Hotspots Detection from Trajectory Data Based on Spatiotemporal Data Field Clustering

K. Qin[a,b] , Q. Zhou[a], T. Wu[c] , Y. Q. Xu[a,*]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China - (qink, whu_zhouqing, xuyq)@whu.edu.cn
[b] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China - qink@whu.edu.cn
[c] School of Information engineering, Lingnan Normal University, Zhanjiang, China - taowu0706@gmail.com

Commission IV, WG IV/3

KEY WORDS: Taxi Trajectory, Spatiotemporal Clustering, Data Field, Hotspots Detection

ABSTRACT:

City hotspots refer to the areas where residents visit frequently, and large traffic flow exist, which reflect the people travel patterns and distribution of urban function area. Taxi trajectory data contain abundant information about urban functions and citizen activities, and extracting interesting city hotspots from them can be of importance in urban planning, traffic command, public travel services etc. To detect city hotspots and discover a variety of changing patterns among them, we introduce a data field-based cluster analysis technique to the pick-up and drop-off points of taxi trajectory data and improve the method by introducing the time weight, which has been normalized to estimate the potential value in data field. Thus, in the light of the new potential function in data field, short distance and short time difference play a powerful role. So the region full of trajectory points, which is regarded as hotspots area, has a higher potential value, while the region with thin trajectory points has a lower potential value. The taxi trajectory data of Wuhan city in China on May 1, 6 and 9, 2015, are taken as the experimental data. From the result, we find the sustaining hotspots area and inconstant hotspots area in Wuhan city based on the spatiotemporal data field method. Further study will focus on optimizing parameter and the interaction among hotspots area.

1. INTRODUCTION

Hotspots refer to the area where some events happen frequently. There are variety of meanings of hotspots, such as hotspots of crime (Malleson and Andrese., 2015; Newton and Felson 2015; Steil and Parrish., 2009), hotspots of incident (Anderson., 2006; Vemulapalli et al., 2016), hotspots of disease (Wanjala et al. 2011; Hu et al., 2013) and hotspots of business (Chen et al., 2016; Turner., 2013). In this paper, the city hotspots refer to the area where city event happen intensively and the function of it would attract more people to travel. Thus, as the departure and destination choice of most people when they trip, city hotspots is a reflection of the vitality and function of the city. Extracting and analysing the spatiotemporal characteristics of city hotspots can help residents to choose the destination and avoid the rush before traveling. Besides that, the study of hotspots can warn the traffic related department of high-traffic locations early, so that they may allocate police and make the emergency plan in advance. For city planning department, the detection of hotspots can be used to evaluate the effect of regional planning and prepare perfect city planning scheme.

With the development of positioning technology, mobile internet technology and ubiquitous computing technology, the trajectory of moving object is accumulated continually. The spatiotemporal trajectory describes the movement patterns and behaviour history of moving objects in the geospatial environment, and reveals the mechanism of urban transportation evolution. So efficient mining analysis at trajectory is not only to provide a new opportunity of improving the transportation service for optimizing urban planning, but also significant for understand the social activities of urban residents and improve the allocation of public resources. At present, the study of trajectory data has been applied in many fields, such as path navigation (Ciscal-Terry et al., 2016; Tang et al., 2010), recommendation of travel (Hoteit et al., 2014; Kang and Qin., 2016), city planning and traffic management (Shen., 2015; Zhang., 2016). Taxi trajectory data contain abundant information about urban functions and citizen activities owing to its long service time, wide coverage of city area and freedom of its motion. The pick-up and drop-off points in the taxi trajectory data depict the spatiotemporal event of passenger or driver behaviour patterns. Thus, we regard the area where the pick-up and drop-off points cluster as the city hotspots area.

Each point in taxi trajectory data indicates the spatiotemporal event of passengers travel behaviour, and the areas where those points cluster can be considered as city hotspots. Thus, clustering such spatiotemporal trajectory points should be beneficial for us to describe the attraction of city area, and discover the behaviour patterns of the passengers or drivers. The data field theory, inspired by the theory of physical field, is proposed to actuate the objects self-organizing clustering by simulating the objects' interaction and motion in virtual data field (Li, 2007). Applying the data field theory in spatiotemporal data clustering analysis can quantify the degree of gathering states of data, recognize clusters with irregular shape and non-uniform density distribution. Besides that, the data filed can use equipotential surface and lines to show the distribution of data object.

In this paper, we introduce a data field-based cluster analysis technique to the pick-up and drop-off points of taxi trajectory data, and present an improved method for spatiotemporal clustering. In addition, to prove the validity of the method we compare the spatiotemporal data field method with other method of hotspots detection. Then, we conduct the experiment on taxi trajectory data in different date. Based on the detection results, we find the sustaining hotspots area and inconstant hotspots area in Wuhan city, China. The results can be used in

* Corresponding author: Yuanquan Xu, Email: xuyq@whu.edu.cn

the traffic management, like congestion warning and traffic dispersion so that we can prevent serious traffic congestion.

2. THE SPATIOTEMPORAL DATA FIELD CLUSTERING

2.1 Methods of hotspots detection

At present, there are three main kinds of methods to detect hotspots, namely spatiotemporal scan statistic, spatial point pattern analysis, the local spatial autocorrelation statistics and spatiotemporal density clustering.

Spatiotemporal scan statistic detects the cluster by comparing the aggregated distribution and random distribution in certain neighbourhood. This method is limited to the fixed time window, and consequently they usually produce the unsatisfied results for the area with uneven density.

Spatial point pattern analysis is an important spatial analysis method of study the data distribution by analysing the location where event happen. The theory divide the data distribution into three types: aggregation distribution, random distribution and uniform distribution. Some geographer developed two point pattern analysis methods. One is based on aggregation, including the quadrat counting method and kernel density estimation method, the other is based on decentralization and analyses the data distribution by measure the nearest distance, including the adjacent index, the *G* function, *F* function and *K*-function method, etc.

The local spatial autocorrelation statistics mainly studies the relationship among the same property from different objects in an area. It uses both global and local metrics to measure the connection among different objects, including Moran's *I*, Geary's *C*, and Getis-Ord *G*. In the case of a given distance, the statistical magnitude of Getis-Ord *G* uses the product between neighbouring values to measure the degree of approximation. It can distinguish the hot spots, which means positive spatial autocorrelation, and cold spots, which means negative spatial autocorrelation.

ST-DBSCAN is a representative method of spatiotemporal density clustering method. It can find the cluster of arbitrary shapes and eliminate the noise of low-density space via dividing the high-density contiguous regions into clusters. But the method is also limited by the fixed time window, which is not easy to find different density of spatiotemporal clusters.

2.2 The data field theory and its spatiotemporal extension

The interaction between objects in the data field can be achieved by potential function of quantitative, its potential value results will become the standard of measure similarity between data objects, so as to realize the classification of the data object detection and accumulation area.

2.2.1 Data field theory: The data field is defined as follow: in the space $\Omega \subseteq R^P$, the data set D consist of n objects: $D=\{x_1, x_2, x_3, \ldots, x_n\}$, and $x_i = (x_i^1, x_i^2, \ldots, x_i^a)$, $i = 1,2,\ldots,n$. Each point in the field is regarded as a field source, which radiates outward and is radiated by others. Such radiation energy decays with the increase of the distance over time and is quantified by potential function, like equation (1) and (2).

$$\varphi_y(x) = m_x \times e^{-\left(\frac{\|x-y\|}{\sigma}\right)^k} \quad (1)$$

$$\varphi_y = \sum_{i=1}^{n} \varphi_y(x_i) \quad (2)$$

where, m = the quality of the object x, in the paper, the quality of the trajectory point is 1. $\|x\text{-}y\|$ = the distance between x and y. $k \in N$ = the distance index. In the paper, $k = 2$. σ is the impact factor, and $\sigma \in (0,+\infty)$. σ is used to control the degree of interaction among objects. In the gravitational field, the potential value would decay rapidly when the distance between two points greater than $\frac{2}{\sqrt{3}}\sigma$. So the object's neighbor radius often equal to $r=\frac{2}{\sqrt{3}}\sigma$. Li propose a method to optimize the σ, that is to calculate the minimum entropy of potential value, as the equation (3).

$$H = -\sum_{i=1}^{n} \frac{\Psi_i}{Z} \ln\left(\frac{\Psi_i}{Z}\right) \quad (3)$$

where, Ψ_i, = the potential value of the object. Z is a normalization factor, and $Z = \sum_{i=1}^{n} \Psi_i$ .For example, calculate the entropy of the potential value of the taxi trajectory data at 8:00-9:00 in May, 1st,2015, we can get the entropy line as the figure 1.

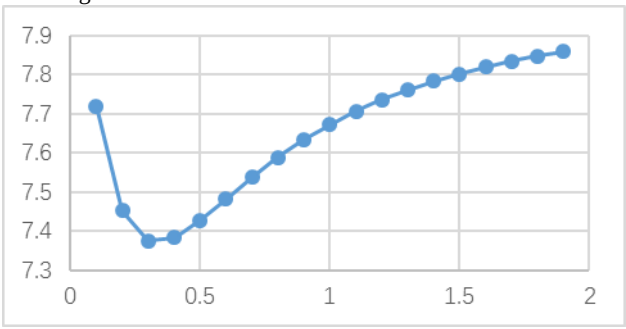
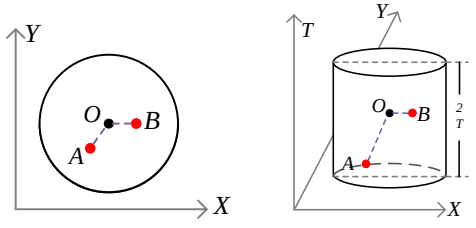


Figure 1. Entropy line

2.2.2 Spatiotemporal data filed theory: The taxi trajectory data, especially the pick-up and drop-off points, are spatiotemporal data with a rich time information. For a trajectory point, its energy is strongest when it is just formed and the energy will decay over time. So the intensity of the interaction between two points has relations with time. As shown in the figure 2, the influence generated by *A* on *O*, is different with the influence generated by *B* on *O*. However, the traditional potential function is none of the time information.



(a) (b)

Figure 2. The comparison of interaction in different dimension. (a) The neighbourhood in traditional data filed, (b) The neighbourhood in spatiotemporal data field

Thus, in this paper, we build a spatiotemporal data field around trajectory data to measure the influence exerted by time on the interaction among data. We improve the data field potential function by introducing the time weight, which has been normalized to estimate the potential value. The spatiotemporal data field is defined as follow: for the point $A = (x_A, y_A, t_A)$ and

the point $B = (x_B, y_B, t_B)$ in the space $\Omega \subseteq R^P$. The potential value between A and B, is calculated by equation (4):

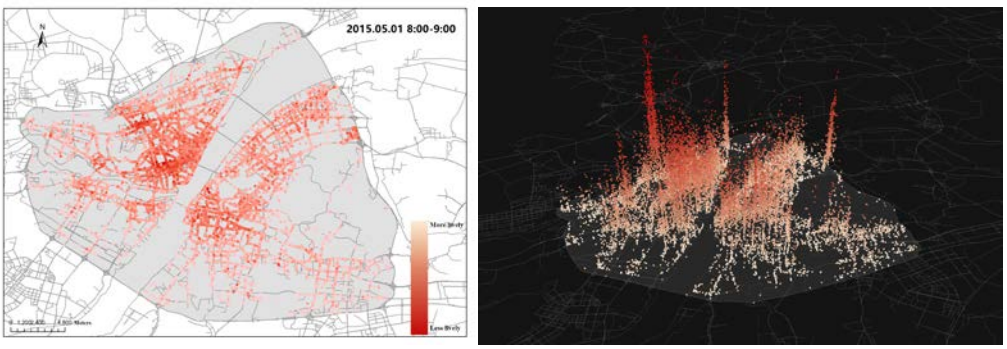
$$\varphi_A(B) = \sum_{i=1}^{n} m_B \times e^{-\left(\frac{\|A-B\|}{\sigma}\right)^2} \times \frac{1}{\Delta T_{AB}} \quad (4)$$

where, Δt_{AB}' is a normalized time parameter, and calculated by equation (5).

$$\Delta t_{AB} = \frac{\Delta t_{AB} - \Delta t_{\min}}{\Delta t_{\max} - \Delta t_{\min}} \quad (5)$$

where, Δt_{AB} is the time difference between A and B, $\Delta t_{\min}$ = the minimum time difference in data set, $\Delta t_{\max}$ is the maximum time difference in data set.

Apply the spatiotemporal data field potential function to the taxi trajectory data, we use the taxi trajectory data at 8:00-9:00 in May 1st, 2015 as the experimental data. The experiment result shown as figure 3. In figure 3 (b), we put the potential value as the height of trajectory point. From the visualization of potential, we can find the areas where gather many points. For example, we can find that Hankou railway station has higher potential value, so it is likely a hotspots area.



(a) (b)

Figure 3. Visualization of potential. (a) Visualization of potential in 2d, (b) Visualization of potential in 3d.

2.3 Clustering method based on spatiotemporal data field

After finishing the calculation of potential value, we need realizing the data division and hotspots detection based on the potential value. Inspired by physics, equipotential line and equipotential surface is introduced to visualization. As shown in figure 4.

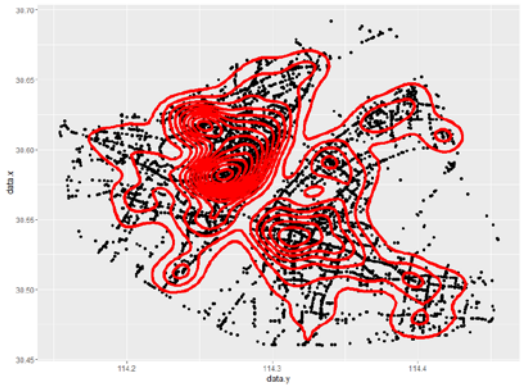


Figure 4. The equipotential line of trajectory data filed

From the distribution of equipotential line, within the neighbourhood of data filed, the smaller difference of potential value between two points, the bigger probability of belonging to the same class. It is a kind of thought of using the nesting structure of the equipotential line to realize the data division. Inspired by this, Yang (Yang, 2012) proposed a clustering method, which uses the averaging difference of potential value as the measurement of similarity among data. Besides that, based on DBSCAN, this method detects the clusters by generating the scanning window and averaging difference of potential value. The method can improve the limit of fixed window, can deal with the data set with the nonuniform density distribution. Based on the algorithm, this paper presents spatiotemporal data field-based clustering method.

The proposed algorithm involves four major steps, including that: 1) generate the spatiotemporal data field around the pick-up and drop-off points and calculate the potential value of each points based on improved data field function. 2) Find the point with extreme potential value, and regard it as current processing object. 3) Determine the radius of neighbourhood according to the current processing object. 4) Obtain the averaging difference value of potential, which is put as the standard of measurement of similarity among points.

2.4 Comparison with other cluster methods

2.4.1 Comparison with traditional data filed methods: We compare the potential result of common data field, which ignore time information and spatiotemporal data field method to elaborate the validity of the improve data field method. As shown in figure 5, we choose the trajectory data in May 6th, and pick out the area in Guang Gu Street to show the result. In the figure 6, the deeper the red colour, the larger the potential value. In the traditional data field result, we can find that the potential of the study area is generally larger, while in the spatiotemporal data field, the potential value is different. Actually, the four points in the rectangular box have different time span although they are close with each other in spatial. Apparently, the spatiotemporal data field method can show the difference while traditional data field can't. In the spatiotemporal data field, the two points having short time span have small gap at potential value, and the two point having long time span have bigger gap at potential value. Therefore, spatiotemporal cluster method can distinguish the difference about time information.

(a) (b) (c)

FID	TIME	Potential
4814	2015-05-06 06:46:43	.951065
24106	2015-05-06 07:07:36	.977142
24883	2015-05-06 11:14:20	.976511
6528	2015-05-06 12:40:16	.954648

(d)

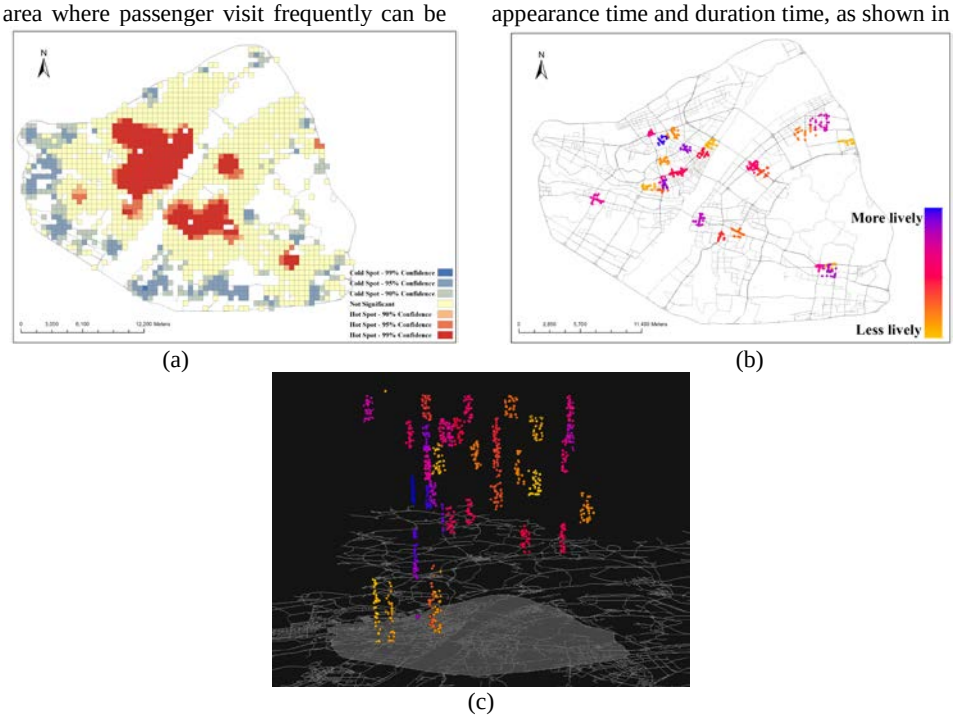
FID	TIME	Potential
4814	2015-05-06 06:46:43	.003040
24106	2015-05-06 07:07:36	.006638
24883	2015-05-06 11:14:20	.018582
6528	2015-05-06 12:40:16	.012681

(e)

Figure 5. The comparision result between traditional data field method and sptio-temporal data field method. (a) The cluster result of traditional data field method, (b) The cluster result of spatiotemporal data field method, (c)3D visualization of four example points, (d)time information and potential value in traditional data field, (e)potential value in spatiotemporal data field.

2.4.2 Comparison with Getis-Ord *G*

Before the calculation by Getis-Ord *G*, we divide the study area into some grids and set the span as 500m. Then we count the number of traffic points in each grid and regard the number as the attribute. Thus, we can use Getis-Ord *G* to detect the area where the pick-up and drop-off event happen frequently, namely the hotspots area. As shown in figure 6, in the result of Getis-Ord *G*, the area where passenger visit frequently can be detected but the hotspots is fuzzy. Such detection result is short of time information, so that it is not easy to mine forward. In the spatiotemporal data field, the detection result is close to the result in Getis-Ord *G* method. However, we can find that the cluster in spatiotemporal data field is clearer than the result in Getis-Ord *G*. Besides that, the improved data field method can show the time span of hotspots area, and we can know the appearance time and duration time, as shown in figure 6(c).



(a) (b)

(c)

Figure 6. Comparision result between spatiotemporal data field and Getis-Ord *G* method. (a)Getis-Ord *G* calculation result. (b)spatiotemporal data field cluster result, (c)3D visualization of spatiotemporal data field.

2.4.3 Comparison with ST-DBSCAN method: As the above, ST-DBSACN is the spatiotemporal extension of DBSCAN. In this paper, the process of cluster division refer to ST-DBSACN. We set the parameter of ST-DBSACN as follow: R = 300, *MinPts* = 20, $\triangle T$ = 10. We set the parameter in spatiotemporal data field as follow: σ = 0.2, *MinPts* = 10. As shown in figure7, the experiment results are similar. Further, the spatiotemporal data field can distinguish the heat difference between the hot clusters based on the potential value.

(a) (b)

Figure 7. Comparision result between spatiotemporal data field and ST-DBSCAN. (a) The cluster result of spatiotemporal data field, (b) the cluster result of ST-DBSCAN.

3. EXPERIMENTS AND ANALYSIS

The taxi trajectory data of Wuhan city in China on May 1, 6 and 9, 2015, are taken as the experimental data as shown in Figure 8. While the object we implement the cluster method is the pick-up and drop-off points which are extracted from taxi trajectory data based on the field named "state". If this field value was "full", the car would carrying passenger at that time. If this field value was "empty", the car would having no passenger at that time. We first sort the trajectory data by the state field. We consider the point whose state became "full" from "empty" as pick-up point, and the point whose state became "empty" from "full" as drop-off points.

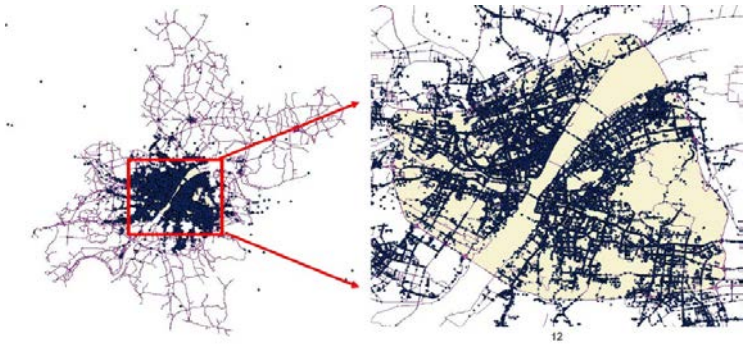
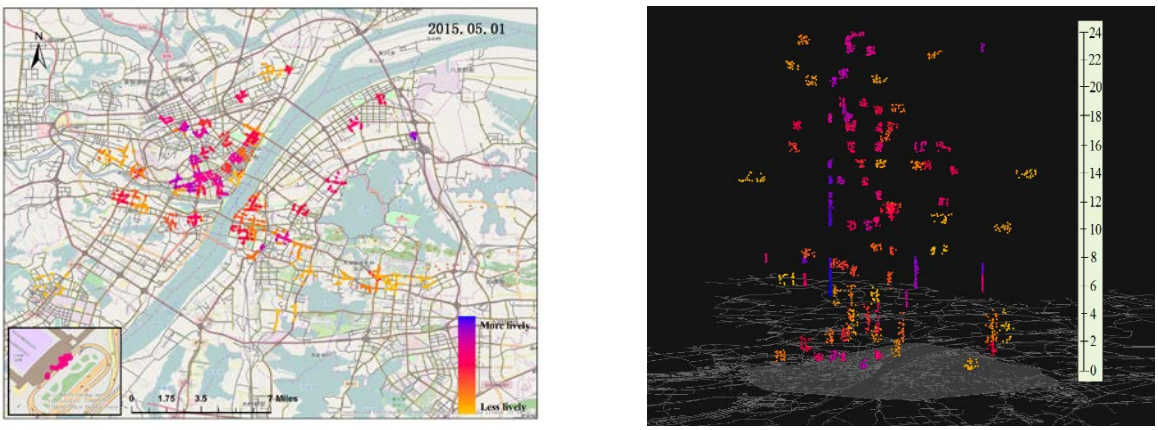


Figure 8. The trajectory data

We use spatiotemporal data field to cluster the pick-up and drop-off point on May 1, 6 and 9, 2015. We get the optimal σ parameter as 0.3 based on the minimum entropy of potential value method and get the optimal *MinPts* parameter as 25. As shown in figure 9, the (a), (c) and (d) are the top-view of the hotspots, while the (b), (d) and (e) are the 3D visualization of hotspots, using time information of trajectory data as the vertical axis. In line with the difference of the sum of the potential value, we can distinguish the degree of vitality of hotspots. The larger the potential value, the more lively the hotspots area.



(a) (b)

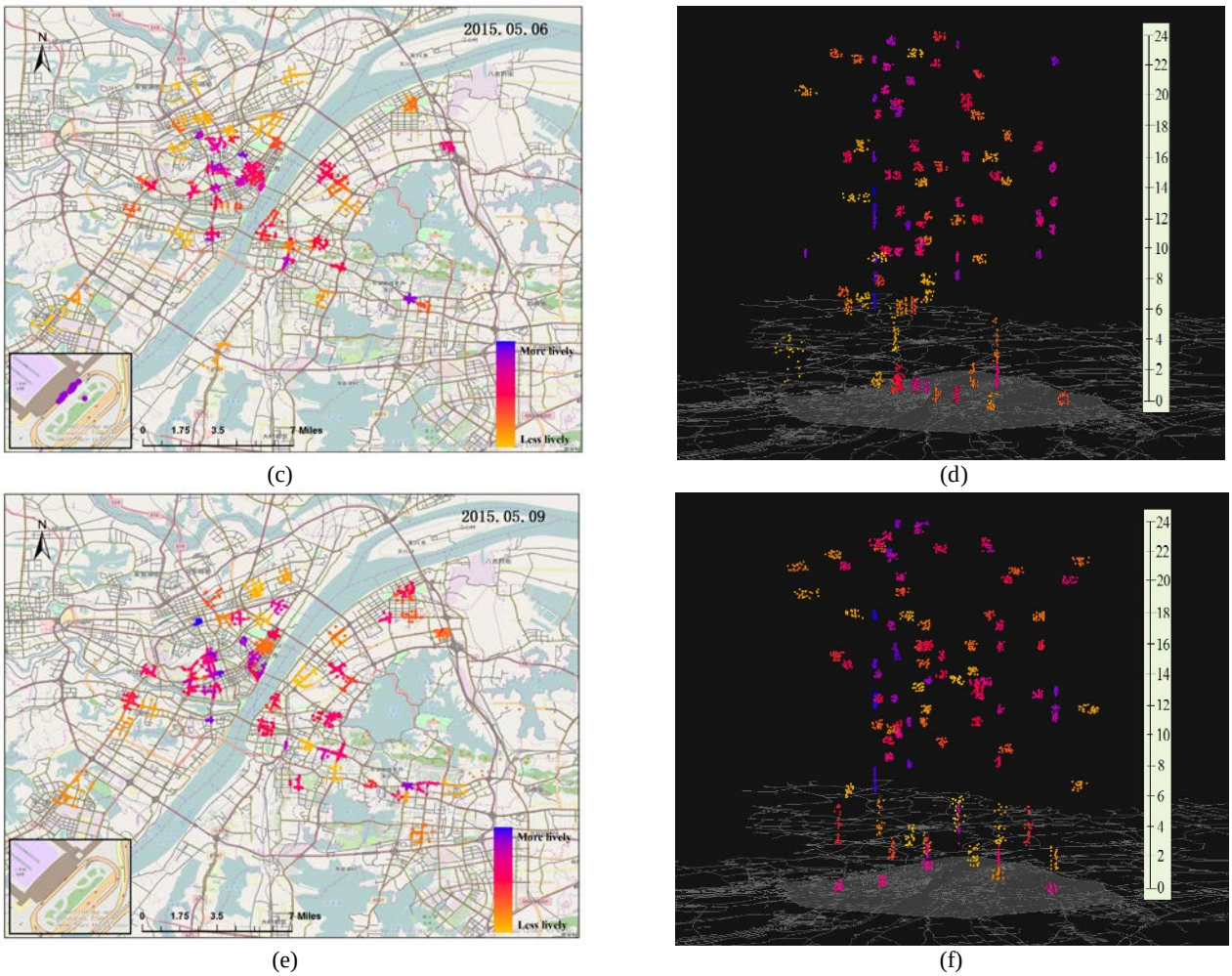


Figure 9. Hotspots detection results. (a) The top-view of hotspots in May 1st, (b) 3D visualization of hotspot in May 1st, (c) the top-view of hotspots in May 6th, (d) 3D visualization of hotspot in May 6th, (e) the top-view of hotspots in May 9th, (f) 3D visualization of hotspot in May 9th.

From the cluster results, we can detect the main hotspots area in Wuhan. There are 105 clusters in May 1st, 92 clusters in May 6th, 98 clusters in May 9th. The main trading areas are the common hotspots in Wuhan city, such as Zhongnan trading area, Wang Jiadun CBD and Jianghan Road etc. From the 3D visualization, we can find that Hankou railway station is the most lively hotspots area because of its largest potential value and long appearance time.

Under closer scrutiny, we further inspected the spatial distribution of hotspots in each time slots over all of the three days. In the morning and noon, the hotspots are mainly distributed around the residential area and commercial block for that it is the time when residents prepare to travel and go to work. Whereas, in the afternoon and evening the hotspots are mainly distributed around the entertainment centre, since this time span is the prime time for residents who get off work to go shopping and for the worker to return home.

Based on the urban functional zoning of Wuhan city, we assigned the derived hotspots into five categories as transport station, entertainment center, residential area and hospital, commercial block and scenic spot. Generally, on the May Day Holiday, the number of hotspots around the transport stations are larger than that of other dates. It is because that resident are preferred to go out for traveling in this long vacation. On holiday, less residents choose to get up early to work, so hotspots at the residential areas and hospital distribute more sparsely than other date relatively.

During 7:00-9:00, citizens are preparing to travel or go to work and hospital, so the city embrace the morning peak time. In this period, the hotspot areas are mainly distributed in residential areas and hospitals. Relative to the working days, in workday, residential areas, hospitals as well as the administrative office class are with less number of hot spots. Considering that, most of office worker would start from home to their work place in this time. While during the holiday season, citizens who do not need to go out early will delay the time they leave home, while reducing the number of trips to the office space. During 11:00-13:00, people begin to have lunch and go shopping. So there are many routes from work place or residential areas to recreation district. In addition, scenic spots embrace peak time for tourist, which causes the hotspots appear around there at this time, however, it less likely become hotspots in usual date, like 6th and 9th, while entertainment centres are the hotspots because it is the time to have dinner. During 18:00-20:00, scenic spots close at 18:00, and residents leave at this time, so hotspots appear around such place on holiday. At night, public transportation is out of service, residents choose take taxi to return.

4. CONCLUSION

In this paper, we detect hotspots and find the sustaining hotspots area and inconstant hotspots area in Wuhan city based on spatiotemporal data field from the taxi trajectory data. The result can be used in the traffic management, like congestion warning and traffic dispersion so that we can prevent serious traffic congestion. The clustering method based on spatiotemporal data field can identify the classes of any shape, and reflect the time influence with getting rid of the fixed time window. In the current state, the performance of this method still depends on the choice of the parameter, σ and *MinPts*, which means minimum number of points within a certain distance calculated with σ. For an optimizing parameter, we should take various experiments using the try and error method. In other words, our method has some limitations, and there are a couple of issues that should be considered in the future research: how to efficiently obtain appropriate parameters is currently under investigation, and will be reported later. Furthermore, the complex network analysis about the interaction among hotspots area is well worth further studying on.

ACKNOWLEDGEMENTS

This research is partial funded by the National Key Research and Development Program of China (No. 2017YFB0503604), National Natural Science Foundation of China (No. 41471326)

REFERENCES

Anderson T., 2007.Comparison of spatial methods for measuring road accident “hotspots”: a case study of London. *Journal of Maps*, 3(1), pp. 55-63.

Chen W. S., Liu L., Liang Y. T., 2016. Retail center recognition and spatial aggregating feature analysis of retail formats in Guangzhou based on POI data. *Geography Research*, 35(4), pp. 703-716.

Ciscal-Terry W., Dell'Amico M., Hadjidimitriou N S., et al ., 2016. An analysis of drivers route choice behaviour using GPS data and optimal alternatives. *Journal of Transport Geography*, 51, pp. 119-129.

Gan W. Y., Li D. Y., Wang J. M., 2006. A hierarchical clustering method based on data fields. *Acta Electronica Sinica*, 34(2), pp. 258-262.

Hoteit S., Secci S., Sobolevsky S., et al., 2014. Estimating human trajectories and hotspots through mobile phone data. *Computer Networks*, 64, pp. 296-307.

Hu B. S., Gong J. H., Zhou J. P., et al., 2013. Spatial-temporal characteristics of epidemic spread in-out flow—Using SARS epidemic in Beijing as a case study. *Science China Earth Sciences*, 56(8), pp. 1380-1397.

Kang C. G., Qin K., 2016. Understanding operation behaviours of taxicabs in cities by matrix factorization. *Computers, Environment and Urban Systems*, 60, pp. 79-88.

Li D. Y., 2007. *Artificial Intelligence with Uncertainty*. Chapman and Hall/CRC.

Malleson N., Andresen M A., 2015. Spatiotemporal crime hotspots and the ambient population. *Crime science*, 4(1), pp. 1-8.

Newton A., Felson M., 2015. Editorial: crime patterns in time and space: the dynamics of crime opportunities in urban areas. *Crime science*, 4(1), pp. 1-5.

Steil D., Parrish A. S., 2009. HIT: A GIS-Based Hotspot Identification Taxonomy. *International Journal of Computers and Their Applications*, 16(2), pp. 81-90.

Shen J. W., Zhou T. G., Zhu X. B., 2015. On the Spatial and Temporal Distribution of Traffic Information Based on GPS Floating Car Data. *Journal of Southwest University (Natural Science Edition)*.37(8), pp. 157-162.

Tang L L., Li Q Q., Chang X. M., et al., 2010. Modelling of taxi drivers’ experience for routing applications. *Science China Technological Sciences*, 53(1), pp. 44-51.

Turner L. M., 2013. Hunting for hotspots in the countryside of Northern Sweden. *Journal of Housing and the Built Environment*, 28(2), pp. 237-255.

Vemulapalli S. S., Ulak M. B, Ozguven E. E, et al., 2016. GIS-based Spatial and Temporal Analysis of Aging-Involved Accidents: a Case Study of Three Counties in Florida. *Applied Spatial Analysis and Policy*, pp. 1-27.

Wanjala C. L., Waitumbi J., Zhou G., et al., 2011. Identification of malaria transmission and epidemic hotspots in the western Kenya highlands: its application to malaria epidemic prediction. *Parasites & vectors*, 4(1), pp. 81-94.

Yang J., Gao J. W., Liang J. Y., et al., 2012. An Improved DBSCAN Clustering Algorithm Based on Data Field. *Journal of Frontiers of Computer Science and Technology*, 6(10), pp. 903-911.

Zhang H., Wang X M., Guo X C., et al., 2016. Application of taxi GPS big trajectory data in intelligent traffic system. *Journal of Lanzhou University of Technology*, 42(1), pp.109-114.