Analysis of Urban Traffic Based on Taxi GPS Data

Li Meng[1,2], Li Ru-tong[1,2,*], Xia Yong[1], and Qin Zhi-guang[1]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[2] Popular High Performance Computers of Guangdong Province / Key Laboratory of Calculation and Application Service of Shenzhen, Shenzhen, 518000, China
Lrutong@gmail.com, {mengli,xiayong,qinzg}@uestc.edu.cn

Abstract. Recently, the problem of traffic jam in major cities is getting worse. By leveraging the taxi GPS data of Shenzhen, this paper analyzes the urban traffic status and proposes rational suggestions for urban traffic management. In particular, this paper firstly presents the get-on and get-off points on GIS map based on taxi GPS data. Secondly, by using K-Means algorithm to allocate urban traffic cells, the hot areas where passenger flow is huge are pointed out. Finally, based on the taxi speed, we locate the crowded area and find the crowded period, then analyze the reasons that cause the traffic jam and propose rational suggestions for urban traffic management.

Keywords: GPS data, clustering algorithm, map matching, traffic analysis.

1 Introduction

Currently, many taxis are equipped with GPS, which can record location, speed, direction and other information of taxis. Taxis, as common vehicle, are becoming an important mobile data source because of its large coverage, accurate allocation and high continuity. Referring to the analysis of taxi GPS data, previous work mainly focused on the travelling characteristics of taxis and passengers [9] [10] [11] [12]. Some researchers proposed that 37% time can be saved for taxi drivers by recommending travelling path based on taxi GPS data. Besides, researchers also did abnormal trajectory detection, identification and other regional function work [1] [2] [19].

Taxi GPS data analysis mainly focused on three key points: (1) Taxi trajectory characteristic analysis, which is about how to extract the reliable trajectories from huge original data and then analyze the moving characteristics of taxis or passengers. (2) As taxis are the public carrier for passengers, its GPS data contains a wealth of passengers' information. How to excavate passengers' information from the GPS data (i.e., flow, density) is a hotspot in research community. (3) Social activity pattern excavation, which is mainly research on exploring the social events and their emerging patterns from passengers' information.

This paper is organized as follows. Section 2 states map matching. By using an open taxi GPS dataset, we presented passengers' get-on and get-off points on the GIS

[*] Corresponding author.

H.-Y. Jeong et al. (eds.), *Advances in Computer Science and Its Applications*, Lecture Notes in Electrical Engineering 279,
DOI: 10.1007/978-3-642-41674-3_140, © Springer-Verlag Berlin Heidelberg 2014

map. Then, we can locate the heavy traffic area on the GIS map. In order to have a better study of the city's traffic flow information, the K-means clustering algorithm is used to divide the city into traffic zones and to locate the hotspots where traffic flow is heavy. In section 3, we divided one day into 12 periods, and defined traffic jam circumstance based on taxis speed. By analyzing the experimental results, we found the congested periods of the corresponding roads and the potential reasons for congestion. Then, based on our results, we can provide reasonable suggestions for city's road traffic management. Major contributions of this paper are as follows: (1) Map-matching taxis data to visualize the passengers' get-on and get-off points; (2) Using K-means algorithm to divide traffic zones and locate hotspots; (3) Locating road congestion for urban road traffic management and providing reasonable suggestions.

2 Road Conditions Analysis Methods

The paper is based on the analysis and processing of Shenzhen taxi GPS data, and mapping the Shenzhen road traffic conditions. In section 2.1, we marked the locations of the taxis at the latitude and longitude coordinates for the horizontal and vertical spatial distribution map, drew a taxis travelling map, and determine taxis driving traces. However, due to the mobility of taxis, taxis dirving traces may deviate from the original tracks. In order to properly represent the taxis moving tracks, we matched passengers’ get-on and get-off points onto the coordinate map, and located the main areas of social activities on the GIS map. In section 2.2, K-means algorithm was used to conduct a traffic cell division that was further used to analyzed the city's road conditions. In section 2.3, we defined traffic jam circumstance and found the peaks of traffic congestion.

2.1 Map Matching

Map matching is mainly used to correct digital maps error, GPS location error and coordinate conversion error. Because of the three kind of errors, there could be a deviation between recorded driving traces and their original traces [15] [16]. Therefore, it is necessary to use map matching to adjust and reposition the taxis traces.

In order to locate passengers' get-on and get-off points and find out the range of latitude and longitude, we matched the GPS data with the coordinate map. Figure 1 shows the map matching result. The green dots are the get-on points and the red dots are the get-off points. The abscissa is longitude and the ordinate is latitude. From the figure, it can be found that passengers’ get-on and get-off points cover the entire coordinate map of Shenzhen, and the distribution of such dots is dense in central area and sparse in surrounding, which fits reality circumstance.

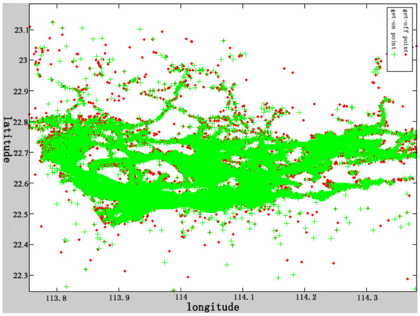

Fig. 1(a). Get-on and get-off points

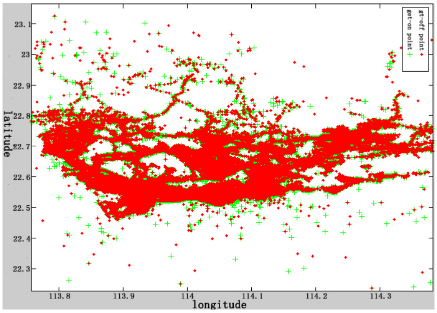

Fig. 1(b) Get-on and get-off points

The basic idea of map matching is as follows. We match the taxis trajectory with vector electronic map, find out current travelling road, and project taxis' current anchor point on the road [17]. Its application is based on two premises: one is precision digital map and the other is taxis on the road, which ensures the positioning error will not breaks away from the original track. Additionally, we only keep the component vector that follows current moving path by using projecting. In this way, the positioning accuracy of the taxis is improved [5] [6]. In this paper, we used traditional map-matching algorithm [7]:

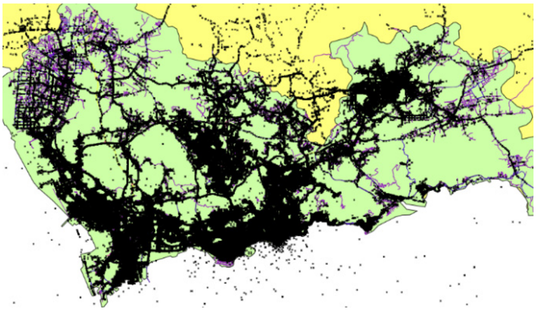

Fig. 2. Map matching

- Step One: Collecting GPS positioning data;
- Step Two: Determining whether the positioning data is invalid or not; if it is, speculating on the historical location data matching, and then turn to step eight;
- Step Three: Judging whether taxis' current state is in a stopped or taxiing state; and if it is stopped, processing it and then goes to step eight;
- Step Four: Within a threshold, if the number of searchable roads is less than 1, it indicates that the taxi is not on the road. Exiting the matching process and considering the GPS data as the taxis' current position;
- Step Five: Within a threshold, if the number of searchable roads equals to 1, it indicates that the taxi is on the road. Projecting directly and considering the taxi's current position is on this road;

- Step Six: Within a threshold, if the number of searchable roads is larger than 1 and the searchable roads are the same road, it indicates that the taxi is nearby this road;
- Step Seven: With a threshold, if the number of searchable roads is larger than 1 and the searchable roads are different roads, it indicates that the taxi is on one of several similar roads;
- Step Eight: the end of this match.

By using the GIS information provided by Mapinfo, passengers' get-on and get-off points are matched to the GIS map. The matching result is shown in Figure 2.

2.2 Traffic Zone Division

The purpose of traffic zone division is to divide the coordinate map into several traffic zones, calculate the throughput of traffic flows for each zone, and compute the taxis' dynamic migration from one zone to another. It can be used to locate hotspot areas and to reflect capacity state [8]. We defined hotspot area as the place that passengers' flow is heavy. Locating hotspot areas can benefit urban traffic management. We defined the accumulated times of passengers' get-on taxis as discharging amount, and the accumulated times of passengers' get-off taxis as absorbing amount. Combining the discharging amount and the absorbing amount can measure the impact of one zone to the city traffic [13] [14]. The zone that has high impact suffers more pressure and it is generally crowded.

We use K-Means algorithm [3] to classify the points on the coordinate map and divide the city into zones to locate hotspot areas. K-Means is a classic data processing algorithm using partitioned clustering methods. The purpose is to find out typical point or central point from mass data and use such point for subsequent processing.

2.3 Road Congestion Detection

We divide one day 24 hours into 12 periods, with each one 2 hours, to calculate the mean speeds of taxis in different periods. The results are shown in figure 3.

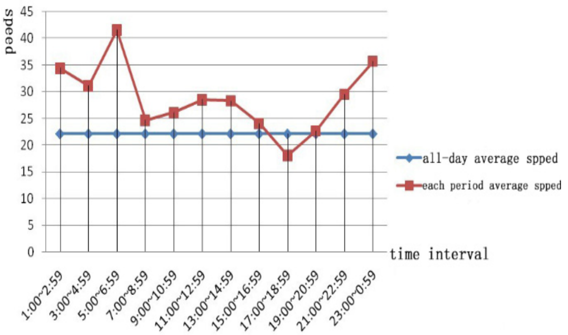

Fig. 3. Speed in each period

It is shown in the figure that two low peaks which indicate the mean speeds of taxis are slow. One of the low peaks is in period 7:00-9:00, and the other in period 17:00-21:00. We can infer that these two periods are associated with high passenger flows. Therefore, we judge these two periods as congested period. Moreover, the mean speed after 19:00 is also low, if we defined the congested path as where the mean speed is below 25 km/h, it is accurately to select 7:00-9:00 (morning peak) and 17:00-19:30 (evening peak) as congested period.

3 Experiment

We used the open dataset, which was collected from April 18, 2011 to April 26, 2011, including 13,799 Shenzhen taxi GPS data. It is workday from April 18 to April 22. And it is weekend from April 23 to April 24. The GPS data contains the taxi license plate number, acquisition time, latitude, longitude, taxis passenger status, instantaneous speed, driving directions and other information. By analyzing the taxi GPS trajectory, we dig out potential valuable information, such as hotspot areas, traffic congestion and so on.

3.1 Data Processing

When processing the taxi GPS data, we firstly eliminated duplication, erroneous and incomplete data, then used the K-Means algorithm to cluster points on the coordinate map, finally divided passenger travelling areas into traffic zones to locate hotspot areas. Table 1 shows the database.

Table 1. Taxi GPS database

Index	Content	Name	Type	Size	Note
0	GPS_ID	ID	int	20	unique identification code
1	car number	CarNumber	varchar	15	Car identification code
2	longitude	GPS_X	float		coordinate
3	latitude	GPS_Y	float		coordinate
4	date	GPS_Date	varchar	16	0 for vacant 1 for laden
5	capacity state	GPS_State	varchar	3	Crowded identification
6	instantaneous speed	GPS_Speed	float		True

3.2 Traffic Monitoring

Since Shenzhen traffic road situation is complicate, firstly we divided the whole area into seven major areas, then for each major area we subdivide it into 50 small blocks, so it is capable of locating exact congested points. Figure 4 shows the result of zone division and the matching with coordinate map. In figure 4, two different colors represent different traffic zones. The points, which are divided into the same traffic zone, have certain correlation and similarity. Each zone reflects the temporal and spatial variation characteristics of urban road network traffic. We found that the center point of each zone also matches the center point of seven administrative districts coordinate of Shenzhen.

As shown in figure 4, based on passengers' get-on and get-off points, traffic zones division approximately matches the seven administrative districts of Shenzhen, although there is some difference. In particular, the left black area matches Guangming District, the left green area matches Bao'an District, the left blue area matches Nanshan District, the red area matches Futian, the right blue area matches Luohu District, the right green area matches Yantian District, and the right black area matches Longgang District. Apparently, the traffic zones division is also closely related to urban population, area, economic characteristics and industrial structure. Next, we can analyze the characteristics of residents' travel traces.

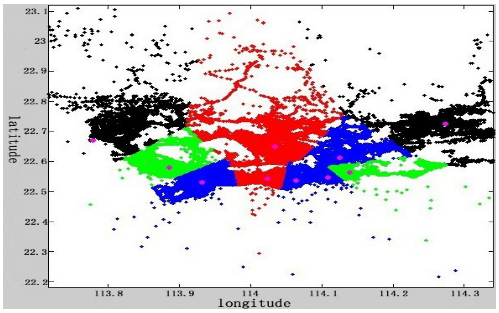

Fig. 4. Passengers' get-on and get-off points divided into traffic zones

In order to verify the accuracy of the zone division, we divide the passengers' get-on and get-off points into 50 blocks with the size of 400*400 square meters, and calculate the taxi throughput in each block, then illustrate the regional function of each zone. Figure 5 shows the divided 50 blocks.

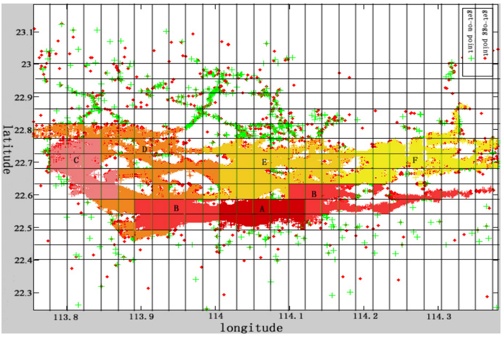

Fig. 5. Flow chart

As shown in figure 5, we use different colors to represent each region's flow. In particular, magenta represents the largest flow area, follows the trend of red, pink, orange, yellow. It shows that the taxi throughput is reducing. Compare with the map of Shenzhen, we infer as follows. Area A is mainly for commercial entertainment. Area B is mainly for resident, government and community, and relatively few for commercial entertainment. Area C is mainly for residential land and industrial land, where the industrial land takes a larger proportion. Area D, E, F are mainly for industrial land, residential land and ecological land, where the ecological land takes a larger proportion.

3.3 Road Congestion Detection

Here, we locate the congested roads. As stated in section 2.3, period 7:00-9:00 and period 17:00-19:30 are congested periods. 7:00-9:30 is the morning peak and 17:00-19:30 is the evening peak. If the mean taxis speed is below 25km/h, we consider roads as crowded. Figure 6 shows the corresponding congested roads in the morning peak and evening peak. The abscissa is the longitude and the ordinate is the latitude. The red dots represent the congested roads where the mean taxis speed is below 25km/h, the green dots vice versa.

As shown in figure 6, we can infer that the major congested roads in morning peak are as follows: Nanping Expressway and Qingping speed intersection, Mei Guan interchange at North Central Avenue Road intersection with Nigang, North Central Avenue East, etc. The major congested roads in evening peak are as follows: Fulong Nanping Road and the intersection of Riverside Avenue and the new Island Road intersection, CaiTian Road and Fuhua Road intersection, Honey Lake Road intersection with Riverside Avenue, Riverside Avenue and Hohai Avenue intersection, etc.

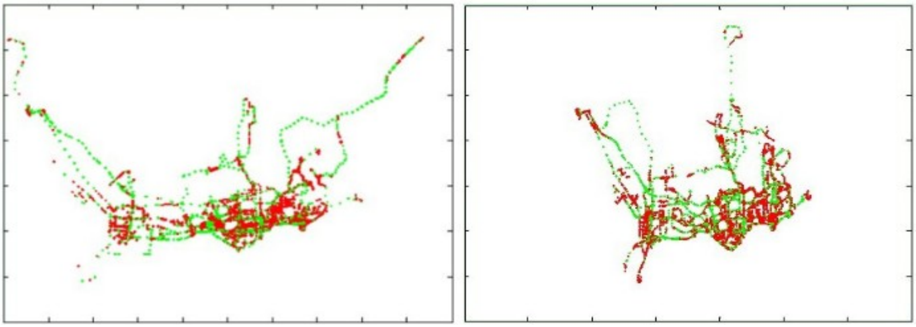

Fig. 6(a) Morning peak road conditions

Fig. 6(b) Evening peak road conditions

4 Conclusion

In this paper, based on taxis GPS data, we presented passengers' get-on and get-off points on coordinate map, and used the K-means cluster algorithm to divide urban traffic zones and locate the hotspot areas. By combining taxis' moving speeds and location information, we can infer the congested periods, congested roads, and further provide rational suggestions to improve urban traffic management.

Acknowledgement. This work is supported in part by the National Science Foundation of China (No.61133016), Ministry of Education - China Mobile Research Foundation (No.MCM20121041), Shenzhen Key Laboratory of Service Computing and Applications (SZU-GDPHPCL-2012-11), and Program of Sichuan Province Applied and Basic Research (No.2013JY0116).

References

1. Yu, Z., Xing, X.: Learning travel recommendations from user-generated GPS traces. ACM Transaction on Intelligent Systems and Technology, 2–19 (2011)
2. Yu, Z., Liu, Y., Jing, Y., Xie, X.: Urban Computing with Taxicabs. In: Proceeding of the 13th ACM International Conference on Ubiquitous Computing, pp. 89–98 (2011)
3. Lee, Y.-J., Vuchic, V.R.: Transit Network Design with Variable Demand. Journal of Transportation Engineering 131(1), l–10 (2005)
4. Tom, V.M., Mohan, S.: Transit route network design using frequency code dgenetic algorithm. Journal of Transportation Engineering 129(2), 186–195 (2003)
5. Fan, W., Machemehl, R.B.: A Tabu Search Based Heuristic Method for the Transit Route Network Design Problem. In: The 9th International Conference on Computer-Aided Scheduling of Public Transport (2004)
6. Fan, W., Machemehl, R.B.: Using a Simulated Annealing Algorithm to Solve the Transit Route Network Design Problem. Journal of Transportation Engineering 132(2), 122–132 (2006)
7. Fan, L., Mumford, C.: A Metaheuristic Approach to the Urban Transit Routing Problem. Journal of Heuristic (2008)
8. Fan, L., Mumford, C., Evans, D.: A simple multi-objective optimization algorithm for the urban transit routing problem. In: The Eleventh Conference on Congress on Evolutionary Computation, pp. 1–7 (2009)
9. Qian, Z., Xu, E., Wang, Z., Yafei, D.: DNA Algorithm on Optimal Path Selection for Bus Travel Network, pp. 245–248 (2009)
10. Liu, L.-Q., Zhang, Y.: Research of Urban Bus Stop Planning based on Optimization Theory, pp. 551–554 (2009)
11. Tang, M., Ren, E., Zhao, C.: Route Optimization for Bus Dispatching Based on Genetic Algorithm-Ant Colony Algorithm, pp. 18–21 (2009)
12. Xu, C., Ji, M., Chen, W., Zhang, Z.: Identifying travel mode from GPS trajectory through fuzzy reasoning. In: Proceeding of the 7th International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, pp. 889–893 (2010)
13. Chen, W., Ji, M., Shi, B., Xu, C., Zhang, B., Deng, Z.: A prompted recall interview platform for GPS-based household travel surveys: design and development. In: Proceeding of International Transport GIS Conference, CDROM, Wuhan (2009)
14. Schaller Consulting. The New York City Taxicab Fact Book, Schaller Consulting, Brooklyn, NY (EB/OL)

15. Yang, H., Wong, S.C., Wong, K.I.: Demand supply equilibrium of taxi services in a network under competition and regulation. Transportation Research: Part B 36(9), 799–819 (2002)
16. Yang, H., Ye, M., Wilson, H.T., et al.: Regulating taxi services in the presence of congestion externality. Transportation Research: Part A 39(1), 17–40 (2005)
17. [19] Jing, Y., Yu, Z., Zhang, L., Xie, X., Sun, G.: Where to Find My Next Passenger? In: Proceeding of the 13th ACM International Conference on Ubiquitous Computing (2011)
18. Jing, Y., Yu, Z., Xing, X.: Discovering regions of different functions in a city using human mobility and POIs. In: Proceeding of the 18th SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 20–29 (2012)
19. Yang, D., Cai, B., Yuan, Y.: An improved map-matching algorithm used in taxis navigation system. In: Proceedings of Intelligent Transportation Systems, Beijing, pp. 1246–1250 (2003)