Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records

Article *in* Transactions in GIS · February 2018

DOI: 10.1111/tgis.12323

6 authors, including:

Liu Zhang
Chinese Academy of Sciences
5 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Ting Ma
State Key Laboratory of Resources and Environmental Information System
46 PUBLICATIONS **538** CITATIONS
SEE PROFILE

Yunyan du
Chinese Academy of Sciences
104 PUBLICATIONS **471** CITATIONS
SEE PROFILE

Tao Pei
Chinese Academy of Sciences
61 PUBLICATIONS **995** CITATIONS
SEE PROFILE

Received: 28 June 2017 | Revised: 30 December 2017 | Accepted: 16 January 2018

DOI: 10.1111/tgis.12323

RESEARCH ARTICLE

WILEY Transactions in GIS

Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records

Zhang Liu[1,2] | Ting Ma[1,2] | Yunyan Du[1,2] | Tao Pei[1,2] | Jiawei Yi[1,2] | Hui Peng[1,2]

[1] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

[2] University of Chinese Academy of Sciences, Beijing, China

Correspondence
Ting Ma, State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A, Datun Road, Chaoyang District, Beijing 100101, China.
Email: mting@lreis.ac.cn

Funding information
National Natural Science Foundation of China, Grant/Award Nos. 4159840011 and 41771418; The Foundation for Innovative Research Groups of the National Natural Science Foundation of China, Grant/Award No. 41421001; The Innovation Program of the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Grant/Award No. 088RA500PA; The Distinguished Young Scholar Program of the Institute of Geographic Sciences and Natural Resources Research, CAS, Grant/Award No. 2014RC102

Abstract

Understanding the spatiotemporal dynamics of urban population is crucial for addressing a wide range of urban planning and management issues. Aggregated geospatial big data have been widely used to quantitatively estimate population distribution at fine spatial scales over a given time period. However, it is still a challenge to estimate population density at a fine temporal resolution over a large geographical space, mainly due to the temporal asynchrony of population movement and the challenges to acquiring a complete individual movement record. In this article, we propose a method to estimate hourly population density by examining the time-series individual trajectories, which were reconstructed from call detail records using BP neural networks. We first used BP neural networks to predict the positions of mobile phone users at an hourly interval and then estimated the hourly population density using log-linear regression at the cell tower level. The estimated population density is linearly correlated with population census data at the sub-district level. Trajectory clustering results show five distinct diurnal dynamic patterns of population movement in the study area, revealing spatially explicit characteristics of the diurnal commuting flows, though the driving forces of the flows need further investigation.

1 | INTRODUCTION

Mapping population dynamics is of great significance to transport and city planning (Becker et al., 2011; De Nadai et al., 2016; Tao, Corcoran, Mateo-Babiano, & Rohde, 2014), public safety warning (Li, Xu, Ma, & Chung, 2015; Traag, Browet, Calabrese, & Morlot, 2011; Zhou, Pei, & Wu, 2016), disaster impact assessments (Bengtsson, Lu, Thorson, Garfield, & Von Schreeb, 2011; Min & Jeong, 2013; Wilson et al., 2016), and epidemic modeling (Faria et al., 2014; Lopez, Gunasekaran, Murugan, Kaur, & Abbas, 2014; Vespignani, 2009). But the acquisition of attribute and location data of human activities is still a challenge for mapping population dynamics at fine temporal resolution. People tend to stay silent for quite a while and then mobilize actively within a short time period (Barabási, 2005; Oliveira & Barabási, 2005). People would generate few records when they sleep at home at night, but would generate many records when they ring around in the workplace during the day. Therefore, location information of human activities is not always available at fine temporal resolution. The spatiotemporal sparseness (Dong et al., 2015) in mobile phone positioning data is ubiquitous due to many restrictions on data recording. A user's location is recorded in the Call Detail Records (CDR) only when a user makes a call or sends a text message. Even a very active user would only generate a few tens of voice or text events per day (Hoteit, Chen, Viana, & Fiore, 2016) and a complete time series of the individual trajectory is usually less available. An intermittent or discontinuous record of users' locations makes it difficult or even impossible to accurately estimate population dynamics at fine temporal scale.

In this study, we proposed an approach for population dynamics mapping based on the time series of individual trajectories reconstructed from CDRs. The proposed method has potential for mapping population dynamics at fine spatial and temporal resolutions. It is also effective at estimating population distribution, even when the users' location information is intermittent or discontinuous. The article is organized as follows: Section 2 reviews related work. Section 3 presents the mobile phone data and the 2010 national census data in Beijing. Section 4 presents the methods for reconstructing time series of users' trajectories and estimating the hourly population distribution. In Section 5, we first evaluate the accuracies of trajectory reconstruction and dynamic population mapping. Then, we compare the mapping results of the method proposed in this study and two other population mapping methods. Finally, we identify five diurnal dynamic patterns of population distribution in Beijing. In Section 6, we summarize the article and present our thoughts on future research.

2 | RELATED WORK

Previous studies of population mapping focused mainly on disaggregating population data from census units into grid cells (Lloyd, Sorichetta, & Tatem, 2017) using simple areal-weighting methods (Balk & Yetman, 2004; Deichmann, Balk, & Yetman, 2001; Doxsey-Whitfield et al., 2015) or the dasymetric approach (Balk et al., 2006; Bhaduri, Bright, Coleman, & Urban, 2007; Eicher & Brewer, 2001; Mennis, 2003; Tatem, Noor, Von Hagen, Di Gregorio, & Hay, 2007). The former group of methods mainly assigns a population count to each grid based on its areal percent within the host areal unit (Mennis, 2003). By contrast, the latter group of methods uses ancillary data such as remote sensing images, land cover, urban extent, and accessibility data (Lloyd et al., 2017) to redistribute population counts within the census units. In light of these studies, Stevens, Gaughan, Linard, and Tatem (2015) combined census population counts with a wide range of geospatial data (e.g., settlement locations, settlement extents, land cover, roads, building maps, health facility locations, satellite nightlights, vegetation, topography, refugee camps) and developed a flexible "Random Forest" estimation technique to more accurately map population density. However, the census and ancillary data used in Stevens et al. (2015), as well as in many previous studies, usually lag behind the time of interest significantly. It is a challenge—or even impossible—to acquire such a wide range of data at a daily or even higher temporal resolution. These datasets are also not very accessible in many developing countries due to the cost (Deville et al., 2014).

In recent years, advancements in information communication technology and increased accessibility of location-aware mobile devices (Lwin, Sugiura, & Zettsu, 2016) have provided multi-source location-based big data about human mobility at very high temporal resolution. Data on taxi trajectories, mobile usage, smart cards, and social media check-in (Agard, Morency, & Trépanier, 2006; Eagle, Pentland, & Lazer, 2009; Liu, Sui, Kang, & Gao, 2014; Pei et al., 2014; Yang, Zhao, & Lu, 2016; Yuan et al., 2010) are able to track human movements over a large area and sense almost the real-time dynamics of urban citizens (Yang et al., 2016). Combined with other ancillary data, such big data have been used successfully in fine-scale population mapping (Deville et al., 2014; Douglass, Meyer, Ram, Rideout, & Song, 2015; Kang, Liu, Ma, & Wu, 2012; Khodabandelou, Gauthier, El-Yacoubi, & Fiore, 2016; Li et al., 2015; Lwin et al., 2016; Patel et al., 2016; Sterly, Hennig, & Dongo, 2013; Yao et al., 2017). For example, Deville et al. (2014) incorporated mobile phone records with a wide range of geospatial data to downscale census population counts into 100 × 100 m grid units in France and Portugal. They used log-linear regression to estimate the relationship between nighttime mobile phone-call records and the census population. Their study showed the potential of mobile phone data in estimating daily, weekly, and seasonal population dynamics. Patel et al. (2016) incorporated the density of geo-located tweets in Indonesia as a covariate layer into the method proposed by Stevens et al. (2015) and successfully disaggregated census population counts to 100 × 100 m grid cells. The density of geo-located tweets helps improve the accuracy of population estimates and the performance of the Random Forest model. Yao et al. (2017) proposed a method to map urban population distribution at the building scale by integrating multisource geospatial big data. They first analyzed the Baidu points-of-interest (POIs) and real-time Tencent user density using a Random Forest algorithm to downscale the street-level population distribution to the grid level. They then used an iterative building-population gravity model to map the population at the building level. They argued that their method performs better than several other popular population mapping methods in terms of mapping accuracy. Lwin et al. (2016) built a spatiotemporal multivariate regression model to examine personal trip survey data by incorporating both mobile call records and geotagged tweets to generate multi-temporal population maps at a spatial scale of 500 × 500 m and a temporal scale of 30 minutes. The results showed that the estimated population distribution at night, especially between 00:00 and 06:30 a.m., is strongly correlated with the national census data except for those grids with railway and subway stations.

Previous works on dynamic population mapping have estimated population distribution at fine spatial scales over a given time period. In recent years, the users' time-series trajectories constructed from multi-source location-based big data on human mobility, such as mobile usage, smart card data, taxi GPS traces, and social media check-in data (Liu et al., 2014; Long, Zhang, & Cui, 2012; Pei et al., 2014; Yuan et al., 2010), make it possible to estimate population distribution at fine temporal resolution. However, it is still a challenge to map the hourly dynamics of an urban population using those trajectories, mainly due to the spatiotemporal sparsity of the data. The users' time-series trajectories at fine spatiotemporal resolution are indispensable for the hourly dynamic population mapping. The trajectory reconstruction that consists of filling the spatiotemporal gaps in the data is a way to mitigate the sparsity of users' time-series trajectories. However, the trajectory reconstruction is an extremely complex and non-linear problem due to the complicated spatiotemporal variation of the trajectories. Backpropagation (BP) neural networks have been used widely to solve such problems (Chen, Chi, Wang, Pang, & Xiao, 2011; Ding, Wang, Wang, & Baumann, 2013; Partsinevelos, Agouris, & Stefanidis, 2005; Xu, Li, & Claramunt, 2018). It was demonstrated in previous studies that BP neural networks have the ability to capture non-linearity, and good prediction capability and flexibility. Here, BP neural networks are used to model and reconstruct users' time-series trajectories in this study.

3 | DATA

This section is organized as follows: the source and properties of mobile phone data are introduced first, and then the spatial distribution and other attributes of population census data are shown.

3.1 | Mobile phone data

In this study, we used anonymized individual CDRs with information on user IDs, time, and base-station locations only (Table 1). The dataset was collected anonymously when the user called or sent a text message, and provided by a

TABLE 1 Description of the mobile phone dataset

FID	Name	Description and coded values
1	UserID	User's ID, which is unique to each individual user
2	PointNum	Number of records
3	Lat.	The latitude of the record
4	Lon.	The longitude of the record
5	Time	The duration of the record

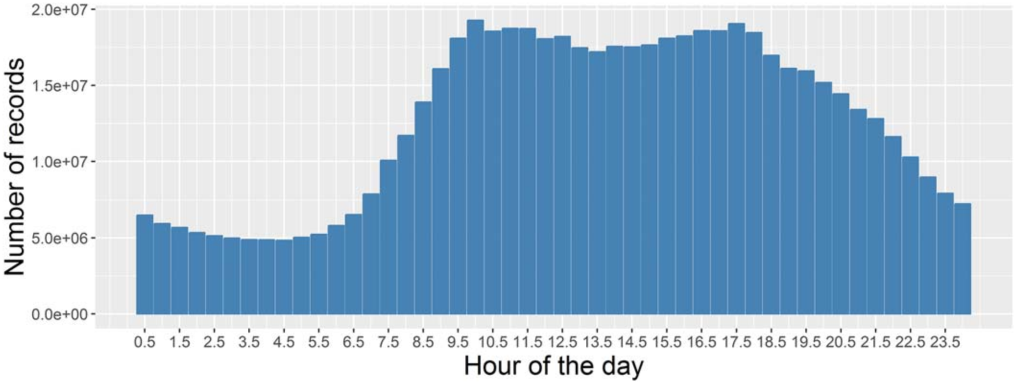

FIGURE 1 Variations in the number of records every half hour

Chinese mobile phone operator for scientific research. All personally identifiable information was masked. The dataset for this study includes over 600 million records of more than 27 million users on December 6, 2015 in Beijing, China. Users are more active during the daytime than at night (Figure 1), and more than 66% of the records were generated from 8 a.m. to 7 p.m.

We also used the locations of all the 20,790 mobile base stations. The coverage area of each mobile base station can be approximated as the Voronoi polygon that is built around it. When a phone is used to make a call or send a text message, its location will be found by the specific mobile base to which the phone is connected. The average distance between base stations in the whole study area is 440 m.

3.2 | Population census data

The latest population census data were obtained from the Beijing Statistical Information Net (http://www.bjstats.gov.cn/; accessed on August 20, 2016). The census counts were reported at sub-district level, with a total number of 315 units across the whole city (Figure 2). This is the best available population data at the time of the study. The average radius of the sub-district units is about 2.9 km within the 6th ring road and increases to 7.3 km beyond the 6th ring road. Population density varies accordingly, with an average of 14,881 and 1,816 people/km^2 within and beyond the 6th ring road, respectively (Figure 2).

4 | METHODOLOGY

This section is organized as follows: we first calculated the radius of gyration (ROG) from the users' trajectories. The ROG was then used to differentiate the mobility patterns of sedentary people, urban residents, and commuters. BP neural networks were then used to reconstruct the time-series trajectories of the users at hourly intervals. The accuracy of trajectory reconstruction was evaluated by testing the influences of two factors: the missing data rate and the numbers of training samples used in the BP neural network analysis. The hourly population density was then estimated by using log-linear regression and the population estimates were compared against the census population density using a 10-fold cross-validation method. Finally, a hierarchical clustering algorithm was used to group the sub-districts based on time-series similarity of the hourly diurnal estimated population.

4.1 | Mobility patterns

We first characterized cellphone users' mobility patterns using the ROG (Hoteit, Secci, Sobolevsky, Ratti, & Pujolle, 2014; Kang et al., 2012), defined as:

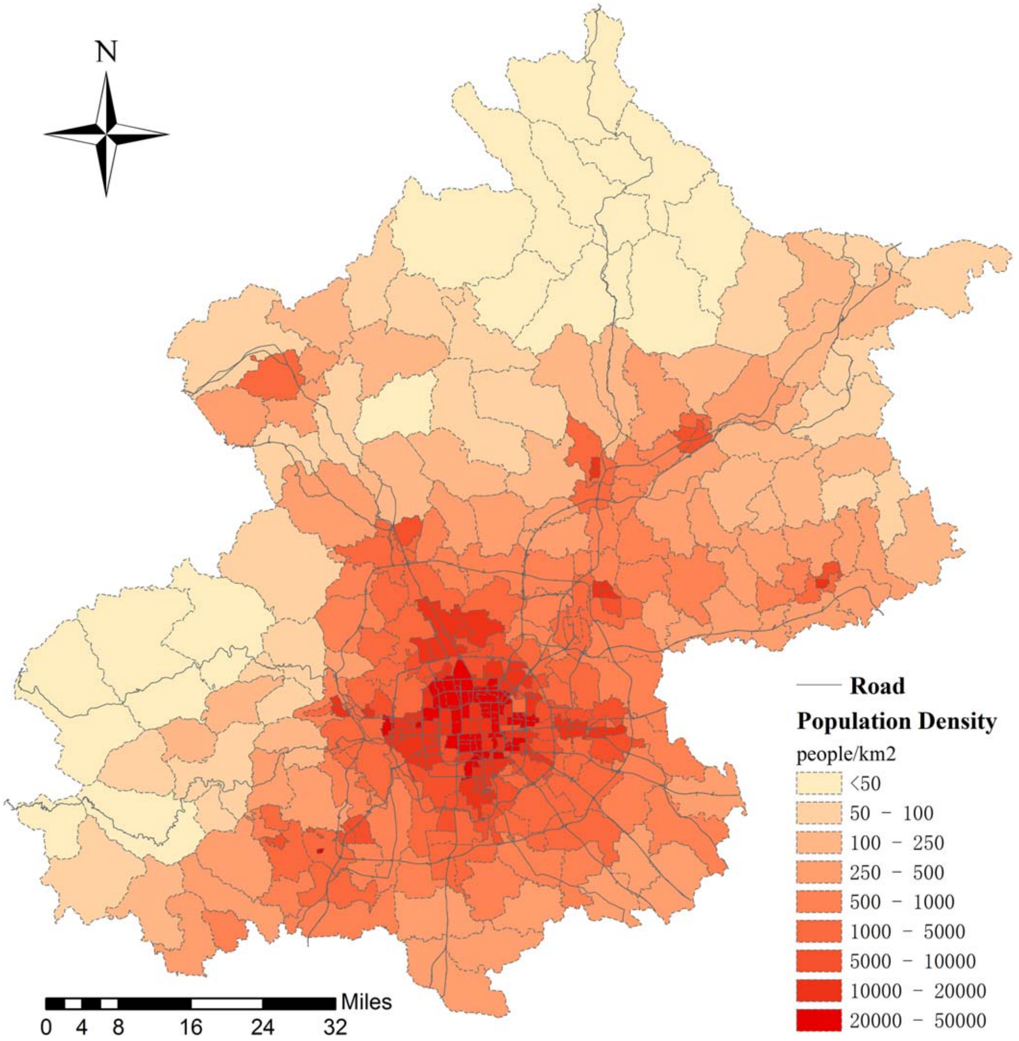

FIGURE 2 Population density at sub-district level in Beijing

$$r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\vec{p}_i - \vec{p}_{centroid}\right)^2} \tag{1}$$

where $\vec{p}_i$ represents the ith position of a user and $\vec{p}_{centroid}$ is the mass center ($\vec{p}_{centroid} = \frac{1}{n}\sum_{i=1}^{n}\vec{p}_i$) of all locations the user visits. A larger ROG would indicate a wider mobile range.

We then categorized users as sedentary people, urban residents, and commuters based on the mobility patterns as indicated by the ROG values. We first constructed the cumulative distribution functions (CDFs) (Hoteit et al., 2014; Levin, 1981) from the ROG values of all user trajectories. We then selected two threshold ROG values of 1.5 and 6.5 km, which correspond to a CDF value of 60 and 90%, respectively (Figure 3). Sedentary people have ROG values no more than 1.5 km (Figure 4a). Urban residents, whose workplaces and residences are within the inner city, have ROG values more than 1.5 km but less than 6.5 km (Figure 4b). 6.5 km is approximately the radius of the 4th ring road in Beijing. The commuters in this study are defined as those who travel a distance no less than 6.5 km one-way daily between their workplaces and residences (Figure 4c).

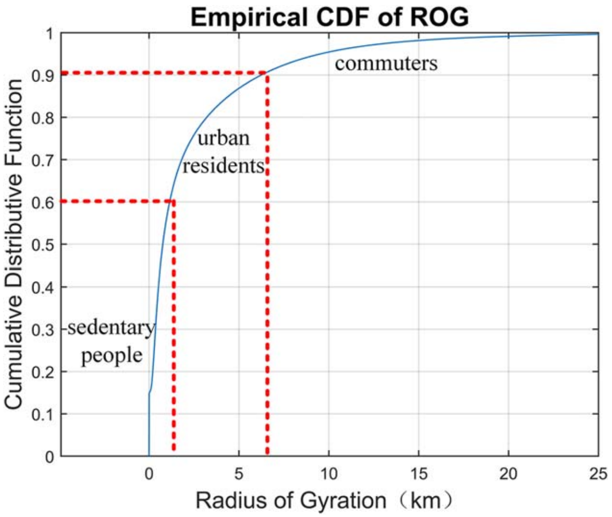

FIGURE 3 The cumulative distribution function of the ROG. The red dashed lines show that when the CDF values reach 0.6 and 0.9, the corresponding ROG values are 1.5 and 6.5 km, respectively. The ROG values of sedentary people are no greater than 1.5 km; the ROG values of urban residents are more than 1.5 km and less than 6.5 km; the ROG values of commuters are no less than 6.5 km

4.2 | Reconstruction of users' time-series trajectories using BP neural networks

We then used non-linear and self-learning BP neural networks (Hecht-Nielsen, 1988; Le Cun, 1988; Zipser & Andersen, 1988) to reconstruct the time-series trajectories of the users with different mobility patterns. The BP neural networks consist of one input layer, two hidden layers, and one output layer (Figure 5). Each hidden layer and the output layer contain 15 and 2 neuron nodes, respectively. The p_{center} in the network structure is the center point of a specific

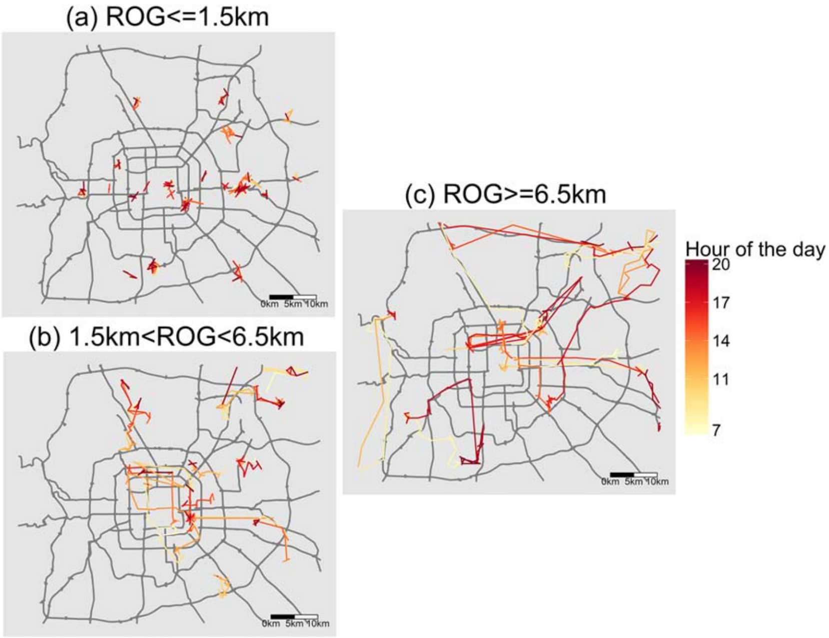

FIGURE 4 Sample trajectories of mobile phone users showing different types of mobility patterns: (a) sedentary people with ROG values no greater than 1.5 km; (b) urban residents with ROG values more than 1.5 km and less than 6.5 km; and (c) commuters with ROG values no less than 6.5 km. The colors represent the times of corresponding points on the trajectories

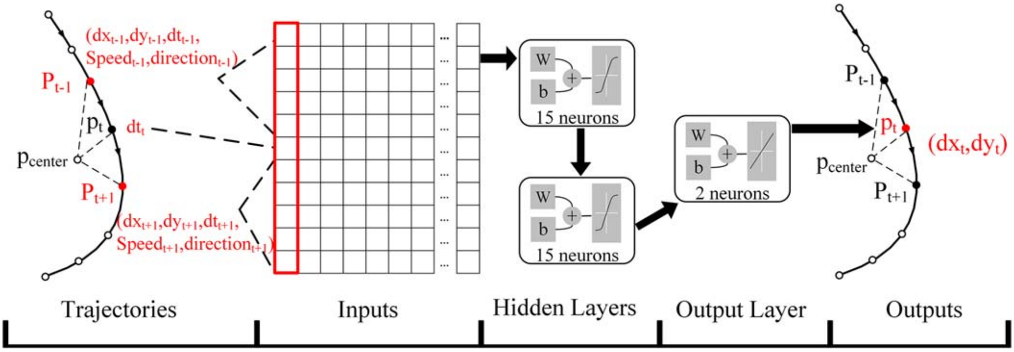

FIGURE 5 The structure of BP neural networks. Point p_{center} is the mass center point of a trajectory. Points p_{t-1}, p_t, p_{t+1} are three consecutive points along a specific trajectory

trajectory and p_{t-1},p_t,p_{t+1} are three consecutive points along the trajectory. The $speed_{t_i}$ and $direction_{t_i}$ are the speed and direction from point p_{t_i} to point $p_{t_{i+1}}$, respectively. These variables are defined as follows:

$$dx_{t_i} = x_{t_i} - x_{center} \tag{2}$$

$$dy_{t_i} = y_{t_i} - y_{center} \tag{3}$$

$$dt_{t_i} = t_{t_i} - 7.0 \tag{4}$$

$$speed_{t_i} = \frac{abs(dx_{t_{i+1}} - dx_{t_i}) + abs(dy_{t_{i+1}} - dy_{t_i})}{dt_{t_{i+1}} - dt_{t_i}} \tag{5}$$

where dx_{t_i} and dy_{t_i} are, respectively, the x-axis and y-axis components of the distance from point p_{t_i} to center point p_{center}; dt_{t_i} is the time difference between t_{t_i} and 7 a.m., which is regarded as the time for getting out of bed and adhering to daily routines; x_{center} and y_{center} are the coordinate values of the center point p_{center} and abs is the function of the absolute value. In this study, the states (location, speed, direction, and time) of point p_t along a specific trajectory are assumed to be the closest point of its two adjacent points (p_{t-1} and p_{t+1}). Therefore, we can use the states of points p_{t-1} and p_{t+1} as input variables of the neural networks to predict the state of point p_t.

We then tested the influences of the missing data rate and the numbers of training samples used in the BP neural network on the accuracy of trajectory reconstruction. On the one hand, a trajectory reconstructed with fewer records (i.e., higher missing data rate) tends to be less accurate. The missing data rate of each trajectory was calculated as the ratio of the numbers of points not sampled to the total number of original points along the trajectory. First, we randomly selected 2,000 trajectories with dense records from each mobility pattern, respectively. Then, we sampled a certain number of points along each trajectory at different intervals by using the inverse transform sampling method (Olver & Townsend, 2013), which generates pseudo-random points based on the time probability function of the trajectory points of a specific user given its CDF. Further, we used BP neural networks to predict the positions of all points along the trajectory based on the states of sampled points. In the end, the mean and variance of the difference between the original points along the trajectory and the predicted points were calculated and used to assess the prediction accuracy. On the other hand, the prediction accuracy can also be affected by the number of training samples. First, we trained the BP neural networks of each mobility pattern with different numbers of user trajectories from 2,000 to 20,000. Then, another 2,000 trajectories in the mobility pattern were further selected as independent test samples, of which all the trajectory points were predicted using the BP neural network. Finally, prediction errors were calculated as the distance between the true and predicted points along each specific trajectory.

The trajectory reconstruction accuracy of the proposed method was then compared with that of the methods proposed by Hoteit et al. (2014). Three trajectory reconstruction methods, including linear, nearest, and cubic interpolation, were first used to reconstruct the trajectories of users with different mobility patterns and various missing data rates. The average errors and the standard deviation for trajectory reconstruction were then calculated for comparison purposes.

4.3 | Hourly dynamic population mapping

Once the hourly time-series user trajectories during the daytime (from 7:00 a.m. to 8:00 p.m.) and at night (from 8:00 p.m. to 7:00 a.m.) were reconstructed, we mapped the hourly population distribution. We first calculated the mobile user density σ_i within the Voronoi polygon v_i of base station i:

$$\sigma_i = {}^{C_i}\!/_{A_i} \tag{6}$$

where C_i is the total number of users located within the Voronoi polygon of base station i and A_i is the area of the Voronoi polygon (v_i) that is constructed right around the base station i.

We then calculated the hourly mobile user density σ_{s_j} for sub-district s_j using the method proposed by Deville et al. (2014):

$$\sigma_{s_j} = \frac{1}{A_{s_j}} \sum_{v_i} \sigma_{v_i} A_{(v_i \cap s_j)} \tag{7}$$

where A_{s_j} is the total area of all Voronoi polygons with active mobile phone users within a sub-district s_j; σ_{v_i} is the hourly user density within Voronoi polygon v_i; and $A_{(v_i \cap s_j)}$ is the intersecting area between the Voronoi polygon v_i and sub-district s_j.

We then modeled the relationship between the hourly active mobile user density σ_{s_j} and the corresponding census population density (ρ_{s_j}) in each sub-district using the following equation:

$$\rho_s = \alpha {\sigma_s}^{\beta} \tag{8}$$

where $\rho_s = [\rho_{s_1}, \rho_{s_2}, \ldots, \rho_{s_n}]$ and $\sigma_s = [\sigma_{s_1}, \sigma_{s_2}, \ldots, \sigma_{s_n}]$, α is the scale ratio, and β is the super-linear effect of population density ρ_s on the hourly active mobile user density σ_s. The linear least-squares method was then used to estimate α and β after the model was transformed into a linear regression equation:

$$\log(\rho_s) = \log(\alpha) + \beta \log(\sigma_s) \tag{9}$$

Finally, we used Equation 7 to estimate the hourly population density $\tilde{\rho}_s$ of all sub-districts and then extracted the total population approximation $\tilde{P}$ within each Voronoi polygon at each time point. Due to the difficulty of obtaining real-time changes in total population during the day in the study area, we assumed that the total estimated population was consistent with the total census population within a day. We finely tuned α and β using Equation 9 to ensure the total estimated population $\tilde{P}$ is consistent with the total census population P :

$$\rho_s = \frac{P}{\tilde{P}} \alpha {\sigma_s}^{\beta} \tag{10}$$

To assess the accuracy of the population estimates, the Pearson product–moment correlation coefficient (r) and the root-mean-square error (RMSE) between the estimated and the census population density were calculated using the 10-fold cross-validation method. First, all 317 administrative units were randomly partitioned into 10 equal-sized sub-samples. Then, a single sub-sample, which contained 31 sub-districts, out of the 10 sub-samples was retained as the validation data for testing the model & calculating r and RMSE. The remaining nine sub-samples, which contained 286 sub-districts, were used as training data to estimate α and β. Further, the cross-validation process was repeated 10 times until each of the 10 sub-samples had been used once as the validation data. We repeated the 10-fold cross-validation 100 times. Finally, RMSE and r were calculated to assess the accuracy of the population estimates. We also calculated the variations in α and β at different times using the 10-fold cross-validation method.

4.4 | Performance test of the proposed method for estimating population

To test the performance of the proposed method for estimating population, we compared the proposed method with the method proposed by Deville et al. (2014), which used active users and active records to estimate population. We first counted the hourly numbers of active records, active users, and predicted users at the cell tower level from our CDR database. Then, the active users and records, and the predicted users, were used to estimate the hourly population based on the method proposed by Deville et al. (2014) and the method proposed in this study, respectively. Finally, we calculated *r* and RMSE of the population estimates using the 10-fold cross-validation method. Two statistics were used to test whether the performance results (*r* and RMSE) among three different methods are significantly different or not. The Shapiro–Wilk test method (Shapiro & Wilk, 1965) was first used to test the normality of the distribution of *r* and RMSE. The Wilcoxon rank sum test (Mann & Whitney, 1947; Wilcoxon, Katti, & Wilcox, 1970) was then used to determine whether the differences in RMSE and *r* among the three different methods are statistically significant or not.

4.5 | Detection of dynamic population patterns

A hierarchical clustering algorithm (Murtagh & Legendre, 2014; Rokach & Maimon, 2005) was used to group the sub-districts based on time-series similarity of the hourly diurnal population estimates. The clustering algorithm is a top-down approach, which first assigns each observation to a cluster. Second, the similarity (e.g., correlation distance) between each pair of clusters is then computed. Third, the two most similar clusters are joined, based on a specific clustering method (e.g., complete linkage). In complete linkage clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. Finally, the second and third steps are repeated until there is only a single cluster left. One of the most important prerequisites of this algorithm is to calculate the proximity matrix containing the distance between each pair of points using the distance function. In this study, we first applied the zero-mean unit-variance normalization (Hyvärinen & Oja, 1997) to the time series of hourly diurnal population estimates at sub-district level. We then created a similarity matrix by calculating the pairwise correlation distance between any pair of normalized time-series hourly population estimates. The correlation distance is used to measure both the linear and non-linear correlation between two paired time series. The correlation distance is zero if and only if two paired time series are independent. In the end, the hierarchical clustering algorithm based on the complete linkage clustering method was applied to detect different diurnal patterns of the population dynamic based on the similarity matrix.

5 | RESULTS AND DISCUSSION

This section is organized as follows: we first show the results of population estimates at different times. The accuracy assessment results of trajectory reconstruction under different missing data rates and numbers of training samples used in BP neural network analysis are then discussed. The accuracy assessment results (*r* and RMSE) of the estimated population are then analyzed. The performance of the proposed method in comparison with other methods in estimating population is then shown. Finally, five distinct diurnal dynamic patterns of population mobility in Beijing are presented.

5.1 | Hourly population estimates

Figure 6 shows the results of population estimates at 10 a.m. (Figures 6a,c) and at night (Figures 6b,d), respectively. Population density shows similar patterns both at 10 a.m. and at night. Outside the 6th ring road, the population density is relatively low, except that a couple of towns such as Miyun (region A in Figures 6a,b), Huairou, Yanqing, Shunyi, and Pinggu show medium population density. The northern and southwestern mountainous areas have extremely low population density. Population density is higher within the 6th ring road, and even higher in areas closer to the center of the city.

There is no significant difference in population density in areas outside the 6th ring road at 10 a.m. and at night, with higher population density around the major towns. Within the 6th ring road, the population density shows a more clustered pattern (Figure 6c) at 10 a.m. and a more dispersed one at night (Figure 6d). In the area between the 6th and 5th ring roads, where major residential communities are located, the population density at night is higher than at 10 a.m. For example, in Tiantongyuan residential community (region C in Figures 6c,d), the population density at night

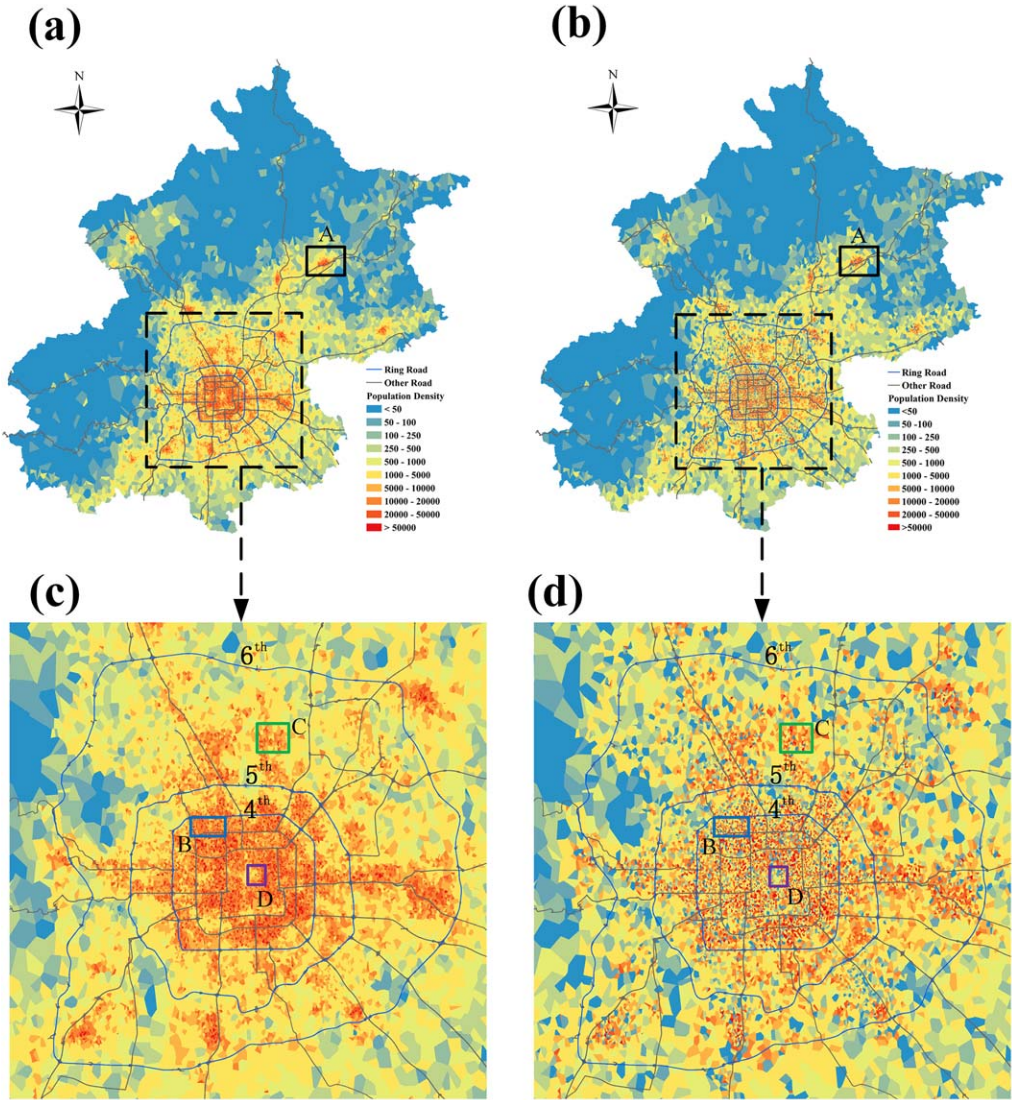

FIGURE 6 Population estimates at the cell tower level at 10 a.m. and at night: (a), (b) The population distribution across our study area at 10 a.m. and at night, respectively; (c), (d) The population distribution within the 6th ring road at 10 a.m. and at night, respectively. The blue rings are the 6th, 5th, and 4th ring roads, respectively. Regions A, B, C, and D are Miyun town, Zhongguancun sub-district, Tiantongyuan community, and the Palace Museum

is 6.6% higher than that at 10 a.m. By contrast, commercial and service areas tend to show an opposite trend of population density. For example, Zhongguancun sub-district (region B in Figures 6c,d), which is known as the Silicon Valley of China, shows a much lower (15.7%) population density at night than at 10 a.m. In areas with sightseeing attractions, such as the Palace Museum (region D in Figures 6c,d), the population density at 10 a.m. was significantly higher than that at night.

5.2 | Accuracy assessment of trajectory reconstruction

Trajectory reconstruction is affected by two factors: the missing data rate and the number of training samples used in BP neural network analysis. The trajectories reconstructed under higher missing data rates tend to be less accurate (Figure 7). The overall error in trajectory reconstruction (Figure 8) gradually increases as fewer points (i.e., increased

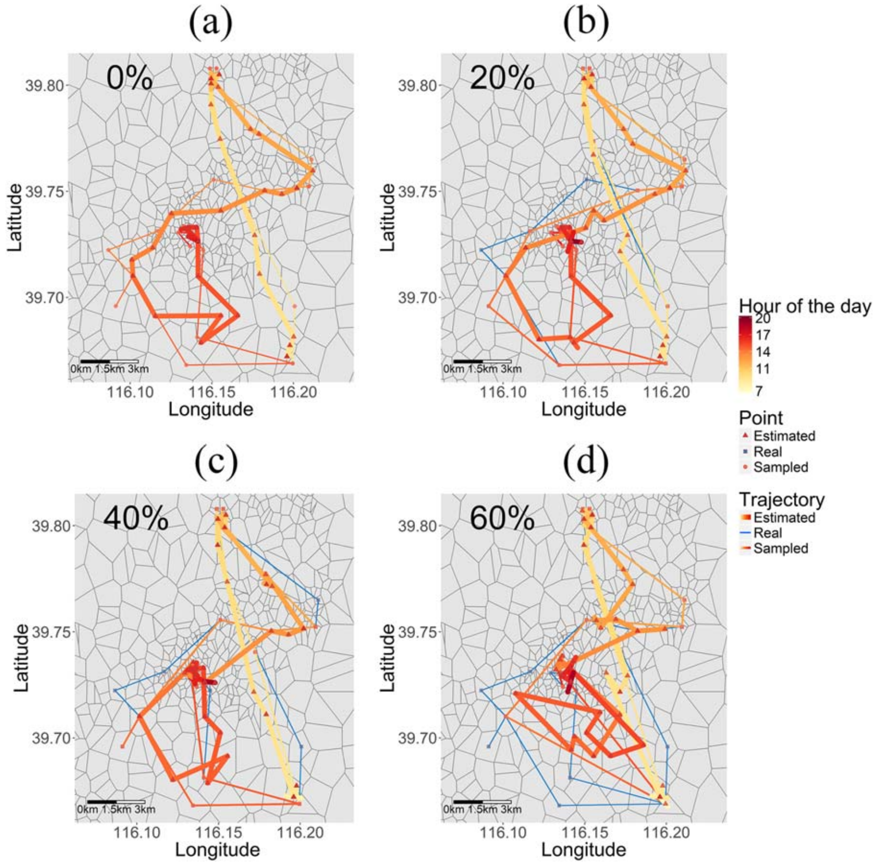

FIGURE 7 Reconstructed trajectories using different missing data rates of 0% (a), 20% (b), 40% (c), and 60% (d). The fine blue lines show the trajectories that were reconstructed from original mobile call records. The red/yellow thick/fine lines represent estimated trajectories and the trajectories after data sampling

missing data rate) were used to reconstruct the trajectories. At the same missing data rate, the commuters' trajectories always show a larger reconstruction error than that of the urban and sedentary residents. There are no significant variations regarding the average reconstruction error of the sedentary residents' trajectories, mainly due to the small ROG. However, the reconstruction errors of the commuters' and urban residents' trajectories gradually increase with increased missing data rate. The overall reconstruction errors in these two groups of people increase significantly once the missing data rate is higher than 40%. Therefore, in this study, we selected the trajectories that were reconstructed with missing data rate no more than 40%. In total, we selected and used 2,826,366 trajectories (10.4% of all user

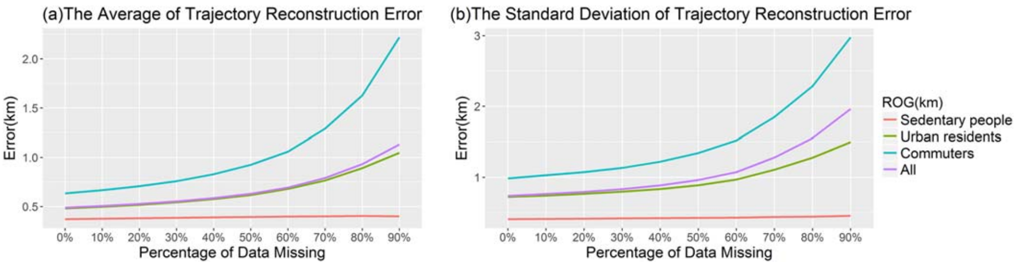

FIGURE 8 (a) The average of trajectory reconstruction errors with different missing data rates; and (b) the standard deviation of trajectory reconstruction errors with different missing data rates. Here, "All" means all the trajectories that were chosen to assess the trajectory reconstruction errors

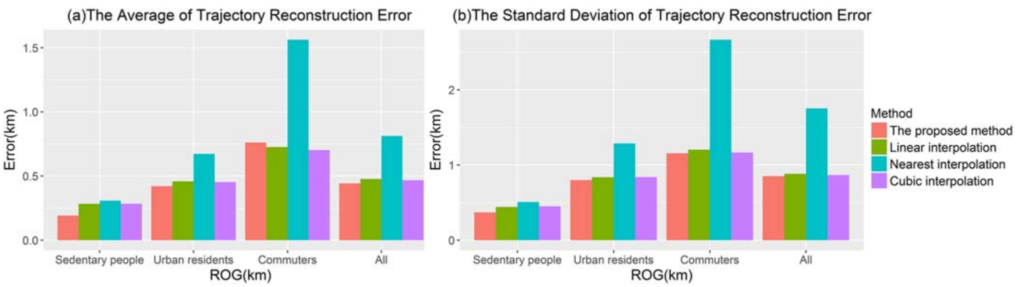

FIGURE 9 (a) The average errors of trajectory reconstruction for different methods; and (b) the standard deviation of trajectory reconstruction for different methods. Here, the missing data rate is set to 40% and a target user is regarded as staying in the same location while two consecutive locations remain unchanged

records) to estimate an hourly population distribution. The trajectory reconstruction accuracy of the proposed method was confirmed using the methods proposed by Hoteit et al. (2014). The proposed method has the best overall performance, except for commuter data where both linear and cubic interpolations give a competitive result (Figure 9).

We also studied how the prediction error was affected by the number of training samples used in the BP neural networks. The average prediction errors of the three mobility patterns fluctuated around 400, 580, and 910 m, respectively. The overall mean error of all three patterns was around 620 m. The standardized *z*-score values of the prediction errors of the commuters' trajectories dropped significantly from 2.8 to 0.1 as the training sample increased from 2,000 to 4,000 and stabilized at −0.4 after the number of training samples reached 6,000 (Figure 10a). By contrast, the standardized *z*-score values of the other two mobility patterns dropped to less than −1.0 when the BP neural networks were trained with 20,000 samples. We also standardized the standard deviation of the prediction errors. The standardized *z*-score values gradually dropped as the number of training samples increased and reached a minimum value of around −1 when the BP neural networks were trained with 20,000 samples (Figure 10b). In short, the standardized *z*-scores of the mean and standard deviation of the prediction errors gradually decreased with increasing training samples and reached their minimum values when the BP neural networks were trained with 20,000 samples. As a result, we chose 20,000 trajectories from each mobility pattern to train the BP neural networks, which were then used to predict the trajectories of sedentary residents, urban residents, and commuters, respectively.

5.3 | Accuracy assessment of population estimates

As expected, the population density reconstructed from nighttime active mobile users and from the census population counts is very similar (Figure 11a). There is a log-linear relationship between nocturnal mobile user density and census population density at sub-district level, and their frequency distribution histograms are very similar. The estimated population density also shows a log-linear relationship with the census population density (Figure 11b). Population estimates are more accurate in the sub-districts with lower population density than those with higher population density, as indicated by the width of the clouds in Figure 11b. Moreover, population densities tend to be underestimated in low-density and high-density areas, while at middle-density areas they tend to be overestimated (Figure 11b).

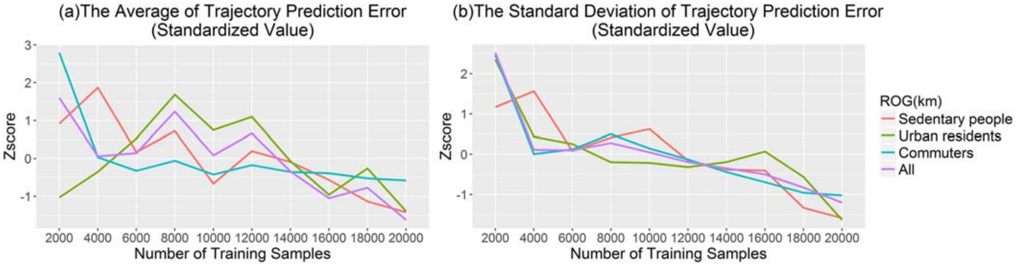

FIGURE 10 (a) Variations in the *z*-score of the average prediction error and (b) variations in the *z*-score of the standard deviation of the prediction error based on different numbers of samples used to train the BP neural networks

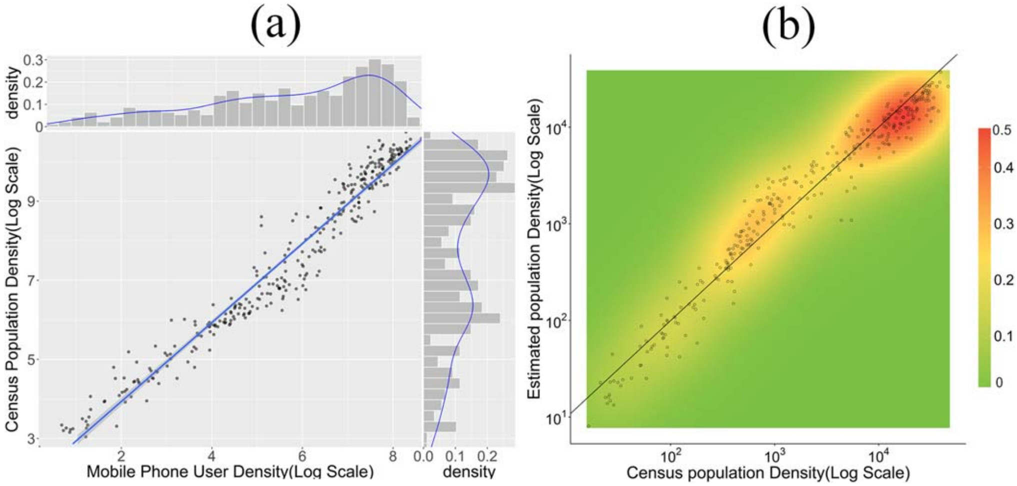

FIGURE 11 (a) The log-linear relationship between nocturnal mobile user density and census population density at the sub-district level; and (b) the log-linear relationship between census population density and estimated population density at the sub-district level. The slope of the black line is 1, and the color represents the density of the data points

Figure 12 shows the accuracy assessment results (*r* and RMSE) of the estimated population using the 10-fold cross-validation method. The figure is a combination of a box plot and a density plot, rotated and placed on each side to show the distribution shape of the data. The thick black bars at the center represent the interquartile range, the thin black lines represent the 95% confidence intervals, and the white dots show the medians. The global RMSE values range between 1,423 and 8,008 with a mean of 4,034, indicating a strong log-linear correlation between the estimated and the census population density at sub-district level. The *r* ranges between 0.82 and 0.99 with a mean of 0.94. There are no significant variations in RMSE and *r* over time, suggesting a similar performance of the model in estimating population over time.

The 10-fold cross-validation method is also used to quantify how the population estimates could be affected by different values of α and β, which are also affected by the training dataset. Figure 13 shows that the α used to model the nighttime population ranged between 6.21 and 7.37 with a mean value of 6.96. The β ranged from 0.98 to 1.01 with a mean value of 0.99. For the daytime population estimates, α and β ranged from 7.3 to 8.1 and from 0.97 to 0.99, respectively. Although α and β are gradually increasing and decreasing from early morning to evening, respectively, there are no significant fluctuations of the corresponding RMSE and *r* values (Figure 12).

5.4 | Performance of the proposed method in estimating population

Performance results, including *r* and RMSE, when estimating the population by different methods are shown in Figure 14. The mean RMSE values of the estimated nighttime population using predicted users, active users, and active

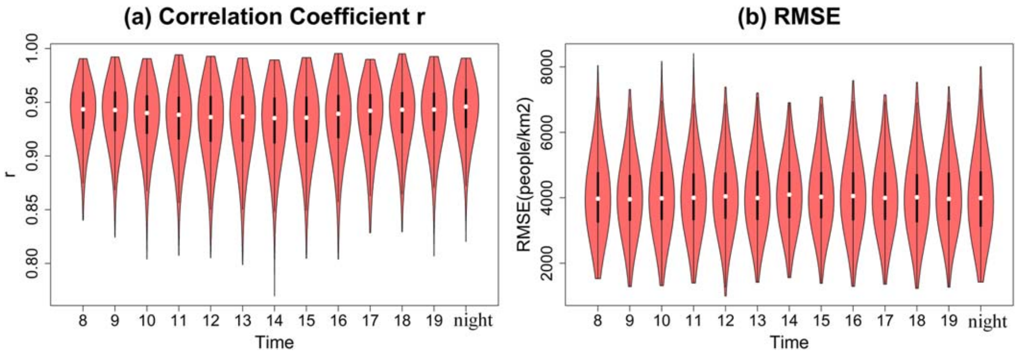

FIGURE 12 The correlation coefficient *r* (a) and RMSE (b) between hourly estimated and census population

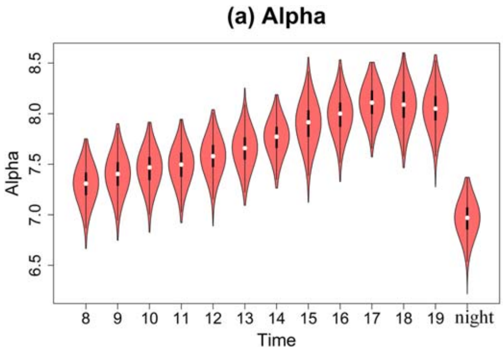

FIGURE 13 Variations in α (a) and β (b) calculated at different times using the 10-fold cross-validation method

records is 4,034, 4,363, and 4,684, respectively. The average values of r between the census population and the estimated population based on the predicted users, active users, and active records are 0.942, 0.911, and 0.934, respectively. The Shapiro–Wilk test results show significantly that the values of RMSE and r calculated by the three different methods are not normally distributed. The Wilcoxon rank sum test shows that the differences in RMSE and r among the three different methods are statistically significant ($p < 0.001$ for all), and the method proposed in this study is superior to methods that use active users and records in terms of estimating the population at sub-district level.

5.5 | Dynamic population patterns

The results of hierarchical clustering show that there are five different patterns of diurnal population dynamic in Beijing (Figures 15 and 16). Pattern 1 is characterized by a gradual increase in population density, as suggested by the z-score values, from morning until 3 p.m., when the z-score value reaches a maximum at 0.85. After 3 p.m., the population density drops slowly. Pattern 3 is totally opposite to pattern 1 (i.e., the population density is gradually decreasing from morning to 3 p.m., when the z-score value reaches its minimum value of -0.76). The population density starts to drop after 3 p.m. Pattern 2 is characterized by a rapid increase in population density from morning to 12 p.m. and then a gradual drop after noon. The population density z-score value reaches its maximum value of 0.95 at 12 p.m. By contrast, pattern 4 is totally opposite to pattern 3 and shows its lowest population density z-score value of -0.69 at 12 p.m. Pattern 5 shows a gradual decrease in population density z-score value from 8 a.m. to 7 p.m. These different patterns of population dynamic probably indicate the various flow patterns of different groups of people living in Beijing. Patterns 1 and 2 indicate that a certain group of people starts to flow into these sub-district areas in the morning and to leave after 12 p.m. or 3 p.m. Patterns 3 and 4 show a totally opposite flow pattern: a specific group of

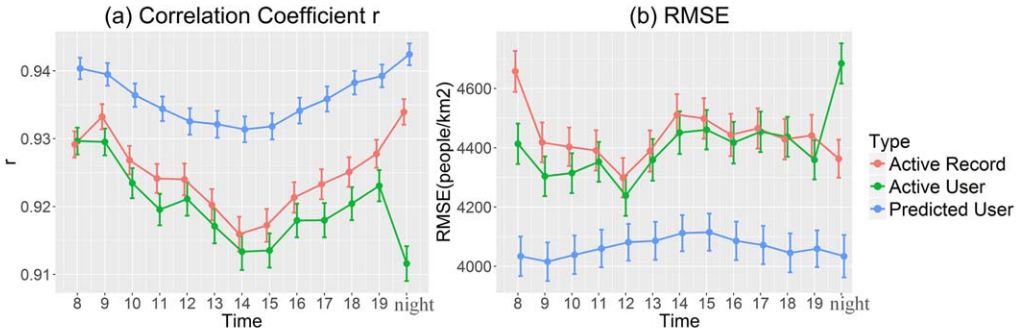

FIGURE 14 The correlation coefficient r (a) and RMSE (b) between census population and hourly population estimated from active users, active records, and predicted users. Better performance is achieved in estimating hourly population using the method proposed in this study based on predicted users, as indicated by a higher r (the blue line in Figure 14a) and lower RMSE (the blue line in Figure 14b) than those of the hourly population estimated using active users (green lines) and active records (red lines)

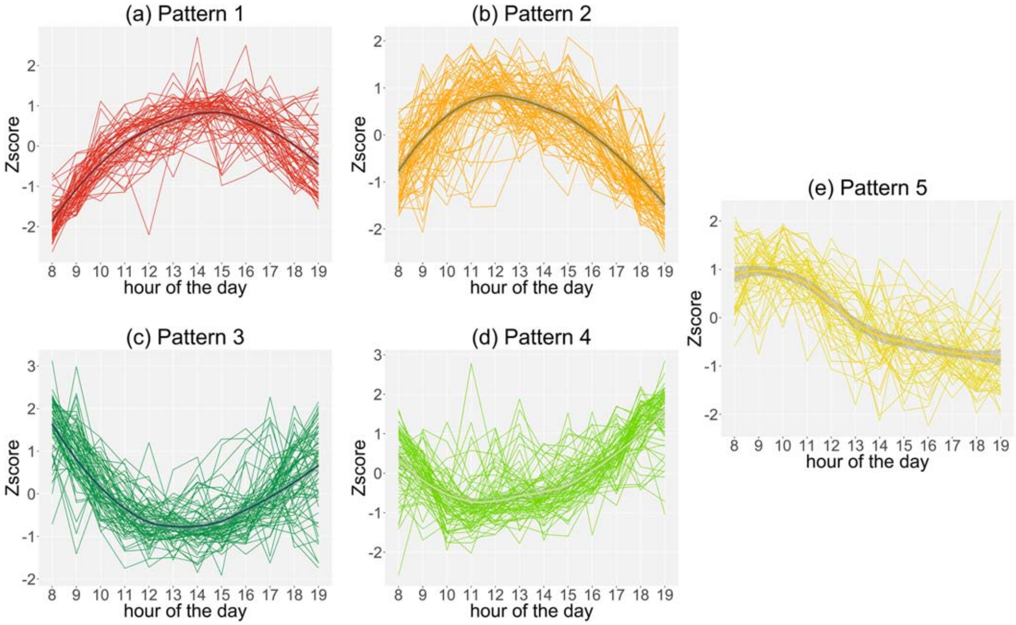

FIGURE 15 The clustering results using the time series of the hourly estimated population distributions at sub-district level: (a)–(e) Five different patterns of diurnal population dynamic in Beijing. The thick lines in the middle show the LOESS curve with 95% confidence interval

people starts to leave from these sub-districts in the morning and to come back in the afternoon. Pattern 5 indicates that people leave these sub-districts slowly during the daytime. A global Moran's *I* value of 0.055 ($p < 0.001$, z-score = 2.87) indicates a spatially clustered distribution of the sub-districts in terms of population density.

The above mentioned five patterns of population flow generally divide the whole city into three regions. The first region is the rural areas, which are far from the 6th ring road, with population density less than 500 people/km^2. The mobility of people living in the rural areas is characterized by patterns 1, 2, and 5. The second region is the suburbs, which are located between the 5th and 6th ring roads, with population density between 500 and 5,000 people/km^2. The mobility of people living in the suburbs is mainly characterized by patterns 3 and 4. The third region is the urban core areas within the 5th ring road, with population density higher than 5,000 people/km^2. The mobility of people living in the urban core areas is mainly characterized by patterns 1 and 2.

Regarding land-use categories, the sub-districts with population mobility characterized by patterns 1 and 2 are likely to be commercial and services land. The areas with population mobility patterns 3 and 4 are likely to be residential areas, whereas the sub-districts characterized by population mobility pattern 5 are likely to be sightseeing areas. Sub-districts within the 5th ring road are mainly characterized by patterns 1 and 2, whereas those between the 5th and 6th ring roads are mainly characterized by patterns 3 and 4, possibly suggesting that most people living in these sub-districts are commuting between their workplaces within the 5th ring road and their residences between the 5th and 6th ring roads. We will investigate further the relationship between people flow and land use in future work.

Previous studies on the dynamic patterns of urban mobility included urban planning and morphology, urban clusters and spread, and urban rhythms (Demissie, Correia, & Bento, 2015; Kang, Ma, Tong, & Liu, 2012; Sagl, Delmelle, & Delmelle, 2014; Xu et al., 2015; Yuan & Raubal, 2016). Particularly relevant for our study in the context of urban mobility patterns were the contributions made by Yuan and Raubal (2012). They first extracted and represented the dynamic mobility patterns in different urban areas using the time series of hourly phone call frequency patterns. A dynamic time warping (DTW) algorithm was then applied to measure the similarity (distance matrix) between these time series. Finally, a hierarchical clustering algorithm was used to classify different urban areas and detect outlier urban areas. All the studies on dynamic patterns of urban mobility using CDR data had to face a common problem: the potential representativeness bias in CDR data. The problem of data representativeness also exists in similar studies

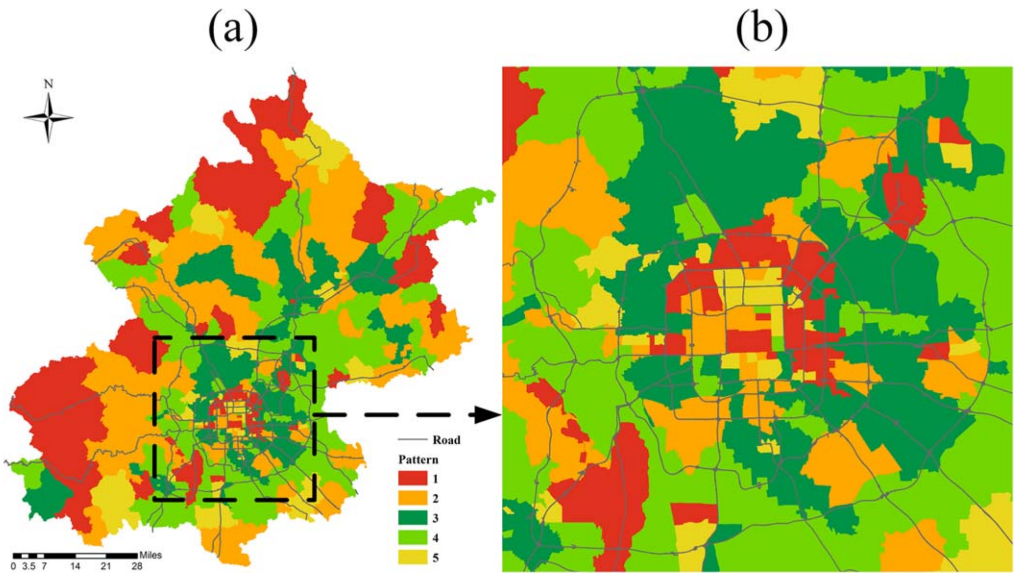

FIGURE 16 The spatial distribution of five diurnal dynamic patterns of estimated population distributions at sub-district level across our study area (a) and within the 6th ring road (b). The color of each pattern corresponds to the color in Figure 15

using social media data (Domínguez, Redondo, Vilas, & Khalifa, 2017). People with limited phone activities are underrepresented in CDR data (Yuan & Raubal, 2016). The intermittent or discontinuous records of users' locations bias the aggregation of human activities based on Voronoi polygons or other spatial units, further resulting in the deviation of urban dynamic mobility patterns. The proposed method provides a possible solution to reduce the potential representativeness bias in CDR data by reconstructing users' time-series trajectories and estimating the dynamic population. Previous studies on identifying the commuting pattern and quantifying the jobs–housing balance (Furletti et al., 2014; Long et al., 2012; Zhang, Zhou, & Zhang, 2017) focused mainly on the spatiotemporal analysis of the commuters' behavior between their homes and workplaces. Such studies essentially examine urban residents' mobility patterns and therefore the population dynamics at a coarse temporal resolution. However, little research has been conducted to explore the diurnal dynamic urban mobility patterns at fine temporal resolutions, mainly due to a lack of appropriate data. The results on hourly population in this study thus fill a gap, and can be used to examine the population dynamics at fine spatial and temporal resolutions. The hourly population dynamic is also useful in studies delineating urban functional areas. Previous research on urban functional areas delineation relied mainly on static information (such as remotely sensed data and POIs) on street blocks (Chen et al., 2017; Yuan et al., 2015), and information about human activity is usually lacking. Human activities tend to show similar spatiotemporal patterns in street blocks with similar functions (Chen et al., 2017). The time series of hourly population estimates can be used to group blocks based on similar patterns and then delineate urban functional areas. Dynamic population estimation can also be beneficial to studies of natural disaster relief and impact assessments, in which population dynamics at high temporal resolution are usually not considered (Bengtsson et al., 2011; Min & Jeong, 2013; Wilson et al., 2016).

6 | CONCLUSIONS

This study proposed a method to estimate hourly population based on the reconstructed time series of individual trajectories using BP neural networks. Evaluation results show that the estimated population is well correlated with the census population at the sub-district level, as indicated by a linear *r* of 0.94. The RMSE between the census population counts and the estimated population based on reconstructed trajectories is 4,034, which is lower than the RMSE values calculated based on active users and active records at night, respectively. The hourly population density at the cell tower level shows significant variation across our study area, and five diurnal dynamic patterns were identified at the sub-district level in Beijing. The patterns clearly reveal the general flow characteristics of the population among sub-districts. The findings in this study will help us to better understand population dynamics at very high spatial and temporal resolution. The driving forces of diurnal population flow will be investigated further in future work.

ACKNOWLEDGMENTS

This research was supported by the Major Program of the National Natural Science Foundation of China (Nos. 4159840011 and 41771418), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 41421001), the Innovation Program of the State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences (No. 088RA500PA), and the Distinguished Young Scholar Program of the Institute of Geographic Sciences and Natural Resources Research, CAS (No. 2014RC102).

CONFLICT OF INTEREST

We declare that we have no financial or personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service, and/or company that could be construed as influencing the position presented in, or the review of, this manuscript.

ORCID

Zhang Liu http://orcid.org/0000-0001-7531-2162

REFERENCES

Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *IFAC Proceedings*, *39*(3), 399–404.

Balk, D., Deichmann, U., Yetman, G., Pozzi, F., Hay, S., & Nelson, A. (2006). Determining global population distribution: Methods, applications and data. *Advances in Parasitology*, *62*, 119–156.

Balk, D., & Yetman, G. (2004). *The global distribution of population: Evaluating the gains in resolution refinement*. New York, NY: Center for International Earth Science Information Network, Columbia University.

Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*, 207–211.

Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A., & Volinsky, C. (2011). A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, *10*(4), 18–26.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Med*, *8*(8), e1001083.

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, *69*(1&2), 103–117.

Chen, X., Chi, R., Wang, J., Pang, C., & Xiao, Q. (2011). Driver safe speed model based on BP neural network for rural curved roads. In D. Li & Y. Chen (Eds.), *Computer and Computing Technologies in Agriculture V: 5th IFIP TC 5/SIG 5.1 Conference, CCTA 2011, Beijing, China, October 29–31, 2011, Proceedings, Part III* (pp. 92–102). Berlin, Germany: Springer.

Chen, Y., Liu, X., Li, X., Liu, X., Yao, Y., Hu, G., . . . Pei, F. (2017). Delineating urban functional areas with building-level social media data: A dynamic time warping (DTW) distance based *k*-medoids method. *Landscape & Urban Planning*, *160*, 48–60.

De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., & Lepri, B. (2016). The death and life of great Italian cities: A mobile phone data perspective. In *Proceedings of the 25th International Conference on World Wide Web*. Montreal, Canada: ACM.

Deichmann, U., Balk, D., & Yetman, G. (2001). *Transforming population data for interdisciplinary usages: From census to grid*. Washington, DC: Center for International Earth Science Information Network.

Demissie, M. G., Correia, G., & Bento, C. (2015). Analysis of the pattern and intensity of urban activities through aggregate cellphone usage. *Transportmetrica A: Transport Science*, *11*(6), 502–524.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., . . . Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, *111* (45), 15888–15893.

Ding, C., Wang, W., Wang, X., & Baumann, M. (2013). A neural network model for driver's lane-changing trajectory prediction in urban traffic flow. *Mathematical Problems in Engineering*, *2013*, 967358.

Domínguez, D. R., Redondo, R. P. D., Vilas, A. F., & Khalifa, M. B. (2017). Sensing the city with instagram: Clustering geolocated data for outlier detection. *Expert Systems with Applications*, *78*, 319–333.

Dong, Y., Pinelli, F., Gkoufas, Y., Nabi, Z., Calabrese, F., & Chawla, N. V. (2015). Inferring unusual crowd events from mobile phone call detail records. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Porto, Portugal: Springer.

Douglass, R. W., Meyer, D. A., Ram, M., Rideout, D., & Song, D. (2015). High resolution population estimates from telecommunications data. *EPJ Data Science*, *4*(1), 4.

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., & Baptista, S. R. (2015). Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography*, *1*(3), 226–234.

Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(36), 15274–15278.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography & Geographic Information Science*, *28*(2), 125–138.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., . . . Pépin, J. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, *346*, 56–61.

Furletti, B., Gabrielli, L., Giannotti, F., Milli, L., Nanni, M., Pedreschi, D., . . . Garofalo, G. (2014). Use of mobile phone data to estimate mobility flows: Measuring urban population and inter-city mobility using big data in an integrated approach. In *Proceedings of the 47th Meeting of the Italian Statistical Society*. Cagliari, Italy.

Hecht-Nielsen, R. (1988). Theory of the backpropagation neural network. *Neural Networks*, *1*(Suppl. 1), 445–448.

Hoteit, S., Chen, G., Viana, A., & Fiore, M. (2016). Filling the gaps: On the completion of sparse call detail records for mobility analysis. ACM Chants. In *Proceedings of the 11th ACM Workshop on Challenged Networks* (pp. 45–50). New York, NY: ACM.

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., & Pujolle, G. (2014). Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, *64*, 296–307.

Hyvärinen, A., & Oja, E. (1997). One-unit learning rules for independent component analysis. In *Proceedings of the 10th Conference on Advances in Neural Information Processing Systems*. Denver, CO.

Kang, C., Liu, Y., Ma, X., & Wu, L. (2012). Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, *19*(4), 3–21.

Kang, C., Ma, X., Tong, D., & Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics & Its Applications*, *391*(4), 1702–1717.

Khodabandelou, G., Gauthier, V., El-Yacoubi, M., & Fiore, M. (2016). Population estimation from mobile network traffic metadata. In *Proceedings of the 17th International IEEE Symposium on the World of Wireless, Mobile and Multimedia Networks*. Coimbra, Portugal: IEEE.

Le Cun, Y. (1988). A theoretical framework for back-propagation. In D. Touresky, G. Hinton, & T. Sejnowski (Eds.), *Proceedings of the 1988 Connectionist Models Summer School* (pp. 21–28). San Mateo, CA: Morgan Kauffman Publishers.

Levin, B. (1981). A representation for multinomial cumulative distribution functions. *Annals of Statistics*, *9*(5), 1123–1126.

Li, Q., Xu, B., Ma, Y., & Chung, T. (2015). Real-time monitoring and forecast of active population density using mobile phone data. In W. Chen, G. Yin, G. Zhao, Q. Han, W. Jing, G. Sun, & Z. Lu (Eds.), *Big Data Technology and Applications: First National Conference, BDTA 2015, Harbin, China, December 25–26, 2015, Proceedings*. Berlin, Germany: Springer.

Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS One*, *9*(1), e86026.

Lloyd, C. T., Sorichetta, A., & Tatem, A. J. (2017). High resolution global gridded data for use in population studies. *Scientific Data*, *4*, 170001.

Long, Y., Zhang, Y., & Cui, C. (2012). Identifying commuting pattern of Beijing using bus smart card data. *Acta Geographica Sinica*, *67*(10), 1339–1352.

Lopez, D., Gunasekaran, M., Murugan, B. S., Kaur, H., & Abbas, K. M. (2014). Spatial big data analytics of influenza epidemic in Vellore, India. In *Proceedings of the 2014 IEEE International Conference on Big Data*. Washington, DC: IEEE.

Lwin, K. K., Sugiura, K., & Zettsu, K. (2016). Space-time multiple regression model for grid-based population estimation in urban areas. *International Journal of Geographical Information Science*, *30*(8), 1579–1593.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*(1), 50–60.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *Professional Geographer*, *55*(1), 31–42.

Min, G. Y., & Jeong, D. H. (2013). Research on assessment of impact of big data attributes to disaster response decision-making process. *Journal of Society for e-Business Studies*, *18*(3), 17–43.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, *31*(3), 274–295.

Oliveira, J. G., & Barabási, A.-L. (2005). Human dynamics: Darwin and Einstein correspondence patterns. *Nature*, *437*, 1251.

Olver, S., & Townsend, A. (2013). Fast inverse transform sampling in one and two dimensions. arXiv preprint, arxiv: 1307.1223.

Partsinevelos, P., Agouris, P., & Stefanidis, A. (2005). Reconstructing spatiotemporal trajectories from sparse data. *ISPRS Journal of Photogrammetry & Remote Sensing*, *60*(1), 3–16.

Patel, N. N., Stevens, F. R., Huang, Z., Gaughan, A. E., Elyazar, I., & Tatem, A. J. (2016). Improving large area population mapping using geotweet densities. *Transactions in GIS*, *21*(2), 317–331.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, *28*(9), 1988–2007.

Rokach, L., & Maimon, O. (2005). Clustering methods. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 321–352). Berlin, Germany: Springer.

Sagl, G., Delmelle, E., & Delmelle, E. (2014). Mapping collective human activity in an urban environment based on mobile phone data. *Cartography & Geographic Information Science*, *41*(3), 272–285.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3&4), 591–611.

Sterly, H., Hennig, B., & Dongo, K. (2013). "Calling Abidjan": Improving population estimations with mobile communication data. In V. Blondel, N. de Cordes, A. Decuyper, P. Deville, J. Raguenez, & Z. Smoreda (Eds.), *Mobile phone data for development: Analysis of mobile phone datasets for the development of Ivory Coast* (pp. 108–114). Cambridge, MA: MIT Media Lab.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS One*, *10*(2), e0107042.

Tao, S., Corcoran, J., Mateo-Babiano, I., & Rohde, D. (2014). Exploring bus rapid transit passenger travel behaviour using big data. *Applied Geography*, *53*, 90–104.

Tatem, A. J., Noor, A. M., Von Hagen, C., Di Gregorio, A., & Hay, S. I. (2007). High resolution population maps for low income nations: Combining land cover and census in East Africa. *PLoS One*, *2*(12), e1298.

Traag, V. A., Browet, A., Calabrese, F., & Morlot, F. (2011). Social event detection in massive mobile phone data using probabilistic location inference. In *Proceedings of the 3rd International Conference on Privacy, Security, Risk and Trust and the 3rd IEEE International Conference on Social Computing*. Boston, MA: IEEE.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science*, *325*(5939), 425–428.

Wilcoxon, F., Katti, S., & Wilcox, R. A. (1970). Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. In H. L. Harter & D. B. Owen (Eds.), *Selected tables in mathematical statistics* (Vol. *I*, pp. 171–259). Providence, RI: American Mathematical Society.

Wilson, R., zu Erbach-Schoenberg, E., Albert, M., Power, D., Tudge, S., Gonzalez, M., … Hughes, C. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake. *PLoS Current Disasters*. https://doi.org/10.1371/currents.dis.d073fbece328e4c39087bc086d694b5c.

Xu, T., Li, X., & Claramunt, C. (2018). Trip-oriented travel time prediction (TOTTP) with historical vehicle trajectories. *Frontiers of Earth Science*, *15*, in press.

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., & Li, Q. (2015). Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. *Transportation*, *42*(4), 625–646.

Yang, X., Zhao, Z., & Lu, S. (2016). Exploring spatial–temporal patterns of urban human mobility hotspots. *Sustainability*, *8*(7), 674.

Yao, Y., Liu, X., Li, X., Zhang, J., Liang, Z., Mai, K., & Zhang, Y. (2017). Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science*, *31*(6), 1220–1244.

Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., & Huang, Y. (2010). T-drive: Driving directions based on taxi trajectories. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. San Jose, CA: ACM.

Yuan, N. J., Zheng, Y., Xie, X., Wang, Y., Zheng, K., & Xiong, H. (2015). Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge & Data Engineering*, *27*(3), 712–725.

Yuan, Y., & Raubal, M. (2012). Extracting dynamic urban mobility patterns from mobile phone data. In *Proceedings of the 7th International Conference on Geographic Information Science*. Columbus, OH.

Yuan, Y., & Raubal, M. (2016). Analyzing the distribution of human activity space from mobile phone usage: An individual and urban-oriented study. *International Journal of Geographical Information Science*, *30*(8), 1594–1621.

Zhang, P., Zhou, J., & Zhang, T. (2017). Quantifying and visualizing jobs–housing balance with big data: A case study of Shanghai. *Cities*, *66*, 10–22.

Zhou, J., Pei, H., & Wu, H. (2016). Early warning of human crowds based on query data from Baidu map: Analysis based on Shanghai stampede. arXiv preprint, axXiv:1603.06780.

Zipser, D., & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*(6158), 679–684.

How to cite this article: Liu Z, Ma T, Du Y, Pei T, Yi J, Peng H. Mapping hourly dynamics of urban population using trajectories reconstructed from mobile phone records. *Transactions in GIS*. 2018;00:1–20. https://doi.org/10.1111/tgis.12323