Exploring Time-dependent Traffic Congestion Patterns from Taxi Trajectory Data

Chengkun Liu[1], Kun Qin[2], Chaogui Kang[3]

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. China

[1]wishchengkun@whu.edu.cn

[2]qink@whu.edu.cn

[3]chaoguikang@pku.edu.cn

***Abstract*— Due to public travel choice, city function zoning and road network structure, urban traffic congestion tends to have strong spatiotemporal correlations. Unveiling the spatiotemporal patterns of urban traffic congestions will provide useful information for urban planning, traffic control, and location based service (LBS). This paper proposes an approach to identify traffic congestion regions and their spatiotemporal distribution from taxi trajectory data. Firstly, slow trajectory sequences are extracted from raw taxi trajectory data. Together with taxi engine states, these sequences are then transferred into congestion events that define the congestion duration and the average speed. Thereafter, highly congestion-prone areas are identified by clustering these congestion events using the DBSCAN clustering method. From the perspective of spatial homogeneity, global aggregation degrees of those identified congestion-prone areas are defined by the Ripley K function. Finally, considering congestions of nearby areas can influence each other and worsen the local traffic condition, the theory of data field is imposed to reveal the interactions between neighbouring congestion events. It also enables the visualization of the congestion intensity distribution from the trajectory potential of trajectory data field. The proposed method is validated by a case study of taxi trajectory data analysis in Wuhan City, China.**

***Keywords*— Taxi trajectory; Congestion event; Spatial clustering; Ripley K function; Data field**

I. INTRODUCTION

With the rapid development of positioning and communication technologies, increasing amount of traffic information has been collected in diverse ways. As a result, traffic patterns, especially congestion patterns, can be more accurately identified to reduce congestion, increase safety, and improve traffic forecasting accuracy. Congestion patterns can also provide useful information for urban planning and location based service (LBS).

Among existing traffic data collection methods, floating car (e.g. taxi) is a relatively reliable and cost-effective tool for gathering traffic data over the road network across a wide area [1]. To generate traffic information for every road segment, speed and position information of floating cars is usually collected at regular time intervals (e.g., 60s). With auxiliary information about the states of passenger and taxi's engine, floating car data enables us to identify slow trajectory sequences caused by congestions in a highly accurate way. Note that a slow trajectory sequence is consisted of consecutive GPS points with low-speed produced by the same car. In this paper, we focus on these slow trajectory sequences and define them as congestion events for further analysis.

To obtain meaningful patterns of urban traffic congestions, this study adopt the DBSCAN (Density-based spatial clustering of applications with noise) clustering method for finding congestion-prone areas based on the density of low-speed points. Thereafter, we quantify the point distribution pattern with the Ripley K function as a measure of the time-dependent aggregation degree of traffic congestions. Additionally, data field is implemented to describe and visualize the congestion intensity.

The remainder of this paper is organized as follows. Section 2 presents the related work about traffic congestion extraction and pattern identification. Section 3 describes the proposed method to extract traffic congestion trajectory and the associated distribution patterns. Experimental results are reported in section 4 as a validation of the proposed approach. Section 5 concludes the paper and indicates the future works.

II. RELATED WORK

Trajectory data are often obtained from location-aware devices that capture locations at specific time intervals, like GPS, mobile phone, Wi-Fi receiver [2]-[3]. Among those trajectory collecting methods, floating car data is well received and widely adopted for providing essential information about the city road network.

Since road traffic status estimation or traffic parameter data acquisition is usually the base of the all kinds of application of floating car data, much work has been done in this field. In the other strand, map-matching is another fundamental research field that gains a lot of attentions due to the lack of positioning accuracy [4]. It is not only of great importance in data pre-processing process, but also for discovering new roads and conducting route recovery [5].

Back to the early stage, traffic analysis mainly aims to compute traffic parameters including traffic volume, speed or occupancy in the network [1]. Studies are then turned to find abnormalities of those traffic parameters as indicators of some traffic incidents [6]. More recently, machine learning algorithms are also adopted to further improve the accuracy of parameter estimation and incident prediction, including fuzzy logic, support vector machine and neural network [7]-[8].

978-1-4799-7749-9 /15/$31.00 ©2015 IEEE

Besides traffic incidents analysis, many researches focused on deriving meaningful information based on trajectory stops, moves and their sequences. Clustering methods are widely used to find places of interest and hot spot areas [9]-[10] and their movement patterns [11]. Great progress has been achieved by building congestion model [12] and conducting spatiotemporal analysis based on mass historical data [13].

Compared to existing methods that focus on the traffic status of road segments or intersections, our study aims to unveil time-dependent congestion patterns at the "area" scale. It concerns more about the entire region to discover the aggregated distribution patterns and to describe the congestion intensity. For this purpose, we mainly adopt density-based methods including clustering, point pattern analysis and the theory of data field to explore the time-dependent congestion pattern in this paper. Details of our approach will be introduced in the next section.

III. Exploring Traffic Congestion Patterns

A. Congestion trajectory extraction and traffic congestion event description

Congestions can occur in various situations, but they usually share a common property of continuous slow driving during a particular time span. Since floating car data contains detail positioning and state information, slow trajectory sequences can be easily extracted from it. Note that actions including picking-up and dropping-off passengers or stoping somewhere to have a rest also result in slow driving, we use the engine and passenger state to filter out them. As a result, we regard those valid slow trajectory sequences as congestion trajectory point sequences.

In this paper, each extracted trajectory point sequence in a time span is described as a congestion event with the following properties:

- car ID
- constituted trajectory point sequence
- average position of the sequence
- time span of the congestion trajectory
- time duration of the congestion
- average speed during the time span

Further experiment and analysis will be conducted based on congestion events.

B. Statistical analysis on congestion event series

Statistics on the property of congestion events can help recognize variations and relationships of these events overtime. Hence, trends on the number of congestion events as well as the average congestion time duration in different time slots are summarized to find the typical time spans for the following analysis.

C. Extracting Congestion-prone areas based on Clustering

Clustering is conducted on traffic congestion events to find the highly congestion-prone areas. Since the number of clusters are unknown beforehand, density-based algorithm is a more adequate tool in this study as compared to partition clustering such as *k*-means which requires a pre-knowledge of cluster number. One of the most common used clustering algorithms is DBSCAN which can find arbitrarily shaped clusters and avoid noise [14].

For the purpose of clustering, DBSCAN algorithm assigns data points as core points, density-reachable points and outliers, as follows [15]:

- A point p is a core point if at least *minPts* points are within distance ε of it, and those points are said to be directly reachable from p. No points are reachable from a non-core point.
- A point q is reachable from p if there is a path p_1, ..., p_n with $p_1 = p$ and $p_n = q$, where each p_{i+1} is directly reachable from p_i. So all the points on the path must be core points, with the possible exception of q.
- All points which are not reachable from any other point are outliers.

Given p as a core point, it forms a cluster together with all points that are reachable from it. Each cluster contains at least one core point and those non-core points outline its "edge" since they cannot reach more points. Since reachability is not a symmetric relation (a non-core point may be reachable, but nothing can be reached from it), a further notion of connectedness is required to formally define the extent of the clusters found by DBSCAN. As a result, a cluster further satisfies two properties:

- All points within the cluster are mutually density-connected.
- If a point is density-reachable from any point of the cluster, it is part of the cluster as well.

Based on aforementioned properties, we can select proper *minPts* and ε to obtain clusters that well represent the highly congestion-prone areas. Though certain studies suggested network constraint as an improvement in trajectory clustering, Euclidean distance is adopted in this study in that our study focuses on congested "area" instead of roads and intersections.

D. Spatial homogeneity detection using Ripley K function

As known, results of the DBSCAN clustering largely depend on the selection of parameters *minPts* and ε. To clarify the time-dependent aggregation degree of congestion events distribution, we address the problem from the perspective of spatial homogeneity. We use the Ripley K function to measure the global aggregation degrees of identified congestion events in different hours of a day.

Ripley's K function [16] is a spatial descriptive statistical method for detecting deviations from spatial homogeneity. The K function is defined as

$$\hat{K}(t) = \lambda^{-1} \sum_{i \neq j} I(d_{ij} < t) / n \qquad (1)$$

where d_{ij} is the Euclidean distance between the i^{th} and j^{th} points in a data set of n points, t is the search radius, λ is the average density of points and I is the indicator function. If the points are approximately homogeneous, $\hat{K}(t)$ should be approximately equal to πt^2.

Since the trajectory points are located on the road network, we also infer the difference between the expected random distribution and that with network constraints.

E. Congestion intensity simulation from trajectory data field

In reality, traffic flows are interconnected with each other. That is, congestions in nearby areas can influence each other and worsen the local traffic condition. This phenomenon can be described by the theory of data field which simulates the interaction among objects. Similarly, the congestion impact distribution can be calculated from the trajectory potential of trajectory data field.

To simulate the congestion intensity influenced by each other, we choose the nuclear data field [17] defined as

$$\varphi(x) = \sum_{i=1}^{n} (e^{-\left(\frac{\|x-x_i\|}{\sigma}\right)^2} \cdot m_i) \quad (2)$$

where m_i is the mass of the object x_i, σ is the influence factor between objects, $\|x - x_i\|$ is the Euclidean distance to object x_i. By selecting proper σ and m_i value, we can simulate the cumulative scalar potential $\varphi(x)$ that represents the congestion intensity in the trajectory data field. Note that, in this paper m_i represents the congested rank value of a congestion event inferred either from the congested time duration or average speed in a congestion to describe congestion for different purpose.

IV. Experiment

The experiment uses 12,000 taxis' trajectories collected on May 5th 2014 (Monday) in Wuhan, a large city in China. The 12,000 taxis transmitted their location and speed data every 60 second with service status, and generated over 14 million trajectory records. We firstly match the trajectory points onto their corresponding road links, and then extract 251,232 congested trajectory points and 21,416 congestion events based on the congestion extraction method in Section III. Fig. 1 gives an overview about the distribution of the extracted congestion trajectory points.

Fig. 1 The distribution of congestion trajectory points

Fig. 2 shows the number of congestion events over time. Obviously, congestion events are largely concentrated within rush hours especially during 8:00~9:00 and 18:00~19:00. While in slack hours fewer congestion events happen, as it reaches the bottom during 13:00~14:00. It also shows that nearly no congestion happens at night.

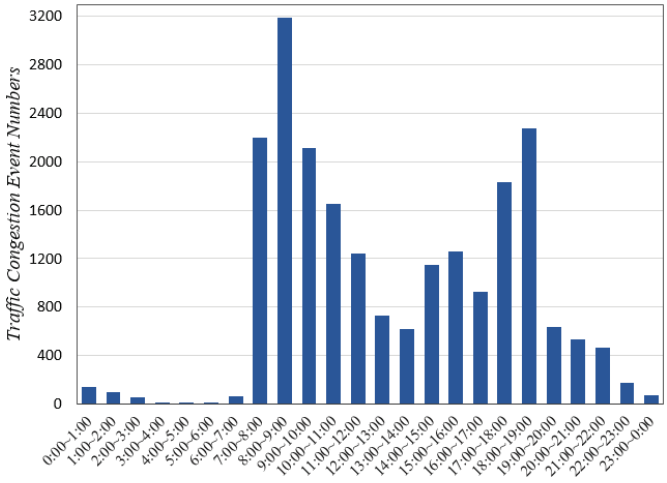


Fig. 2 Statistics on traffic congestion event occurrences numbers

The average congestion time duration over time is shown in Fig. 3. The result reveals that congestion time durations are generally between 400~500 seconds which seems relatively stable over the day. Under closer scrutiny, we find that the morning congestion time durations are longer than those in the afternoon and evening.

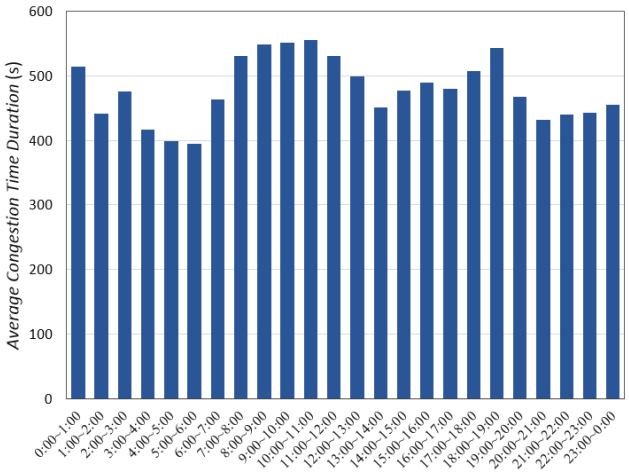


Fig. 3 Statistics on average congestion time duration

Based on the statistical analysis, we choose to focus on five time spans in typical rush hours and slack hours for the clustering process as follows: 8:00am-9:00am, 12:00am-13:00pm, 13:00pm-14:00pm, 18:00pm-19:00pm, and 21:00pm-22:00pm. Each time span corresponds to a cluster result as summarized in Table I. To get congestion-prone areas, we choose those clusters which have more than 100 congestion events happen in the range over 400 meters. The 100 congestion events can ensure that almost 2 congestions

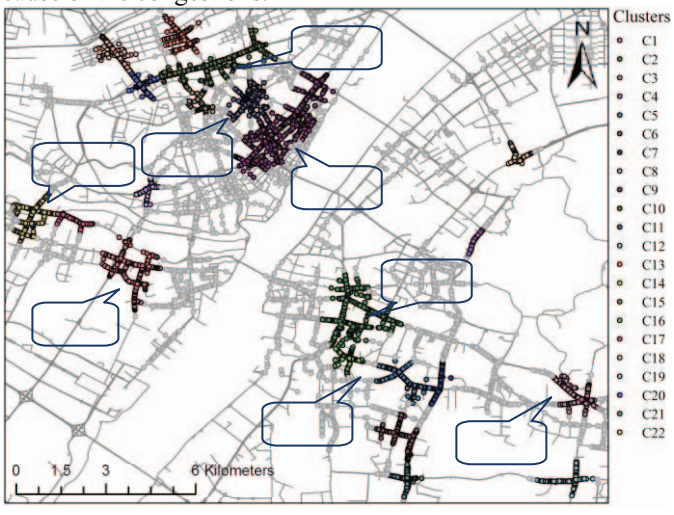
happen every minute, which guarantee the identified area to be highly congestion-prone.

TABLE I
CLUSTERING RESULTS ON FIVE TIME SPANS

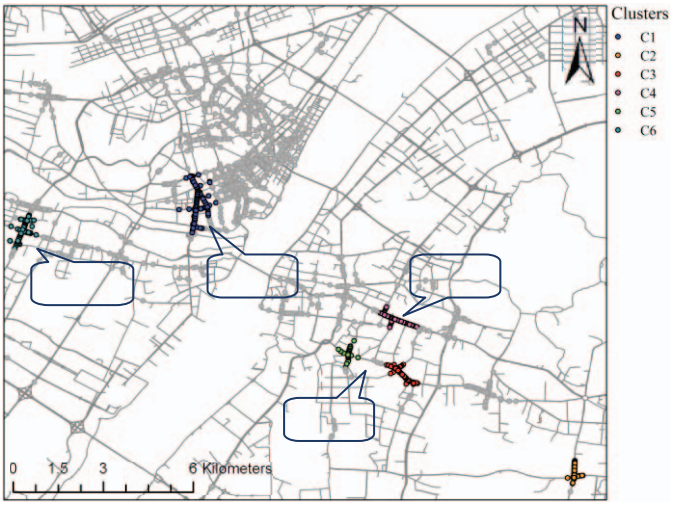
Time span	No. of slow trajectory points	No. of congestion events	No. of clusters (n>100)
8:00~9:00	26418	3186	22
12:00~13:00	5523	727	6
13:00~14:00	4321	620	3
18:00~19:00	18933	2273	22
21:00~22:00	3144	461	2

Table 1 demonstrates that more clusters are found during rush hours than those during slack hours, which reveals the potential consistency between the number of congestion events and the aggregated clusters. Spatial distributions of the clustering results are shown in Fig. 4. Note that complete congestion trajectory sequences are shown in this figure to give a clear overview of the congestion distribution on road network.

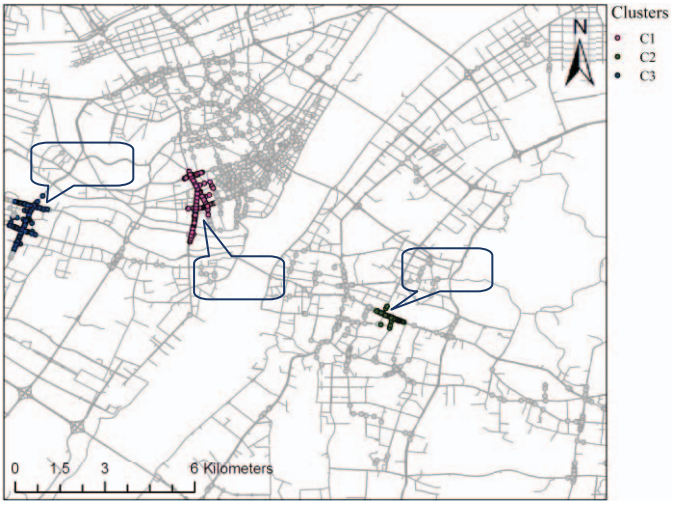
The congested areas during different hours are indicated in Fig.4 from (a) to (e). Among these figures, Fig. 4(a) and 4(d) indicate that congested areas during rush hours mainly distribute around main roads/avenues or intersections, including Jiefang Road along the river segment, Jianshe Avenue, Fazhan Avenue, Hanyang Avenue, Xiongchu Avenue, Luoyu Road, Jiedaokou Area and so on. In contrast, fewer clusters are formed during slack hours (Figure 4(b-c, e)), mainly appearing around Jiedaokou Area, Xiongchu Avenue and Wangjiawan Area. We argue that ongoing road construction in these area during the study period as a possibly cause of the congestions.



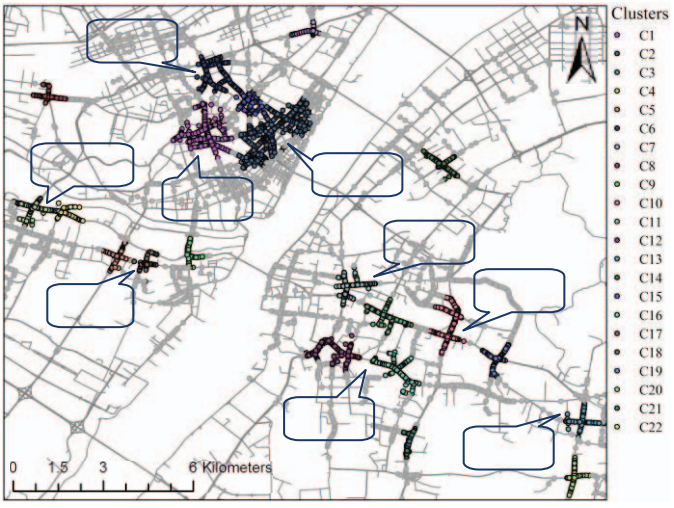
(a) Clustering result (8:00~9:00)



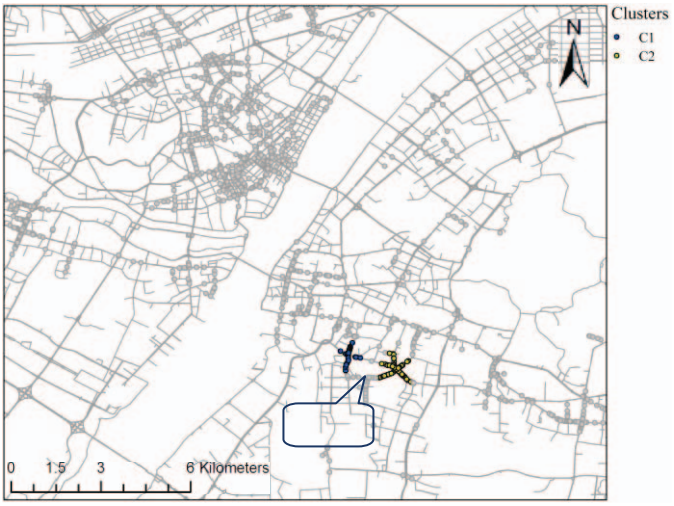
(b) Clustering result (12:00~13:00)



(c) Clustering result (13:00~14:00)



(d) Clustering result (18:00~19:00)



(e) Clustering result (21:00~22:00)

Fig. 4 Clustering results in different time spans

The clustering process partly reveals the aggregated degrees overtime based on the number of clusters. However, the results can be fairly inconsistent since it largely depends on the selection of parameters. Therefore, Ripley K function is adopted to describe the spatial homogeneity of traffic congestions.

By conducting experiments from distance between 300~3000 meters with an increment of 300 meters, the time-dependent spatial distribution patterns are given in Fig. 5. As mentioned in Section III, we compare two expected patterns and five observed patterns. To be clear, the expected random spatial pattern is calculated from those points randomly picked in the flat area, while the expected road network spatial pattern can be approximately calculated from the original trajectory points. The observed spatial distribution patterns are investigated within the same time spans as the DBSCAN clustering. The result shows that with road network restriction, the distribution is more aggregated than that in unconstraint space, while the observed congestion spatial distribution patterns are generally more aggregated the expected ones. Additionally, the distribution is less aggregated during rush hours, implying congestion mainly occurs in specific areas during slack hours.

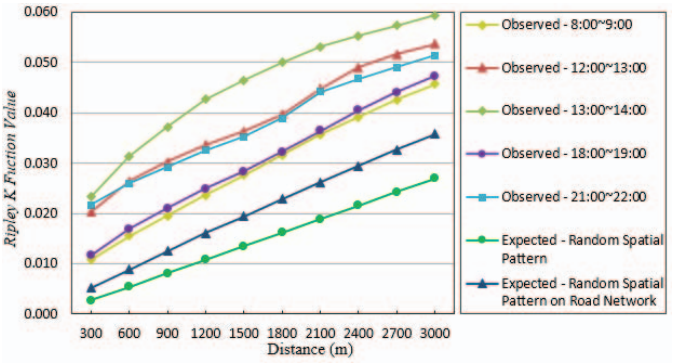


Fig. 5 Ripley K function value during different time periods

Congestion clustering considers congestion-prone areas as independent clusters, and ignores the interaction and influence between them. Nevertheless, traffic flows are interconnected with each other in reality. To describe the congestion intensity caused by the interaction, the experiment based on trajectory data field is conducted. Due to page limitation, here we only use the morning rush hours 8:00~9:00 for illustration.

To measure the congestion rank of each congestion event, the property of congestion time duration is used for frequency statistics, as shown in Fig. 6. Considering that it takes 3~4 minutes for one pass in complex traffic intersections, we defined that as one round and then classify the congestion rank based on average waiting rounds, as the low congestion lasts for 2 rounds and time increases with the congestion degree. The cumulative frequency also demonstrates its rationality as 75% congestion events are below high rank level, while only 10% congestion events reach the top rank. Detailed classification of the four congestion ranks is shown in Tab. II.

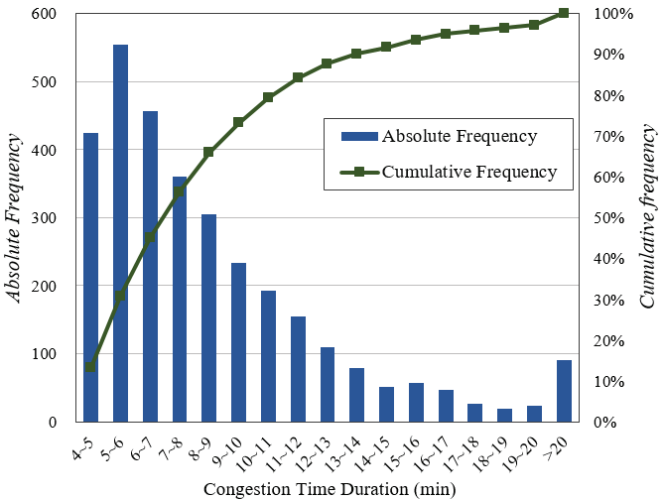


Fig. 6 Congestion time duration statistics of morning rush hours

TABLE II
CONGESTION RANK CLASSIFICATION FOR TRAJECTORY DATA FIELD

Congestion degree	Congestion rank value	Cumulative frequency	Congestion time duration (*min.*)
Low	1	50%	4~7
Medium	2	75%	7~10
High	3	90%	10~15
Very High	4	100%	>15

The cumulative scalar potentials are then calculated based on rank classification of the congestion events. By classifying the scalar potentials of congestion intensity, a heat map is generated and visualized in Fig. 7 to show the congestion intensity distribution. The green color indicates low congestion intensity while the red color represents high risk of congestion.

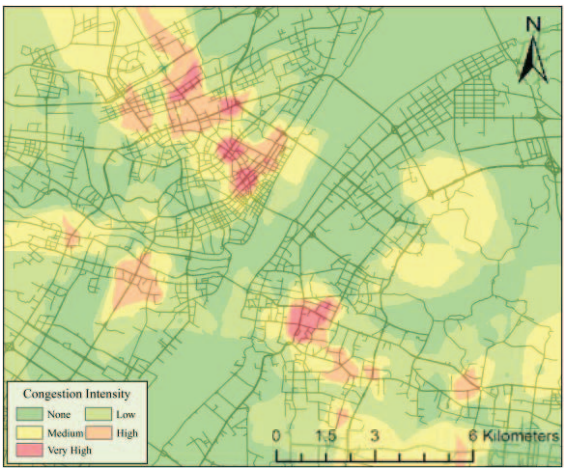


Fig. 7 Congestion intensity visualization base on Trajectory Data Field

V. Conclusions

In this paper, a method is proposed to identify congestion-prone areas and to explore their associated distribution patterns from taxi trajectory data. Clustering method is adopted to detect highly congestion-prone areas, and congestion aggregation degree and interaction between different congestion clusters are used to describe and visualize the congestion intensity.

The extracted congestion trajectory can accurately reflect the traffic conditions in that the proposed method takes additional useful information from the states of passenger and taxi's engine into full consideration. The analysis on congestion distribution pattern identification and congestion intensity visualization also demonstrate their superiority and convenience due to the notion of congestion event and the application of data field.

However, a couple of difficulties still exist in congestion influence measurement and classification for that we cannot tell the exact congestion influential factor between congestion areas thus the generated trajectory data field for congestion intensity may lack reasonable classification basis. Moreover, our congestion pattern analysis method remains to further verification since this study only consider the scale of "area" which may lack the usage of road network in the road or intersection scale.

In the future research, we would like to adopt different analysis methods aimed to "area", "road" and "road segment" scales for multi-scale description of congestion, and expect more experiments based on datasets from holidays and non-holidays, weekends and weekdays to obtain more concise and reasonable time-dependent traffic congestion patterns.

Acknowledgment

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

[1] C. Fabritiis, R. Ragona and G. Valenti, "Traffic Estimation And Prediction Based On Real Time Floating Car Data," in 11th International IEEE Conference on Intelligent Transportation Systems, 2008, pp. 197-203.

[2] B. Guc, M. May, Y. Saygin, and C. Korner, "Semantic Annotation of GPS Trajectories," in 11th AGILE International Conference on Geographic Information Science Spain, 2008, pp. 1-9.

[3] Y. Ishikawa, "Spatio-temporal Data Mining from Moving Object Trajectories," in 4th Symposium on Intelligent Media Integration for Social Information Infrastructure: Intelligent Media Integration Nagoya University, 2006, pp. 125-126.

[4] S. M. Turner, W. L. Eisele, R. J. Benz and D. J. Holdener, Travel Time Data Collection Handbook, Texas Transportation Institute, 1998.

[5] B. S. Kerner, C. Demir, R. G. Herrtwich and S. L. Klenov. "Traffic state detection with floating car data in road networks," in Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria, 2005, pp. 44-49.

[6] L. M. Hall, "Congestion identification aspects of the McMaster incident detection algorithm", Transportation Research Record, 1990:167-175.

[7] J. Lu and L. Cao, "Congestion evaluation from traffic flow information based on fuzzy logic," in IEEE Proceedings of Intelligent Transportation Systems, 2013, pp. 50-53.

[8] F. Porikli and X. Li, "Traffic congestion estimation using HMM models without vehicle tracking," in IEEE Intelligent Vehicles Symposium, 2004, pp. 188-193.

[9] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A Clustering based Approach for Discovering Interesting Places in Trajectories," in Instituto de Informática. Programa de Pós-Graduação em Computação, 2008, pp. 863-863.

[10] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining Interesting Locations and Travel Sequences from GPS Trajectories," in WWW 2009 Madrid, Spain, 2009.

[11] Y. Yue, Y. Zhuang, Q. Q. Li, Q. Z. Mao, "Mining time-dependent attractive areas and movement patterns from taxi trajectory data," in 17th International Conference on Geoinformatics, 2009, pp. 1-6.

[12] S. Lin, B. D. Schutter, Y. G. Xi and H. Hellendoorn, "Efficient network-wide model-based predictive control for urban traffic networks", Transportation Research Part C, 2012 (24): 122-140.

[13] L. Xu, Y. Yue and Q. Q. Li, "Identifying Urban Traffic Congestion Pattern from Historical Floating Car Data," in 13th COTA International Conference of Transportation Professionals, 2013, pp. 2084-2095.

[14] J. Sander, M. Ester, H. P. Kriegel and X. W. Xu, "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications", Data Mining and Knowledge Discovery, 1998, 2(2): 169-194.

[15] M. Ester, H. Kriegel and X. W. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining 1980, pp. 226-231.

[16] T. Lagache, V. M. Yedid, J. C. Olivo-Marin, "A statistical analysis of spatial colocalization using Ripley's K function," in IEEE 10th International Symposium on Biomedical Imaging (ISBI), 2013, pp. 896-901.

[17] S. L. Wang, W. Y. Gan and D. Y. Li, "Data Field for Hierarchical Clustering", International Journal of Data Warehousing and Mining, 2011, 7(4): 43-63.