Prediction of Intra-urban Human Mobility by Integrating Regional Functions and Trip Intentions

Shuyang Shi, Lin Wang, *Senior Member, IEEE*, Shuangdie Xu and Xiaofan Wang, *Senior Member, IEEE*

Abstract—Understanding intra-urban human mobility patterns and their potential driving forces are vital to city planning and commercial site selection. In this paper, we first investigate the functions of urban regions and how different region types dynamically influence people's trip decisions. Furthermore, we characterize urban circadian rhythms by time-vary inter-regional transition probabilities between these regions with different functions, and integrate them into intervening opportunity model to predict human mobility. Public transportation card data in Shanghai are used to demonstrate the effectiveness of the model in terms of station passenger flows, travel time and trip flux. By taking regional function into consideration, the proposed model significantly improved the prediction accuracy. Quantitative analysis ulteriorly indicates that trip intentions and regional features are critical elements in trip flux prediction, especially in the afternoon and evening when people have an abundance of opportunities to travel by their own volition. When the function of a certain region changes, our model is able to make reasonable predictions accordingly. The results indicate the importance of considering individual travel motivation and regional function in modeling human mobility. The proposed model could serve as a guide for popularity and trip flux prediction in urban planning and reconstruction.

Index Terms—Human Mobility, Urban Structure, Mobility Predictions, Social Data Analysis

1 Introduction

City is a complex spatial system composed of regions with distinctive features and interregional citizen trip fluxes. Identifying regional features and their impact on intra-urban human movement is crucial to the understanding of urban dynamics and the driving mechanism of travel choice behavior. In recent years, ubiquitous social signal sensors have enabled us to obtain large-scale human mobility data more conveniently and cheaply, so as to quantitatively study people's travel choices in the microcosmic aspect and determine the characteristics of population mobility considering individual travel choices and regional features in the macrocosmic aspect. Currently, human mobility data and models [1] have been widely applied to urban planning [2], [3], [4], traffic flow prediction and optimization [5], [6], [7], [8], [9], early warning and risk evaluation systems [10], [11], [12], [13], and reflecting the socioeconomic characteristics of cities [14], [15], [16].

Researches in the past decade have achieved satisfying results in human mobility pattern and regional feature identification. However, to the best of our knowledge, exploring the connections between these two aspects and combining them dynamically to predict large-scale intra-urban human flow is still an open issue. In real life, people's travel decisions highly depend on travel purposes and departure time. An office worker is more likely to go to a social place rather than another office building after work at 17 o'clock. He/she would also incline to go to a social occasion close to his/her workplace with rich social activities, since it's human nature to visit a place not too far, and/or a place with more choices. This kind of travel characteristics of citizens in different time periods of a day are worthy of further study. Motivated by the above travel strategies, a model based on regional features and intervening opportunity theory is proposed to simulate the intra-urban human mobility observed from more than 100 million pieces of metro passage trip records in Shanghai, China. By analyzing the passengers' entrance/exit fluxes of a certain metro station in different time periods of the day, the region features around the station can be characterized. All sorts of across-region trips constitute the circadian rhythm of the city. The rhythm of the city and the above natural characteristics of human travel constitute the starting point of our model. By comparing the model results with the large-scale trip data of Shanghai Metro, we validate that the proposed model can better predict urban population flow, and can respond to the changes of the major sources of population flow caused by the change of regional functions. This analyzation provides potential guidance for predicting the popularity and impact on the transportation network of regional function changes or new urban area development. More concretely, the major contributions of this work are summarized as follows.

- We capture the circadian rhythms of Shanghai and investigate the regional functions via clustering stations with similar entrance/exit flow features by

- *Shuyang Shi and Lin Wang are with the Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China. E-mail: mail@shishuyang.cn; wanglin@sjtu.edu.cn.*
- *Shuangdie Xu is with State Grid Zhejiang Electric Power Company Hangzhou Power Supply Company, Zhejiang 310026, China. E-mail: xushuangdie@126.com.*
- *Xiaofan Wang is with the Department of Automation, Shanghai Jiao Tong University and the Department of Automation, Shanghai University, Shanghai 200444, China. E-mail: xfwang@sjtu.edu.cn.*
- *Lin Wang and Xiaofan Wang are the corresponding authors of this paper.*

Manuscript received February 27, 2020; revised August 15, 2020.

applying K-Means algorithm. Distinct regional functions are characterized by clusters, which consist of residential areas, comprehensive regions and working regions.

- In order to illustrate the effects of regional features, we propose a combined model by integrating inter-regional transition probability into the intervening opportunity model to predict intra-urban human mobility on an hourly basis. Based on more than 100 million pieces of metro passage trip records in Shanghai, the new model is tested under different cluster numbers. Extensive comparisons show that the model coincide well with the empirical data in multiple aspects, including destination station passenger flow, travel time distribution, trip fluxes and its distribution. The prediction accuracy is significantly improved by introducing regional functions, especially in the afternoon and evening with higher flux entropy when people have more opportunities of autonomous choices.
- We further investigate the effectiveness of the proposed model by exploring the major sources of human flow before and after functional changes. Longhua Station, the function of which changed from "comprehensive" to "working" between the two studied time windows, is taken as a typical example to illustrate how our model can predict the major change in fluxes. Results show that our model can adjust the prediction results according to the change of regional function to conform to the actual situation.

2 RELATED WORK

With the development of big data and IoT Technology, numerous outstanding researches on human mobility and urban structure have emerged in the past decade. In this section, we review relevant literature from the following three aspects.

Human mobility. Classical researches on human mobility mostly focus on population distribution and departure-destination distance. In terms of geographical heterogeneity, classical gravity model performed well on various datasets [17], [18], [19]. Inspired by the job-applying scenario, [20] pointed out that the number of trips from location i to location j are affected by both the population of the two locations and the total population of the circular area centered at location i with a radius of the distance between i and j. The results of this radiation model is different approaches but equally satisfactory results with that of classical intervening opportunities model [21]. Motivated by this finding, several improved population movement prediction models based on diverse spatial scales ranging from city scale to national scale have been proposed [22], [23], [24], [25]. Meanwhile, taking individual mobility characteristics such as memory effect [26] and driving force underlying individual movement [27], [28], [29], [30] into account is of great benefit to our understanding of macro mobility patterns. It is also critical to consider the unique nature of human travel strategy when studying human mobility, since citizens seldom walk randomly like animals [31], [32]. The circadian rhythm of a city is deeply influenced by these travel strategies.

Another way to explore people's travel patterns is from the perspective of statistics and knowledge learning. [33] proposed a multidimensional Markov model and realized precise trajectory prediction. By learning the mobility behaviors of different users, [34] used a Bayesian mixture model to describe users' mobility patterns.

Regional features and functions. The features and characteristics of regions can be extracted and learned from large-scale population flow data. Using pick-up/set-down data extracted from taxi traces at different time intervals in the selected areas as training input, [35] compared several common machine learning algorithms in predicting the functions of areas in Hangzhou, China. [36] analyzed the node number of individual mobility motifs to describe the relationship between motifs and urban land use. Using geolocated social media data, [37], [38] characterized urban landscapes through the K-Means clustering algorithm.

Human mobility patterns and geographical features. Recent technology development has facilitated the obtaining of multi-dimensional urban data. Some researches focused on describing and predicting human mobility patterns by combining various geographical data sources. [39] set up a model based on individual and collectivity's past trajectory and the geographical features of the area to predict individual travel destination. [40], [41] combined the conceptions of functional distance and land-use function complementarity indices based on regional functions with improved gravity model and gave out decent regression results. Using smart card data and POI data, [42] analyzed the regional mobility patterns and used ANN to link the mobility patterns with the regional properties. [43] learned the urban forms of residential communities from heterogeneous human mobility data and POI, and applied it to real estate ranking and restaurant popularity prediction. [44] used the land-use transition matrix to calculate population distribution, and verified their results by gravity model. In this paper, besides geographical features, we also consider departure time and trip intentions to predict human mobility fluxes. We further analyze the fluctuations in prediction accuracy at different times of a day, and explore the reasons behind.

3 DATA ANALYSIS

Shanghai Public Transportation Card can be used on metro, bus, taxi and ferry in Shanghai. The card data could serve as a sample to analyze human mobility in urban areas. We adopt Shanghai Public Transportation Card data provided by the SODA(Shanghai Open Data Apps) as the data source, among which only metro records contain information of both the entry station and the exiting station. Thus we extracted 129,964,604 metro trip records in September, 2016 from the dataset. Each record contains card ID, station name and entering/exiting time. During that period, Shanghai metro system had a total of 14 lines and 289 stations and bore nearly half of the public transportation flow of the whole city. Card IDs are ignored, since we are only interested in the trip fluxes between stations in different time periods rather than individual identity. The regional flow around a

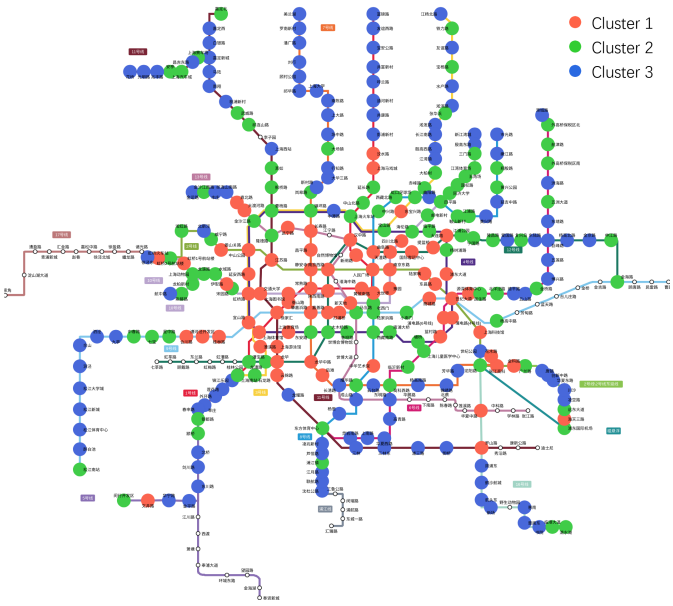


Fig. 1. Metro stations in Shanghai. Stations in the same color are clustered in the same cluster. Newly built stations which haven't been color-filled are beyond the scope of this article. This map is a simplified presentation and partially reflects real geographic location information.

station can be represented by the flow of this station. For simplicity, we refer to the region around station i as *region* i.

Shanghai metro begins operation at around 5 o'clock and stops operation at around 24 o'clock, while the operating period of each station may vary. Thus, we only consider travels starting between 7 and 22 o'clock (a 15-hour-duration) when all stations are open with a non-negligible passenger flow.

We first cluster the stations with similar features according to the entrance/exit dynamics. For each station, the entrance/exit dynamic is defined as a normalized temporal sequence consisting of entrance/exit flow values in the time interval of an hour. Since travel behavior in weekdays and weekends can be rather different [29], [45], [46], we discuss these two cases respectively. Based on the above assumptions, the passenger flow dynamic of a station can be described by a 60-dimensional vector comprised of two 15-dimensional normalized entrance flow subvectors and two 15-dimensional normalized exit flow subvectors on weekdays and weekends, respectively. Specifically, we define $F_\tau^{exit}(d,S)$ as the exit flow in time interval τ of day d in station S and D_1 as the set of the studied weekdays. $\tau \in \{[t,t+1) : t \in [7,21] \cap \mathrm{Z}\}$ represents the one-hour duration that begins at a certain hour on the hour and ends at the next hour on the hour. A normalized element of the exit flow vector on weekdays is thus defined as

$$v_\tau^{exit}(S) = \frac{\sum_{d\in D_1} F_\tau^{exit}(d,S)}{\sum_{d\in D_1}\sum_{\tau'=[7,8)}^{[21,22)} F_{\tau'}^{exit}(d,S)}, \tag{1}$$

and elements in the other three subvectors can be calculated similarly. Together, the four subvectors compose the vector that describes the dynamic features of a metro station.

The K-Means algorithm is applied to cluster stations according to the above 60-dimensional vector. On our dataset, K-Means has the best performance among common unsupervised clustering algorithms regarding 3 clustering evaluation indexes (see Appendix for details). Three clusters are derived and Fig. 1 shows the geographical location of all metro stations in Shanghai and their corresponding clustering results. The city center is covered by stations in clusters 1 and 2. Fig. 2 demonstrates the dynamics of stations in each cluster. Fig. 2(a, b) show that regions in cluster 1 fit the characteristics of working regions, with population inflow peaks in the morning and outflow peaks in the evening, while regions in cluster 2 may be comprehensive regions, with crowd inflow and outflow peaks both in the morning and evening. These comprehensive and working regions are roughly included in the Shanghai Middle Ring Road, in accordance with urban function and traffic planning. Most stations in cluster 3 are located in suburban areas. As is seen in Fig. 2(c), among the passengers departing from stations in cluster 3 on weekdays, there are 2 obvious peaks departing in the morning and arriving after work hours, indicating that these suburban regions are of residential functions. It's worth noticing that the suburban areas also own several stations in cluster 1 and cluster 2. These stations are the subcentrals of the corresponding regions, including suburban government institutions and high-tech zones. This kind of polycentric structure [47], [48] is the nature of a metropolis like Shanghai.

The transition probability between clusters in different time intervals can be easily computed from the empirical data, as shown in Fig. 3. This transition probability reflects the urban circadian rhythms and the motivation of people's travelling patterns in different time periods. For instance, at 8 o'clock, more than 60% of the trips end at stations in cluster 1, regardless of the departure station, indicating that people tend to go downtown in this duration. That is because people tend to go to working places, usually located in downtown areas, in this period. In the afternoon or after work, however, people incline to departure for places of leisure and recreation, indicating a significantly different driving force with the early peak. This driving force is considered a prime motivation in our prediction model.

4 Model

In this section, we combine intervening opportunity theory and transition probability calculated above to build a model to predict trip fluxes. According to the theory of intervening opportunities [21], the probability of an individual going to a certain region is directly proportional to the number of opportunities at that region and inversely proportional to the number of intervening opportunities. Following and extending this theory, we assume that the attractiveness of region j to region i is determined by the number of opportunities O_j in j, and the *rank value*. Thus, this attraction can be expressed as:

$$A(j \leftarrow i) \propto O_j rank_i(j)^\alpha, \tag{2}$$

where $\alpha = -0.84$ is obtained in [49], for it fits well in 34 cities despite the variations in political and economic situations, and

$$rank_i(j) = |\{k : d(i,k) \le d(i,j), k \in \mathcal{C}_{\lambda_j}\}| \tag{3}$$

in which $|\cdot|$ represents the element number in a set. $\mathcal{C}_{\lambda_j}$ is the cluster of stations partitioned by K-Means algorithm

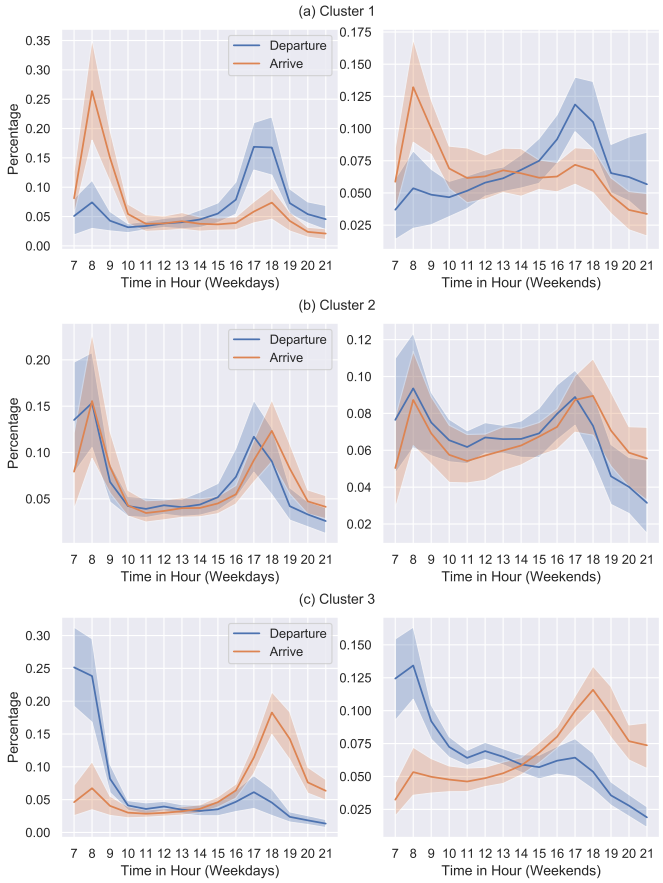


Fig. 2. The average and the 91% confidence interval of the entrance/exit dynamics of stations when the clustering number is 3. The dynamics of these three types of stations obtained by clustering have completely different pattern and explicit rational explanation: (a) The dynamics of stations in working regions. (b) The dynamics of stations in comprehensive regions. (c) The dynamics of stations in residential areas. In the cluster of residential regions, there exists a pronounced peak of departure between 7 and 9 o'clock. Meanwhile in the cluster of the working regions, a remarkable peak appears between 8 and 9 o'clock. This is a typical commuting pattern in the morning rush hour, indicating the distinct functions of different regions in city and the interaction between these regions at the aggregation level.

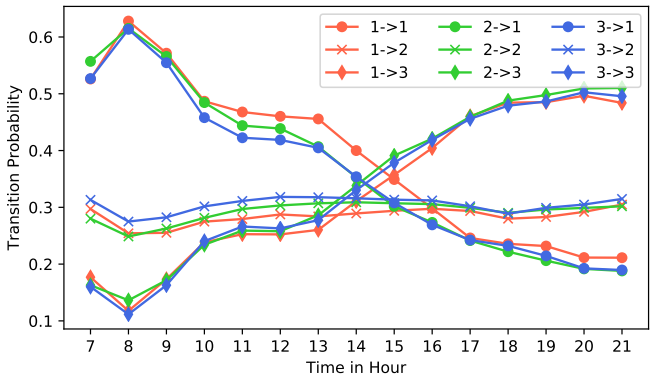


Fig. 3. The transition probability of all clusters when $K = 3$ on weekdays.

which contains station j. $d(i, j)$ is the cost of travelling from station i to station j. Generally, Euclidean distance or trip distance obtained by shortest path algorithm is selected as the cost. In the circumstances of metro case, the interchange behavior [45], [50] and varied train speeds can hardly be reflected by the conception of physical distance, but they do influence passengers' behavior. Therefore, we adopt the average travel time between pairs of stations estimated from the empirical data as the travel cost. Following the assumption in [22], [26], we assume that the number of opportunities O_j at region j is proportional to the total daily passenger flow in station j throughout the day, which can easily be calculated from the dataset. The $rank_i(j)$ reflects the ranking of the trip cost from station i to station j in the costs of all trips starting at station i to all the other stations. The higher the cost, the lower j ranks against i.

Considering the transition probability from one region to another (or itself) as a known premise, the transition flux of travels from i to region j in τ can be modeled as

$$T_\tau(i \to j) = T_\tau(i) p_\tau(\mathcal{C}_{\lambda_i} \to \mathcal{C}_{\lambda_j}) \frac{A(j \leftarrow i)}{\sum_{k \in \mathcal{C}_{\lambda_j}} A(k \leftarrow i)}, \quad (4)$$

where $T_\tau(i)$ is the total number of passengers departing from station i in τ, and p_τ is the transition probability between clusters.

$\frac{A(j \leftarrow i)}{\sum_{k \in \mathcal{C}_{\lambda_j}} A(k \leftarrow i)}$ reflects the effect of competition of attractions. For a passenger at departure station i whose destination is one of the regions in $\mathcal{C}_{\lambda_j}$, all stations in $\mathcal{C}_{\lambda_j}$ have attraction to the passenger. The competitiveness of potential destination j will be weakened if there exist a certain number of stations with opportunities closer to i than j, resulting in a lower possibility of traveling from i to j. Relatively speaking, if a potential destination j has a large number of attractive "opportunities", then the competitiveness of region j is stronger. Once the departure station cluster $\mathcal{C}_{\lambda_i}$ and destination station cluster $\mathcal{C}_{\lambda_j}$ are fixed, all competitive regions must belong to the same cluster, for such competition won't exist between regions with different features.

5 Results

5.1 Model Validation

To validate the proposed mobility model, we compare the computational results of our model with the empirical data. We concentrate on four critical aspects: trip fluxes between stations, distribution of trip fluxes, exit passenger flow of each station, and distribution of passenger travel time. We analyze the model performance under different cluster numbers $K = 1$, 3, 5 and 9, representing different degrees of regional function segmentation. When the clustering number is 1, our model degenerate to the simple intervening opportunities model. 9 clusters are substantial enough to effectively classify regional functions comparing with related literatures: [27] classifies travel purposes into 6 categories, and 8 regional functions are extracted in [35].

Trips starting between 17 o'clock and 18 o'clock on weekdays are taken as an example. Fig. 4 demonstrates that the our proposed model can effectively and accurately predict trip fluxes and travel time patterns. We notice that

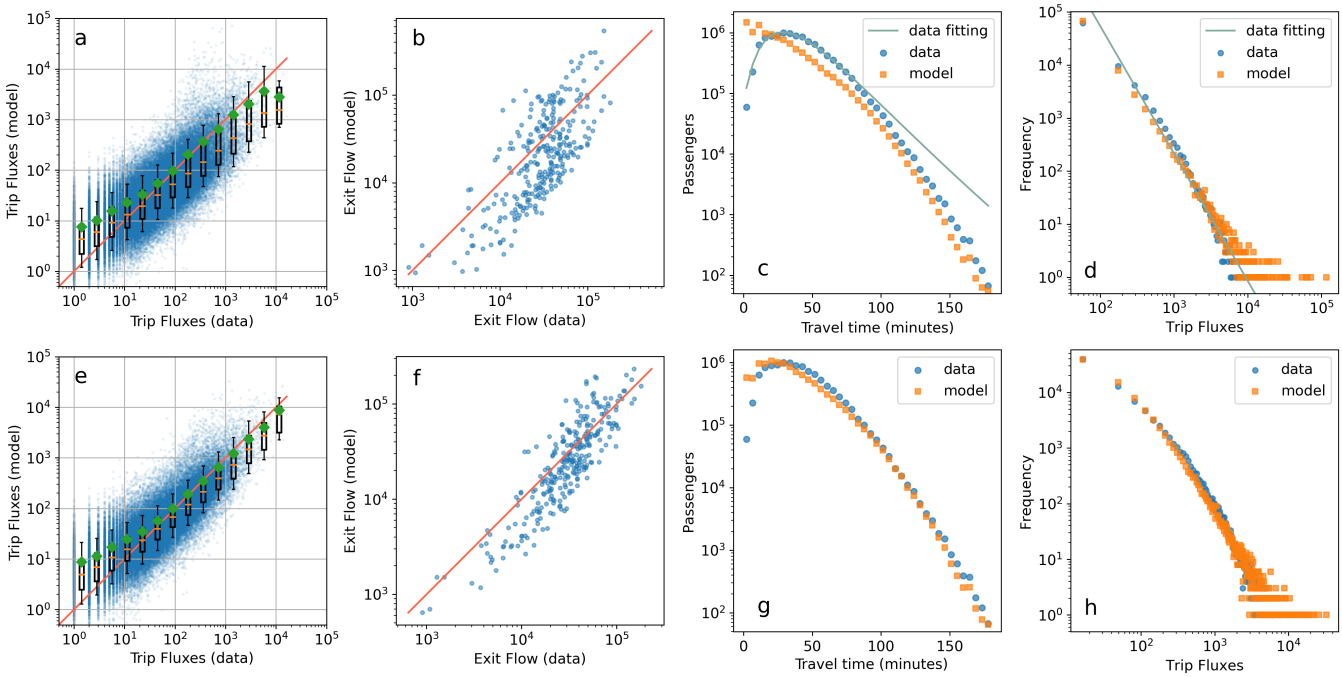


Fig. 4. **Comparing the predictions of our model and the empirical data.** Each row represents the indicators to be compared under the same clustering number. The clustering number is one in (a-d) and three in (e-h). (a, e) Comparing the observed trip fluxes with the predicted trip fluxes. The translucent blue points are scatter plot for fluxes between each pair of stations. The red line, as a guide of eyes, is of slope 1. The boxplots are used to describe the distribution of the number of predicted trip fluxes in different bins of the number of actual trip fluxes. The green diamond correspond to the mean number of predicted travelers in that bin. The ends of the whiskers represent the 9th percentile and the 91st percentile. (b, f) Comparing the observed and predicted numbers of exit people. The red line is $y = x$. (c, g) Travel time distribution. The empirical travel time approximately obeys log-normal distribution with $\mu = 4.006$ and $\sigma = 0.3937$, shown in (c). (d, h) Trip fluxes distribution. The empirical trip fluxes distribution is well approximated by power-law with power exponent $\gamma = 2.0329$, shown in (d).

the results follow a similar pattern when $K \geq 3$, thus we only compare the case of $K = 1$ with $K = 3$ as illustrated in Fig. 4. Fig. 4(a, e) show that for the clustering number 1, 3 respectively, the trip fluxes predicted by our model coincide well with actual data. By observing the whiskers of boxplots, it is easy to notice that when $K = 3$, the range of confidence interval is smaller, meaning that the predicted fluxes between stations are closer to real data in statistical sense. Fig. 4(b, f) illustrate the predicted and actual numbers of exit passengers in all stations. It can be noticed that our model, with the clustering number $K = 3$, offers a more authentic approximation than the pure intervening opportunities model. Fig. 4(c, g) are the comparisons of passengers' travel time distributions. The empirical travel time approximately obeys log-normal distribution. If all stations are in one cluster, the travel time follows an approximate exponential distribution which is monotonically decreasing. When $K = 1$, namely when all stations are in one cluster, the travel time of the intervening opportunities model approximately obeys exponential distribution, which is monotonically decreasing; While when taking different regional features into account, the passenger number shows a trend of decrease after increase with the increasing of travel time. There exists a peak travel time bin which is consistent with the statistical result of the actual data. The existence of such difference is fully reasonable. In the simple intervening opportunities model, individual travel behavior is solely influenced by the number of intervening opportunities and trip cost, resulting in that one tends to travel to regions with lower travel time costs. In our model, on the other hand, when K is larger than one, an individual necessitates some additional consideration on trip intentions. The candidate destinations may all be far from the departing location, resulting in longer travel time. Fig. 4(d, h) show the predicted and actual trip fluxes distributions. The empirical trip fluxes approximately obey power-law distribution, which agrees with the previous finding in [47]. The model prediction again fits ideally with the empirical data.

5.2 Quantitative Analysis of the Prediction Accuracy

In this subsection, we quantitatively analyze the prediction accuracy and discuss how clustering number affects the accuracy of the proposed model. The Sørensen similarity index (SSI) and the root mean square error ($RMSE$) are adopted to compare the difference between actual and model fluxes, and the Hellinger coefficient is used to compare the difference regarding travel time distribution and inter-station trip flux distribution.

Sørensen similarity index (SSI) is a statistic used for comparing the similarity of two samples. A modified version of the index is applied in this paper to measure the extent of consistence between trip fluxes and exit passenger flow of the actual data and our mobility prediction model:

$$SSI_{flux}(\tau) = \frac{1}{N^2} \sum_{i}^{N} \sum_{j}^{N} \frac{2\min(T'_\tau(i \to j), T_\tau(i \to j))}{T'_\tau(i \to j) + T_\tau(i \to j)}; \tag{5}$$

$$SSI_{exit}(\tau) = \frac{1}{N}\sum_{i}^{N}\frac{2\min\left(S'_\tau(i), S_\tau(i)\right)}{S'_\tau(i) + S_\tau(i)}, \tag{6}$$

where $N = 289$ is the station number in all, $T_\tau(i \to j)$ is the trip flux predicted by our model from station i to station j in time period τ and $T'_\tau(i \to j)$ is the actual trip flux, $S_\tau(i)$ and $S'_\tau(i)$ are the predicted and actual exit flow of station i in τ. The SSI ranges between 0 and 1, and $SSI = 1$ indicates that the two samples match perfectly.

Root Mean Squared Error($RMSE$) is used to measure the differences between values predicted by a model and the values observed. In our scenario, it has the following two forms:

$$RMSE_{flux}(\tau) = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N^2}(T'_\tau(i \to j) + T_\tau(i \to j))^2}, \tag{7}$$

$$RMSE_{exit}(\tau) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(S'_\tau(i) - S_\tau(i))^2}, \tag{8}$$

The smaller these two values, the higher the prediction accuracy of the model.

Hellinger coefficient is used to quantify the similarity between two probability distributions. Denote $p(k)$ and $q(k)$ as the probability density function of two discrete distributions within the same domain D ,the Hellinger coefficient is defined as:

$$R_H = \sum_{k \in D}\sqrt{p(k)q(k)}. \tag{9}$$

This coefficient is used to compare the model and the empirical distribution of the travel time and the trip fluxes. All indicators in equation 5 to 9 are time varying in our case.

SSI is an index commonly used in human flux predicting accuracy evaluation. It emphasizes on the **proportion** of the difference between the predicted and actual values. Consider two fluxes between different metro stations. Assume that the actual value of the flux between two metro stations, named flux A, is 1, and the predicted flux is 50 passengers. While the predicted and the actual flux between another two stations, named flux B, is 2400 and 2000 respectively. From the perspective of prediction accuracy, the prediction result of B is more acceptable. From the perspective of metro operating, however, the **absolute number** of passengers means the actual load on the line, and the influence of the predicting error of flux A on the operation of the whole metro line is obviously less than that of B. Therefore, SSI and $RMSE$ are both used to quantitatively describe the prediction accuracy of human fluxes from the perspective of proportion and absolute value.

In Fig. 5, we demonstrate the quantitative calculation results of the above three indices between the empirical data and the prediction results of our model in various time intervals on both weekdays and weekends. Generally speaking, there are similar patterns between weekends and weekdays. In all subfigures, the three cases containing inter-cluster transitions(the orange, green and red lines representing cluster numbers of 3, 5 and 9, respectively) have a similar pattern, whereas results of the pure intervening opportunities model (the blue line representing cluster number of 1) have distinct patterns. In most cases, a larger clustering

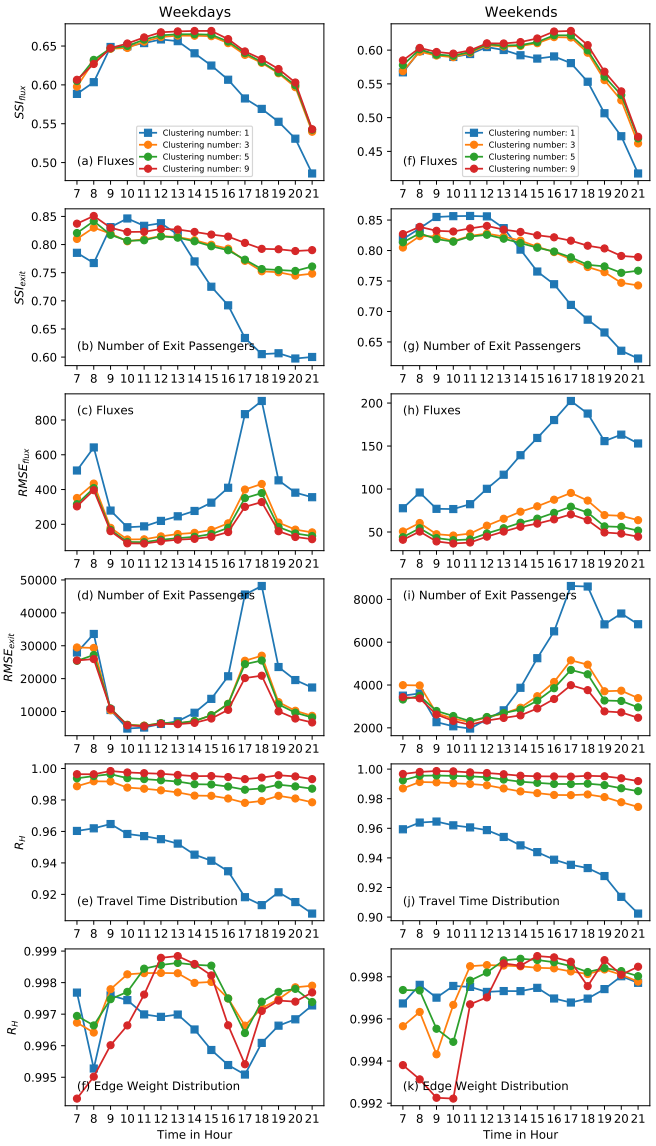


Fig. 5. **Quantitative comparisons of the empirical data and the results of our model in various time interval.** The first column is the comparisons of the weekdays and the second column is the comparisons of the weekends. (a, b, f, g) The SSI of the fluxes and the numbers of exit passengers. (c, d, h, i) The $RMSE$ of the fluxes and the numbers of exit passengers. (e, f, j, k) The Hellinger coefficient of the travel time distribution and the edge weight distribution.

number indicates a more accurate result. Furthermore, in the afternoon and evening, our model has especially huge advantages against the pure intervening opportunities in prediction accuracy. Additionally, when $RMSE$ is used as the evaluation index, our model is obviously better than pure intervening opportunities model in all time periods; while for the SSI index, our model and pure intervening opportunity model have very close evaluation results in the morning. Accordingly, our model performs better in estimating the absolute number of human fluxes, which might because that the prediction bias of our model mainly falls on station-pairs with less human fluxes (see Fig. 4(e)). Similar results are derived in predicting the number of arriving passengers at different stations. In the statistical

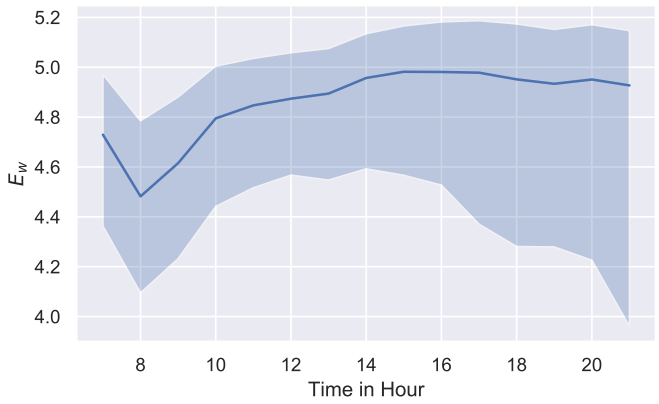


Fig. 6. Dynamic entropy of the stations on weekdays. The shadow is the 91% confidence interval.

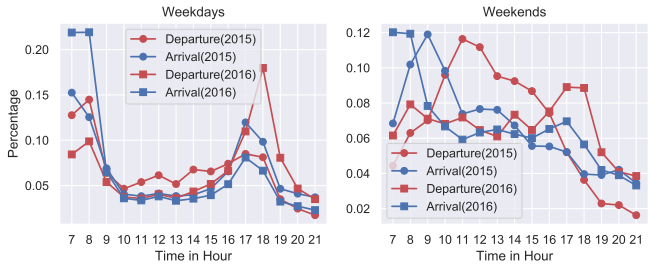


Fig. 7. The changes of entrance/exit dynamics of Longhua Station in April 2015 and September 2016. When the clustering number is 3, the cluster label of Longhua Station changes from 2 (comprehensive region) to 1 (working region).

sense, as shown in Fig. 5(e-k), all R_H indexes are close to 1, indicating that the model has an excellent performance throughout the whole day.

As mentioned above, Fig. 5 shows that our model has a better performance than the pure intervening opportunities model especially in the afternoon and evening. This phenomenon may due to the fact that people have a larger chance to travel by their own volition after noon, lading to greater competition in regions in the same cluster. To verify this hypothesis, we calculate the weighted average entropy of stations' entrance flows and demonstrate the result in Fig. 6. A higher entropy reflects a more balanced choice of travel destinations, indicating that people have more freedom to choose the place of their next activity. The entropy of station i in τ is $E_i(\tau) = -\sum_j p_\tau(i \to j) \log(p_\tau(i \to j))$, where $p_\tau(i \to j) = \frac{T'_\tau(i \to j)}{\sum_k T'_\tau(i \to k)}$ is the actual probability of a person going from i to j in τ. The entropy of a station indicates the diversity of travel destinations and can further infer people's ability to make autonomous traveling. The weighted average entropy $E_w(\tau) = \frac{\sum_i [E_i(\tau) \sum_j T'_\tau(i \to j)]}{\sum_i \sum_j T'_\tau(i \to j)}$ is used to describe the global diversity of travel purposes. In the morning peak, this entropy is obviously low, since it is impossible for people to choose their workplace arbitrarily every day. After 14 o'clock, the weighted average entropies are higher, showing that people make more abundant travel choices. During this period, people have a larger chance of travelling by their own volition, so the competition introduced by our model happens. Our model captures this kind of competition in regions in the same cluster, so it has achieved better prediction results especially in the afternoon and evening.

5.3 Model predictability after the changes of regional functions

With the pace of urban construction, the functions of some regions may change, and the trip fluxes of these regions may change accordingly. In order to further verify the validity and the predictability in functional change-induced flux changes of our model, we extract the trip data of April 2015 from Shanghai Public Transportation Card data, and compare it with the data of September 2016. The topological structures of Shanghai metro network are the same in these two periods. The functions of most regions remain unchanged, and most functionally changed regions are in the suburbs where the topological structure of the metro line is simple, thus the changes of regional functions have little effect on actual trip fluxes in these suburban regions. As an exception, Longhua Station is located near the city center, as labeled by a green circle in Fig. 8. During the studied period, a renovation project in the area around Longhua Station is underway [51]. When $K = 3$, its function changes from "comprehensive region" to "working region", as shown in Fig. 7. Taking 18:00-19:00 as an example, we calculated the passenger fluxes that ends at Longhua Station in 2015 and 2016, respectively. During the two periods, the top 30 stations in the number of departures are recorded, defined as "major source stations" here. As shown in Fig. 8, the major source stations of Longhua Station are obviously different in these two periods. Only 13 stations are on the top 30 lists of both periods, indicating that the change of regional function may bring about changes in human fluxes. In the meantime, while our model can correctly predict 20 of the top 30 stations in 2015, it still can successfully identify 19 stations in 2016, proving that our model can adjust the prediction results according to the change of regional function to conform to the actual situation. It's remarkable that 5 stations marked by a blue circle in Fig. 8(b) turn into the major source stations in 2016, and our model successfully capture the change. The total departure number of these 5 stations increased from 1.75 million to 2.20 million in the whole month, while the average number of trips to Longhua Station significantly increased from 7465 to 42242. The slight increase of T_i can hardly cause this kind of burst. We believe that it should be attributed to functional changes of region Longhua, which is reflected by our model.

Further, we study the changes of the major source stations of all stations between these two month, and examine to what extent our model can predict these differences. Let the set of major source stations in 2015 be S_{2015} and S_{2016} in 2016, and the set of major source stations predicted by our model is S'_{2016}. Then the station set that is not included in S_{2015} but can be captured by our model in S_{2016} can be expressed as $\check{S} = (S_{2016} - S_{2015}) \bigcap S'_{2016}$. Fig. 9 shows the relationship between $|S_{2016} - S_{2015}|$ and $|\check{S}|$, and compare it with the expectations of random selected station set that is in the set $S_{2016} - S_{2015}$. From this figure we can see that although the major source stations changes with time, more than one-third of these changes can be predicted.

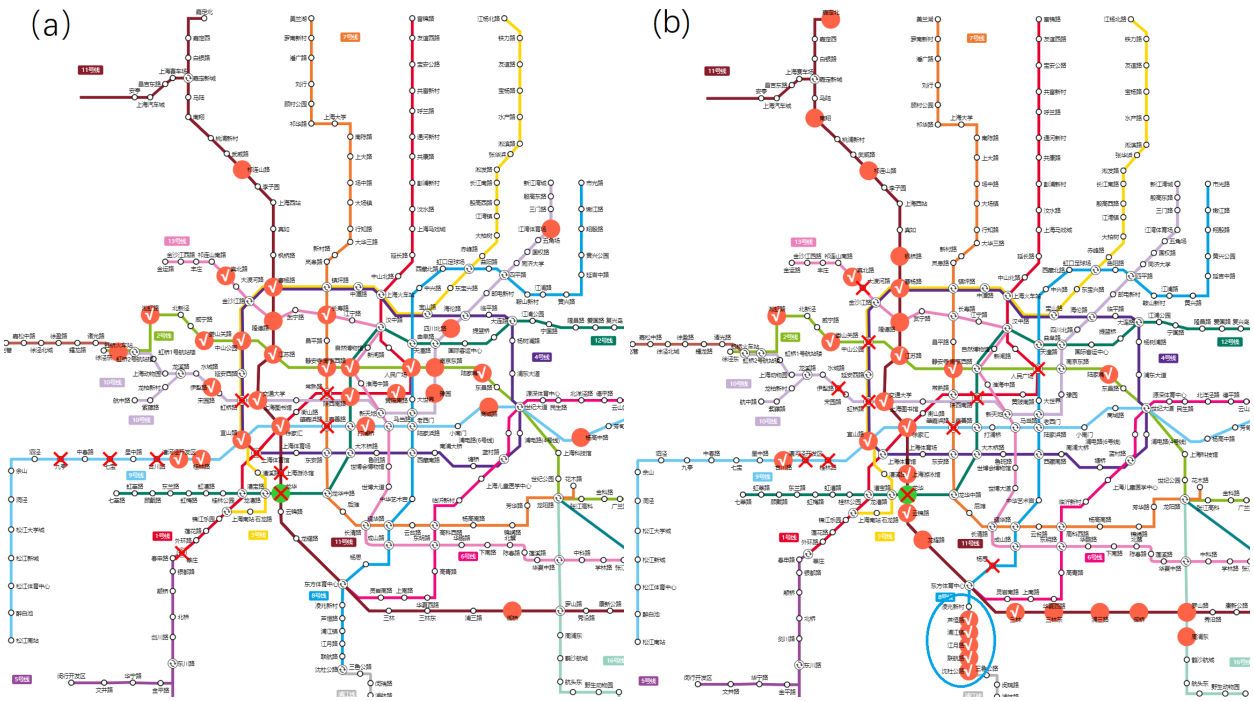


Fig. 8. The actual and predicted major source stations of Longhua Station in (a) April 2015 and (b) September 2016. Longhua Station is labeled in green. The actual major source stations are labeled in orange. If our results match the practical data, white ticks are marked, or else red crosses are marked.

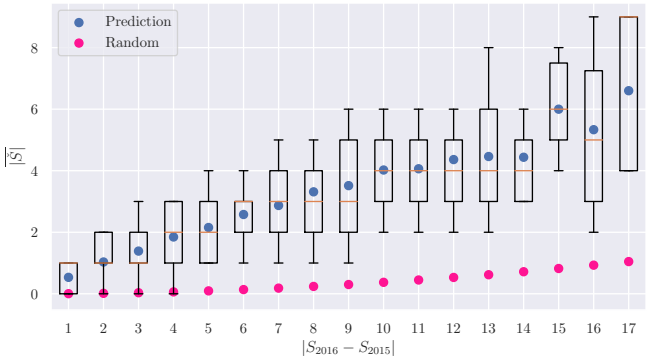


Fig. 9. The changes of the major source stations and the ability of the model to predict these changes. The pink dots are the expectations of random selected station set that in the set of $S_{2016} - S_{2015}$. The boxplot expresses the same meanings as what expressed in Fig. 4. There is only one $|S_{2016} - S_{2015}|$ value greater than 17, which has no statistical significance.

5.4 Discussions

We find from the prediction results that when the number of clusters is greater than or equal to 3, the prediction accuracy-time curves of the model have very similar patterns, and the prediction accuracy increases slightly with the increase of the number of clusters. What's more, when the clustering number is 3, we can deduce the practical functions of the 3 kinds of regions: working, residential and comprehensive. We also notice a similar result in [52], in which the non-negative matrix factorization and optimization methods were employed, discovering that only when the number of bases is 3, the factorization results can be stable. These three bases have similar practical significance: commuting between home and workplace, workplace to workplace, and others such as leisure activities. Therefore, when considering clustering number, choosing $K = 3$ leads to both clear practical significance and sufficient accuracy. Furthermore, for urban planners who have detailed geographic information data, they can also aggregate point of interests [24], [53] in different regions in order to acquire more detailed prediction results and clearer practical significance.

Individual travel motivation including trip purpose and intervening opportunities is key to studying the macroscopic human mobility fluxes from the microscopic perspective. Such microscopic perspectives concerning individual mobility laws [1] will further assist us in modeling trip fluxes and detecting urban characteristics. SSI_{flux} in Fig. 5 shows a trend of first increase and then decrease with time, indicating that our model has a good performance in the afternoon. It is probably because individual travel in the morning and evening are confined to two fixed location: working place and home, so citizens are not able to choose their travel destination entirely through competition mechanism in our model. When an individual applies for jobs, salary is the major consideration rather than trip cost [20], [54], which is beyond the discussion of competitiveness in our model, affecting the accuracy of our model in the morning. Similarly, one has to go to his own home after sunset without considering the attractions of others' home. Some previous studies have considered home as a special spot [27], [30], being a way to exploit human traits and dispositions.

6 Conclusions

Human movement between regions with different functions and within regions with the same function constitutes the

circadian rhythm of the city. In this paper, we cluster the regions with similar functions and investigate how people's trip decisions are dynamically influenced by these different kinds of regions. We focus on human travel motivation and take both intervening opportunities and interregional transition probability into consideration to predict human mobility. When travel motivation is introduced, the prediction accuracy is significantly improved, especially in the afternoon and evening. The proposed model can successfully predict popularity of regions with different features, which is an important element for estate developers and individual enterprises to decide on the location of business premises and projects. Our model can also capture the fluxes between regions, which is critical for urban planners to understand whether the regional orientation of a new developing region is reasonable and its future impact on urban traffic. From the perspective of travel time distribution and trip flux distribution, the prediction results of our model conform ideally with empirical data. When the function of a region changes, our model can predict fluxes in the new scene corresponding to this change. It has become a new thought in human mobility modelling and urban structure analyzing to take both regional features and individual mobility laws into account. The development and interoperability of the IoT [55] can provide massive data in multiple dimensions, enabling us to describe and predict individual and urban dynamics more precisely. Ultimately, it will facilitate the construction of smart city and help improve the convenience and comfortability of city life.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant Nos 61773255 and 61873167. We thank Shanghai Open Data Apps for providing us with data support. The metro maps are plotted by Amap.

REFERENCES

[1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018.

[2] R. Louf and M. Barthelemy, "How congestion shapes cities: from mobility patterns to scaling," *Scientific reports*, vol. 4, no. 1, pp. 1–9, 2014.

[3] C. Zhong, S. M. Arisona, X. Huang, M. Batty, and G. Schmitt, "Detecting the dynamics of urban structure through spatial network analysis," *International Journal of Geographical Information Science*, vol. 28, no. 11, pp. 2178–2199, 2014.

[4] E. Barbour, C. C. Davila, S. Gupta, C. Reinhart, J. Kaur, and M. C. González, "Planning for sustainable cities by estimating building occupancy with mobile phones," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.

[5] J. Wang, D. Wei, K. He, H. Gong, and P. Wang, "Encapsulating urban traffic rhythms into road networks," *Scientific reports*, vol. 4, no. 1, pp. 1–7, 2014.

[6] K. He, Z. Xu, P. Wang, L. Deng, and L. Tu, "Congestion avoidance routing based on large-scale social signals," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 9, pp. 2613–2626, 2015.

[7] S. Çolak, A. Lima, and M. C. González, "Understanding congested travel in urban areas," *Nature communications*, vol. 7, no. 1, pp. 1–8, 2016.

[8] X. Yang, A. Chen, B. Ning, and T. Tang, "Measuring route diversity for urban rail transit networks: A case study of the beijing metro network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 2, pp. 259–268, 2016.

[9] S. Sarkar, S. Chawla, S. Ahmad, J. Srivastava, H. Hammady, F. Filali, W. Znaidi, and J. Borge-Holthoefer, "Effective urban structure inference from traffic flow dynamics," *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 181–193, 2017.

[10] R. Silva, S. M. Kang, and E. M. Airoldi, "Predicting traffic volumes and estimating the effects of shocks in massive transportation systems," *Proceedings of the National Academy of Sciences*, vol. 112, no. 18, pp. 5643–5648, 2015.

[11] Z. Huang, P. Wang, F. Zhang, J. Gao, and M. Schich, "A mobility network approach to identify and anticipate large crowd gatherings," *Transportation research part B: methodological*, vol. 114, pp. 147–170, 2018.

[12] S. A. Shah, D. Z. Seker, M. M. Rathore, S. Hameed, S. B. Yahia, and D. Draheim, "Towards disaster resilient smart cities: Can internet of things and big data analytics be the game changers?" *IEEE Access*, vol. 7, pp. 91 885–91 903, 2019.

[13] B. Du, C. Liu, W. Zhou, Z. Hou, and H. Xiong, "Detecting pickpocket suspects from large-scale public transit records," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 3, pp. 465–478, 2018.

[14] Y. Xu, A. Belyi, I. Bojic, and C. Ratti, "Human mobility and socioeconomic status: Analysis of singapore and boston," *Computers, Environment and Urban Systems*, vol. 72, pp. 51–67, 2018.

[15] E. Thuillier, L. Moalic, S. Lamrous, and A. Caminada, "Clustering weekly patterns of human mobility through mobile phone data," *IEEE Transactions on Mobile Computing*, vol. 17, no. 4, pp. 817–830, 2017.

[16] J. Gao, Y.-C. Zhang, and T. Zhou, "Computational socioeconomics," *Physics Reports*, vol. 817, pp. 1–104, 2019.

[17] X. Liang, J. Zhao, L. Dong, and K. Xu, "Unraveling the origin of exponential law in intra-urban human mobility," *Scientific reports*, vol. 3, p. 2983, 2013.

[18] Y. Liu, Z. Sui, C. Kang, and Y. Gao, "Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data," *PloS one*, vol. 9, no. 1, p. e86026, 2014.

[19] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility," *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.

[20] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, 2012.

[21] S. A. Stouffer, "Intervening opportunities: a theory relating mobility and distance," *American sociological review*, vol. 5, no. 6, pp. 845–867, 1940.

[22] X.-Y. Yan, C. Zhao, Y. Fan, Z. Di, and W.-X. Wang, "Universal predictability of mobility patterns in cities," *Journal of The Royal Society Interface*, vol. 11, no. 100, p. 20140834, 2014.

[23] C. Kang, Y. Liu, D. Guo, and K. Qin, "A generalized radiation model for human mobility: spatial scale, searching direction and trip constraint," *PloS one*, vol. 10, no. 11, p. e0143500, 2015.

[24] K. K. Jahromi, M. Zignani, S. Gaito, and G. P. Rossi, "Simulating human mobility patterns in urban areas," *Simulation Modelling Practice and Theory*, vol. 62, pp. 137–156, 2016.

[25] T. Ma, R. Zhu, J. Wang, N. Zhao, T. Pei, Y. Du, C. Zhou, and J. Chen, "A proportional odds model of human mobility and migration patterns," *International Journal of Geographical Information Science*, vol. 33, no. 1, pp. 81–98, 2019.

[26] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, pp. 818–823, 2010.

[27] L. Wu, Y. Zhi, Z. Sui, and Y. Liu, "Intra-urban human mobility and activity transition: Evidence from social media check-in data," *PloS one*, vol. 9, no. 5, p. e97010, 2014.

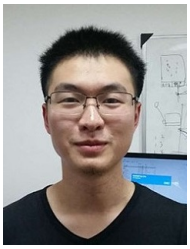
[28] X.-Y. Yan, W.-X. Wang, Z.-Y. Gao, and Y.-C. Lai, "Universal model of individual and population mobility on diverse spatial scales," *Nature communications*, vol. 8, no. 1, pp. 1–9, 2017.

[29] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, no. 84, p. 20130246, 2013.

[30] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González, "The timegeo modeling framework for urban mobility without travel surveys," *Proceedings of the National Academy of Sciences*, vol. 113, no. 37, pp. E5370–E5378, 2016.

[31] Y. Hu, J. Zhang, D. Huan, and Z. Di, "Toward a general understanding of the scaling laws in human and animal mobility," *EPL (Europhysics Letters)*, vol. 96, no. 3, p. 38006, 2011.

[32] M. G. Meekan, C. M. Duarte, J. Fernández-Gracia, M. Thums, A. M. Sequeira, R. Harcourt, and V. M. Eguíluz, "The ecology of human mobility," *Trends in Ecology & Evolution*, vol. 32, no. 3, pp. 198–210, 2017.
[33] J. Ding, H. Liu, L. T. Yang, T. Yao, and W. Zuo, "Multiuser multivariate multiorder markov-based multimodal user mobility pattern prediction," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4519–4531, 2019.
[34] H. Wang, H. Zeng, Y. Li, P. Zhang, and D. Jin, "Human mobility prediction using sparse trajectory data," *IEEE Transactions on Vehicular Technology*, 2020.
[35] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li, "Land-use classification using taxi gps traces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 113–123, 2012.
[36] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore," *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 208–219, 2017.
[37] V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez, "Characterizing urban landscapes using geolocated tweets," in *2012 International conference on privacy, security, risk and trust and 2012 international confernece on social computing*. IEEE, 2012, pp. 239–248.
[38] Y. Zhi, H. Li, D. Wang, M. Deng, S. Wang, J. Gao, Z. Duan, and Y. Liu, "Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data," *Geo-spatial Information Science*, vol. 19, no. 2, pp. 94–105, 2016.
[39] F. Calabrese, G. Di Lorenzo, and C. Ratti, "Human mobility prediction based on individual and collective geographical preferences," in *13th international IEEE conference on intelligent transportation systems*. IEEE, 2010, pp. 312–317.
[40] J. Kim, J. Park, and W. Lee, "Why do people move? enhancing human mobility prediction using local functions based on public records and sns data," *PloS one*, vol. 13, no. 2, p. e0192698, 2018.
[41] M. Ren, Y. Lin, M. Jin, Z. Duan, Y. Gong, and Y. Liu, "Examining the effect of land-use function complementarity on intra-urban spatial interactions using metro smart card records," *Transportation*, pp. 1–23, 2019.
[42] G. Qi, A. Huang, W. Guan, and L. Fan, "Analysis and prediction of regional mobility patterns of bus travellers using smart card data and points of interest data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1197–1214, 2018.
[43] Y. Fu, G. Liu, Y. Ge, P. Wang, H. Zhu, C. Li, and H. Xiong, "Representing urban forms: A collective learning model with heterogeneous human mobility data," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 3, pp. 535–548, 2018.
[44] M. Lee and P. Holme, "Relating land use and human intra-city mobility," *PloS one*, vol. 10, no. 10, p. e0140152, 2015.
[45] M. Lenormand, M. Picornell, O. G. Cantú-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frias-Martinez, and J. J. Ramasco, "Cross-checking different sources of mobility information," *PloS one*, vol. 9, no. 8, p. e105184, 2014.
[46] Z. Huang, P. Wang, F. Zhang, J. Gao, and M. Schich, "A mobility network approach to identify and anticipate large crowd gatherings," *Transportation research part B: methodological*, vol. 114, pp. 147–170, 2018.
[47] C. Roth, S. M. Kang, M. Batty, and M. Barthélemy, "Structure of urban movements: polycentric activity and entangled hierarchical flows," *PloS one*, vol. 6, no. 1, p. e15923, 2011.
[48] R. Louf and M. Barthelemy, "Modeling the polycentric transition of cities," *Physical review letters*, vol. 111, no. 19, p. 198702, 2013.
[49] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, "A tale of many cities: universal patterns in human urban mobility," *PloS one*, vol. 7, no. 5, p. e37027, 2012.
[50] Z. Guo and N. H. Wilson, "Assessing the cost of transfer inconvenience in public transport systems: A case study of the london underground," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 2, pp. 91–104, 2011.
[51] http://www.soupu.com/news/589439.
[52] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò, "Collective human mobility pattern from taxi trips in urban area," *PloS one*, vol. 7, no. 4, p. e34487, 2012.
[53] X. Zheng, W. Chen, P. Wang, D. Shen, S. Chen, X. Wang, Q. Zhang, and L. Yang, "Big data for social transportation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 3, pp. 620–630, 2015.
[54] G. Carra, I. Mulalic, M. Fosgerau, and M. Barthelemy, "Modelling the relation between income and commuting distance," *Journal of the Royal Society Interface*, vol. 13, no. 119, p. 20160306, 2016.
[55] B. Ahlgren, M. Hidell, and E. C.-H. Ngai, "Internet of things for smart cities: Interoperability and open data," *IEEE Internet Computing*, vol. 20, no. 6, pp. 52–56, 2016.



Shuyang Shi was born in Shanghai, China, in 1992. He received the B.S. degree from Department of Automation, Shanghai Jiao Tong University, Shanghai, China in 2015. He is currently working toward the Ph.D. degree in control theory and control engineering with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. His research interests include the human mobility networks, data science and urban science.

Lin Wang (Senior Member, IEEE) received the B.S. and M.S. degrees from the School of Mathematical Sciences, Shandong Normal University, Jinan, China, in 2003 and 2006, respectively, and the Ph.D. degree in operations research and control theory from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, in 2009. She is currently a Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include multi-agent systems, adaptive complex networks, and coordination of multiple manipulators.

Shuangdie Xu was born in Zhejiang, China, in 1995. She received the B.S. and M.S. degrees from the Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2017 and 2020, respectively.She is currently an assistant engineer with State Grid Zhejiang Electric Power Company Hangzhou Power Supply Company, Zhejiang, China. Her current research interests include power system reliability analysis, power system uncertainty analysis and integration of distributed generation.

Xiaofan Wang (Senior Member, IEEE) received the Ph.D. degree from Southeast University, China in 1996. He has been a Professor with Shanghai Jiao Tong University (SJTU) since 2002 and a Distinguished Professor of SJTU since 2008. He is now vice-president of Shanghai University. He received the 2002 National Science Foundation for Distinguished Young Scholars of P. R. China, the 2005 Guillemin-Cauer Best Transactions Paper Award from the IEEE Circuits and Systems Society, the 2008 Distinguished Professor of the Chang Jiang Scholars Program, Ministry of Education, and the 2015 Second Class Prize of the State Natural Science Award. His current research interests include analysis and control of complex dynamical networks. He is currently vice-chair of IFAC Coordinating Committee and vice-chair of Systems Engineering Society of China.