DETECTING HOTSPOTS FROM TAXI TRAJECTORY DATA USING SPATIAL CLUSTER ANALYSIS

P. X. Zhao[1], K. Qin*[1], Q. Zhou[1], C. K. Liu[1], Y. X. Chen[2]

[1]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China – (pxzhao, qink, whu_zhouqing, wishchengkun)@whu.edu.cn
[2]College of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing, China - chenyixiang@njupt.edu.cn

KEY WORDS: Taxi Trajectory, Decision Graph, Data Field, Trajectory Clustering, Urban Hotspots

ABSTRACT:

A method of trajectory clustering based on decision graph and data field is proposed in this paper. The method utilizes data field to describe spatial distribution of trajectory points, and uses decision graph to discover cluster centres. It can automatically determine cluster parameters and is suitable to trajectory clustering. The method is applied to trajectory clustering on taxi trajectory data, which are on the holiday (May 1st, 2014), weekday (Wednesday, May 7th, 2014) and weekend (Saturday, May 10th, 2014) respectively, in Wuhan City, China. The hotspots in four hours (8:00-9:00, 12:00-13:00, 18:00-19:00 and 23:00-24:00) for three days are discovered and visualized in heat maps. In the future, we will further research the spatiotemporal distribution and laws of these hotspots, and use more data to carry out the experiments.

1. INTRODUCTION

Hotspot detection is significant for many applications such as city infrastructure construction, urban transportation planning and management, location-based service, and so on. In recent years, spatial clustering methods are widely used to discover hotspots from trajectory data. Lee et al. used k-means clustering to analyze pick-up patterns of taxi service, and conducted location recommendation for taxies (Lee et al., 2008). Chang et al. proposed a four-step approach to handle the problem of taxi demand analysis, and the performances of three clustering algorithms were compared, including k-means, agglomerative hierarchical clustering and DBSCAN (Chang et al., 2010). Yue et al. used single-linkage clustering to explore time-dependent attractive areas based on taxi trajectory data in Wuhan City, China (Yue et al., 2009). Zheng et al. proposed a tree-based hierarchical structure to model the trajectories of multiple users and used a density-based clustering algorithm to discover interesting locations of different spatial scales, which can facilitate travel and friend recommendation (Zheng et al., 2009). Gui et al. put forward a parallel executed DBSCAN algorithm on the time-focused block data to discover traffic hotspots in different periods (Gui et al., 2012). In a word, clustering methods have been widely applied to trajectory-based hotspot detection. However, the existing clustering algorithms for hotspot discovery have some difficulties in meeting requirement of trajectory data for their heterogeneous spatial distribution, which brings demands to research new methods of spatial clustering.

On the basis of clustering method of Rodriguez and Laio (2014) and the theory of data field (Li and Du, 2007), this paper proposes a method of trajectory clustering based on decision graph and data field. It can automatically determine cluster parameters, and can be effectively applied to trajectory-based hotspot detection for its adaptability to the uneven spatial distribution of trajectory data. The pick-up and drop-off points in taxi trajectory data represent origins and destinations of passengers, so trajectory clustering analysis can be used to discover urban hotspots effectively.

The rest of this paper is organized as follows. Section 2 expounds the proposed method of trajectory clustering based on decision graph and data field. Experiments of trajectory clustering based on the method are carried out to discover urban hotspots in section 3. Section 4 summarizes the contributions of this paper, and analyses the future research directions.

2. METHOD OF TRAJECTORY CLUSTERING BASED ON DECISION GRAPH AND DATA FIELD

Inspired by the field theory in physics, Li put forward data field (Li and Du, 2007), and introduced field theory to data space, which can be used to analyze the interaction among data objects. Based on the clustering method of decision graph (Rodriguez and Laio, 2014) and the theory of data field, the paper put forward a method of trajectory clustering based on decision graph and data field.

2.1 Trajectory data field

Suppose $P=\{P_1,P_2,\ldots,P_n\}$ is a data set consisting of n trajectory points, each point is regarded as a particle with mass, and there exists a virtual field around it. Any trajectory point in this field will receive mutual interaction from other points. Thereby, a trajectory data field forms in this trajectory space. Potential value of P_i is represented as:

$$\varphi(P_i)=\sum_{j=1}^{n}\left(m_j\times e^{-\left(\frac{d_{ij}}{\sigma}\right)^k}\right) \quad (1)$$

Where m_j = mass of trajectory point $P_j\ (j=1,\ldots,n)$

d_{ij} = distance between P_i and P_j

* Corresponding author: qink@whu.edu.cn

$\sigma \in (0, +\infty)$ = range of interaction between points
$k \in N$ = distance index

Many researches (Li and Du, 2007; Wang et al., 2011) have proved that spatial distribution of data field mainly depends on σ and is irrelevant with the specific form of potential function. When $k = 2$, the potential function corresponds to the Gaussian function which has favourable mathematical property. Thus, we fix $k = 2$.

2.2 Decision graph

As a visualized method to select cluster centres, decision graph is proposed by Rodriguez and Laio (Rodriguez and Laio, 2014), which includes two quantities: local density ρ_i and distance δ_i from points of higher density. Those points with both higher local density ρ_i and greater δ_i can be considered as cluster centres.

For data point i, its local density ρ_i can be defined as follows:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \quad (2)$$

Where $\chi(x) = 1$, if $x < 0$; or $\chi(x) = 0$
d_c = cutoff distance
d_{ij} = distance between point i and j

δ_i is the minimum distance between the point i and any other point with higher density, which can be calculated as follows:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \quad (3)$$

For the point with highest density, we define $\max_j (d_{ij})$ as its δ_i.

2.3 Trajectory clustering algorithm

The algorithm of trajectory clustering based on data graph and data field is as follows:

(1) Randomly select several values for σ, and calculate potential value corresponding to each σ according to eq. (1).

(2) Calculate optimal value for impact factor σ. According to the method proposed in literature (Li and Du, 2007), the optimal σ is obtained when the potential entropy reaches the minimum.

(3) Based on the optimal factor σ selected in step 2, compute potential value for each trajectory point with eq. (1). The influential strength of every point is generally considered to be the same. Thus, the mass of each data object is fixed as 1.

(4) Compute δ_i value for each trajectory point. The value of δ_i is defined as the minimum distance between the point i and any other points with higher potential:

$$\delta_i = \min_{j:\psi_j > \psi_i} (d_{ij}) \quad (4)$$

As for the point with the highest potential value, the value of δ_i is set as the maximum distance between itself and any other point, that is $\delta_i = \max_j (d_{ij})$.

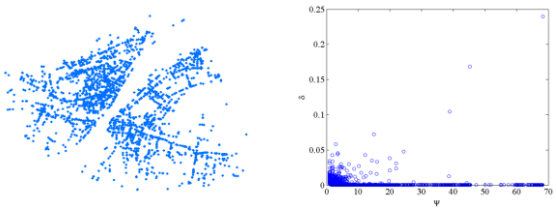
(5) Select cluster centres. As anticipated, cluster centres are usually the points with local maximum potential value. Hence, those points with relatively higher potential value ψ_i and higher δ_i can be regarded as centres.

(6) Identify noise points. Since noise points usually scatter in data field and receive weak mutual interaction, they have lower potential values. Thus, we employ threshold method to recognize noise points.

(7) Partition classes. After cleaning the noise points, for each normal data object, it is assigned to the same cluster as its nearest neighbour of higher potential value. Clustering is finally accomplished by executing this step for every normal data object.

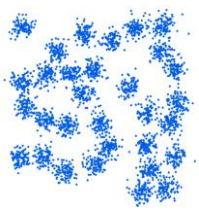
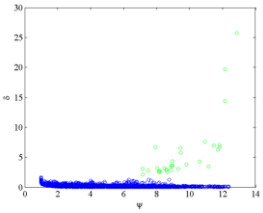
The key of the algorithm lies in selecting cluster centres and recognizing noise points. Here we put emphasis on depicting step 5 and 6.

In literature (Rodriguez and Laio, 2014), an index $\gamma_i = \rho_i \delta_i$ for choosing the number of centres is provided. Though this index works well for those aggregately distributed data, it poorly distinguishes centres when data present as random distribution instead. For a random distribution, one observes a continuous distribution in ρ_i and δ_i values. Figure. 1(a) displays taxi trajectory data in a time span, and Figure. 1(b) illustrates the corresponding decision graph generated by computing ψ and δ. Figure. 1(c) displays synthetic dataset, and Figure. 1(d) illustrates its decision graph. Compared with Figure.1 (d), the decision graph in Figure. 1(c) can hardly recognize cluster centres for some points are mixed together especially in the low left corner. Therefore, this paper makes further improvement while selecting cluster centres based on decision graph, and gives a quantitative method for center selection by computing thresholds for potential value ψ_i and distance δ_i respectively.



(a) Trajectory dataset. (b) Decision graph of the trajectory dataset.

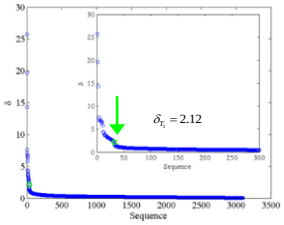
This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
doi:10.5194/isprsannals-II-4-W2-131-2015

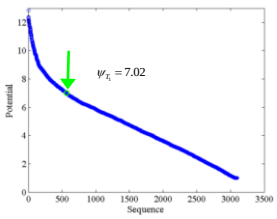



(c) Synthetic dataset. (d) Decision graph of the synthetic dataset.

Figure 1. Experimental datasets and decision graphs.

We adopt the method in literature (Yuan and Raubal, 2014) to ascertain thresholds by searching 'elbow point'. Take the dataset in Figure. 1(c) as an example, the obtained threshold of δ satisfies $\delta_{T_1} = 2.12$, labeled by the green arrow in the embedded graph in Figure. 2(a). Similarly, threshold of potential values is selected as $\psi_{T_1} = 7.02$ shown in Figure. 2(b). Therefore, points satisfying $\delta > \delta_{T_1}$ and $\psi > \psi_{T_1}$ correspond to cluster centres, labeled by those green points in Figure. 1(d).

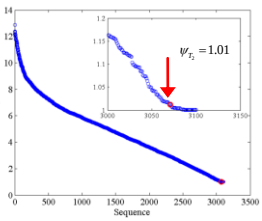


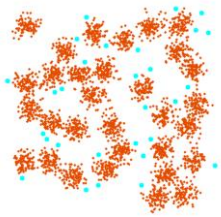


(a) Threshold of δ. (b) Threshold of potential value ψ.

Figure 2. Cluster center selection.

In the same way, we obtain the threshold $\psi_{T_2} = 1.01$ for noise points, which is labeled by the red arrow in the embedded graph in Figure. 3(a). In Figure. 3(b), potential values ψ of all the highlighted blue points are lower than ψ_{T_2} and they are recognized as noise points.





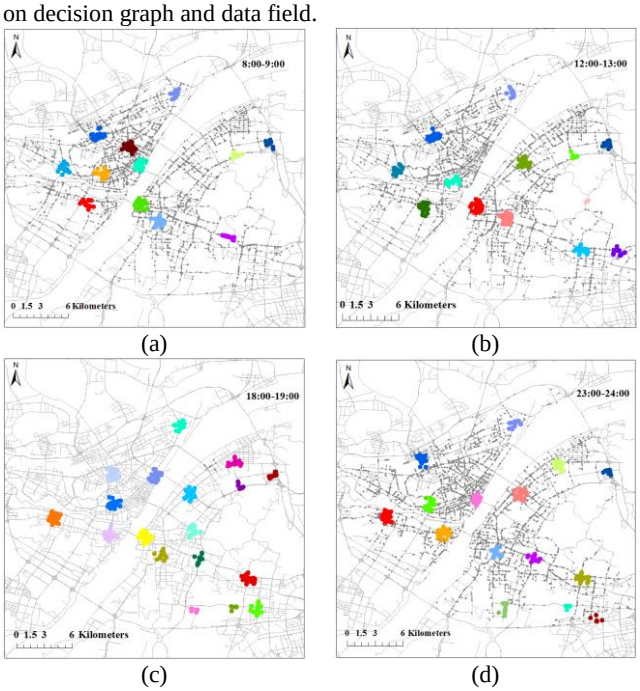
(a) Threshold for discerning noise points. (b) The recognized noise points.

Figure 3. Recognition of noise points.

3. EXPERIMENTAL RESULTS

With the method described in 2.2, we adopt taxi trajectory data of Wuhan City to detect hotspots. Furthermore, distribution and dynamics of the hotspots with respect to holiday, weekday and weekend are analyzed and compared. The experiments datasets are the trajectory data of 3000 taxis on holiday (May 1st, 2014), weekday (Wednesday, May 7th, 2014) and weekend (Saturday, May 10th, 2014) in Wuhan City, China. The study area is located within the 3rd ring road of Wuhan City for citizens mainly travelling within downtown. Data preprocessing steps for these datasets mainly include data extraction with respect to time slices, map matching and pick-up/drop-off points extraction.

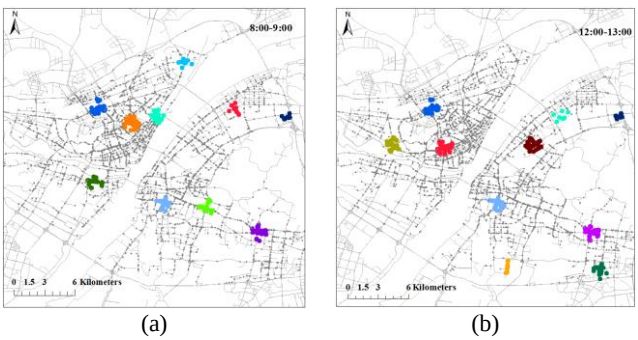
In these experiments, we put the focus on four typical time spans for hotspots detection, namely 8:00-9:00, 12:00-13:00, 18:00-19:00 and 23:00-24:00, which facilitates further analysis of hotspot changes in the morning, noon, afternoon and night respectively. Considering that taxi passengers tend to get off in a small scope around service facilities, and then walk across a road intersection or go through a street to destination, thus hotspots with dense pick-up/drop-off points can be detected within a scope. In this work, we select 800m as a search radius while detecting hotspots with pick-up/drop-off points. Regions over 800m away from cluster centres no longer belong to hotspot scope. The experimental results are illustrated in Figure. 4, Figure. 5, and Figure. 6, which are obtained through the clustering of pick-up/drop-off points with respect to the 4 selected time spans by the method of trajectory clustering based on decision graph and data field.



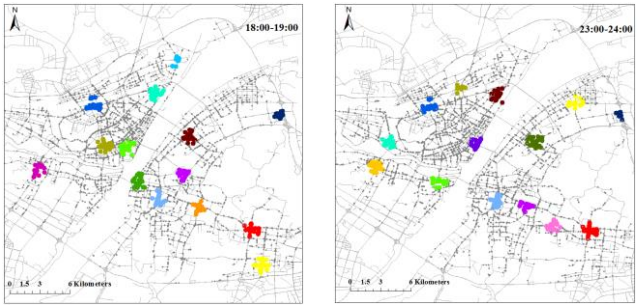
(a) (b)

(c) (d)

Figure 4. Hotspots on holiday (May 1st, 2014).

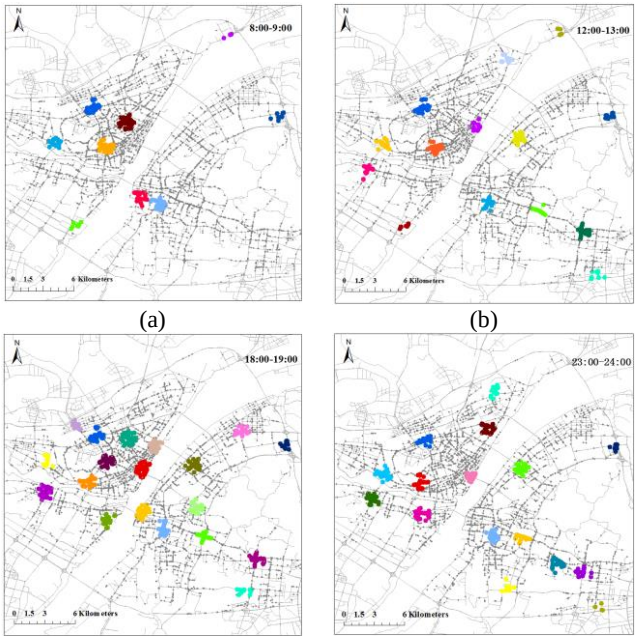


(a) (b)



(c) (d)

Figure 5. Hotspots on weekday (May 7th, 2014).



(a) (b)

(c) (d)

Figure 6. Hotspots on weekend (May 10th, 2014).

Comparing and analyzing the hotspot distribution maps on holiday, weekday and weekend, we find that the distribution patterns of hotspots during selected four hours are similar. For instance, some regions are constant hotspots and seldom vary with time. Represented as blue areas in Figure.4-6, the constant hotspots mainly locate on Hankou Railway Station (corresponding to the blue area in Figure. 4-6), Wuchang Railway Station (the light blue area), Wuhan Railway Station (the dark blue area), and so on. The constant hotspots mainly depend on passenger flow volume with respect to different time slices. As main places of transferring passengers between cities, railway stations load huge volume of passenger flow. With further analysis we find that residents travel intensively during 8:00-9:00 while sparsely during 18:00-19:00.

However, other hotspots only appear in some particular time spans. Moreover, differences of their spatial distribution and the varieties are largely influenced by holiday, weekday and weekend. During the May Day Holiday (May 1st, 2014), plenty of travellers come to Wuhan and parts of citizens also go out to enjoy leisure. So the hotspots focus on the stations, entertainment venues (such as Hubu Alley, the river beach, etc.), business centres, universities and communities, as displayed in Figure.4. During the weekday, individuals mainly shuttle between dwellings and work sites. Thus, hotspots mainly locate on business centres (such as the Optics Valley, Jianghan Road, Xudong Road, etc.), as represented in Figure.5. Hotspot distribution on weekend is similar with that on holiday for weekend can be taken as a short holiday and the hotspots mainly lie in entertainment and business centres. However, some lower-level hotspots (such as zoo, the Happy Valley, etc.) on holiday no longer appear as hotspots on weekend, as shown in Figure.6.

4. CONCLUSIONS

The paper proposes a method of trajectory clustering based on decision graph and data field. Compared with common clustering methods, it can automatically ascertain parameter instead of doing that by experience and is suitable to trajectory clustering. Furthermore, we apply it to trajectory-based urban hotspot discovery in Wuhan City, China. Distribution and dynamics of the hotspots are analyzed by employing taxi trajectory data with respect to holiday, weekday and weekend. However, similar to most of the existing clustering algorithms, the proposed method only considers spatial information, measuring similarity of points with distance between them. In the future research, we will consider the abundant attribute information related to trajectory data, especially some time information, and pay more attention to the high-dimensional properties of trajectory data, and expand this method to spatial-temporal domain.

ACKNOWLEDGEMENTS

We would like to thank the constructive comments from the anonymous referees, and we appreciate the financial supports from the National Natural Science Foundation of China (No. 41471326 and 61172175), and Fundamental Research Funds for the Central Universities (No. 2042015kf0183).

REFERENCES

Chang H, Tai Y, Hsu J Y, 2010. Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, 5(1), pp. 3-18.

Gui Z, Xiang Y, Li Y, 2012. Parallel discovering of city hot spot based on taxi trajectories. Journal of Huazhong University of Science and Technology (Natural Science Edition), 40, pp. 187-190.

Li D, Du Y, 2007. Artificial Intelligent with Uncertainty, National Defence Industry Press: Beijing. pp. 193-211.

Lee J, Shin I, Park G L, 2008. Analysis of the Passenger Pick-Up Pattern for Taxi Location Recommendation. *Proceedings of the 2008 Fourth International Conference on Networked Computing and Advanced Information Management, IEEE Computer Society*, 1, pp.199 - 204.

Rodriguez A, Laio A, 2014. Clustering by fast search and find of density peaks. Science, 344(6191), pp. 1492-1496.

Wang S, Gan W, Li D, et al, 2011. Data field for hierarchical clustering. *International Journal of Data Warehousing and Mining (IJDWM)*, 7(4), pp. 43-63.

Yue Y, Zhuang Y, Li Q, et al, 2009. Mining time-dependent attractive areas and movement patterns from taxi trajectory data, *Geoinformatics*, 17th International Conference on IEEE, pp. 1-6.

Yuan Y, Raubal M, 2014. Measuring similarity of mobile phone user trajectories–a spatiotemporal edit distance method. *International Journal of Geographical Information Science*, 28(3), pp. 496-520.

Zheng Y, Zhang L, Xie X, et al, 2009. Mining interesting locations and travel sequences from GPS trajectories, *Proceedings of the 18th international conference on World Wide Web*. ACM, pp. 791-800.