Urban Population Distribution Characteristics Analysis Method based on Mobile Phone Data

Wu Dongdong[1, a], Shi Ruixuan[1, b], Wang Jiachuan[1, c] and Wu Shuqing[2, d]

[1] Beijing Transportation Information Center, Beijing, 100073, China

[2]Beijing Tongtu Permanent Technology Co. Ltd., Beijing, 100161, China

[a]email: wudongdong@bjjtw.gov.cn, [b]email: shiruixuan@bjjtw.gov.cn

[c]email: wangjiachuan@bjjtw.gov.cn, [d]email: wusq@tongtusoft.com.cn

Keywords: Mobile Phone, OD, Information Entropy, DBSCAN.

Abstract. Traditional way of artificial survey statistics for urban population distribution is too time-consuming to adapt to the rapidly changing city traffic flow. In this paper, we propose a method to find out the workplace and residence, according to mobile phone location data and historical trajectory mining. This method is low cost with high accurate statistical results. It is verified by real data in Beijing.

Introduction

Traditional approach to collection of traffic OD matrix needs a lot of personal trip surveys, which will expend a lot of manpower and material resources and renovate data by inches. Because of the limited number of samples and the long collection interval, it is difficult to get the OD information dynamically. In practice, traditional way of artificial survey statistics is hard to obtain high-quality distribution data of urban population, to reflect the actual conditions of transportation accurately, and to draw up transportation plan reasonably. Moreover, in China, rapid economic growth changes the urban population distribution, so it is impossible to investigate traffic continually.

With the development of wireless mobile positioning technology, the location data based on mobile communication base station has been widely concerned by ITS researchers. It has effectively compensated the disadvantages of traditional data acquisition mode with its wide coverage, low cost and strong applicability. Locating the target, map matching its tracks, and performing data processing and modeling, then the population distribution information can be handled. It can be used to obtain the real-time city traffic status in intelligent transportation System (ITS) [1]. Other researchers use cellular phone data to obtain the city's OD matrix [2] and analyze user's mobility pattern [3].Domestic and foreign researchers have already done a lot of work. However, the existing methods cannot meet the demands of large-scale practical application due to its position accuracy, privacy and cost.

Many institutions integrate different type data and dis-cover the relationship of these data which indicates human mobility pattern. Marta C. González studied on the pattern of human mobility, mainly on the distribution of travel length based on mobile data [4]. The Senseable City Laboratory of MIT has done a series of research on the application of mobile data: Santi Phithakkitnukoon use mobile data to identify human activity pattern by clustering based on people's telephone using intensity [5]. They also use Boston's mobile phone data to analyze the crowd mobility during special events [6]. Carlo Ratti discussed the potential application using cellphone data for urban analysis [7]. Researches above are all based on individual trajectory discover. Few researchers have tried to observe the data based on cellular towers' communication intensity. Richard A.Becker, Ramon Caceres from AT&T lab use cellphone data to analyze people flow in and out of city [8], they studied the cellular tower's communication intensity, but only give a brief description about the relation cellular towers communication intensity and the environment surrounds. Montoliu R et al.2010 [9]discovers place-of-interest (POIs) from multimodal mobile phone data, such as GPS, Wi-Fi, GSM,

accelerometer sensors, etc. Yue Y et al.2012 [10]extracts users' semantic features from mobile phone trace data, POIs and real estate price data.

This paper proposes a calculation method of urban population distribution characteristics based on the positioning data of mobile phone base stations, with studying the strengths and weaknesses of existing research results and the actual mobile phone data characteristic analysis. It includes an algorithm of user trajectory calculation based on mobile phone positioning data, a method of finding residents worksite and residence, and verification of actual mobile phone data in Beijing.

Dataset Characteristics

In this paper, mobile phone location data sets are provided by China's largest mobile communications operator. This data set was collected from Beijing, and there have been around 17,000,000 users every day. If the signal is available, our correspondence will record the processes of the occurrence when the wireless is available. Such as, the location of phone base station, the time of triggered action, the communication events, etc. which make up users' signaling interaction. The data sheets are running into about 700 million per day. The average number of daily data was 45 per user. Figure 1 shows statistical data of mobile phone locaion by the hour, the results show that dataset volume distribution conforms to the certain circumstance.

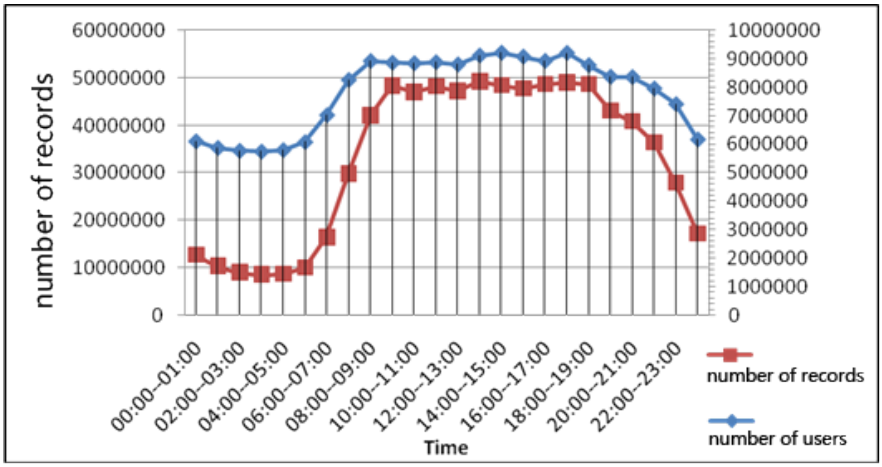

Fig.1. One-day mobile phone location data distribution

a. Coverage Area of Base Station

In this paper, we estimate user's location by analyzing the coordinates of the base station, so the coverage area of base station can directly affect data precision. The data of base station include two parts: urban areas and suburbs in Beijing. The coverage area of base station is range from 500 meters to 2000 meters, and the test result showed that larger coverage areas are lower data precision.

b. Date Sampling Cycle

The sampling cycle of phone signaling data refers to the interval between the time of each recording. The less time interval, the higher the density of sampling is , and the data quantity is enough, which makes the chain of travel spotting more accurate. But phone signaling data use the technology of network location, so this technology is passive. It means when the wireless is available, it will generate phone signaling data. Therefore, the interval of sample data is usually random.

c. Information Entropy Analysis

In 1948, Shannon puts forward the concept of "entropy", which solves the problem on quantitative measures of the information. Information quantity of the message is related directly to uncertainty. So, the element can be thought that the metrics of information is directly proportional to uncertainty in estimates. According to definition of the Shannon Entropy, it can be used to measure the phone

signaling data. The Shannon Entropy is proportional to uncertainty of the phone signaling data and random with the position.

$$S^U = -\sum_{i=1}^{N} P_i log_2 P_i \quad (1)$$

where, P_i is the position probability in N places all day.

Figure 2 shows the cumulative frequency distributing of Shannon Entropy in the whole sample. From frequency distribution, the trend of Shannon Entropy corresponds to normal distribution, and the average of Shannon Entropy is 2.23. Among them, the Shannon Entropy SU is 0 in some bit of user.. This is due to the precision of phone signaling is lower, and the action is not recognized in coverage area of base station .

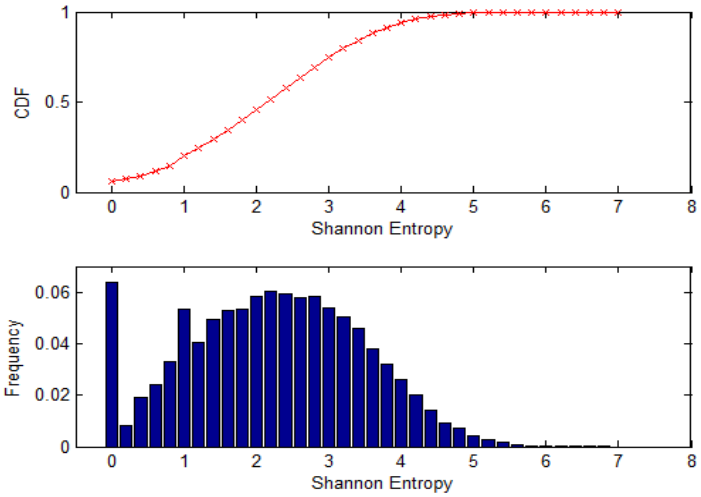

Fig.2. Shannon entropy cumulative distribution and frequency distribution

d. Gyration radius Analysis

Gyration radius, known as inertial radius, is a physical quantity which can be used to calculate rotary inertia. Its physicals meaning is that object will rotate as rotary inertia with a torque working with it. In this paper, the active range of mobile phone users will be quantified by turning radius.

$$R_G = \sqrt{\frac{1}{L}\left(\sum_{i=1}^{L} \overrightarrow{R_i} - \overrightarrow{R_{cm}}\right)^2} \quad (2)$$

where, $\overrightarrow{R_i}$ is position of the i^{th} record of user, $\overrightarrow{R_{cm}}$ is the average of user all day position.

Gyration radius cumulative distribution and frequency distribution is shown in Figure 3. The average of gyration radius is 1.9 km. 70% of users' radiuses cumulative are below 2 km. It shows that daily activities of most users are within 2 km. This feature accords with the daily experience.

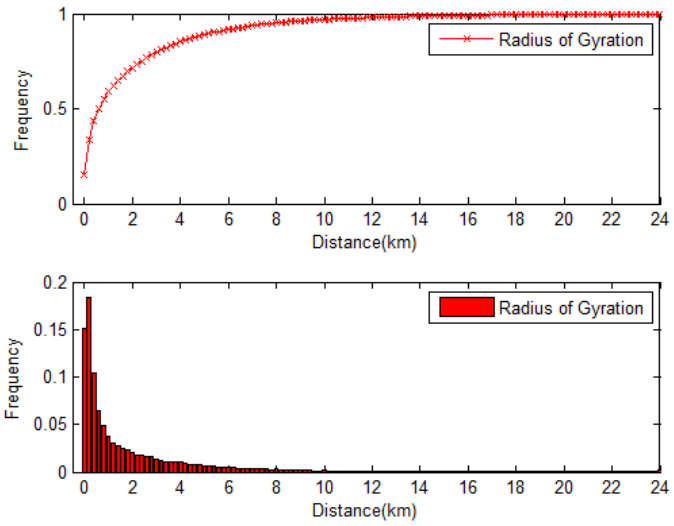

Fig.3. Gyration radius cumulative distribution and frequency distribution

e. Summary of Dataset Characteristics

Comparative study shows that, because of the impact of the coverage area or the base station re-election or the sample cycle, the quality of phone signaling data has lots of disadvantages, such as limited precision and low quality. But phone signaling data as traffic data sources is considerably larger than other data, and high stability. In the meantime, Quantitative Shannon Entropy has been proven phone signaling data is relatively few uncertainty. And it's also truly reflected in the factual use about radius of gyration. Although it is low quality of phone signaling data, it reflects key characteristics of travel.

Trajectory identification

a. The Definition of Subscribers Trajectory

In order to infer the trajectory, we define user's state as stop and move.

Roughly speaking, a stop state implies user's location does not vary for a long time. We introduce gradient to describe the state. If user stops at one place, both gradients of latitude and longitude at the time must be less than a threshold. Particularly, pedestrian moving of a long distance may also result in small latitude and longitude gradient. To distinguish pedestrian moving, a record will not be defined as stop unless its distance of movement is less than a distance threshold.

Expressed as formula, a stop state can be defined as follows:

$$\begin{aligned} Stop(r_i) = \{ & | g(r_k^i.loc.lat, r_{k+1}^i.loc.lat) | < g_{thresh} \cap \\ & | g(r_k^i.loc.lon, r_{k+1}^i.loc.lon) | < g_{thresh} \cap \\ & dis(r_k^i.loc., r_{k+1}^i.loc) < d_{thresh} \} \end{aligned} \tag{3}$$

where, d_{thresh} denotes the threshold of distance between the location of continuous records. *t*

when a sequence of records of a user *i* $seq_i=\{r^i_1, r^i_2,\ldots,r^i_n\}$ can be gathered in a position, we build a stop record, whose location set as average of above location s, start time as the time of the first record, end time as the time of the last record. A stop record *s* can be defined as follow:

$$\begin{cases} s = (loc,\ ts,\ te) \\ loc = avg(r^i_1.loc, r^i_2.loc, \ldots..r^i_n.loc) \\ ts = min(r^i_1.t, r^i_2.t, \ldots..r^i_n.t) \\ te = max(r^i_1.t, r^i_2.t, \ldots..r^i_n.t) \end{cases} \tag{4}$$

It means a user stops at *loc(x,y)* from *ts* to *te*.

A moving state implies user is traveling, either latitude or longitude must change. In this paper, we represent a moving as:

$$moving:m(loc,t) \tag{5}$$

which means a user travel through $loc(x, y)$ at t.

With stop and moving definition, a trajectory can be represented by an ordered sequence of stops and movings.

b. The Method of Subscribers Trajectory Identification

DBSCAN is a spatial clustering algorithm based on density, which can discover clusters of arbitrary shape and resist noise disturbs. The resultant clustering with high density suits cluster to data characteristics.

For users to stay state recognition based on density clustering algorithm, purpose is to find the maximum of density connected point set as soon as possible. Algorithm thinks characteristic of stay points is a series of point set of high density. When you select an anchor point *r*, and it has more than $MinPts$ points within a radius of d_{thresh}, which indicates that a user is considered stay condition in in all period of time which is covered by these points. This algorithm divides anchor points between important stopover point or noisy points, and eliminate the disturbing of outlier and random error for stop and go judgment. Even if anchor of stay condition was abnormal, and this algorithm could still conclude stopover point, if density of the front-and-back registration is high,

The clustering algorithm looked for the characteristic of anchor point, which agrees with the data characteristics of mobile location-based in users' stay condition. By eliminating the disturbing of outlier and random error, rigorous clustering algorithms should be applied while performing cluster analysis to increase the recognition accuracy

c. Mining Workplace and Residence

This article defines the evening until midnight as Home Time, and working elapsed time as Work Time. For each user, settling stopover point of the range of Home Time as stopover point of potential residence, and the range of Work Time as stopover point of potential work.

User behavior is periodic. With commuters, for instance, most of commuters work during 9 a.m. to 5 p.m. from Monday to Friday， and rest from 0 a.m. to 6 a.m. and 7 p.m. to 0 p.m.in residence. According to the laws of the residents' behavior, during work-time, the longest position of user retention time is seen as job location. Instead, in the rest of the time, the longest position of user retention time is seen as residence. Hence, we put forward the algorithms of work or residence based on DBSCAN algorithm, which includes:

We could extract the data of the residence time distribution from users' multi-day data, and then clustered them for all data. For each result of clustering, which to calculate home time domain D_{ht} and work time domain D_{wt} . , we will define home time domain D_{ht} as the intersection pace times of all data and home time in these clusters, then work time domain D_{wt} as the intersection pace times of all data and work time.

$$D_{ht} = \sum_{stop \in C} (stop.sojourn \cap HomeTime) \tag{6}$$

$$D_{wt} = \sum_{stop \in C} (stop.sojourn \cap WorkTime) \tag{7}$$

Clustering results in the cluster is a key location of the user. The nature of key location is decided by home time domain D_{ht} and working time domain Dwt. When home time domain D_{ht} is higher than working time domain D_{wt}, the location of the cluster is speculated to home place. Otherwise, when working time domain D_{wt} is higher than home time domain D_{ht}, the location of the cluster is speculated to working place.

Test results

a. Permanent Population Analysis

The population of permanent is the residents living in Beijing for more than six months, and the average spend days is greater than or equal to 4 days a week. The calculation results are compared with those of the sixth census data as shown in Table 1. Linear regression results show that both average correlation coefficients are above 0.9, which means that both of them are in good agreement.

Table.1. Calculation results of permanent population

Region and County	Phone Population Data (Ten Thousand)	The Sixth Census Data(Ten Thousand)
Chaoyang	367.31	354.5
Haidian	342.01	328.1
Fengtai	215.22	211.2
Changping	169.65	166.1
Daxing	138.73	136.5
Xicheng	121.48	124.3
Tongzhou	124.74	118.4
Fangshan	95.88	94.5
Dongcheng	91.41	91.9
Shunyi	89.84	87.7
Shijingshan	62.39	61.6
Miyun	47.63	46.8
Pinggu	42.50	41.6
Huairou	38.04	37.3
Yanqing	32.03	31.7
Mentougou	28.01	29.0
Total	2006.87	1961.20

b. Space Distribution of Workplace and Residence

The distribution of residents' work place and residence in Beijing is obtained from the model analysis, as shown in Figure 4 and Figure 5. The work place is distributed mainly between the second ring road and the fourth ring road. The density is higher in the east than in the west, and is higher in the south than in the north

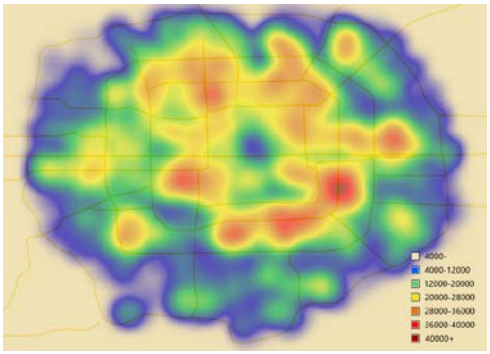

Fig.4. space distribution of residence

The range of work place distribution is more centralized. The main work place is in three business gathering area, CBD, Wangfujing area, Financial Street and Zhongguancun Area.

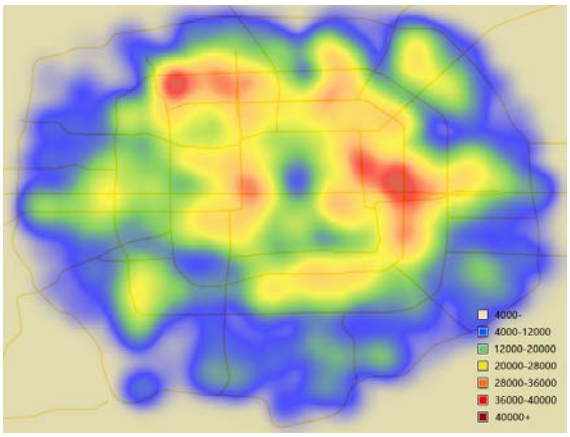

Fig.5. space distribution of workplace

Space distribution relationship between resident population and work population is shown in Figure 6. The area between second ring road and third ring road has the highest concentration of working population and living population, and the density of working population is about 1.5 times of living population which means this area attracts a large number of work commuting trips, formed conspicuous urban tide traffic situation. Inside the fourth ring road, the quantity and the density of working people are both higher than those of living people. Otherwise, outside the fourth ring road, the quantity of working people is less than that of living people, formed radial traffic.

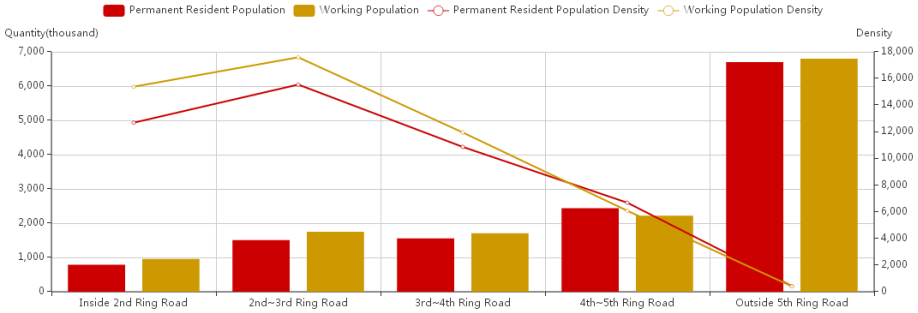

Fig.6. density analysis of population inside the fifth ring road

Conclusion

This paper provides a method to find out the work place and residence of the residents based on statistics and spatial clustering, according to mobile phone location data and historical trace. This method is verified by China Mobile users data in Beijing. Comparing this result with the data from Bureau of Statistics, the result of the model analysis has relatively high accuracy.

Acknowledgement

In this paper, the research was sponsored by Beijing Transportation Information Center and Beijing Key Laboratory for Integrated Transportation Operation Monitoring and Service.

References

[1] Poolsawat, A., W. Pattara-Atikom, and B. Ngamwongwattana, Acquiring road traffic information through mobile phones, pp.170-174, 2008: IEEE.

[2] Chung, E. and M. Kuwahara, Mapping personal trip OD from probe data, International Journal of Intelligent Transportation Systems Research, vol. 5(1), pp. 1-6, 2007.

[3] Laasonen, K., Clustering and prediction of mobile user routes from cellular data, Knowledge Discovery in Databases: PKDD 2005, pp.569-576, 2005.

[4] Gonzalez, M.C., C.A. Hidalgo, and A.L. Barabási, "Understanding individual human mobility patterns," Nature, vol. 453(7196), pp.779-782, 2008.

[5] Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: Identifying human daily activity pattern using mobile phone data," Human Behavior Understanding, pp. 14-25, 2010.

[6] Calabrese, F., F. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti, "The geography of taste: analyzing cell-phone mobility and social events," Pervasive Computing, vol., pp. 22-37, 2010.

[7] Ratti, C., S. Williams, D. Frenchman, and R. Pulselli, "Mobile Landscapes: using location data from cell phones for urban analysis," ENVIRONMENT AND PLANNING B PLANNING AND DESIGN, vol. 33(5), pp. 727, 2006.

[8] Becker, R.A., et al., "A tale of one city: Using cellular network data for urban planning," IEEE Pervasive Computing, pp. 18-26, 2011

[9] Montoliu R, Gatica-Perez D. Discovering human places of interest from multimodal mobile phone da-ta[C]//Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia. ACM, 2010: 12.

[10] Yue Y, Chen J, Hu B, et al. Labeling Personal Charac-teristics from Mobile Phone Traces[C]//the 2nd Interna-tional workshop on mobile sensing, IPSN'12 and CPSWeek. 2012.