Exploring the distribution and dynamics of functional regions using mobile phone data and social media data

Jinzhou CAO, Wei TU, Qingquan LI, Meng ZHOU and Rui CAO

Abstract

How different functional regions in urban space are distributed and dynamically changing is determined by how their residents interact with them, which is crucial for urban managers to make urban planning decisions, respond to emergency quickly. Based on these, this paper proposed a novel approach for the probability based labelling individual activities which can be further used to explore the distribution of social land use at base tower station (BTS) level using a combination of multi-source spatiotemporal data, namely, call data and check-in data. We applied an experiment in Shenzhen, China, and the result is compared to Tencent Street View to demonstrate the effectiveness of the proposed approach to infer urban functional regions.

Jinzhou CAO
State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, P. R. China
Email: caojinzhou@whu.edu.cn

Wei TU (Corresponding author)
Key Laboratory for Geo-Environment Monitoring of Coastal Zone of the National Administration of Surveying, Mapping and GeoInformation, Shenzhen University, Shenzhen 518060, P.R. China
Email: tuwei@szu.edu.cn

Qingquan LI (Corresponding author)
Key Laboratory for Geo-Environment Monitoring of Coastal Zone of the National Administration of Surveying, Mapping and GeoInformation, Shenzhen University, Shenzhen 518060, P.R. China. Email: liqq@szu.edu.cn

Meng ZHOU
Department of Geography, Hong Kong Baptist University, Hong Kong, P.R. China. Email: zhoumeng@life.hkbu.edu.hk

Rui CAO
School of Geodesy and Geometics, Wuhan University, Wuhan 430079, P. R. China
Email: cr@whu.edu.cn

1. Introduction

The thriving rise of mobile Internet technology and the widespread use of GPS-embedded smart phones have offered opportunities to gather huge amounts of spatiotemporal data that can be used to analyze and model various aspects of human mobility and activity patterns (Csáji et al. 2013; González et al. 2008; Song, Qu, et al. 2010). For example, people are localized when conducting a call; many social network platforms allow users to add their location to their profiles. These allow to a full analysis of statistical laws on human behavior at temporal and spatial dimensions (Noulas et al. 2012; Song, Koren, et al. 2010), or understanding the regularities (Noulas et al. 2012), similarities between people (Li et al. 2008), predictability(Chen et al. 2014; Song, Qu, et al. 2010) based on their location trajectories. In general, these studies provide a unique perspective on general features of human mobility pattern.

However, despite the disclosure of these general features, previous studies do not provide further insights on motivation or preferences of daily activities which are reflected in these data. The study of underlying activities performed at the locations that motivate the movements are still on a less-explored stage (Liu et al. 2013). Therefore, how to build a bridge between spatiotemporal data and activity knowledge, capable of supporting management decisions that are related to activity behavior is the main concerned issue in this paper.

Urban space is quite complex, which is highly concentrated and complex composed of people flow, material flow, energy flow, and information flow (Jacobs 1961). The interaction between humans and urban space occurs at any moment. People move between different functional regions and engage in social activities. As a result, urban functional regions are mixed, dynamic and social. In describing the definition of "land use", (Rodrigue et al. 2013)stated that functional region is a higher level of dynamic spatiotemporal change, which refers to its socio-economic description in space, compared to formal land use, which refers to its physical form and pattern. In other words, urban space can be partitioned into different functional regions based on individual daily activities such as residential areas, business areas etc., and property and proportion of functional regions can be temporal dynamically changed as well, which is the biggest difference with conventional method on land use classification. Clearly, how different functional regions are distributed and dynamically changed is determined, mostly, by how their residents move within, use and interact with them. Therefore, it's crucial to understand the distribution and dynamics of functional regions from the perspective of individual social activities, which helps to make the rational urban planning decisions, optimize spatial structure(Frias-Martinez et al. 2012), and harmonize urban society(Toole et al. 2012). To solve these problems, the use of multi-source spatiotemporal data is a potential solution.

The motivation of this research is to explore the distribution and dynamics of functional region by combining multi-source spatiotemporal data, namely, call data and check-in data in light of the close correlation of real urban functions and individual activity patterns.

The main contributions of this paper lie as follows. First, we develop a model for rigorously inferring the real social activities behind their trajectories only using spatiotemporal information. The core element of the model is an Improved Hidden Markov Model (I-HMM) which intend to make the model more compatible for spatiotemporal data, increasing the inference accuracy. Second, we propose a framework to explore the distribution of functional regions in a city from the point view of individual social activity and mobility.

The remainder of this paper is organized as follows. In section 2, the proposed methodology is presented, including data description, framework, and details of model. Section 3 discusses the case study demonstrating the proposed method in the whole city of Shenzhen. Section 4 concludes the paper and discusses further research.

2. Methodology

In this section, we first give an overview of the two spatiotemporal datasets employed in this methodology. Next, we introduce the framework of the proposed model. Finally, step-by-step detail descriptions of the model are presented.

2.1 Datasets Analysis

2.1.1 Call Data

Call data are collected by telecommunication providers, triggered by initiating or receiving a call. A record contains the timestamp, anonymized id representing identified user, and the code of cell which refers to the location information where the communication started.

Mobile phones have a high penetration, which is the reason for the popularity in research on human behavior. According to MIIT of China, penetration rate of mobile phones reached 94.5 per hundred person as of December 2014 in china.

The dataset in this paper was collected in August 2011 spanning a 12 days' duration by China Unicom, the world's third-biggest mobile provider. The collection area is the whole city of Shenzhen, an emerging metropolitan area at the south China. We have processed thirty three million call records, which were generated by 1.6 million users. Further, 2841 distinct base station towers (BTS) were detected. Fig.2.1 depicts the spatial distribution of density of call data at BTS level, and the complementary cumulative distribution functions (CCDF) of the number of calling per day for individuals is illustrated in Fig.2.2, as well. About 90% of individuals calls less than 5 times with 99% above 10. The two figures well reflect the heterogeneity in the volume and frequency of call behavior**.**

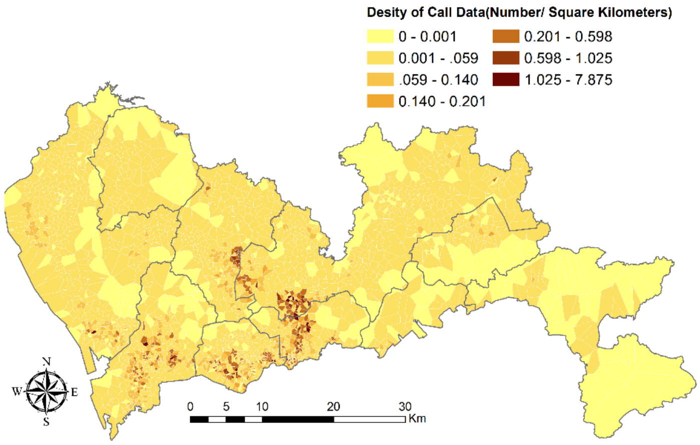

Fig. 2.1. Spatial distribution of density of call data at BTS level

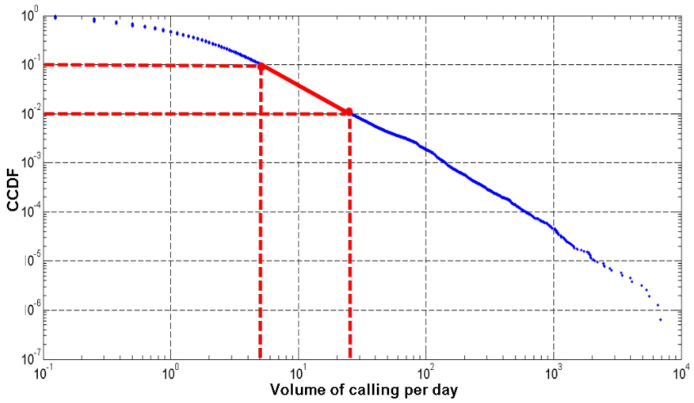

Fig. 2.2. Complementary Cumulative Distribution Functions(CCDF) of the volume of calling per day for individuals

2.1.2 Check-in Data

The information we use as a prior knowledge of individual activity and its transition pattern is sourced from Weibo check-in data. Weibo is a Chinese microblogging platform, similar to Twitter with a high market penetration in the USA. It has more than 0.5 billion registered users as of 2013 with frequent information update. Users check-in at places through a dedicated mobile device using GPS and other sensing technologies to automatically detect their location and post on Weibo platform. Characteristics of users' activities, therefore, can be expressed by places they've visited. In fact, the platform and its aggregate data link the physical activities and cyber behaviors.

Check-in dataset crawled by free API contains five million check-in records, each of which corresponds to a check-in at one of the 0.14 million POIs with its geographic coordinates. Activity category information about each POI has also been assigned by Weibo Co., Ltd. A spatial distribution of collected dataset is depicted in Fig.2.3. A merge and split method is conducted on activity category information for simplification purpose and thus eight activity categories are used in this paper. The statistical analysis on the popularity of check-ins per category at 24 timeslots in a day is conducted, as shown in Fig.2.4.

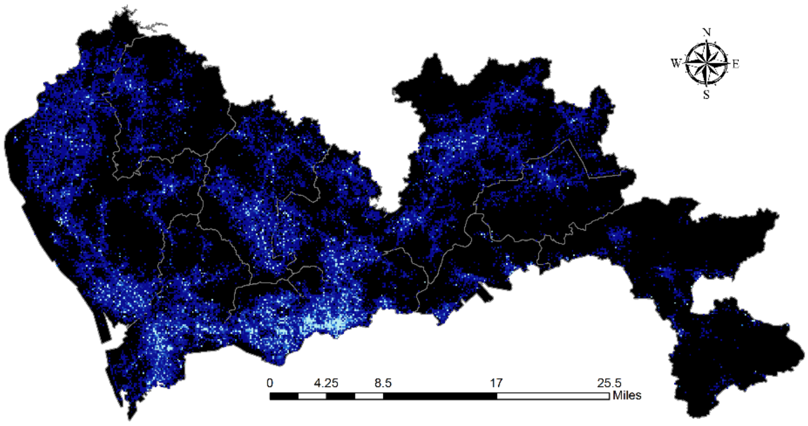

Fig. 2.3. Spatial distribution of collected check-in dataset in Shenzhen

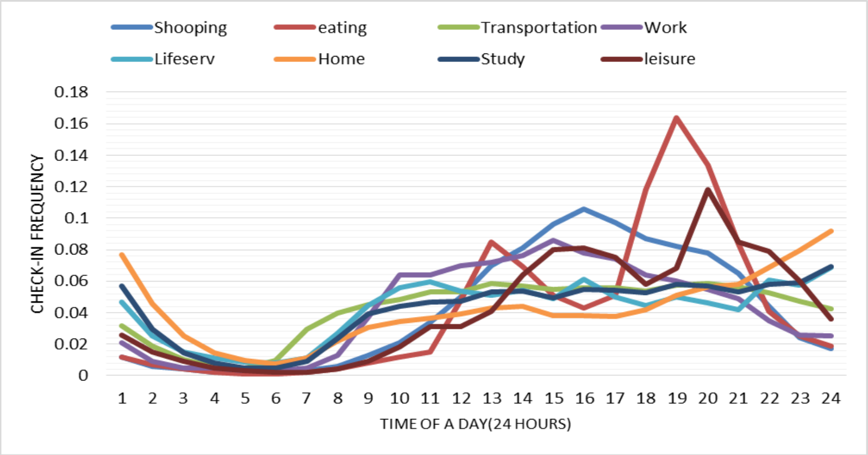

Fig. 2.4. The Check-in frequency for each category at 24 timeslots in Weibo dataset

2.2 The Framework

This paper proposes a novel framework for the probability based labelling of individual activities which can be further used to explore the distribution of functional region at BTS level. The proposed framework is consisted of three steps, as implied in Fig.2.5. First, pre-processing is applied to the two kinds of raw data independently. These two data are used for different purposes. Call data is used for extracting individual trips and providing information that reflects peoples' daily movements. Check-in data is used as a prior knowledge of activity information and its transition pattern. Therefore, the two sub-steps include the identification of trip and extraction of activity transition pattern, the result of which are the inputs of the next step. Next, activity labelling model is conducted on the individual trips to identify the hidden real activity information based on prior activity knowledge. The result of this step is a probability distribution of daily activities linked to each individual. In the last step, these daily activities are mapped in the functional regions to obtain the distributions and spatiotemporal dynamics of different functional regions.

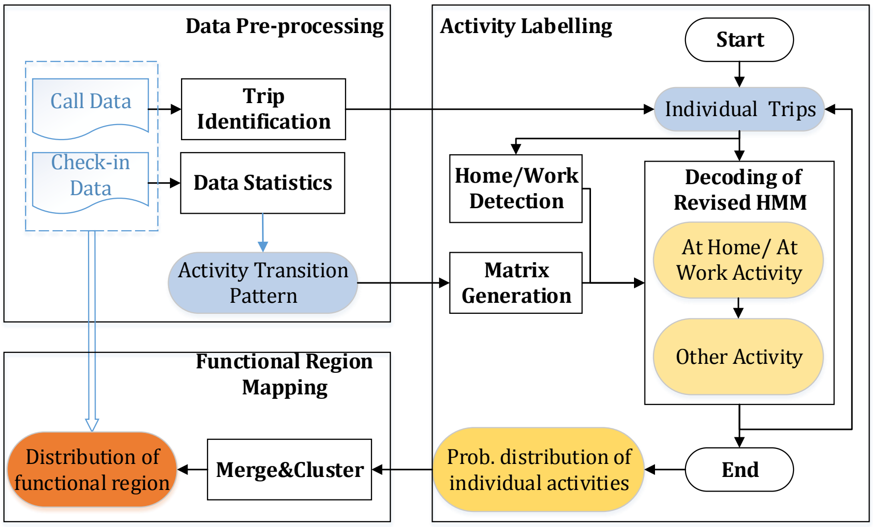

Fig. 2.5. Overview of the framework for exploring the distribution and dynamics of functional regions

2.3 Data Pre-processing

Before proceeding with the next analysis, a filter method of raw data was necessary to eliminate data that are fake and invalid. Three criteria of filter can be as follows: (i) the location of data is not in the study area; (ii) the user who has only one check-in or one call record; (iii) records by the same user have huge quantities in a short period ,which is contrary to common sense. After eliminating both check-in dataset and call dataset, the two datasets practically used are obtained.

Then, the call data and the check-in data are independently processed. For call data, individual trips are extracted according to some space and time criteria. A trip defined in this study is a sequence of consecutive call records that are accord with the spatial and temporal constraints:(i)the time interval between two successive records is less than 12 hours; if the time interval is greater than 12 hours, the two successive records are regarded as low correlation and should be segmented into different trips(Wu et al. 2014); (ii)the stop point should be detected if the locations of successive records remain the same; that is successive records which are at the same location should be combined and replaced with a stop point in a trip. For check-in data, statistical analysis is applied to find the activity transition pattern, which are used to generate three matrices for the next step.

2.4 Activity Labelling

The objective of this step is to label the most likely types of daily activities for individual trips extracted from the previous step. An improved Hidden Markov Model (HMM) is designed for conducting this work.

A HMM is a statistical Markov model in which the system being modeled is assumed to be a Markov process with hidden states(Rabiner 1989).This model can solve the problem such as how do we best explain the hidden meaning of the observations, given a sequence of observations? The advantage of the model is that the states in HMM can well model the sparse trajectory data where users only call at certain places or timestamps, rather than consecutive points as in GPS trajectories(Cheng et al. 2013). Therefore, we define individual trips as the observed states, while the sequence of corresponding activities are the hidden states. Our goal is to identify the hidden states and label them with certain activities.

Individual daily activities exhibit a strong temporal regularity, however, it is necessary to improve the regular HMM to make the model more compatible with spatiotemporal data. Hence, two improvements in HMM, renamed as I-HMM, are proposed in this study.

2.4.1 Home and Work Pre-detection

As we know, detecting homologous home and work locations for a regular person using his or her call trips is a tractable issue due to the spatial and temporal recurrence of these two locations(Isaacman et al. 2011; Schneider et al. 2013). Then information of home and work locations are input into the HMM in advance, where a higher accuracy of the model could be promoted. Inspired by(Calabrese et al. 2013), we developed a statistical algorithm that calculate a score for each location when there is a call at night in the time window 10pm-6am (for home location), or during the day in the time window 9am-12pm and 3pm-5pm (for work location), and select as home or work location with the highest score respectively. The highest score ranging from 0 to 1 must be more than 0.5.

Of course, it is notable to understand that not everyone will have distinct home or work locations. The reasons lie in two aspects: one is the data probably do not show significant recurrence pattern for a portion of persons, the other is some people work at home or they have no fixed work place. We calculated about 10% of people having been detected about their home or work locations at least half of the monitored days based on this call dataset. Nevertheless, the method of home and work pre-detection is meaningful to improve the HMM and produce higher approximations.

2.4.2 Time-inhomogeneous Extension

Unlike the regular HMM that has the assumption that the state transition is independent of time(Rabiner 1989), in other words, it is a time-homogenous process. However, heterogeneity of daily activity patterns have been observed for different time periods of a day(Schlich and Axhausen 2003), adding the time dimension to the state transition along with the observation sequence advancing, called time-inhomogeneous process, is an effective extension for the model. It means the probability of transition between two states at different timeslot is different, except for in a different order. Therefore, the time-inhomogeneous extension copies with the dynamic probabilities effectively and follows the temporal and spatial constraints, stemming from the essence of daily activity patterns, as well. In this case, 24 hours of a day was segmented into three periods: (i) 0am-8am (night period), (ii) 8am-16pm (daytime period), (iii) 16pm-24pm (evening period).

2.4.3 Modeling of I-HMM

We define the I-HMM λ as *(N, M, A, B, π)* that is adapted for our purpose:

- N is the number of hidden states.In our model, the hidden states are activities performed by an individual at certain times. The individual states are denoted as $C = \{C_1, C_2, \dots, C_N\}$, the states at observation t as C_t. The daily activities that we define cover the most population, and have been abstracted to a more generalized level. In this study, eight categories are used, i.e.,*working, at home, transportation, shopping, eating, leisure, life service, studying*. N=8, therefore.
- M is the number of observed states. The observation states correspond to the physical output of the modeled system. For our model, these are distinct locations of BTS, namely, M=2841, representing all possible locations one can appear in cities theoretically. Let us denote the observed locations as $O = \{O_1, O_2, \dots, O_M\}$.
- A is state transition probability matrix $A = \{a_{i_p j_q}\}$, where $a_{i_p j_q} = Pr(S_q = C_j | S_p = C_i)$ for $1 \leqslant i, j \leqslant N$; that is, the probability to engage in type of activity C_j at timeslot q given the previous engaged activity type $C_i$at timeslot p. As detailed in previous section, p, q refer to corresponding number of timeslots of a day.
- B is confution probability matrix $B = \{b_j(k)\}$, where $b_j(k) = Pr(O_k\ at\ p | S_p = C_j)$ for $1 \leqslant k \leqslant N$ and $1 \leqslant j \leqslant M$ representing the probability that an individual is observed at the k_{th} location of BTS at the timeslot p, caused by his or her motivation of the C_j activity.
- π is the initial state matrix $\pi = \{\pi_i\}$, where $\pi_i = P_r(S_1 = C_i)$.

The parameter estimation of the model above is a sophisticated task in this step. A possible solution is that three matrices are generated from the information of activity pattern learned from a statistical analysis of the check-in data as a prior knowledge. Specifically, we approximate the initial state matrix as the percentage of check-in records belonging to each activity class. The state transition probability matrix can be estimated using the activity transition pattern in check-in dataset. Also, based on the fact that the probability of an individual engaging in C_j activity at the k_{th} location of BTS is proportional to the popularity of check-ins belonging to activity class C_j nearby, confusion probability matrix is approximated by calculating the percentage of check-ins for each activity class at different BTS regions.

Given the estimated values of parameter in I-HMM, a complete form model $\lambda = (N, M, A, B, \pi)$ can be used to label the hidden states $HS = \{HS_1, HS_2, \dots, HS_n\}$ from the individual trips $S = \{S_1, S_2, \dots, S_n\}$.

2.4.4 Labelling types of activities

The solution of labelling the types of activities is a dynamic programming problem, namely finding the “optimal” state sequence associated with given observation sequence(Rabiner 1989), which is called the decoding of I-HMM as well.

In this study, We define the $\delta_t(i)$ as the highest probability, at timeslot *t*, which was caused due to hidden states C_i, namely, activity engaged in.

$$\delta_t(i) = \max_{HS_1, HS_2, \dots, HS_{t-1}} \Pr(HS_1, HS_2, \dots, HS_t = C_i, S_1, S_2, \dots, S_t \mid \lambda) \tag{2.1}$$

Further, by induction at timeslot *t+1*, we have $\delta_{t+1}(j)$ to represent the highest probability for activity classC_j, via state transition probability.

$$\delta_{t+1}(j) = \{\max_i \delta_t(i) a_{i_{t+1} j_t}\} \cdot b_j(S_{t+1}) \tag{2.2}$$

Obviously, we need to record the state, that maximized (2.2), for each time t and activity class C_j via $\psi_{t+1}(j)$.In other words, we need to backtrack to the previous state recursively which make the final observation state with the highest probability.

$$\psi_{t+1}(j) = \operatorname*{argmax}_{1 \le i \le N} \delta_t(i)\, a_{i_{t+1} j_t} \tag{2.3}$$

We used the Viterbi algorithm(Forney 1973; Viterbi 1967)to solve the problem. It’s worth noting that the algorithm label other activities except for working or at home only if information of home or work locations were known. Of course, the problem turn to a regular one if the home and work location information is not available.

The result of this step is sequences of activities for all individuals, and then all those are aggregated and the probability distributions of different activities is deduced.

3. Case Study

We conducted our method in Shenzhen as a case study. Shenzhen is a special economic zone and an emerging metropolitan area at the south China, which has the fourth largest resident population in all cities of China, and which contains multiple functional regions such as commercial area, residential area, etc.

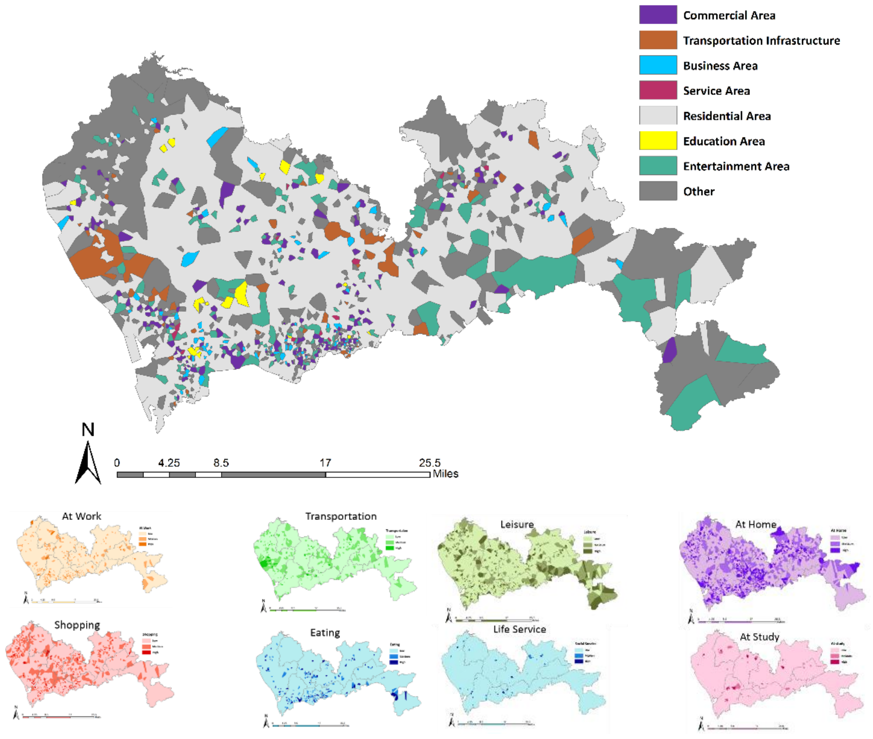

Fig. 3.1. Spatial distribution of urban functional regions in Shenzhen (Top). Spatial intensity distributions of 8 activities at BTS level. (Bottom)

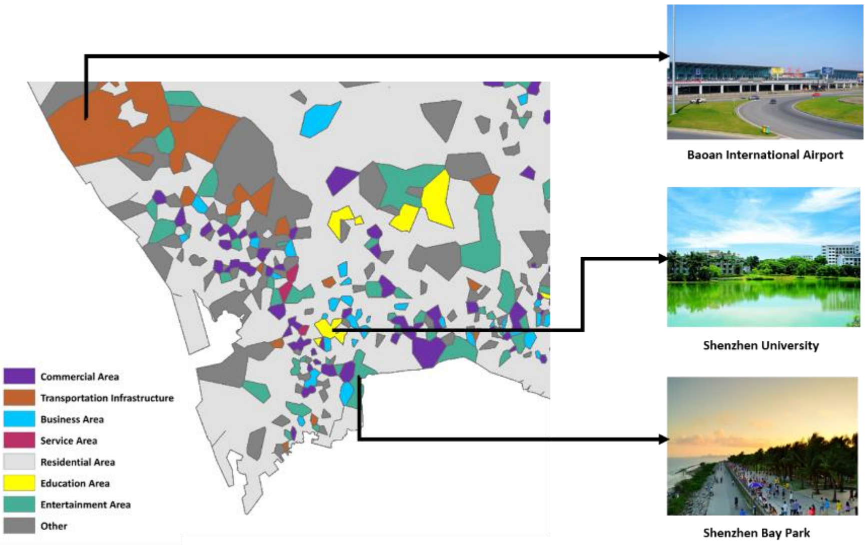

Fig. 3.2. Selected three landmarks compared to Tencent Street View

Fig.3.1 presents the final result of the distribution of urban functional regions, with different colors indicating the most likely types of functional regions. We could see that the distribution is spatially changed. The cores of different functional regions are also different. Residential areas occupy the most urban space, whereas the other functional regions are within the small scopes. The intensity distributions of each activity category are also presented. For example, the distribution of shopping activity shows the regions colored in deeper red having a higher intensity of having a shopping mall or other shopping places than those in lighter red. Statistics on the classification results are shown in Table 3.1.

To evaluate the accuracy, we selected three landmarks for the comparison to Tencent Street View, as shown in Fig.3.2. These landmarks include a university, a coastal park and an airport. All of the obtained functional regions in the corresponding places are in accord with the ground truth, which demonstrates that the proposed method can effectively infer social functional regions at BTS level.

Table 3.1. Statistics on Classification Results

Social Land use	Cell
Commercial Area	221
Transportation Infrastructure	82
Business Area	78
Service Area	13
Residential Area	1631
Education Area	25
Entertainment Area	182
Other	609
Total	2841

4. Conclusion

This paper has presented a novel approach combining mobile phone data and social media data to infer individual activities and to explore the distribution of functional regions. Specifically, we proposed an improved Hidden Markov Model to make the model compatible with the spatiotemporal data. The experiment demonstrated the effectiveness of the proposed approach. The obtained result deepens the understanding of the spatiotemporal dynamic of functional regions. Applying our approach for urban planners can help to obtain a sufficient understanding of urban spaces and its dynamics quickly, capable of supporting management decisions support and emergency response. It is noted that other data sources which can well reflect individual activity and mobility pattern can also be input in this model.

In terms of future work, future validation of the I-HMM in individual level and more in-depth analysis of spatiotemporal pattern of functional regions are necessary.

Acknowledgments. The research was supported in part by the National Sc ience Foundation of China (No. 41371377, 414401444), the Shenzhen Ded icated Funding of Strategic Emerging Industry Development Program (No. JCYJ20121019111128765), the China Postdoctoral Science Foundation Fu nded Project (No.2014M560671), and the Open Research Fund of State Ke y Laboratory of Information Engineering in Surveying, Mapping and Rem ote Sensing (13S02).

References

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr., J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, *26*, 301–313. doi:10.1016/j.trc.2012.09.009

Chen, C., Bian, L., & Ma, J. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, *46*, 326–337. doi:10.1016/j.trc.2014.07.001

Cheng, H., Ye, J., & Zhu, Z. (2013). What's Your Next Move: User Activity Prediction in Location-based Social Networks. In *SDM'13* (pp. 171–179).

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., et al. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, *392*(6), 1459–1473. doi:10.1016/j.physa.2012.11.040

Forney, J., G.D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268–278. doi:10.1109/PROC.1973.9030

Frias-Martinez, V., Soto, V., Hohwald, H., & Frias-Martinez, E. (2012). Characterizing Urban Landscapes Using Geolocated Tweets. *2012 International Conference on Social Computing (SocialCom)* (pp. 239–248). doi:10.1109/SocialCom-PASSAT.2012.19

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. doi:10.1038/nature06958

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., & Varshavsky, A. (2011). Identifying Important Places in People's Lives from Cellular Network Data. In K. Lyons, J. Hightower, & E. M.

Huang (Eds.), *Pervasive Computing* (pp. 133–151). Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-21726-5_9. Accessed 6 October 2014

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Random House.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., & Ma, W.-Y. (2008). Mining User Similarity Based on Location History. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 34:1–34:10). New York, NY, USA: ACM. doi:10.1145/1463434.1463477

Liu, F., Janssens, D., Wets, G., & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications*, *40*(8), 3299–3311. doi:10.1016/j.eswa.2012.12.100

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PLoS ONE*, *7*(5), e37027. doi:10.1371/journal.pone.0037027

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi:10.1109/5.18626

Rodrigue, J.-P., Comtois, C., & Slack, B. (2013). *The geography of transport systems* (Third Edition.). New York: Routledge. http://people.hofstra.edu/geotrans

Schlich, R., & Axhausen, K. W. (2003). Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, *30*(1), 13–36. doi:10.1023/A:1021230507071

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of the Royal Society, Interface / the Royal Society*, *10*(84), 20130246. doi:10.1098/rsif.2013.0246

Song, C., Koren, T., Wang, P., & Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature Physics*, *6*(10), 818–823. doi:10.1038/nphys1760

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of Predictability in Human Mobility. *Science*, *327*(5968), 1018–1021. doi:10.1126/science.1177170

Toole, J. L., Ulm, M., González, M. C., & Bauer, D. (2012). Inferring Land Use from Mobile Phone Activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing* (pp. 1–8). New York, NY, USA: ACM. doi:10.1145/2346496.2346498

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. doi:10.1109/TIT.1967.1054010

Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data. *PLoS ONE*, *9*(5), e97010. doi:10.1371/journal.pone.0097010