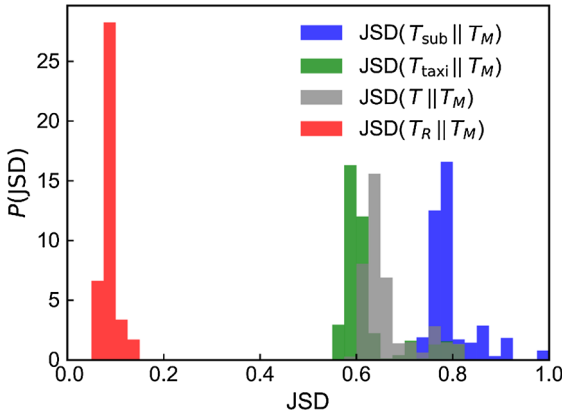


**Fig. 11.** Distributions of JSD values between mobility flows $T_{sub}$, $T_{taxi}$, $T$, $T_R$ and $T_M$.

flows, we validated the proposed model by comparing the similarity of mobility flow distributions via the JSD algorithm, demonstrating that the predicted mobility flows preserve the distribution of travel demands obtained from the mobile phone signaling data. In addition, we also indirectly validated our results via reports of large crowd gatherings occurring in the city, and we found that the proposed human mobility model can well capture the abrupt and anomalous growth of zone population. Finally, as an application of the model, we proposed a machine-learning-based model to predict the dynamical population in traffic zones, which may warn transportation agencies in advance and allow them to prevent severe traffic jams or dangerous crowding situations.

One limitation of the proposed human mobility model results from the quality of urban transportation data. Simple up-scaling of subway and taxi passenger flows is conducted because the two sets of transportation data have reasonably large penetration rates. If

data for subway passengers using tickets or number of passengers in each taxi trip are available in the future, the estimation of passenger mobility flows can be further improved. Another limitation results from the time difference of data collections. Despite the fact that the subway smartcard data and taxi GPS data were collected in the same period of 2014, the mobile phone signaling data were collected in 2012. Given the fast development of Shenzhen, land use may change in a few regions. If mobile phone signaling data of the same year or mobile phone signaling data covering more days were used, the human mobility model will be more accurate. However, given that structural change of travel demands is a relatively slow process, combining mobile phone signaling data and urban transportation data in upcoming years will not have a qualitative influence on the model. This can be seen from the successful application of the proposed model using subway smartcard data and taxi GPS data collected in 2016 (Appendix A).

More data can be incorporated into the present modeling framework. For example, results will be further improved when more urban transportation data (i.e., radio-frequency identification (RFID) data, bus smartcard data, and traffic count data) are included. This also inspires an interesting research direction: how to infer or calibrate mobility or travel demand of a transportation mode without enough records (i.e., private cars and walking) using the mobility data of transportation modes with accurate and abundant data (i.e., subway, buses, and taxis). With the availability of overall mobility flows from mobile phone signaling data, it is possible to infer unknown mobility flows from known mobility flows. Finally, we also expect that improvements of the proposed human mobility model will attract wide attention from both the human mobility and OD estimation communities.

## Acknowledgements

P.W. is supported by the National Natural Science Foundation of China (Nos. 61473320 and 71871224), the Project of Innovation-driven Plan in Central South University (No. 2016CSX014), the Fok Ying Tong Education Foundation (No. 141075), and the Hunan Science and Technology Plan Project (No. 2015RS4011).

## Appendix A

To further validate our findings, we obtain another three-month subway smartcard data and taxi GPS data of Shenzhen. The three-month data were collected during August, September, and October of 2016. There were 133 million subway trips and 9.93 million taxi trips collected during the three months. We use the same method as described in Section II to process the subway smartcard data and the taxi GPS data. Taking the time window 09:00–10:00 as an example, we analyzed the mobility flow patterns of OD pairs with two ranges of scaling factors $\beta(i, j, t) < \delta$ and $\beta(i, j, t) \geq \delta$.

In Fig. 12a, each OD pair with the scaling factor $0 < \beta(i, j, t) < 0.05$ is represented by a link connecting the origin zone and destination zone, whereas the mobility flow volume $T_M(i, j, t)$ between the two zones is illustrated by the width of the link. This group of links are usually distributed between two suburban zones or traffic zones that are close to each other. The small $\beta$ between suburban zones can be explained by the lack of subway and taxi service, whereas the small $\beta$ between neighboring zones can be

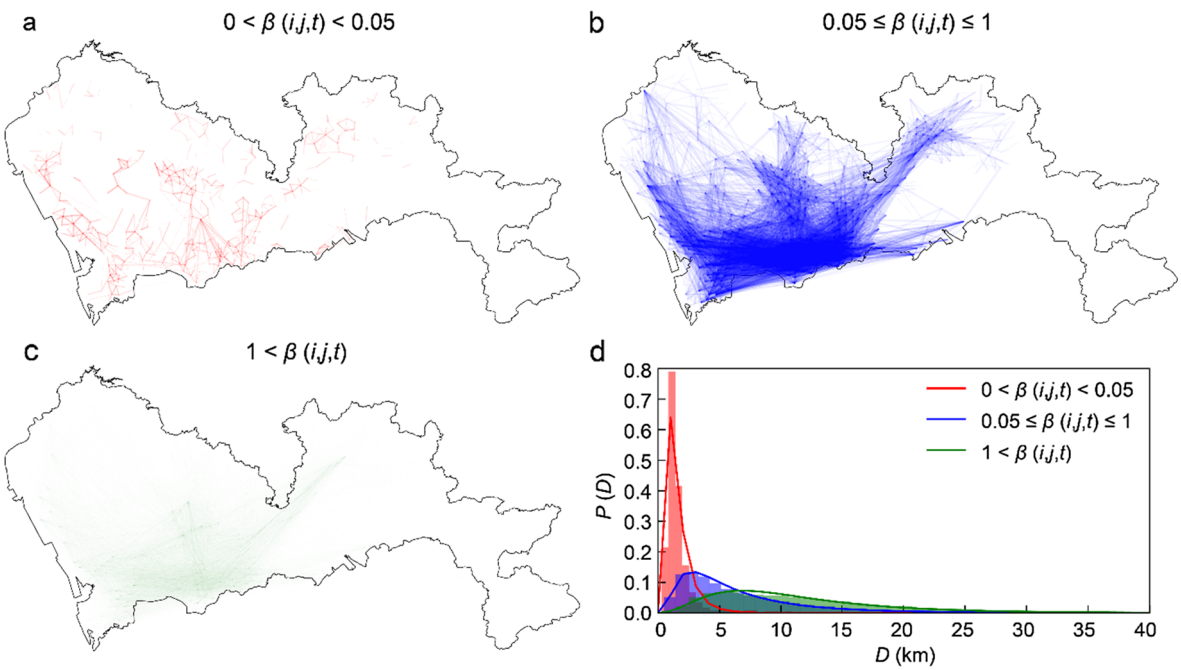


**Fig. 12.** Spatial distributions and trip distances of OD pairs with scaling factors $0 < \beta(i, j, t) < 0.05$, $0.05 \leq \beta(i, j, t) \leq 1$, and $\beta(i, j, t) > 1$.

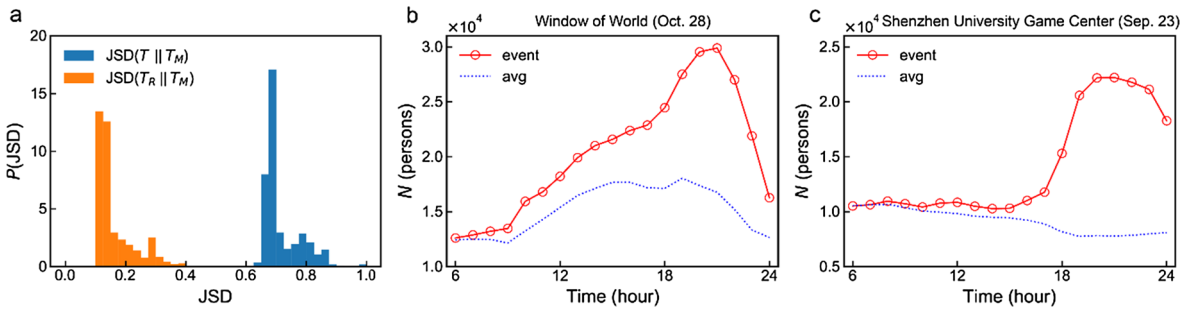


**Fig. 13.** (a) Distribution of JSD values between passenger mobility flow $T$ and user mobility flow $T_M$ (blue bars), and distribution of JSD values between estimated mobility flow $T_R$ and $T_M$ (orange bars). (b) and (c) Hourly populations of traffic zones during crowd events (red solid curves), and during ordinary days (blue dashed curves). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

attributed to trips using the walk mode. For OD pairs with moderate scaling factors $0.05 \le \beta(i, j, t) \le 1$, mobility flows of subway and taxi passengers represent a moderate portion of overall mobility flows (Fig. 12b). The average passenger mobility flows $\langle T(i, j, t)\rangle$ sometimes are larger than the user mobility flows $T_M(i, j, t)$, and these large $\beta(i, j, t)$ are usually found between two far-away traffic zones with a subway connection (Fig. 12c). Fig. 12d shows the trip distances of the three groups of OD pairs with $0 < \beta(i, j, t) < 0.05$, $0.05 \le \beta(i, j, t) \le 1$, and $\beta(i, j, t) > 1$ respectively. Trip distances of OD pairs with $\beta(i, j, t) < 0.05$ are mostly smaller than 3 km, whereas OD pairs with $\beta(i, j, t) > 1$ are dominated by long-distance trips.

We use the same method as described in Section III to build the human mobility model. When enough penetrations of transportation data are obtained ($\beta(i, j, t) \ge 0.05$), the scaling factor $\beta(i, j, t)$ is used to calibrate the passenger mobility flow $T(i, j, t)$:

$$T_R(i, j, t) = T(i, j, t) \times (1/\beta(i, j, t)) \tag{A1}$$

When the penetrations of transportation data are low ($\beta(i, j, t) < 0.05$), the real-time mobility flow $T_R(i, j, t)$ is estimated by adjusting $T_M(i, j, t)$ using mobility activity index $\gamma(i, j, t)$:

$$T_R(i, j, t) = \gamma(i, j, t) \times T_M(i, j, t). \tag{A2}$$

We used $P(x)$ to represent the probability distribution of normalized estimated mobility flows $T_R(i, j, t)/\sum_{i,j} T_R(i, j, t)$ of a randomly selected weekday, and $Q(x)$ to represent the probability distribution of normalized user mobility flows $T_M(i, j, t)/\sum_{i,j} T_M(i, j, t)$. The JSD values between $P(x)$ and $Q(x)$ are distributed around 0.15, indicating that the distribution of travel demands is well preserved by the proposed model (Fig. 13a). In addition, we found another two large crowding events occurred in weekdays during the studied three months of 2016. The first crowding event occurred in October 28, 2016 and was caused by individuals gathered at the park "Window of Word" to celebrate Halloween. The second crowding event occurred in September 23, 2016 and was a result of the "Mayday" concert held at Shenzhen University. As shown in Fig. 13b and c, the estimated hourly population $N(z, t)$ in each of the two event zones shows anomalous growth during the same time periods when crowd events occurred, implying that abrupt burst of mobility flows is well captured by the proposed model.

## Appendix B

Another way to determine the trips is to consider a user movement as a user trip. We also analyzed our major findings using this method of trip extraction. Taking time window 09:00–10:00 as an example, we analyzed the mobility flow patterns for OD pairs with two ranges of scaling factors $\beta(i, j, t) < \delta$ and $\beta(i, j, t) \ge \delta$ ($\delta = 0.05$).

For roughly 49% of the detected OD pairs, no taxi or subway passenger trips were observed during the three-month data collection. These OD pairs contain roughly 20% of all trips obtained from the mobile phone signaling data. In Fig. 14, each OD pair with $0 < \beta(i, j, t) < 0.05$ is represented by a link connecting the origin zone and destination zone, whereas the mobility flow volume $T_M(i, j, t)$ between the two zones is illustrated by the width of the link. This group of links are usually distributed between two suburban zones or traffic zones that are close to each other. The small $\beta$ between suburban zones can be explained by the lack of subway and taxi service, whereas the small $\beta$ between neighboring zones can be attributed to trips using the walk mode. For OD pairs with moderate scaling factors $0.05 \le \beta(i, j, t) \le 1$, mobility flows of subway and taxi passengers represent a considerable or dominant portion of overall mobility flows (Fig. 14b). Surprisingly, the average passenger mobility flows $\langle T(i, j, t)\rangle$ sometimes are even larger than the user mobility flows $T_M(i, j, t)$, and these large $\beta(i, j, t)$ are usually found between two far-away traffic zones with a subway connection (Fig. 14c). Fig. 14d shows the trip distances of the three groups of OD pairs with $0 < \beta(i, j, t) < 0.05$, $0.05 \le \beta(i, j, t) \le 1$, and $\beta(i, j, t) > 1$ respectively. Trip distances of OD pairs with $\beta(i, j, t) < 0.05$ are mostly smaller than 3 km, whereas OD pairs with $\beta(i, j, t) > 1$ are dominated by long-distance trips.

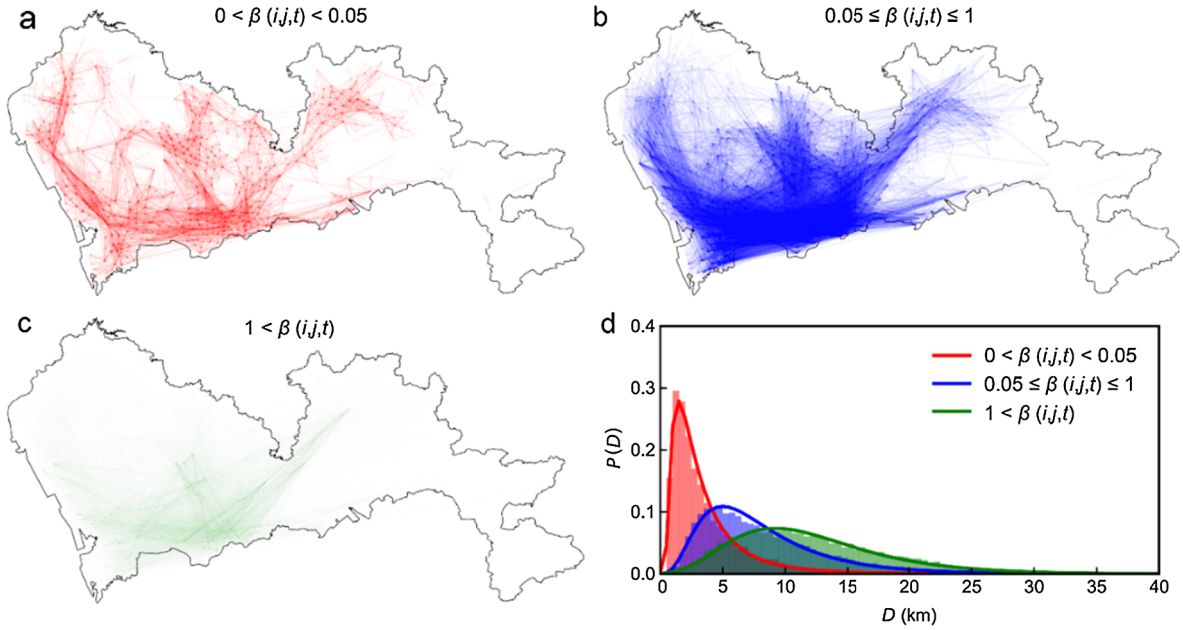


**Fig. 14.** Spatial distributions and trip distances of OD pairs with scaling factors $0 < \beta(i, j, t) < 0.05$, $0.05 \leq \beta(i, j, t) \leq 1$, and $\beta(i, j, t) > 1$.

We used $P(x)$ to represent the probability distribution of normalized passenger mobility flows $T(i, j, t)/\sum_{i,j} T(i, j, t)$ in a randomly selected weekday, and $Q(x)$ to represent the probability distribution of normalized user mobility flows $T_M(i, j, t)/\sum_{i,j} T_M(i, j, t)$. The JSD values between $P(x)$ and $Q(x)$ are distributed around 0.7, indicating that $T(i, j, t)$ shows a considerable difference from $T_M(i, j, t)$ (Fig. 15a). However, as the threshold $\delta$ increases, the JSD values between $P(x)$ and $Q(x)$ first drop quickly and then saturate to a stationary value JSD ~ 0.25 after $\delta > 0.05$ (Fig. 15b). This indicates that threshold $\delta = 0.05$ is also a good selection to determine the conditions for using two data-fusing methods (Eqs. (4) and (5)).

Using the proposed human mobility model (Eqs. (4) and (5)), we estimated the mobility flow $T_R(i, j, t)$. The estimated mobility flows $T_R(i, j, t)$ are used to infer the time-varying population in each traffic zone (Eqs. (8) and (9)). In the present study, nine large crowd events were observed in seven traffic zones. The hourly population $N(z, t)$ in each of these seven traffic zones shows anomalous growth during the same time periods that crowd events occurred, implying that the estimated mobility flows $T_R(i, j, t)$ captured the real-time mobility fluxes of travelers (Fig. 16).

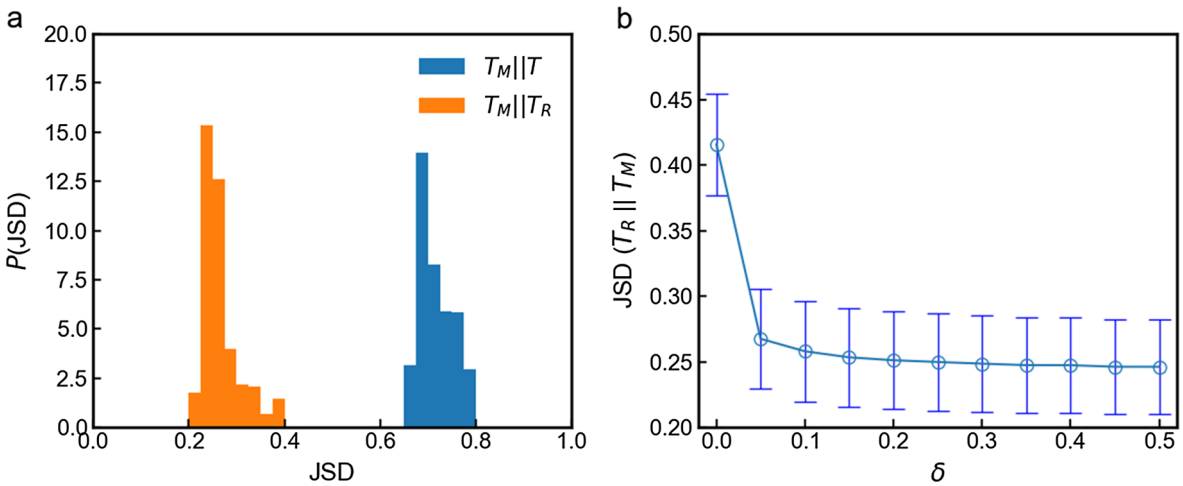


**Fig. 15.** (a) Distribution of JSD values between passenger mobility flows $T$ and user mobility flows $T_M$ (blue bars), and distribution of JSD values between estimated mobility flow $T_R$ and $T_M$ (orange bars). (b) JSD values between $T_R$ and $T_M$ under different selections of $\delta$. Error bars represent the standard deviations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

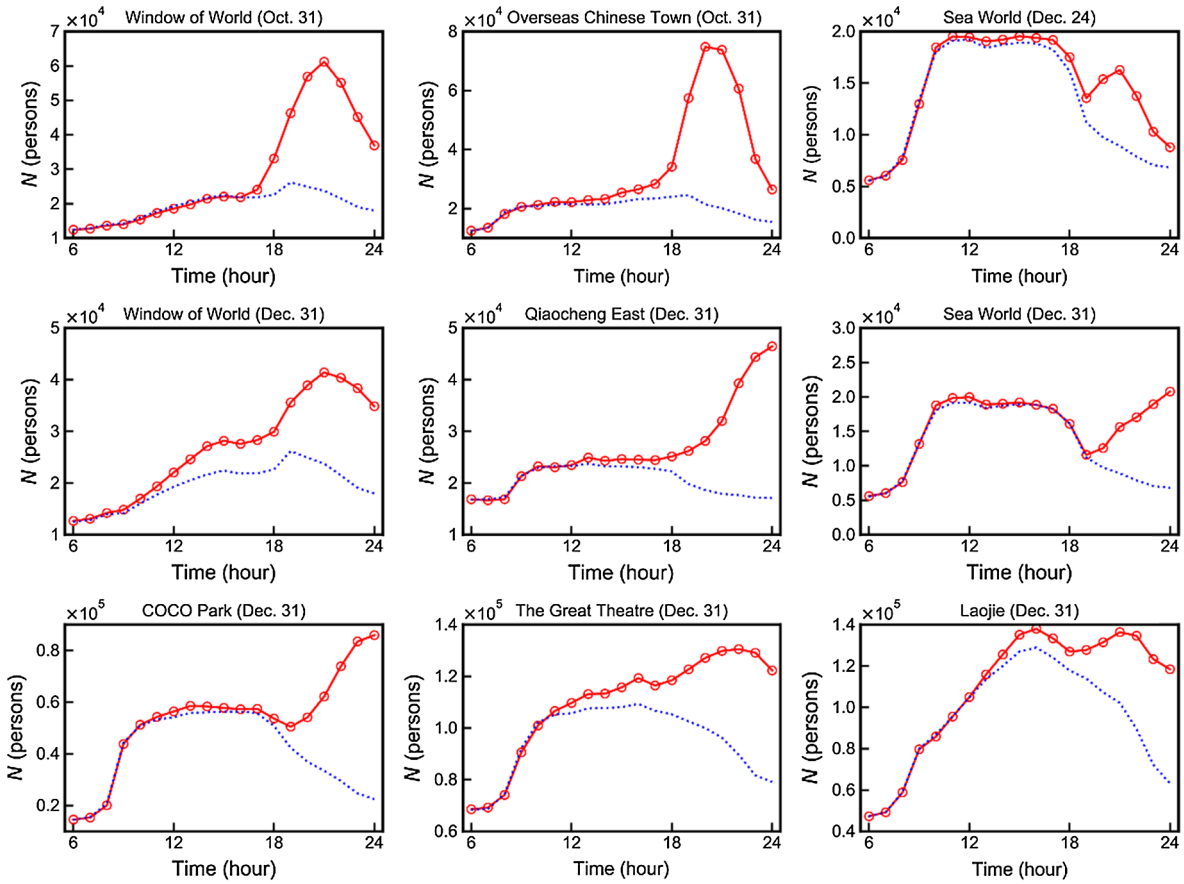


**Fig. 16.** The dashed line shows the average zone population estimated by combining three datasets. The red circles represent the estimated zone population during a crowding event. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

## References


Ahas, R., Aasa, A., Silm, S., Tiru, M., 2010b. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. Transport. Res. Part C: Emerg. Technol. 18 (1), 45–54.

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010a. Using mobile positioning data to model locations meaningful to users of mobile phones. J. Urban Technol. 17 (1), 3–27.

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transport. Res. Part C: Emerg. Technol. 58, 240–250.

Briand, A.S., Côme, E., Mohamed, K., Oukhellou, L., 2016. A mixture model clustering approach for temporal passenger pattern characterization in public transport. Int. J. Data Sci. Anal. 1 (1), 37–50.

Briand, A.S., Côme, E., Trépanier, M., Oukhellou, L., 2017. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transport. Res. Part C: Emerg. Technol. 79, 274–289.

Brockmann, D., Hufnagel, L., Geisel, T., 2006. The scaling laws of human travel. Nature 439 (7075), 462–465.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transport. Res. Part C: Emerg. Technol. 68, 285–299.

Cheng, X., Li, W., Jia, F., Yang, D., Duan, Z., 2013. Analyzing human activity patterns using cellular phone data: case study of Jinhe New Town in Shanghai, China. In: Transportation Research Board 92nd Annual Meeting (No. 13-1782).

China Daily, 2013. China's mobile phone users reach 1.11 billion. http://www.chinadaily.com.cn/china/2013-01/25/content_16172589.htm.

Clarke, J.I., 1979. Population geography. Progr. Geogr. 3 (2), 261–266.

Deichmann, U., 1996. A Review of Spatial Population Database Design and Modeling. National Center for Geographic Information and Analysis.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J., 2014. Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. 111 (45), 15888–15893.

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., Worley, B.A., 2000. LandScan: a global population database for estimating populations at risk. Photogramm. Eng. Remote Sens. 66 (7), 849–857.

Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., Zhou, X., 2015. Traffic zone division based on big data from mobile phone base stations. Transport. Res. Part C: Emerg. Technol. 58, 278–291.

Duan, Z., Wang, C., Zhang, H.M., Lei, Z., Li, H., Yang, D., 2017. Using longitudinal mobile phone data to understand the stability of individual travel patterns: case study of three communities in Shanghai, China. Transport. Res. Rec.: J. Transport. Res. Board 2643, 166–177.

Eagle, N., Pentland, A.S., 2009. Eigenbehaviors: Identifying structure in routine. Behav. Ecol. Sociobiol. 63 (7), 1057–1066.

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., Trasarti, R., 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J.—Int. J. Very Large Data Bases 20 (5), 695–719.

Gómez, P., Menéndez, M., Mérida-Casermeiro, E., 2015. Evaluation of trade-offs between two data sources for the accurate estimation of origin–destination matrices. Transportmet. B: Transp. Dyn. 3 (3), 222–245.
González, M.C., Hidalgo, C.A., Barabasi, A.L., 2008. Understanding individual human mobility patterns. Nature 453 (7196), 779–782.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (March), 1157–1182.
Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C., 2014. Geo-located Twitter as proxy for global mobility patterns. Cartogr. Geogr. Inf. Sci. 41 (3), 260–271.
He, K., Xu, Z., Wang, P., Deng, L., Tu, L., 2016. Congestion avoidance routing based on large-scale social signals. IEEE Trans. Intell. Transp. Syst. 17 (9), 2613–2626.
Huang, Z., Wang, P., Zhang, F., Gao, J., Schich, M., 2018. A mobility network approach to identify and anticipate large crowd gatherings. Transp. Res. Part B Methodol. 114, 147–170.
Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transport. Res. Part C: Emerg. Technol. 40, 63–74.
Järv, O., Ahas, R., Witlox, F., 2014. Understanding monthly variability in human activity spaces: a twelve-month study using mobile phone call detail records. Transport. Res. Part C: Emerg. Technol. 38, 122–135.
Järv, O., Tenkanen, H., Toivonen, T., 2017. Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. Int. J. Geograph. Inf. Sci. 31 (8), 1630–1651.
Leng, B., Du, H., Wang, J., Li, L., Xiong, Z., 2016. Analysis of taxi drivers' behaviors within a battle between two taxi apps. IEEE Trans. Intell. Transp. Syst. 17 (1), 296–300.
Li, L., Lin, Y., Zheng, N., Wang, F.Y., 2017. Parallel learning: a perspective and a framework. IEEE/CAA J. Autom. Sin. 4 (3), 389–395.
Lin, J., 1991. Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 37 (1), 145–151.
Ling, X., Huang, Z., Wang, C., Zhang, F., Wang, P., 2018. Predicting subway passenger flows under different traffic conditions. PLoS ONE 13 (8), e0202707.
Liu, S., Qu, Q., 2016. Dynamic collective routing using crowdsourcing data. Transport. Res. Part B: Methodol. 93, 450–469.
Lu, X., Bengtsson, L., Holme, P., 2012. Predictability of population displacement after the 2010 Haiti earthquake. Proc. Natl. Acad. Sci. 109 (29), 11576–11581.
Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y., 2015. Traffic flow prediction with big data: a deep learning approach. IEEE Trans. Intell. Transp. Syst. 16 (2), 865–873.
Ma, C., He, R., Zhang, W., 2018. Path optimization of taxi carpooling. PLoS ONE 13 (8), e0203221.
Ma, X., Wu, Y.J., Wang, Y., Chen, F., Liu, J., 2013. Mining smart card data for transit riders' travel patterns. Transport. Res. Part C: Emerg. Technol. 36, 1–12.
Munizaga, M., Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. Transport. Res. Part C: Emerg. Technol. 24, 9–18.
Nantes, A., Ngoduy, D., Bhaskar, A., Miska, M., Chung, E., 2016. Real-time traffic state estimation in urban corridors from heterogeneous data. Transport. Res. Part C: Emerg. Technol. 66, 99–118.
Ni, L., Wang, X.C., Chen, X.M., 2018. A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. Transport. Res. Part C: Emerg. Technol. 86, 510–526.
Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P., 2014. Supporting large-scale travel surveys with smartphones–a practical approach. Transport. Res. Part C: Emerg. Technol. 43, 212–221.
Palubinskas, G., Kurz, F., Reinartz, P., 2008. Detection of traffic congestion in optical remote sensing imagery. In: IEEE International on Geoscience and Remote Sensing Symposium, IGARSS 2008, vol. 2, pp. II-426.
Pan, B., Zheng, Y., Wilkie, D., Shahabi, C., 2013. Crowd sensing of traffic anomalies based on human mobility and social media. In: Proc. 21st ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst., pp. 344–353.
Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.L., 2015. Returners and explorers dichotomy in human mobility. Nat. Commun. 6.
Ren, Y., Ercsey-Ravasz, M., Wang, P., González, M.C., Toroczkai, Z., 2014. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nat. Commun. 5 (5), 5347.
SA, R., Karim, M.A., Qiu, T.Z., Kim, A., 2015. Origin-destination trip estimation from anonymous cell phone and foursquare data. In: Transportation Research Board 94th Annual Meeting (No. 15-2379).
Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C., 2013. Unravelling daily human mobility motifs. J. R. Soc. Interface 10 (84), 20130246.
Semanjski, I., Gautama, S., 2016. Crowdsourcing mobility insights–Reflection of attitude based segments on high resolution mobility behaviour data. Transport. Res. Part C: Emerg. Technol. 71, 434–446.
Shenzhen Transportation Committee, 2015. Annual report of Shenzhen Transportation Authority. http://www.sztb.gov.cn/xxgk/ztzl/wgk/jggk/lsqkgk/201503/t20150306_5270811.htm.
Simini, F., González, M.C., Maritan, A., Barabási, A.L., 2012. A universal model for mobility and migration patterns. Nature 484 (7392), 96–100.
Song, C., Qu, Z., Blumm, N., Barabási, A.L., 2010a. Limits of predictability in human mobility. Science 327 (5968), 1018–1021.
Song, C., Koren, T., Wang, P., Barabási, A.L., 2010b. Modelling the scaling properties of human mobility. Nat. Phys. 6 (10), 818–823.
Stopher, P.R., Greaves, S.P., 2007. Household travel surveys: Where are we going? Transport. Res. Part A: Policy Pract. 41 (5), 367–381.
Tang, J., Jiang, H., Li, Z., Li, M., Liu, F., Wang, Y., 2016. A two-layer model for taxi customer searching behaviors using GPS trajectory data. IEEE Trans. Intell. Transp. Syst. 17 (11), 3318–3324.
Tobler, W., Deichmann, U., Gottsegen, J., Maloy, K., 1997. World population in a grid of spherical quadrilaterals. Int. J. Populat. Geogr. 3 (3), 203–225.
Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: travel demand estimation using big data resources. Transport. Res. Part C: Emerg. Technol. 58, 162–177.
Wang, F., Chen, C., 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. Transport. Res. Part C: Emerg. Technol. 87, 58–74.
Wang, P., Hunter, T., Bayen, A.M., Schechtner, K., González, M.C., 2012. Understanding road usage patterns in urban areas. Sci. Rep. 2, 1001.
Wang, M., Schrock, S., Vander Broek, N., Mulinazzi, T., 2013. Estimating dynamic origin-destination data and travel demand using cell phone network data. Int. J. Intell. Transp. Syst. Res. 11 (2), 76–86.
Wang, J., Wei, D., He, K., Gong, H., Wang, P., 2014. Encapsulating urban traffic rhythms into road networks. Sci. Rep. 4.
Wang, S., Zhang, X., Cao, J., He, L., Stenneth, L., Yu, P.S., Li, Z., Huang, Z., 2017. Computing urban traffic congestions by incorporating sparse GPS probe data and social media data. ACM Trans. Inf. Syst. (TOIS) 35 (4), 40.
Wu, C., Thai, J., Yadlowsky, S., Pozdnoukhov, A., Bayen, A., 2015. Cellpath: fusion of cellular and traffic sensor data for route flow estimation via convex optimization. Transport. Res. Part C: Emerg. Technol. 59, 111–128.
Yan, X., Wang, W., Gao, Z., Lai, Y., 2017. Universal model of individual and population mobility on diverse spatial scales. Nat. Commun. 8 (1), 1639 2017.
Yuan, N.J., Zheng, Y., Zhang, L., Xie, X., 2013. T-finder: a recommender system for finding passengers and vacant taxis. IEEE Trans. Knowl. Data Eng. 25 (10), 2390–2403.
Zhang, Y., Qin, X., Dong, S., Ran, B., 2010. Daily OD matrix estimation using cellular probe data. In: 89th Annual Meeting Transportation Research Board, vol. 9.
Zhang, D., Huang, J., Li, Y., Zhang, F., Xu, C., He, T., 2014, September. Exploring human mobility with multi-source data at extremely large metropolitan scales. In: Proceedings of the 20th Annual International Conference on Mobile Computing and Networking. ACM, pp. 201–212.
Zhang, J., Zheng, Y., Qi, D., 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In: AAAI, pp. 1655–1661.
Zhao, J., Qu, Q., Zhang, F., Xu, C., Liu, S., 2017. Spatio-temporal analysis of passenger travel patterns in massive smart card data. IEEE Trans. Intell. Transp. Syst. 18 (11), 3135–3146.
Zhong, G., Wan, X., Zhang, J., Yin, T., Ran, B., 2017. Characterizing passenger flow for a transportation hub based on mobile phone data. IEEE Trans. Intell. Transp. Syst. 18 (6), 1507–1518.
Zhou, X., Mahmassani, H.S., 2006. Dynamic origin-destination demand estimation using automatic vehicle identification data. IEEE Trans. Intell. Transp. Syst. 7 (1), 105–114.