Chapter 9
Mobility Pattern Identification Based on Mobile Phone Data

Chao Yang, Yuliang Zhang, Satish V. Ukkusuri, and Rongrong Zhu

9.1 Introduction

Understanding human mobility pattern is a crucial component of urban planning and has applications in analyzing the dynamics of cities, land use changes, and epidemic control. With economic growth and rapid advances in sensing technology, mobile phone ownership and usage is increasing. In China, the number of mobile phone users is close to 1.35 billion by April 2017. Many researches realized that the mobile phone data can be used as an important complement of the existing traffic data collection technology [1–3] in human mobility study. In trip origin–destination (OD) matrix generation, White and Wells [4] obtained the OD matrix with the MOLA data (OD matrix which was obtained from a roadside survey conducted in 1992) and phone calls cost data. Combining mobile phone signaling data with vehicle detection data, Friedrich et al. [5] obtained vehicles OD matrix by identifying the vehicle on the roads using fuzzy algorithm and generating the travel path using Kalman filter. This method can be used for continuous monitoring of road service and network traffic demand. Recently, many researchers focus on the mobility pattern and activity model of human. For frequency trajectory identification, there is significant work on the pattern mining using GPS data, and development of indices such as distance, slope, and spatial similarity to measure the

C. Yang (✉) · Y. Zhang · R. Zhu
School of Transportation Engineering, Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai, China
e-mail: tongjiyc@tongji.edu.cn

S. V. Ukkusuri
School of Transportation Engineering, Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai, China

School of Civil Engineering, Purdue University, West Lafayette, IN, USA

S. V. Ukkusuri, C. Yang (eds.), *Transportation Analytics in the Era of Big Data*,
Complex Networks and Dynamic Systems 4,
https://doi.org/10.1007/978-3-319-75862-6_9

frequency patterns [6–9]. But, they only consider the spatial–temporal trajectory and have not considered the meaning of location for the users. Song et al. [10] calculate the chaotic degree of personal mobile trajectory (entropy) by the anonymous mobile phone users and find that 93% of the users are predictable. This research gives us confidence that it is possible to predict the users' future travel using historical data. Ahas et al. [11] propose a way to use passive mobile phone data (data generated by call and message) to define meaningful points (home point, work point, and secondary point) for users. Phithakkitnukoon et al. [12] use "activity-aware map" to estimate the activities most likely related to the specific space and then build a simple model to describe the activity type of the users. Hasan and Ukkusuri [13] use check-in data to classify the urban activity pattern using topic models. Kung et al. [14] identify the home/work location and analyze the commute mobility using the mobile phone record data and compare the results of different cities. Farrahi and Gatica-Perez [15] use latent Dirichlet allocation (LDA) model to discover the location (home, work, and other) routines of the 97 mobile phone users. They build location sequence bag to represent the mobility information of days, and use topics to explain the mobility patterns. But, 200 topics of their model are too many to explain and it is hard to model the mobility by single topic because the mobility pattern of the day is represented by the distribution of all the topics even though some days only show one topic.

In this research, we first identify home and work locations for the mobile phone users. Following the work of Farrahi and Gatica-Perez [15] and Shih et al. [9], we build the bag of location sequence for all days of users. Then, we develop an LDA model to analyze the location sequence information of the users. We cluster the model results to decipher the mobility patterns of the users and compare the different mobility patterns of the users on weekday and weekend. Finally, representative daily location sequence is captured for each pattern and by measuring the accuracy of the representative feature, we find that the representative mobility feature of cluster can describe the main mobility of the users to a big degree.

9.2 Data and Methodologies

In this study, we use 60 days of the call record data (CRD) of Shenzhen city, China in August, September, and October in 2013. Data of few days were missing. The base station regions (BSRs) defined by the Voronoi diagram are illustrated in Fig. 9.1, and BSRs are used to locate users [16]. Positioning accuracy ranges from 100 m to 2000 m depending on the density of the base station. There are totally 3884 BSRs in Shenzhen city. Samples of the CRD are listed in Table 9.1. Data cleaning is conducted due to lack of field information, matching error with BSRs, wrong IMSI, and duplicate records.

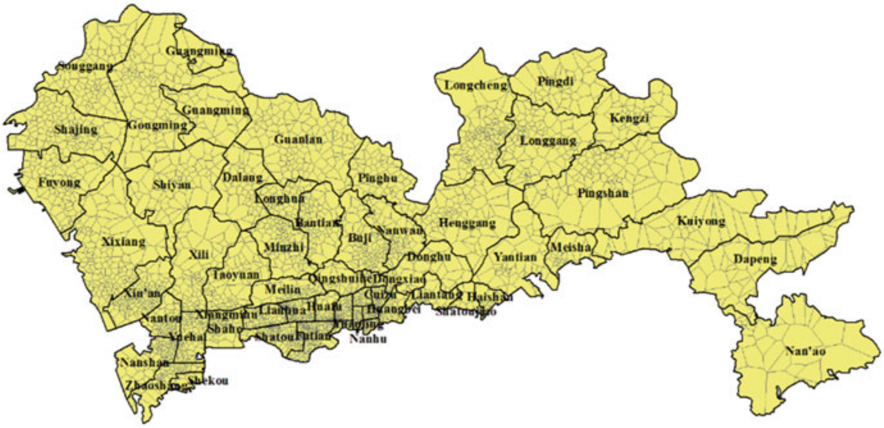


Fig. 9.1 Base station region (BSR) by Voronoi. Fine line is the boundary of the BSR and coarse line is the boundary of the district region in Shenzhen

Table 9.1 Sample of the mobile phone call record data (CRD)

IMSI	BSC	Cellular ID	Sector ID	Call sign	Data and time
4600357544****4	13	200	2	0	2013/08/19 05:45:58
4600357544****1	18	1009	2	0	2013/08/19 23:58:04
4600357544****0	14	131	1	0	2013/08/19 18:50:34

IMSI is the unique sim card ID of user, BSC refers to base station controller. With BSC, cellular ID, and sector ID, the BSR of user location can be identified. For call sign, 0 means dialing, 1 means incoming call, 2 means hard handover, and 3 means null value

9.2.1 Identification of Home/Work Locations

To identify the home and work location, we only choose the data on weekday (41 days). It is considered that the span of 2 weeks (for weekday, 10 days) can relatively show user's mobility pattern rule well in the general case, so we choose the users who both have call records during home time (8 pm–8 am) and working time (9 am–6 pm) for more than 10 weekdays.

Identification of the home/work locations is based on our daily behavior habits. Residents' activity starts from home and ends at home. During the daytime, residents (for commuters) are more likely to stay in their work locations, so most of the calls in working time are made at work locations. The rules of identifying one user's home and work locations are presented below (some definitions are showed in Fig. 9.2):

- For home location:
  - Days are demarcated by 3:00.
  - Record the earliest call (after 3:00) and the latest call (before 3:00) of each day.

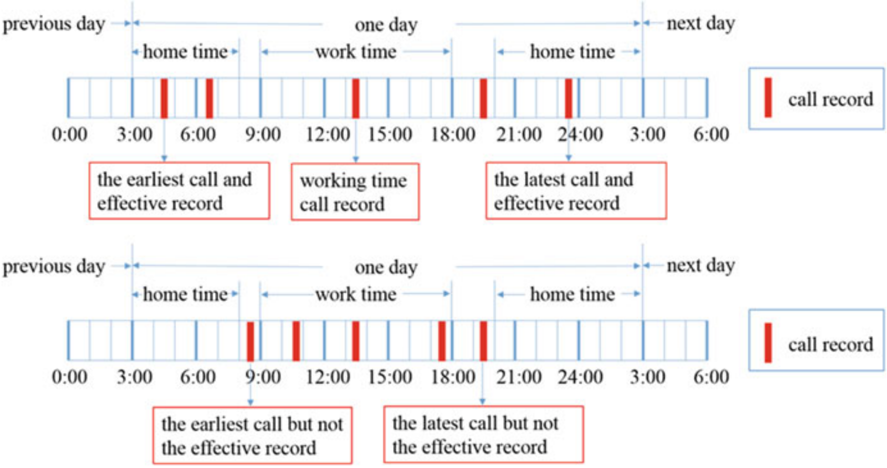


Fig. 9.2 Definitions of working time call record and effective record

  - Define the earliest call records before 8:00 and the latest call records after 20:00 as effective records because during 3:00–8:00 and 20:00–3:00 people are more likely to stay home.
  - Count the effective record frequency of the different BSRs and get the most frequent BSR.
  - If the frequency is more than 10 (at least 1 record per day), we set this BSR as home location for the user.

- For work location:
  - Count the frequency of the working time (9:00–18:00) call records in different BSRs and obtain the most frequent BSR.
  - For the BSR above, count the number of days that have at least one working time call record.
  - If the number is larger than 10, we set this BSR as work location for the user.

We get 10,790,048 call records from 12,846 users whose home location and work location can be identified by the above rules.

9.2.2 *Latent Dirichlet Allocation Model*

Topic model is a type of statistical model for discovering the abstract “topics” that occur in a collection of documents. LDA, introduced by Blei in 2003, is the most common topic model currently used for collections of discrete data. It was originally used for text analysis which can identify the latent topics for documents with a set of words [17, 18]. We can get topic distribution of each given document and

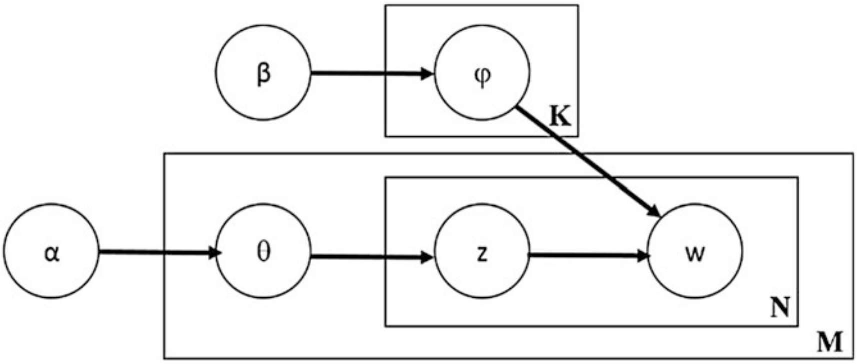


Fig. 9.3 Graphical models of latent Dirichlet allocation (LDA)

word distribution of each topic through word distribution of documents. Nowadays, the LDA model is widely used in the analysis of image, video, and so on. In this paper, LDA model is performed to find the latent mobility topic behind the location information sequence.

Figure 9.3 shows the generative process of the LDA model. Let α and β be the hyper parameters for Dirichlet document-topic distribution and topic-word distribution, respectively. θ is an $M \times K$ matrix of topic proportions for the K topics drawn from Dirichlet(α) and φ is a $V \times K$ matrix of distribution over vocabulary for the K topics drawn from Dirichlet(β). The topic assignments for a given document are $Z = (Z_1, Z_2, \ldots, Z_K)$ drawn from multinomial distribution with parameter θ. The words of the document are $W = (W_1, W_2, \ldots, W_N)$ drawn from multinomial distribution with parameter φ.

The main objective of LDA is to obtain topic distribution of every given document and word distribution of every topic.

Parameter perplexity is used to acquire the best latent topic number of the model [18]. The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance. More formally, for a test set of M documents, the perplexity is:

$$\text{perplexity}\left(D_{\text{test}}\right) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p\left(W_d\right)}{\sum_{d=1}^{M} N_d}\right\} \tag{9.1}$$

where d is document, W_d is a sequence of word in document d, and N_d is the number of words in document d.

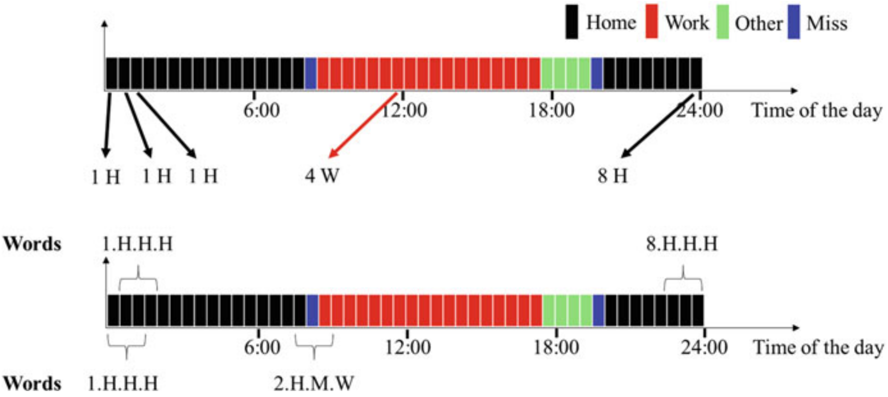


Fig. 9.4 Words generation. One day (one document) consists of 46 words

9.2.3 *Bag of Location Sequences*

We divide each day into 48 timeslots (Farrahi and Gatica-Perez [15]), where every timeslot is 30 min and at the same time, location information for every timeslots (H—Home, W—Work, O—Other, and M—No record) is also labeled. Then, every three consecutive timeslots is considered to be a sequence. Lastly, we add the coarse-grain timeslots label to every sequence (one day can be divided into eight coarse-grain timeslots, (1) 1–7 am, (2) 7–9 am, (3) 9–11 am, (4) 11 am–2 pm, (5) 2–5 pm, (6) 5–7 pm, (7) 7–9 pm, and (8) 9–12 pm). Thus, the words of the LDA model have been generated (see details in Fig. 9.4). By calculating the words frequency vector of the users, we obtain the bag of location sequences, which can be the input of the LDA model.

9.2.4 *Clustering Algorithm: Affinity Propagation*

Affinity propagation (AP) adopts the measures of similarity between pairs of data points to determine the cluster. The number of clusters need not be pre-specified and all the data points are thought to be the cluster centers in the algorithm, named “exemplars.” Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters occur. Affinity propagation found clusters with much lower error than other methods [19].

In this study, we use AP clustering to extract mobility pattern. The input feature is the topic distribution generated by LDA model. The similarity of two mobility topic distribution is measured by Euclidean distance.

9.3 Results and Discussions

For every user, calculate the average number of effective timeslots (timeslots with call records) per day. Figure 9.5 is the user distribution of effective timeslots number. To describe the location information better, we use parameter q, which means the fraction of noneffective timeslots (timeslots with no call records) [10], and select days' of users with $q < 0.8$. Furthermore, we select the days in which the location information of 48 timeslots can be totally and clearly identified with H, W, O, and M (H—Home, W—Work, O—Other, and M—No record). After cleaning, we use 3371 days' records of 287 users on weekday and 1014 days' records of 275 users on weekend.

9.3.1 Finding the Best Latent Topic Number of the Model

The perplexity of different number of topics is showed in Fig. 9.6. The perplexity tends to be stable when k reaches 80 for weekday and 35 for weekend. So, we set k equal to 80 and 35 for weekday and weekend, respectively.

9.3.2 Model Results

We choose $k = 80$ for weekday and $k = 35$ for weekend to calculate the results of the LDA model, and we obtain the daily probability distribution matrix of topics and probability distribution matrix of words for every topic (part of results is shown

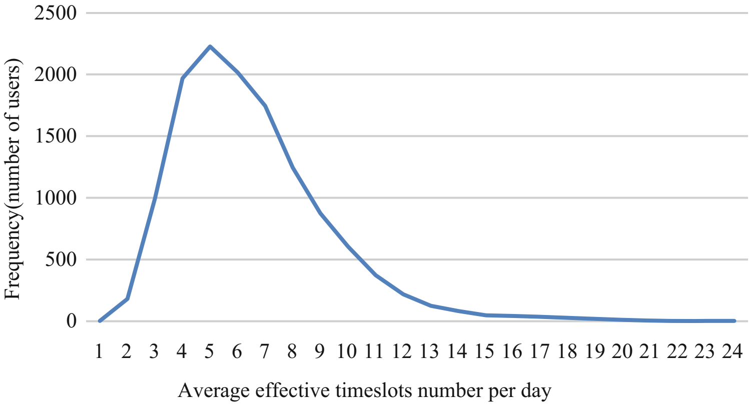


Fig. 9.5 User distribution of effective timeslots number. Number of effective timeslots for most users is less than 10 because the sparsity of the mobile phone record data

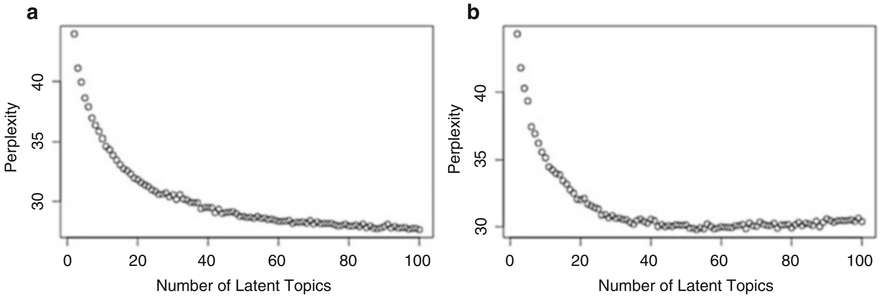


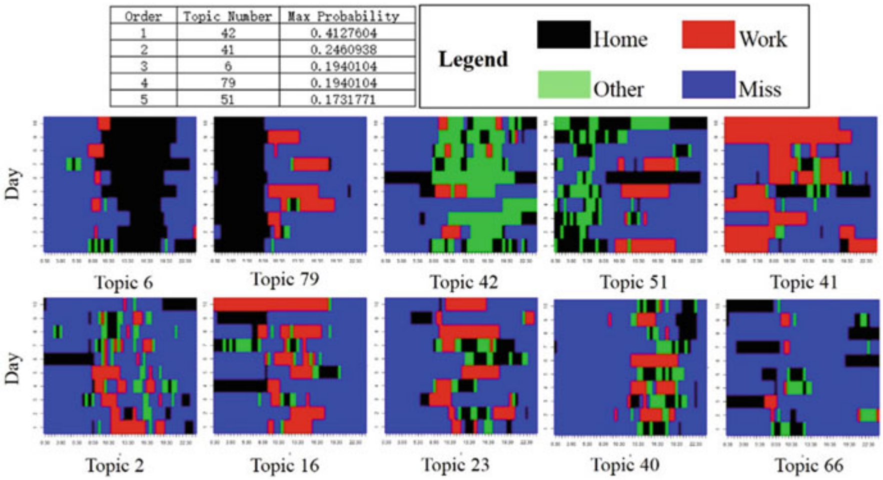
Fig. 9.6 Perplexity of different topics. (**a**) represents weekday and (**b**) as weekend. Perplexity decreases and then gradually becomes stable with the increase of topic number

Table 9.2 Some results of weekday by LDA model

	Topic						
Day	1	2	3	4	5	6	7
1	0.0065	0.0065	0.0065	0.0065	0.0065	0.0065	0.0378
2	0.0065	0.0065	0.0065	0.0065	0.0065	0.0065	0.0065
3	0.0065	0.0065	0.0065	0.0065	0.0065	0.0065	0.0378
4	0.0065	0.0169	0.0065	0.0065	0.0065	0.0065	0.0898
5	0.0065	0.0065	0.0065	0.0065	0.0065	0.0065	0.0169
	Word						
Topic	1.H.H.H	1.H.H.M	1.H.H.O	1.H.H.W	1.H.M.H	1.H.M.M	1.H.M.O
1	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
2	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
3	0.0001	0.0006	0.0001	0.0006	0.0001	0.0012	0.0001
4	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
5	0.7421	0.1077	0.0000	0.0000	0.0000	0.0812	0.0008

For weekday, we get distributions of 80 topics for 3371 days and distributions of 512 words for 80 topics. For weekend, distributions of 35 topics for 1014 days and distributions of 512 words for 35 topics are obtained

in Table 9.2). We observe that a single topic cannot explain the mobility well in our case. As shown in Fig. 9.7, top 5 probability topics are chosen from the topic distribution matrix and the top 10 probability days' location information for each of the 5 topics are plotted. The higher the probability of a topic in a day, the more evident is the mobility related to the topic for this day. Thus, we get the highlighted 5 topics and the days which show the topics most obviously. In this way, we are able to know what mobility the topic means. The reason why results of the topics are not satisfactory will be discussed in next section. Finally, clustering has been performed to get a better outcome.



Order	Topic Number	Max Probability
1	42	0.4127604
2	41	0.2460938
3	6	0.1940104
4	79	0.1940104
5	51	0.1731771

Fig. 9.7 Top 10 probability days' mobility for some topics. Different colors mean different locations of users. The horizontal axis is time of the day and vertical axis shows different days. Each row of the figure represents 1 day's mobility and by this way the mobility of different days can be specifically described. The five figures above are the top five probability topics among the topic distribution matrix and the five figures below are some topics without expectation. Topic 6 and topic 79 are mobility about home location but at different times, topic 42 and topic 51 are about other locations; however, obvious regularity cannot be captured from other topics in this figure

9.3.3 Cluster Results and Analysis

Before clustering the results of the LDA model, two issues ought to be addressed. The reason for expending a lot of effort on clustering the LDA model results instead of using the topics directly to explain mobility patterns and why we do not cluster the location information directly. For the first question, on the one hand, it is hard to model the mobility by single topic because the mobility pattern of the day is represented by the distribution of all the topics even though some day obviously shows one topic and it is impossible to explain all the days' patterns by topics directly. On the other hand, with the sparsity of the CRD (showed in Fig. 9.5), the topics about "M" (missing location information) will account for a large part of the results (e.g., blue part of topic 16, 23, and 40 showed in Fig. 9.7) and the significant information (topics about "H," "W," and "O") may easily be ignored if we just consider single topic. For the second question, performing a cluster first does not yield satisfactory results. The main reason can be attributed to the sparsity of our data. When we calculate the distance matrix of our data, a lot of "M" will adversely impact the final results of the cluster and thus weakens the real information of interest to us ("H," "W," and "O"). But, by considering the similarity of the distribution of the topics, this problem can be addressed. Some results by clustering the location information directly are shown in Fig. 9.8 using the same cluster algorithm and the same data.

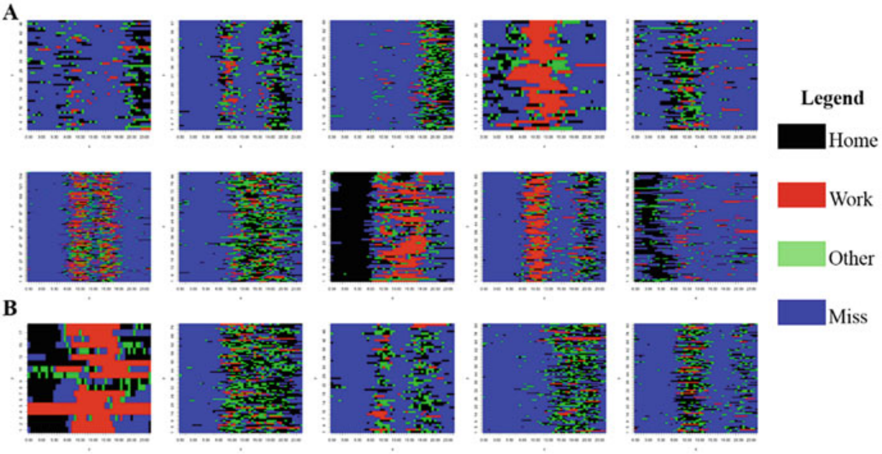


Fig. 9.8 Results of clustering the location information directly. (**a**) is for weekday and (**b**) is for weekend. Details about the figure are shown in Fig. 9.7. Much noise makes it difficult to classify mobility patterns in this way

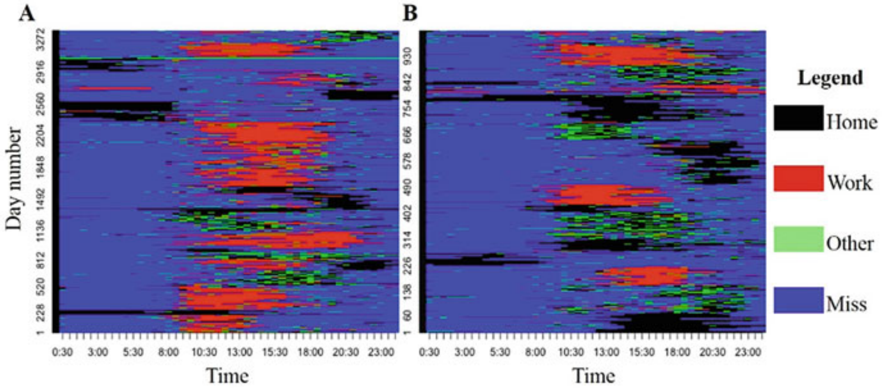


Fig. 9.9 Cluster results of affinity propagation. (**a**) is for weekday and (**b**) is for weekend. Details about the figure are showed in Fig. 9.7. We can see several main patterns in this figure and the different density of the color shows the different mobility on weekday and weekend

We use affinity propagation (*19*) to cluster the topic distribution matrix and we get 25 clusters for weekday data and 17 clusters for weekend data. The location information of all the users are plotted in Fig. 9.9 and users in the same cluster are put together.

Figure 9.9 shows the different classification of the mobility patterns of the days on weekday and weekend. We can easily get the location information during daytime because people usually make calls during this time. But, from 0:00 to 6:00, people make few calls. Concerning this situation, we label the timeslots between two consecutive calls with home or work location when the interval between two consecutive calls at the same location (home or work) is less than 8 h. For other locations, we use 4 h. But, not all users make calls before they go to sleep and after they wake up at home location and this is the reason why there still are a lot of blue grids in the figure. Generally speaking, activities of users at late night are few, and if they have some entertainment or work behaviors at that time, the probability they make calls is higher than usual and we are more likely to capture the location information. Above all, the missing location information does not affect the results too much when we extract the main travel behaviors of the users.

Left part of the Fig. 9.9 is about weekday and users of our data are commute users whose home location and work location can be identified, so the cluster result shows high density for working activities. Most of the users are working between 9 am and 6 pm, which is in accordance with the result of Shenzhen travel survey data in 2010, and some users do not finish their work until 9 pm. In [20], we also conclude that the evening peak extends to 3–4 h in Shenzhen. Few users work on the night shift. And some users are not at the work spots all day, for example, jobs like watchman may work for some days and have some days off. Right part of the Fig. 9.9 is for weekend and we can see lower density of working than weekday. But, some users still work on the weekends and even have higher work intensity in some days. This situation is normal for some jobs like services. Some users stay home all day on the weekends and most users do not leave far away from home (because of the black and green stripes). Few users go out all day, they may go to some remote places to spend their weekends.

Results of different clusters are plotted, respectively, in Fig. 9.10. We can see a very significant regularity on the whole, but there still remains a lot of noise from the micro perspective. Because of the limitation of the CRD, the same travel behavior of a user may show different results by mobile phone data, and sometimes even two calls in the same location may be recorded in different BSRs due to the signal intensity of the base station. This limitation can be showed by the jagged shape of location switching boundaries in the figure. As mentioned before, sparsity of the CRD may result in the loss of some location information of the users, which is another limitation. Hence, the results cannot be used to assess microlevel locational analysis but can be used to aggregate patterns of travel. We have changed the parameters of the AP algorithm to put all the information of the data into consideration. However, too many clusters (e.g., weekday data has 190 clusters) makes our result too specialized to make sense. Regarding the noise as the part of results will make it difficult to capture the real and main mobility patterns of the users. Thus, like the fuzzy processing showed in Fig. 9.11, the mobility patterns are more distinct when we make the streaks fuzzy and negative effects of the noise can be reduced.

In order to better describe the characteristics of each cluster, we regard repetitive behavior patterns of most days of the cluster as the type of mobility pattern for this cluster. If $\varphi_l^i(j)$ represents the count of location label l (H, W, O, M) on the day of number i in the cluster number n during timeslot j ($\varphi_l^i(j) = 0, 1$), then

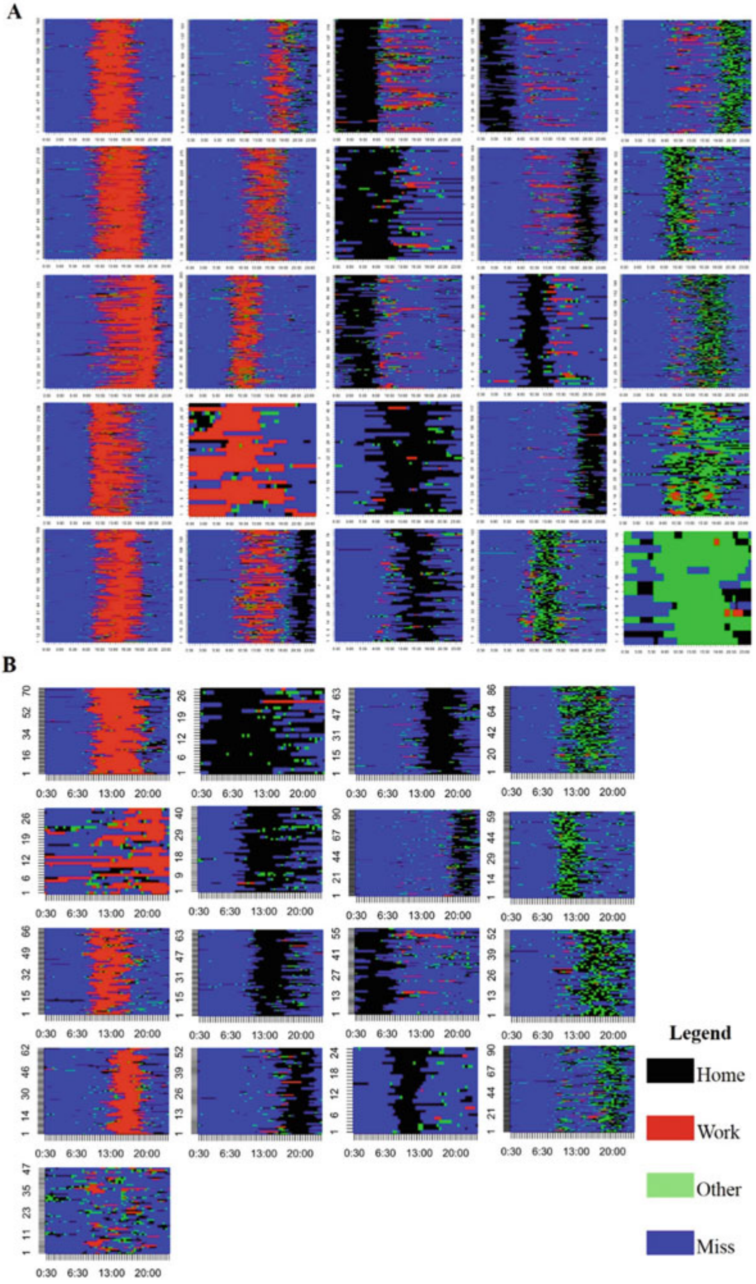


Fig. 9.10 Results of different clusters. (**a**) is for weekday and (**b**) is for weekend. Details about the figure are showed in Fig. 9.7. There are 25 clusters on weekday and 17 clusters on weekend

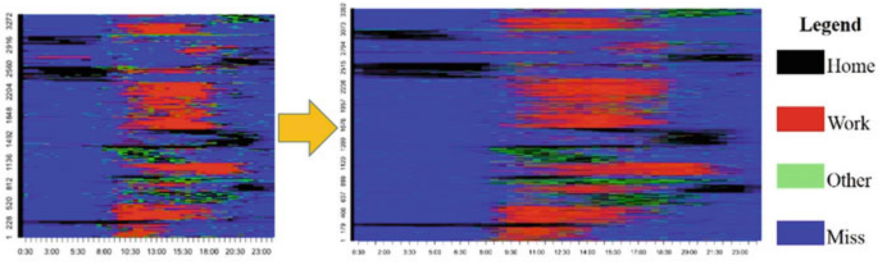


Fig. 9.11 Fuzzy processing of the figure (for weekday). After fuzzy processing, the color becomes more regular and the mobility patterns are more distinct

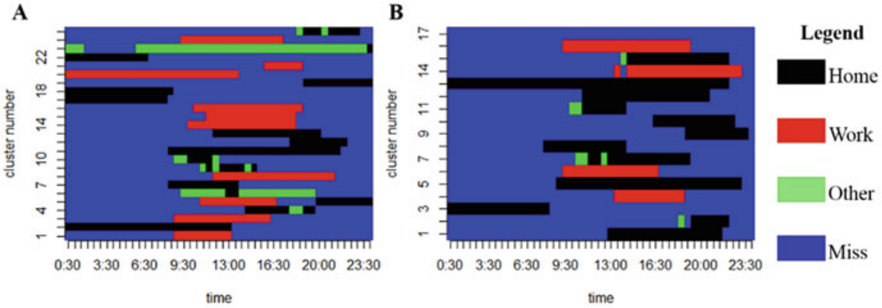


Fig. 9.12 Mobility feature of different clusters. (**a**) is for weekday and (**b**) is for weekend. Different colors represent different locations for users. The horizontal axis is time of the day and vertical axis shows different clusters. Each row of the figure is the mobility feature of one cluster and by this way mobility features of all clusters can be clearly described. It is worth noting that the mobility of cluster 17 on weekend does not have any location information because the cluster shows less regularity (we can see the last figure of Fig. 9.10b)

$$f_n(j) = \arg\max_{l} \sum\nolimits_{i}^{D_n} \varphi_l^i(j) \tag{9.2}$$

where $f_n(j)$ is the assigned characteristic location label for timeslot j of cluster number n and $D_n$is the total day number of cluster n. Thus, we can get the representative mobility feature of every cluster in Fig. 9.12 and we can know what pattern the cluster represents. It can be found that people's main mobility is very regular and focuses on several patterns which can be easily explained by our life experience. Particularly, the mobility patterns are just generated by CRD and we have no idea about the location information for the timeslots with no call. As mentioned above, most of the timeslots with unknown location information are in 0:00–6:00, which is the time for sleeping and with low activity intensity, but if they have some entertainment or work behaviors at this time, it also has a greater possibility to make a call. So, we believe that our results can capture the main mobility patterns of the users. It needs to be noted that all the classes are not completely independent. For example, as illustrated in Fig. 9.10a-1-4 (Figure in the first row and fourth column of the Fig. 9.10a) and Fig. 9.10a-2-4, the main mobility feature of the two clusters is at home before 6:30 (cluster 22) and at home after 17:30 (cluster 12). These two daily behavior patterns of weekdays are not contradicting. As pointed out before, the same behavior pattern may generate different results because of limitation of the CRD and these two patterns maybe two mobility pieces of one behavior captured by our data. Through extracting the behavior characteristics of several days, the complete mobility patterns of users can be acquired.

We use the cluster feature to explain the mobility and the parameter φ is defined to describe the accuracy of the results. Parameter φ is used to measure the similarity between real mobility and the mobility of representative cluster feature. If L_{ij} is the timeslots' sequence for a given day i of cluster number j (the label "M" is not effective location information), and F_j is timeslots' sequence for a given representative feature of cluster number j, then

$$\varphi_j = \frac{1}{D_j}\sum\nolimits_{i=1}^{D_j} \frac{S\left(L_{ij} \cap F_j\right)}{N\left(L_{ij} \cap F_j\right)} \tag{9.3}$$

where φ_j is the accuracy of cluster j, and $L_{ij} \cap F_j$ is the timeslot sequence both have the real location information and cluster feature information in a given day i, and $N(L_{ij} \cap F_j)$ is the number of effective timeslots in the $L_{ij} \cap F_j$, and $S(L_{ij} \cap F_j)$ is the number of the timeslots which have the same location information in the $L_{ij} \cap F_j$ about real and cluster situation, and D_j is the days' number of cluster j. The results of φ_j are showed in Fig. 9.13. The mean value of φ is 0.841 for weekday and 0.837 for weekend. Almost all clusters have reached high accuracy. Low accuracy of some

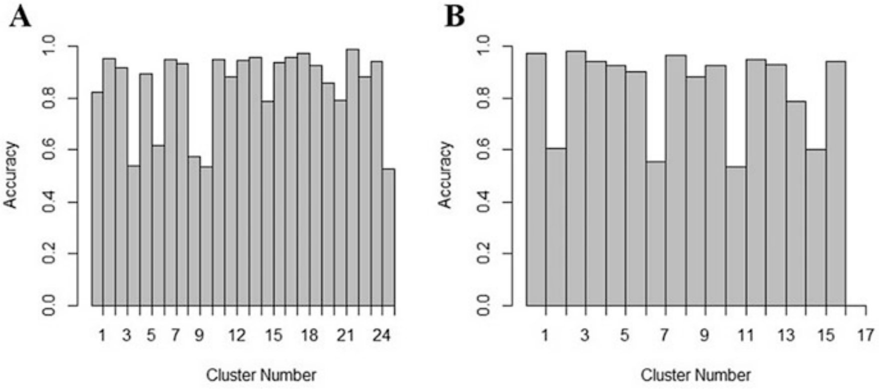


Fig. 9.13 Accuracy of cluster feature. (**a**) is for weekday and (**b**) is for weekend. Almost all clusters have reached high accuracy. Low accuracy of some clusters can be subject to the fluctuation of home and other places (black and green stripes in Fig. 9.10)

clusters can be subject to the fluctuation of home and other places (black and green stripes in Fig. 9.10). On the whole, our cluster results can show one's mobility to a great degree.

9.4 Conclusions

In this study, we simplify users' travel destination by home, work, and other to describe their mobility. By clustering the output of the LDA model, we reduce the negative impact of missing information to the cluster results and identify the mobility patterns of the mobile phone users (25 classes on weekday and 17 classes on weekend). The results are explainable and consistent with our life experience.

Locating users using CRD has its own advantages and limitations. On the one hand, it has a large sample which can be accessed continuously for a very long time and with little deviation (almost everyone has a mobile phone and most traffic modes can be covered, while people have various usage habits). On the other hand, localization error and data sparsity are the main limitation of the methodology, and we are unable to obtain the social attributes of the users. Hence, we would get stuck if we excessively pursuit the accurate result. It is very likely to treat the location error as a part of results, which will make it difficult to capture the real and main mobility patterns of the users. For data sparsity, we just consider the mobility which is shown by the CRD and by combining the behavior characteristics of several days, the relatively complete mobility patterns of users can be acquired. For localization noise, we use LDA model, AP algorithm, and representative feature extraction to capture the main mobility patterns of the users, finding that almost all the mobility of users gathered in several classes.

Many researches indicate that humans follow simple reproducible patterns with high predictability [10, 21], and our results support this conclusion from the perspective of the activity sequence based on home and work location. For commuter, home and work make up a high proportion of daily mobility and show different density on weekday and weekend.

Through learning the mobility of the users, we can predict the future travel behaviors for various users. There still remain doubts about how to choose classifier and how to deal with the situation that all the classes are not completely independent with each other. Therefore, multi-label classification will be an alternative choice in our future research.

Acknowledgments This research was sponsored by National Natural Science Foundation of China (71171147) and Fundamental Research Funds for the Central Universities.

References

1. Z. Wang, S.Y. He, Y. Leung, Applying mobile phone data to travel behaviour research: a literature review. Travel Behav. Soc. 11 (2018)
2. S. Lu, Z. Fang, X. Zhang, S.-L. Shaw, L. Yin, Z. Zhao, X. Yang, Understanding the representativeness of mobile phone location data in characterizing human mobility indicators. ISPRS Int. J. Geo-Inf. **6**(1), 7 (2017)
3. S. Jiang, J. Ferreira, M.C. Gonzales, Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. IEEE Trans. Big Data **3**, 208 (2016)
4. J. White, I. Wells, Extracting origin destination information from mobile phone data, in *Eleventh International Conference on Road Transport Information and Control* (IET, 2002)
5. M. Friedrich, P. Jehlicka, T. Otterstätter, J. Schlaich, M. Friedrich, P. Jehlicka, T. Otterstätter, J. Schlaich, Monitoring travel behaviour and service quality in transport networks with floating phone data, in *Proceedings of the 4th International Symposium Networks for Mobility* (Stuttgart University, Stuttgart, 2008), pp. 1–7
6. A.J. Lee, Y.-A. Chen, W.-C. Ip, Mining frequent trajectory patterns in spatial–temporal databases. Inf. Sci. **179**(13), 2218–2231 (2009)
7. A.A. Shaw, N. Gopalan, Frequent pattern mining of trajectory coordinates using Apriori algorithm. Int. J. Comput. Appl. **22**(9), 1 (2011)
8. S. Abraham, P.S. Lal, Spatio-temporal similarity of network-constrained moving object trajectories using sequence alignment of travel locations. Transp. Res. C **23**, 109–123 (2012)
9. D.-H. Shih, M.-H. Shih, D.C. Yen, J.-H. Hsu, Personal mobility pattern mining and anomaly detection in the GPS era. Cartogr. Geogr. Inf. Sci. **43**(1), 55–67 (2016)
10. C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility. Science **327**(5968), 1018–1021 (2010)
11. R. Ahas, S. Silm, O. Järv, E. Saluveer, M. Tiru, Using mobile positioning data to model locations meaningful to users of mobile phones. J. Urban Technol. **17**(1), 3–27 (2010)
12. S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, C. Ratti, Activity-aware map: Identifying human daily activity pattern using mobile phone data, in *International Workshop on Human Behavior Understanding* (Springer, Berlin, 2010), pp. 14–25
13. S. Hasan, S.V. Ukkusuri, Urban activity pattern classification using topic models from online geo-location data. Transp. Res. C **44**, 363–381 (2014)
14. K.S. Kung, K. Greco, S. Sobolevsky, C. Ratti, Exploring universal patterns in human home-work commuting from mobile phone data. PLoS One **9**(6), e96180 (2014)
15. K. Farrahi, D. Gatica-Perez, Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans. Intell. Syst. Technol. **2**(1), 3 (2011)
16. J.E. Spinney, Mobile positioning and LBS applications. Geography **88**, 256–265 (2003)
17. D.M. Blei, Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
18. D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
19. B.J. Frey, D. Dueck, Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)
20. C. Yang, Y.L. Zhang, F. Zhang, Commute feature analysis based on mobile phone data: case for Shenzhen. Urban Transp. China **14**(1), 30–36 (2016.) (in Chinese)
21. M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns. Nature **453**(7196), 779–782 (2008)