Network-constrained Bivariate Clustering Method for Detecting Urban Black Holes and Volcanoes

Z.H. Wu and Q.L. Liu

Department of Geo-informatics, Central South University, Hunan Province, China
Email: zhihui.wu@csu.edu.cn; qiliang.liu@csu.edu.cn

Abstract

Urban black holes and volcanoes are typical traffic anomalies in cities. The discovery of urban black holes and volcanoes has played an important role in urban planning and public safety. It is still challenging to detect arbitrarily shaped urban black holes and volcanoes considering the network constraints with less prior knowledge. In this study, a network-constrained bivariate clustering method is proposed to detect statistically significant urban black holes and volcanoes with irregular shapes. First, an edge-expansion strategy is used to construct the network-constrained neighbourhoods without the time-consuming calculation of the network distance between each pair of objects. Then, a network-constrained spatial scan statistic is constructed to identify candidate sub-areas of urban black holes and volcanoes, which are then combined to form arbitrarily shaped urban black holes and volcanoes based on the multidirectional optimization method. Finally, the statistical significance of each detected urban black hole and volcano is evaluated using Monte Carlo simulation. The simulations demonstrate that proposed method is more effective and stable than the three state-of-the-art methods in detecting urban black holes and volcanoes. The empirical analysis of the Beijing taxicab spatial trajectory data also shows that the proposed method is useful for detecting the spatiotemporal variations of traffic anomalies.

Keywords: bivariate clustering, network-constraints, urban black hole, urban volcano

1. Introduction

An urban black hole is defined as a subgraph of the road network whose overall inflow is significantly greater than the overall outflow during a certain time interval. In contrast, an urban volcano is defined as a subgraph of the road network whose overall outflow is significantly greater than the overall inflow during a certain time interval (Hong *et al*., 2015). Urban black holes and volcanoes usually reflect disasters, catastrophic accidents, and traffic congestion in the city. The detection of urban black holes and volcanoes helps to maintain public safety and optimize urban planning.

In the era of big data, the massive spatial trajectory data can be used to represent the traffic flow in a city, and inflow and outflow in a region can be calculated by using the origin and destination points recorded in the trajectories (Zheng, 2015). Urban black holes and volcanoes can be defined as bivariate clusters, i.e., groups containing the origin and destination points, and forming hot and cold spots. Although many spatial clustering methods are currently available, most of them are only designed for detecting clusters from a univariate spatial point process (Grubesic *et al*., 2014). Only a few methods can be used to detect blackholes and volcanoes, e.g. graph clustering (Li *et al*., 2012; Hong *et al*., 2015) and spatial point clustering (Kulldorff, 1997; Pei *et al*., 2015). However, there is no method that considers the network-constrained spatial trajectory data and the statistically significant urban

blackholes and volcanoes with arbitrary shapes. To overcome this limitation, a network-constrained bivariate clustering method is developed in this study.

2. The network-constrained bivariate clustering method

2.1 An edge-expansion method for searching network-constrained neighbours

The spatial trajectory data is first matched onto the corresponding edges of a road network using the map-matching method. If e is the nearest edge of point p_i, $Dist_{start}$ represents the distance between p_i and the start node of e, and $Dist_{end}$ represents the distance between p_i and the end node of e. Then, for a given point p_i and radius of the neighbourhood eps, the network-constrained neighbourhood of p_i ($\boldsymbol{NN}_{eps}(p_i)$) can be identified in three steps.

Step 1: Find the edge e where p_i is located on; if $Dist_{start} \geqslant eps$ and $Dist_{end} \geqslant eps$, $NN_{eps}(p_i) = \{p_j | d_E(p_i, p_j) \leq eps, p_j \in p_{obs}\}$, where $d_E(p_i, p_j)$ represents the Euclidean distance between p_i and p_j, $\boldsymbol{P}_{obs}$ is the set of points matched onto e, else go to Step 2;

Step 2: Search neighbouring edges of p_i:

- If $Dist_{start} < eps$ and $Dist_{end} \geqslant eps$, search all possible paths $\boldsymbol{P}$ from the start node of e until the cumulative length of the edges of each path is equal to or larger than eps.
- If $Dist_{start} \geqslant eps$ and $Dist_{end} < eps$, search all possible paths $\boldsymbol{P}$ from the end node of e until the cumulative length of the edges of each path is equal to or larger than eps.
- If $Dist_{start} < eps$ and $Dist_{end} < eps$, search all possible paths $\boldsymbol{P}$ from the nodes of e until the cumulative length of the edges of each path is equal to or larger than eps.

Step 3: Identify network-constrained neighbours of p_i: for each path $path_i$ in $\boldsymbol{P}$, each object p_j on $path_i$ will be added to $\boldsymbol{NN}_{eps}(p_i)$ if $d_p(p_i, p_j) \leqslant eps$, where $d_p(p_i, p_j)$ represents the distance between p_i and p_j on the path.

The complexity of the edge expansion method is approximately linear.

2.2 Test statistic for the detection of urban black holes and volcanoes

The null hypothesis is that there are no urban black holes or volcanoes within the study area:

$$H_0 : p_O^r = q_O^r \text{ and } p_D^r = q_D^r, \quad (1)$$

where p_O^r and p_D^r are the probabilities of a point being an origin point and a destination point inside region r, respectively. q_O^r and q_D^r are the probabilities of a point being an origin and a destination point outside region r, respectively.

If $p_O^r < q_O^r$ *and* $p_D^r > q_D^r$, then the edges in r may form an urban black hole; if $p_O^r > q_O^r$ *and* $p_D^r < q_D^r$, the edges in r may form an urban volcano. For region r, the Bernoulli-based log-likelihood ratio test statistic (Kulldorff, 1997) can be used to detect urban black holes and volcanoes. Monte Carlo simulation is used to test whether $\boldsymbol{NN}_{eps}(p_i)$ is a subarea of an urban black hole or volcano. The points are randomly generated on the road network following complete spatial randomness R times, and the p-value of $\boldsymbol{NN}_{eps}(p_i)$ is calculated as:

$$p\text{ - }value(p_i) = \frac{\sum_{j=1}^{R} I(\log\lambda_{eps}^{j} > \log\lambda_{eps}^{obs})}{1 + R} \quad (2)$$

where I is an indicator function that takes the value of 1 if the statement is true, and the value of 0 if the statement is false. $\log\lambda_{eps}^{obs}$ and $\log\lambda_{eps}^{j}$ are the test statistics calculated based on the observed and random points in $\boldsymbol{NN}_{eps}(p_i)$. Given a significance level α, for each point p_i whose *p-value* is smaller than α, if the number of destination points is larger than that of origin points in $\boldsymbol{NN}_{eps}(p_i)$, then $\boldsymbol{NN}_{eps}(p_i)$ will be identified as a subarea of an urban black hole, else $\boldsymbol{NN}_{eps}(p_i)$ will be identified as a subarea of an urban volcano.

2.3 Discovery of arbitrarily shaped urban black holes and volcanoes

Because the construction of arbitrarily shaped urban black holes and volcanoes is the same, we only take the construction of urban black holes as an example to introduce the multidirectional optimization method.

Step 1: Start from an unvisited point p_i whose neighbourhood $\boldsymbol{NN}_{eps}(p_i)$ is a subarea of an urban black hole and select $\boldsymbol{NN}_{eps}(p_i)$ as the seed of the urban black hole.

Step 2: Identify all the subareas overlapped with $\boldsymbol{NN}_{eps}(p_i)$ and sort these subareas in descending order based on their log-likelihood ratio test statistic values, represented as $\boldsymbol{A}_{overlap}$=[$\boldsymbol{NN}_{eps}(p_1)$, $\boldsymbol{NN}_{eps}(p_2)$,…, $\boldsymbol{NN}_{eps}(p_{max})$].

Step 3: Combine the first subarea $\boldsymbol{NN}_{eps}(p_1)$ in $\boldsymbol{A}_{overlap}$ with $\boldsymbol{NN}_{eps}(p_i)$; if the log-likelihood ratio test statistic value of the newly-built urban black hole decreases compared to the value of the urban black hole detected in the previous step, then $\boldsymbol{NN}_{eps}(p_i)$ is identified as an urban black hole and then go to Step 1; else, $\boldsymbol{NN}_{eps}(p_1)$ and $\boldsymbol{NN}_{eps}(p_i)$ are combined to form a new urban black hole, and other subareas in $\boldsymbol{A}_{overlap}$ are sequentially combined with $\boldsymbol{NN}_{eps}(p_i)$ using the same procedure until the log-likelihood ratio test statistic value of the newly-built urban black hole decreases compared to the value of the urban black hole detected in the previous step.

Step 4: Select the new urban black hole detected in Step 3 as the new seed, go to Step 2. The iterative procedure stops when the log-likelihood ratio test statistic value of the newly-built urban black hole decreases compared to the value of the urban black hole detected in the previous step.

Step 5: Iteratively implement Step 1 to Step 4 until all the points have been visited. A number of overlapped urban black holes will be obtained.

Step 6: Find the urban black hole with the highest log-likelihood ratio test statistic value UB_{max} and delete all the urban black holes overlapped with UB_{max}.

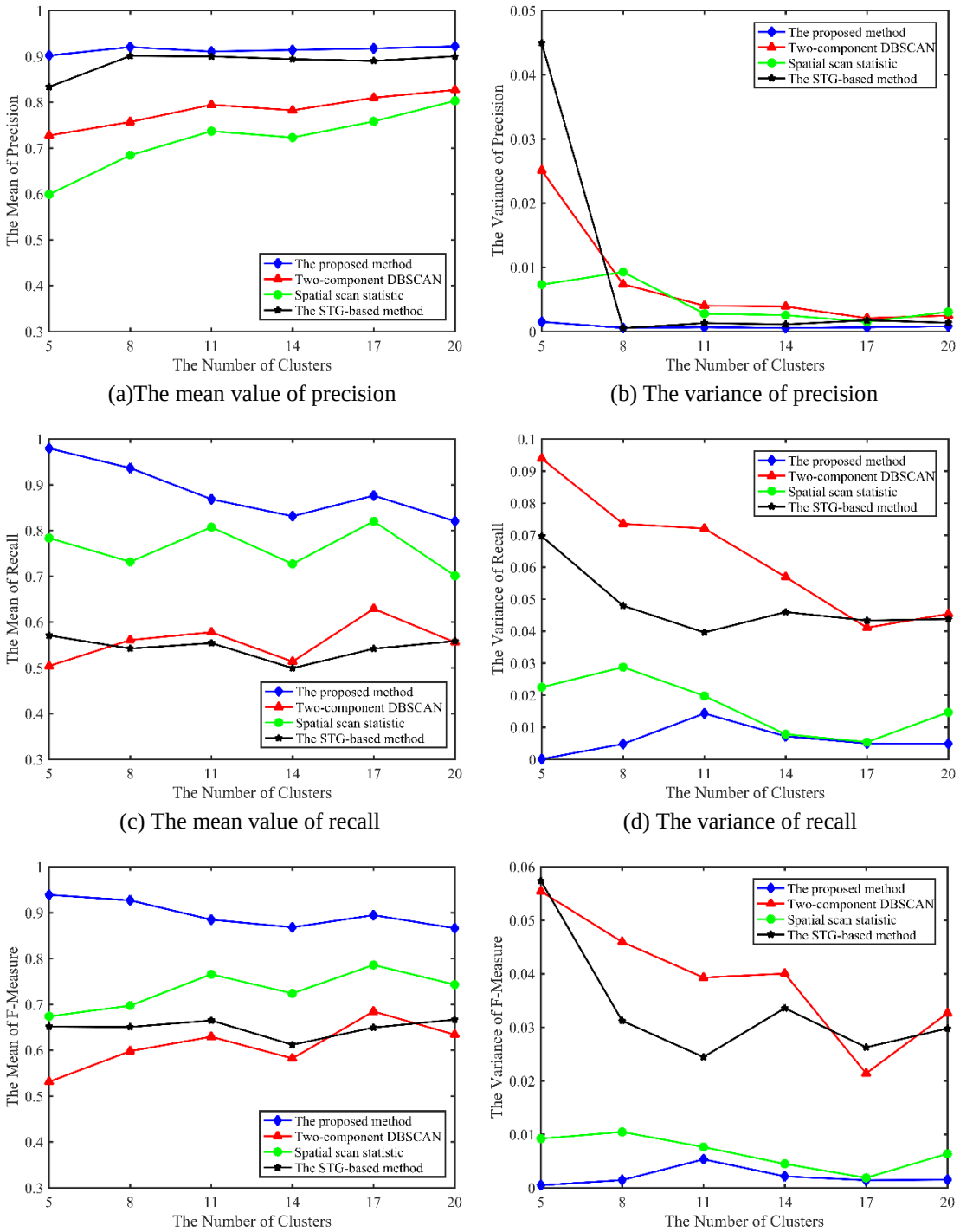
Step 7: Repeat Step 6 until no overlapping urban black holes occurs.

Step 8: Evaluate the statistical significance of detected urban black holes by using the method introduced in Section 2.2. The false discovery rate approach (Benjamini and Yekutieli 2001) was used to solve the multiple and dependent testing problem.

3. Experimental analysis

Six groups of simulated datasets that contain 5, 8, 11, 14, 17, and 20 urban black holes and volcanoes were generated on a real road network in Beijing. In each group, ten datasets with the same number of urban black holes and volcanoes were generated. The proposed method was compared with three state-of-the- art methods, i.e., the spatiotemporal graph (STG)-based method (Hong *et al*., 2015), network-constrained spatial scan statistic (Shiode, 2011), and two-component DBSCAN (Pei *et al*., 2015). The performance of different methods was quantitatively evaluated using precision, recall, and F-measure. The results are shown in Figure 1. It can be found that the mean values of precision, recall, and F-

measure usually exceed 0.9; therefore the urban black holes and volcanoes discovered by the proposed method are indeed the most accurate and complete.



(a)The mean value of precision (b) The variance of precision

(c) The mean value of recall (d) The variance of recall

(e) The mean value of F-Measure (f) The variance of F-Measure

Figure 1. Performance of the four methods on simulated

The proposed network-constrained bivariate clustering method was further applied to detect urban black holes and volcanoes from the taxi GPS trajectory data generated by approximately 30,000 taxis from May 23 to 29, 2016 in Beijing, China. From the mining results, we found that :

(i) Most urban black holes and volcanoes were detected during the rush hour and at night, and urban black holes and volcanoes always occur near the train stations.

(ii) Urban black holes and volcanoes discovered on the weekdays and weekends are different. The commuting patterns of Beijing residents can be well revealed from the mining results.

(iii) The mixed urban functions in some commercial areas do not relieve the traffic anomalies in Beijing because people who work in these areas usually do not live there due to the high housing prices.

4. Conclusion

The experimental results on simulated datasets show that the urban black holes and volcanoes detected by the proposed bivariate clustering method are more accurate and complete than those detected by existing methods. The case study on Beijing taxi trajectory data show that the proposed bivariate clustering method is able to describe the detailed traffic anomaly patterns. The mining results provide a new insight into traffic anomaly patterns, and will be useful for understanding the characteristics of urban traffic operation and optimizing urban planning.

5. References

Benjamini, Y and Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29 (4):1165-1188.

Grubesic, T. H., Wei, R., Murray, A.T. 2014. Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers*, 104(6):1134-1156.

Hong, L., Zheng, Y., Yung, D., Shang, J., Zou, L. 2015. Detecting urban black holes based on human mobility data. In: *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seattle, Washington, USA: ACM, 35-44.

Kulldorff, M. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481-1496.

Li, Z., Xiong, H., Liu, Y. 2012. Mining blackhole and volcano patterns in directed graphs: a general approach. *Data Mining & Knowledge Discovery*, 25 (3):577-602.

Pei, T., Wang, W., Zhang, H., Ma, T., Du, Y., Zhou, C. 2015. Density-based clustering for data containing two types of points. *International Journal of Geographical Information Science*, 29(2):175-193.

Shiode, S. 2011. Street-Level Spatial Scan Statistic and STAC for Analyzing Street Crime Concentrations. *Transactions in GIS*, 15 (3):365-383.

Zheng, Y. 2015. Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology*, 6 (3):29.