Taxi Origin-Destination Areas of Interest Discovering Based on Functional Region Division

Xuejin Wan, Mengdan Gao, Jianchu Kang, Jianling Zhao
School of Computer Science and Engineering
Beihang University
Haidian, Beijing, China
vannadal90@gmail.com
gao_md@nlsde.buaa.edu.cn
kang@nlsde.buaa.edu.cn
zhao_jl@nlsde.buaa.edu.cn

ABSTRACT: *This paper introduced a new method to discover the taxi Origin-Destination areas of interest based on functional region division to resolve the problem that taxis distribution is unreasonable. The problem occurs due to dispatching t axis is not in line with the taxi demand in different functional regions during different time periods. The method proposed in this paper clustered t axi Origin-Destination data points of the functional region with the improved DB-SCAN algorithm and yielded the areas of interest with the statistical method named Chi-square distribution. The accuracy and credibility of the method was evaluated through two indicators called coverage and hit ratio and a comparison analysis on situations of different regions. The experiment proved that the proposed method can effectively discover and predicted the taxi Origin-Destination areas of interest in different functional regions in different time periods.*

Keywords: Functional Region, Taxi OD Areas of Interest, Taxi Dispatch, Chi-Square Distribution

Received: 13 September 2016, Revised 12 October 2016, Accepted 22 October 2016

1. Introduction

With the accelerating process of urbanization, modernization and motorization, traffic conditions are increasingly serious. More and more citizens travel by public transport means. Taxi is a kind of fast and convenient floating car. However, passengers is not easy to hail a taxi which is manifested by waiting a long time and by the situation that passengers do not know where to hail a taxi in a strange place. This problem has become one of the main challenges which taxi companies and the public are facing which is caused by the following three reasons. The first one is that taxi drivers are not entirely clear to the taxi demand of the whole city in the peak. They will park to have a rest for the serious traffic jam[1][2]. The second one is in the peak, due to concentration of human mobility time and intensive activity areas in some functional regions like commercial region, workspace and railway stations, taxi demand is too great, as a result it is really difficult for passengers to hail a taxi[1][2]. The last one is that passengers are not clear to the nearby road network distribution which makes it hard to discover the best hailing taxi locations[17].

All the situations above can be summed up as the problem that the taxis are dispatched unreasonably which leads to unreasonably taxi distribution in various time periods and functional regions. A series of studies have carried out in the past years to find a fast and reliable method which can meet this objective. Hu, K., Z. He, et[3] proposed a method to resolve the problem. Their main idea is involved of dividing the city into several administrative regions. In a certain period time in a region, according to the ratio of the number of empty taxis to that of all taxis in the region, they can compute the regional empty car ratio. With the ratio of each region, taxi distribution can achieve a balance by transferring taxis from the region with high empty ratio to the region with low empty regions. This method can realize a balance of empty taxi distribution indeed. However, it cannot resolve the problem due to these reasons that dividing administrative region cannot reflect the characteristics of human mobility in the region and the taxi demand is considerably different in regions with different travel characteristics during different time periods. Lee, J., I. Shin, et al.[4] proposed a user-oriented method to discover AOIs. They made a cluster of analysis based on K-means algorithm. However, the disadvantages of the method to resolve the problem are that the algorithm is based on the distance of clustering algorithm and its clustering results are greatly affected by the initial clustering center of the selected class clusters[5]. Taxis parking locations are distributed along the roads and it is unable to determine the center of the intensive area. In addition, the algorithm is applicable not to clusters that are with non-convex shape and vary greatly in size but to the clustering result of convex sets of data. Taxi travels along the main road of zonal distribution within the region[6]. Therefore, the algorithm's defects will seriously affect the accuracy of the final AOIs (using AOIs instead of areas of interesting in the next) discovering.

Yuan, J., Y. Zheng, et al[7]. proposed a method of functional regions division to divide Beijing City into many functional regions depending on human mobility and POIs. The functional region division helps people clearly understand the complex characteristics of urban region and is conducive to a variety of applications such as urban planning, corporate site selection, user journey analysis and social recommendations. The differences between functional regions and administrative regions are that administrative regions are divided in accordance with the political and administrative management while functional regions are divided according to population travel characteristics. So the advantage of the functional regions division is being able to reflect the population travel characteristics, and help scholars to make a clearer analysis of the characteristics ofliving, working and entertainment etc.

In order to resolve the problem of unreasonable taxi dispatching, this paper supposed a method to lead to a reasonable taxi distribution according to functional regions division and travel time periods classifying. The rest of this paper is organized as followings: Section 2 describes the overview of the method, including architecture and modules of proposed method. Section 3 describes the method and validates the method in terms of the indicators coverage, hit ratio and comparison tests. In section 4, a brief conclusion was made.

2. AOIs Extracting

Due to the problem that taxi dispatching is unreasonable, this paper innovatively puts forward discovering taxi OD AOIs based on functional region division.

2.1 Overview

The method can be split into the following three steps. The first step is filtering taxi OD data points in the selected functional region. The second step is grid division of the selected region and computing the weight of each grid. The last step is filtering the grids with the improved DB-SCAN method and merging grids with chi-square distribution. The architecture of the method is showed in Fig.1.

2.2 Definitions

- **Time period Set** D **[1]:** Due to the temporal correlation of urban transport, in order to effectively extract points of interest, this paper divided 24 hours a day into 8 time periods. Suppose that regarding 24 hours as a set, can be written by $D = \{h_i = [3i, 3i + 3) \mid 0 \leq i \leq 7\}$

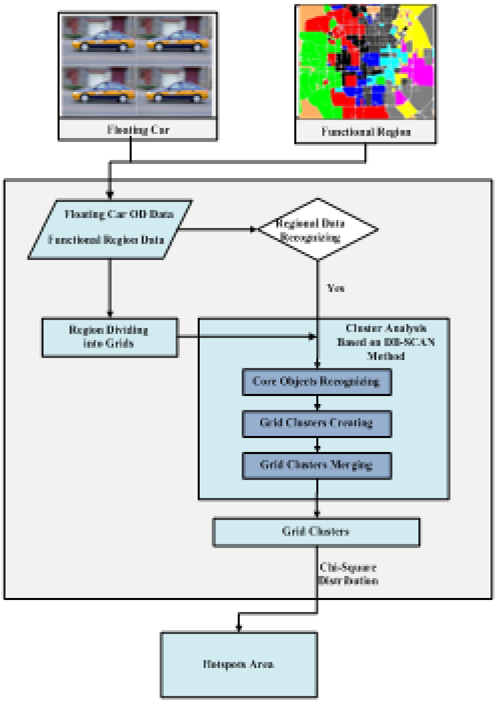


Figure 1. Architecture of our method

- **Characteristic Day Set W:** Regard one week as a set W, *W = {Mon, Tue, Wed, Thu, Fri, Sat, Sun }* and D_0 *= {Sat, Sun }*, D_1 *= {Mon, Fri}, D2 = {Tue, Wed, Thu },* then W can be described as $W = \{ D_i \mid 0 \leq i \leq 2 \}$

2.3 Data Processing Framework

There are three steps to discover the taxi OD AOIs based on improved DB-SCAN algorithm. Firstly, extract OD data from taxi GPS Data. Secondly, compute OD points clusters on different days using improved DB-SCAN algorithm. At last, discover AOIs based on OD clusters using statistical method.

Yuan, J., Y. Zheng, et al.[7] proposed a method of discovering regions of different functions in Beijing which divides the city into work regions, commercial regions, living regions and stations according to the activity regions of the residents. They found, in different function regions, passengers have different regularity of life with periodicity. For instance, passengers in commercial region tend to go out after 12:00. On the basis of the functional regions, this paper focused on the analysis of OD AOIs distribution at different time periods in each region. The OD data of each region were extracted from taxi GPS data which was uploaded from taxi on-board equipment to traffic information center.

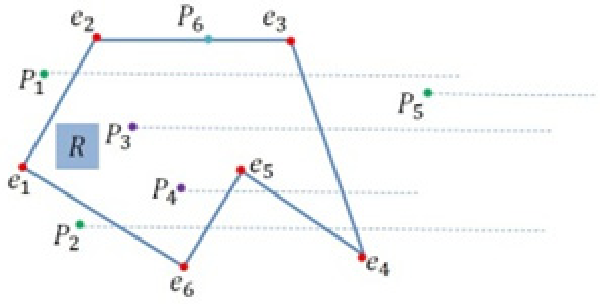


Figure 2. Schematic diagram of the discriminant method

Due to that taxi driving track covers the entire Beijing city road network, this paper used the discriminant method of horizontal/ vertical cross points[8] to extract the OD data which fall in the functional region. As shown in Fig.2, $R = \{e_1, e_2,...,e_7\}$ is a functional region. P_i (i = 1,...,6) is the data point in the same plane with R. First, a ray is drawn from P_i to its right. If P_i is inside R, the ray and the edge of intersect R in odd points, such as P_3, P_4. If P_i is outside of R, the number of intersection points will be an even number (including 0), such as P_1, P_2, P_5. For the points on the boundary of R, such as P_6, whether it belongs to R can be judged according to the total number of intersection points of its rays with each edge of R.

According to the method above, the OD data points can be easily extracted in the functional region. Then, the OD points were clustered using the improved DB-SCAN algorithm[9][10][11][12][18], which eliminated the outliers and for medan area of each cluster, as the candidate for AOIs.

- Split the functional region into $r \times r$ grids. The parameter r varied with the road width and the terrain of the functional region. The value of the parameter r was in table 1. After that, the density of each grid was computed according to the number of OD points falling within the grid. The grid density was initialized to 0 and all the GPS points in the cluster were projected to the grids. If the GPS point belonged to a grid, density of the grid will plus 1. So the problem of discovering AOIs was converted into the problem of clustering grids according to their density.

r	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
Value	15	20	25	30	35	40	45	50

Table1. Parameter R Selected Values In Our Method

- Due to the small taxi quantity and the low taxi demand, this paper did not consider the situation during the time periods which were h_0 and h_1.
- On the premise of the time periods divided above, AOIs were determined through the clustering analysis method and the statistical analysis method. First, at each time period, the grids were clustered by improved DB-SCAN method. Then, the results of the clusters in the same season were put together and analyzed. Because the data in the region followed the chi - square distribution[13][14], the confidence interval of 95%[15][16]was selected to filter the grids. The grids falling within the scope composed AOIs.

3. Experiment and Result Analysis

3.1 Data

This paper selected the GPS data of 12000 vehicles traveling in Beijing road network, which was related to the period of time from Jan 1, 2012 to Dec 30, 2012, a total of 366 days. Each vehicle equipped a GPS device which uploaded the GPS data[19] with the frequency of 30s-70s.Each piece of data was made up of the following fields: company ID, vehicle ID timestamp, timestamp date longitude, latitude, speed, direction, state and event. These data were a collection of all the GPS data collected by the taxi company. For one day, the number of the data records is around 16,128,000. In addition, four dimensional diagram of newly generated 2012 autumn edition in Beijing was chosen as the test map. Beijing west railway station, Guomao Region, Wangjing Region and Xidan Region were selected as regions to be analyzed.

3.2 AOI Coverage

This section analyzed the coverage of the OD AOIs. Taking Xidan region as an example, the region was analyzed using the method based on the data of three months (January to March, 2012).The Fig.3 shows the raw taxi OD data distribution and AOIs distribution respectively in h_6. The green points represent the real OD data points of the taxi passengers. And the red areas represent AOIs which are clustered using the method. Obviously, AOIs cover a majority of the OD points, with the coverage ratio of 78.53%. AOIs, during this time period, are mainly along the gateways and their surrounding roads of the business circles, presenting in a banding shape. Fig.4 shows the three-dimensional distribution of AOIs on the map. The higher the pillar is, the more passengers have got on or got off on the grid, which means a higher density.

Fig.5 shows the AOI coverage to raw OD data points in each time period. Fig.5 (a) (b) (c) represents the data in D_0, D_1, D_2. In

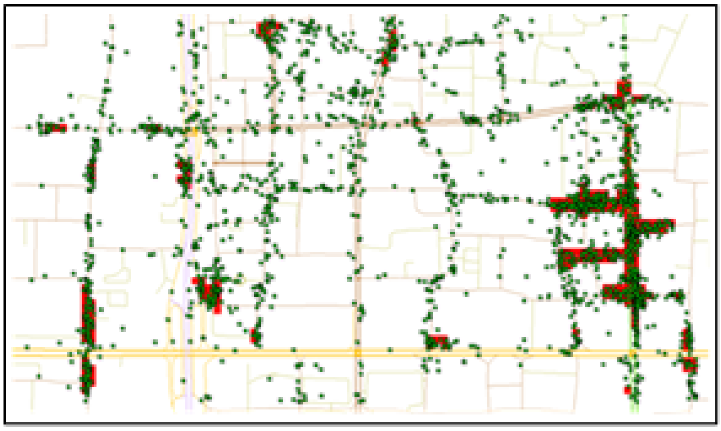


Figure 3. Raw OD data distribution and AOIs distribution in h_6

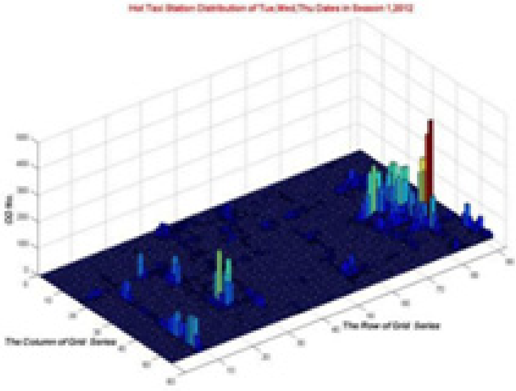
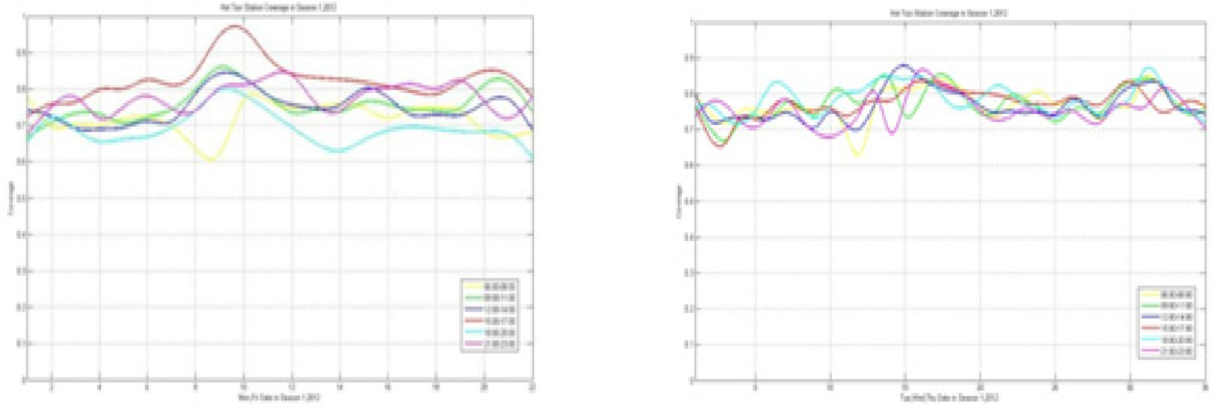


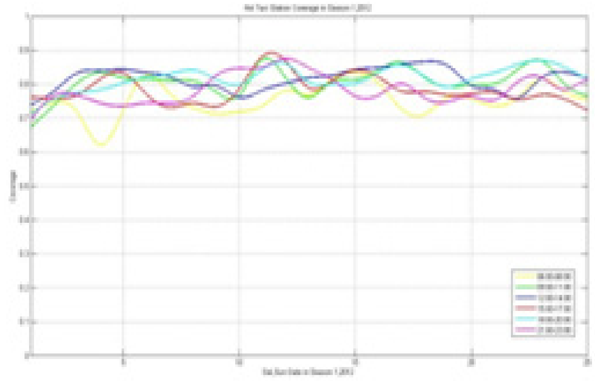
Figure 4. Three-dimensional distribution of the AOIs



(a) Monday and Friday

(b) Tuesday, Wednesday and Thursday

each figure, there are six curves which indicate the coverage of AOIs in $D = \{h_i = [3i, 3i + 3) \mid 2 \le i \le 7\}$. The AOI coverage mainly ranges from 75% to 85%.In Fig.5 (b), the coverage in $h_i\,(2 \le i \le 7)$ in D_2 range from 70% to 85%.



(c) Saturday, Sunday

Figure 5. Different characteristic days coverage in season 1, 2012

Because the numbers of taxi OD data points are various in different time periods, the AOI coverage presents different degree of fluctuation. Fig.6 shows the standard deviation of AOI coverage in season 1. As shown in the figure, the standard deviation of the coverage in h_2 and h_7 is larger than those in other time periods, which was caused by fewer taxi numbers in h_2. This is consistent with the regular pattern of hailing taxi in rush hour in this region.

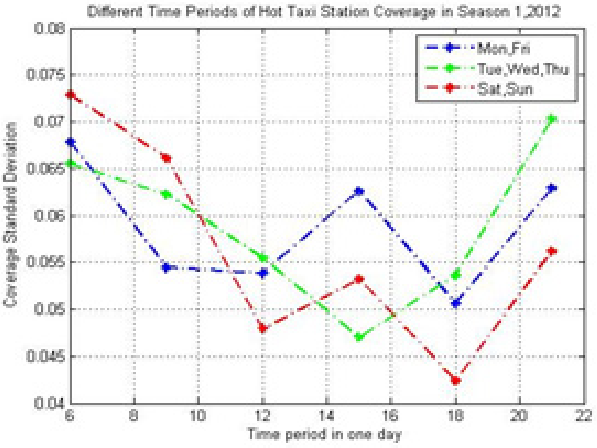


Figure 6. Standard deviation of AOI coverage in season 1, 2012

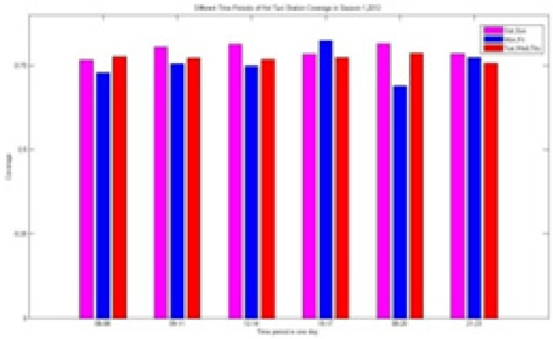
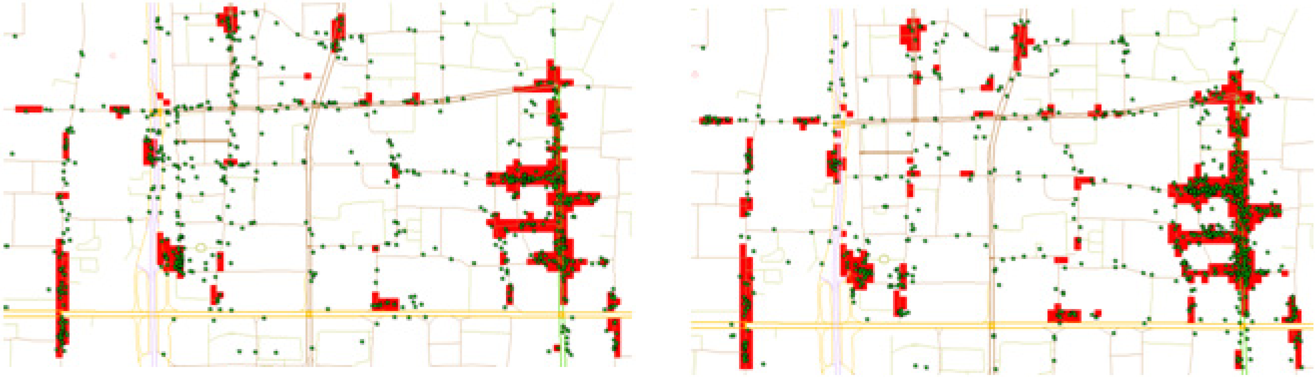


Figure 7. Different time periods of AOI average coverage in season 1, 2012

Fig.7 shows the average AOI coverage in season 1. As shown in the figure, the AOI coverage is about 77.40% in h_6 in D_1, which can also be verified by the distribution of raw OD data points and AOIs in Fig.4.

3.3 Hit Ratio of AOIs

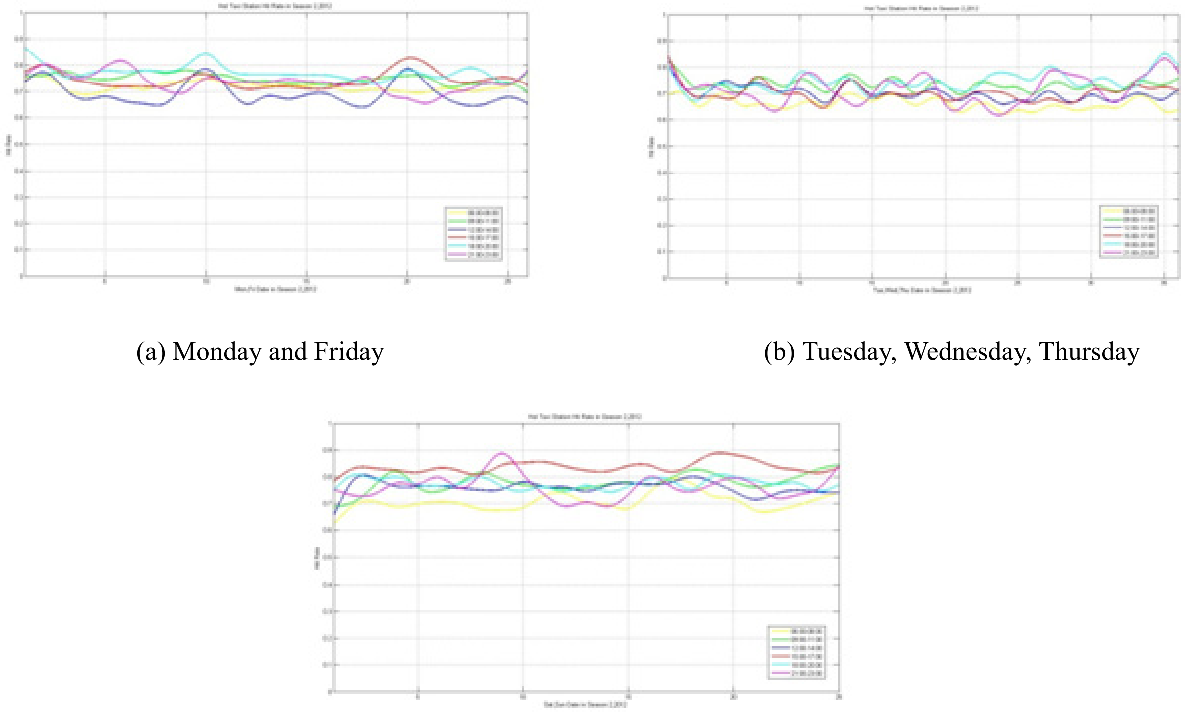
This section analyzes the hit ratio of AOIs. The hit ratio of AOIs means the degree to which OD data points located in the AOIs which was extracted with the above method fall within all the collected OD data points in the region in one day. The OD data points of Xidan region were also selected in the first season to extract AOIs and the data of April were selected to compute the hit ratio of AOIs. Parts of the results are shown in Fig.8. Fig.8 (a) shows the hit ratio of the OD data at h_6 on April 4th 2012, which is 72.03%. Fig.8 (b) shows the hit ratio of the OD data in h_6 on April 7th 2012 is 70.22%.



(a) April 4th 2012

(b) April 7th 2012

Figure 8. Hit ratio of AOIs in Xidan region in different days



(a) Monday and Friday

(b) Tuesday, Wednesday, Thursday

(c)Saturday, Sunday

Figure 9. Hit ratio of different characteristic days in season 2, 2012

The data of season 2 was selected to analyze the hit ratio of AOIs in each time period. The results are shown in Fig.9 (a) which shows the hit ratio of AOIs to the raw OD point prediction in h_i $(2 \leq i \leq 7)$ in D_1 in the second season.Fig.9 (b) ,(c) shows the hit ratio of AOIs to the raw OD points in D_2 and D_0 respectively. As shown in following three figures, AOIs which are computed using the data of season 1 have a high hit ratio (more than 70%) to the raw OD points of season 2.

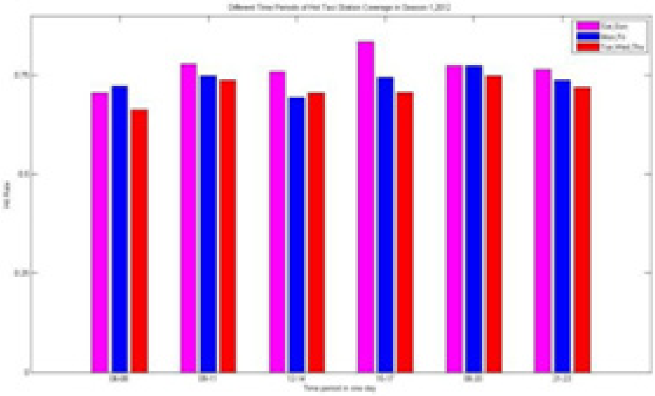


Figure 10. Hit ratio of AOIs in season 2,2012

Fig.10 shows the hit ratio of AOIs to raw OD point prediction in six time periods in one day and three types of days in one week in season 2. As shown in the figure, the hit ratio fluctuated around 70%. The average hit ratios of the three types of days in one week are 76.34%, 73.73% and 71.89% respectively, as shown in Fig.11, which indicates that AOIs determined by the proposed method can be used to discover AOIs where taxi passengers always get on or get off.

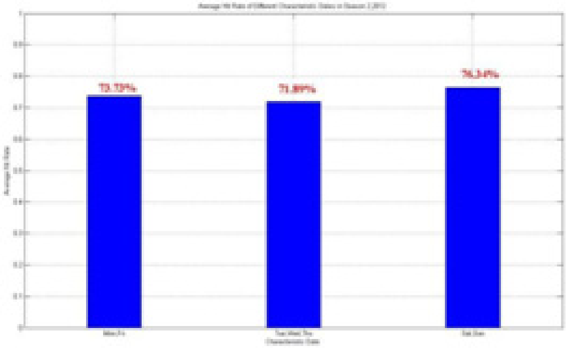


Figure 11. Average hit ratio of different characteristic dates in season 2,2012

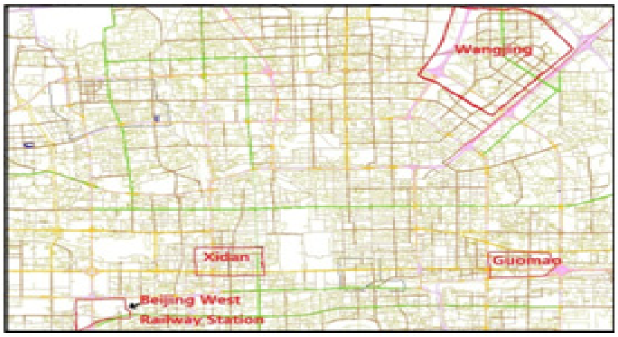


Figure 12. The distribution of four regions on the map

3.4 Comparison Analysis

This section compared AOIs distribution of different functional regions.Four typical functional regions were selected to analyze. They were Xidan region, Guomao region, Wangjing region and Beijing West Railway Station which represented commercial region, workspace, residential region and stations respectively. The distribution of the four regions is shown in Fig.12.

AOI coverage of the regions is shown in the Fig.13, in which the vertical axis is the grid number AOIs covered. Although Wangjing region, which belongs to the residential region, is the biggest one among these regions, its AOI coverage is the lowest. This is because in residential regions, a community is a unit.The gateway of the community is the concentration area of the taxi OD points. However, the number of gateways is small in the residential region which leads to that a small ratio area can cover most of the taxi OD points. AOI coverage of Xidan region, which belongs to the commercial region, is the largest. This is because AOIs of the commercial region are concentrated on surroundings and intersections of the business center, or near the subway stations and the bus stations, etc. and that the demand for taxi in the business region is great.

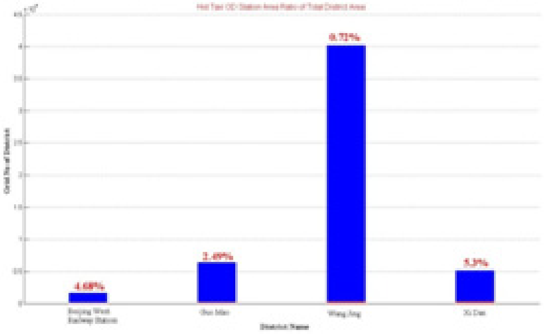


Figure 13. AOI ratio of the total region area

Fig.14 shows AOI coverage of Xidan region in h_6 in D_2 of the four seasons. A conclusion can be made that at the same time period of different seasons, AOI coverage of the region is relatively stable, changing by less than 1%. It is easy to find that from the figure taxi number is various at different time periods, which leads to different AOI coverage. Besides, it is also showing that in Xidan Region, the taxi demand is low in h_2 and h_6. So AOI coverage is low, too. In other time periods, the number of taxi passengers is large in Xidan Region.

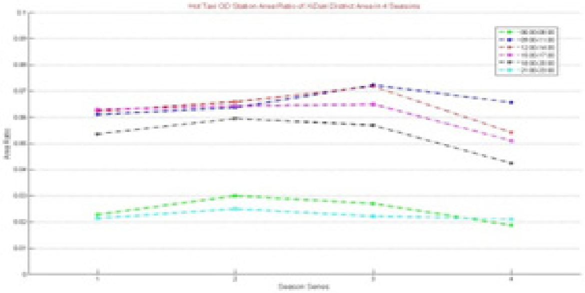


Figure 14. AOIs area ratio of the Xidan region area in four seasons

4. Conclusion

This paper proposed a method to discover taxi OD AOIs in functional regions to make taxi dispatching reasonable. An improved DB-SCAN algorithm was used to make a cluster analysis and the Chi-square distribution was used to make a final AOIs merging analysis. The method was evaluated with the raw GPS data generated by recording over 12,000 taxis in 2012. According to human mobility in different functional regions during different time periods, AOIs of the regions were extracted finally. The result of AOIs was evaluated through two indicators called coverage and hit ratio. Through comparison analysis, the correctness of AOIs depending on human mobility was verified in different functional regions.

In the future, keeping on optimizing this method should be done since the method is not perfect so far. For example, AOIs can be merged according to the road distribution conditions and the actual distance. Besides, the method will also be optimizedcombining other data such as bus station data and subway station data.

5. Acknowledgment

I would like to thank the experts who take their time to check the paper and give me valuable suggestions. At the same time, I would like to thank my partner to help me find out the digital map.

References

[1] Lee, J. (2008). Analysis on the waiting time of empty taxis for the Taxi telematics system. Convergence and Hybrid Information Technology, 2008. ICCIT'08. Third International Conference on, IEEE. 2008

[2] Jian-cheng, W., Ya-qiao, Z. (2009). Floating car data based taxi operation characteristics analysis in beijing. Computer Science and Information Engineering, 2009 WRI World Congress on, IEEE. 2009

[3] Hu, K., He, Z. (2010). Taxi-Viewer: Around the Corner Taxis Are! *In:* Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing (UIC/ATC), 2010 7th International Conference on, IEEE. 2010

[4] Lee, J., Shin, I. (2008). Analysis of the passenger pick-up pattern for taxi location recommendation. Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on, IEEE. 2008

[5] Hartigan, J. A., Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm, *Journal of the Royal Statistical Society.* Series C (Applied Statistics), 28 (1) 100-108.

[6] Wagstaff, K., Cardie, C. (2001). Constrained k-means clustering with background knowledge. Machine Learning- Interntional Workshop, 2001.

[7] Yuan, J., Zheng, Y. (2012). Discovering regions of different functions in a city using human mobility and POIs, *In:* Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data mining, ACM. 2012

[8] Nabighian, M. N. (1972). The analytic signal of two-dimensional magnetic bodies with polygonal cross-section: its properties and use for automated anomaly interpretation, *Geophysics,* 37 (3) 507-517.

[9] Ester, M., H., Kriegel, P. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise, Kdd.

[10] Hinneburg, A., Keim, D. A. (1999).Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering, 1999.

[11] Karlsson, C. (2007). Clusters, functional regions and cluster policies, JIBS and CESIS Electronic Working Paper Series (84) 2007.

[12] Wagstaff, K., C. Cardie, et al. Constrained k-means clustering with background knowledge. Machine Learning- Interntional Workshop, 2001.

[13] Miller, R., Siegmund, D. (1982).Maximally selected chi square statistics, Biometrics, 1011-1016.

[14] Bearden, W. O., Sharma,S. (1982). Sample size effects on chi square and other statistics used in evaluating causal models, Journal of Marketing Research, 425-430.

[15] Lancaster, H. O., Seneta, E (1969). Chi Square Distribution,1969, Wiley Online Library.

[16] Deng, Z., Ji, M. (2011). Spatiotemporal structure of taxi services in Shanghai: Using exploratory spatial data analysis. Geoinformatics, 2011 19th International Conference on, IEEE. 2011.

[17] Chen, G., Jin, X. (2010). Study on spatial and temporal mobility pattern of urban taxi services. Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on, IEEE. 2010

[18] Lee, J., Park, G.-L. (2007). A telematics service system based on the Linux cluster. Computational Science–ICCS 2007, Springer, 660-667.

[19] Liao, Z. (2003). Real-time taxi dispatching using global positioning systems, *Communications of the ACM*, 46 (5) 81-83.