Ant Colony Optimization-based Spatial Scan Statistic for Detecting Statistically Significant Spatial Communities in Vehicle Movements

S.C. Zhu, Q.L. Liu, Z.H. Wu

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

Abstract

In the era of big data, discovery of spatial communities in vehicle movements plays a key role in understanding the urban structures and functions. While a number of community detection methods can be used to detect spatial communities in vehicle movements, these methods are usually designed without considering the network-constraint of vehicles and testing the significance of spatial communities. On that account, this study develops an ant colony optimization-based spatial scan statistic to detect statistically significant spatial communities in vehicle movements on urban road network. Road segments are used as basic units to represent the moving paths of vehicles. The spatial scan statistic is generalized to weighted spatially embedded graph to provide quantitative assessment for spatial communities, and the generalized spatial scan statistic and ant colony optimization are combined to detect arbitrarily shaped spatial communities. A Monte Carlo simulation method is developed to estimate the statistical significance of each detected spatial community. The effectiveness of the proposed method is evaluated by using both simulated and taxi GPS trajectory data sets.

Keywords: spatial community, significance test, trajectory, spatial data mining

1. Introduction

With the development of location-aware technologies (e.g. Global Navigation Satellite System), a large amount of individual vehicle trajectory data (e.g. taxi GPS trajectories) have become increasingly available. These trajectory data can reveal the spatial interactions among different regions, and make it possible to analyse spatial communities in a city. A spatial community refers to a sub-graph of the spatially embedded graph constructed based on the interactions among different regions, where nodes within the sub-graph have significantly more internal connections than connections to other nodes (Guo et al., 2018). Discovery of such spatial communities in vehicle movements plays a key role in understanding the urban structures and functions (Liu et al., 2019).

Most of existing studies for discovery of spatial communities first map the origin and destination (OD) points of vehicle trajectories onto certain areal units (e.g. spatial grids or traffic analysis zones), and then detect spatial communities by using two kind of methods, i.e. community detection methods ignoring geographic constraint and spatial community detection methods (Guo et al., 2018). The former methods first define an objective function (e.g. modularity), and then use an optimization method to find the best partition of the spatially embedded graph to maximize the objective function (Clauset et al., 2004; Rosvall and Bergstrom, 2008). However, without the consideration of geographic constraint, the community detection methods mainly identify communities strongly determined by geographical factors and usually fail to detect communities determined by other underlying factors

(Expert et al., 2011). To overcome this limitation, spatial community detection methods have been developed by integrating spatial factors into the objective function of a traditional community detection method (Expert et al., 2011; Gao et al., 2013) or enforcing a spatial contiguity constraint in a traditional community detection method (Wan et al., 2018; Guo et al., 2018). Although spatial community detection methods are more robust in discovering underlying spatial community structures, they usually neglect that vehicles in urban space are strongly constrained by road network. The spatial communities detected based on the areal units will be seriously influenced by the aggregation problem (Zhu et al., 2017). Recently, Zhu et al. (2017) and Liu et al. (2019) chose road segment or "stroke" as geographic unit to detect spatial communities, however, the statistical significance of discovered communities cannot be evaluated (they can always discover spatial communities in a spatially embedded graph, even though the graph has no natural spatial community structure) (Zhang and Moore, 2014). Although Wang et al. (2008) extended the spatial scan statistic (Kulldorff, 1997) to identify statistically significant communities, it cannot be used to detect arbitrarily shaped spatial communities from a weighed graph.

Based on the above analysis, it can be seen that existing spatial community detection methods are usually designed without considering the network-constraint of vehicles and testing the significance of spatial communities; therefore, the identified spatial communities are very likely to be unreliable, even spurious. To overcome this limitation, this study aims to develop an ant colony optimization-based spatial scan statistic for detecting statistically significant spatial communities with irregular shapes in vehicle movements.

2. Spatial scan statistic for weighted spatially embedded graph

Vehicle trajectories are first matched onto the corresponding streets using a simple map-matching method: choosing the nearest road segment of a sampled location as its matched road segment (Zhu et al., 2017). Then, a moving path of a vehicle can be represented as: path= [road segment$_1$, road segment$_2$, …, road segment$_n$]. Further, the spatially embedded graph $\boldsymbol{G}$ = ($\boldsymbol{V}$, $\boldsymbol{E}$) is constructed as follows:

(i) $\boldsymbol{V}$ is the vertex set, and each road segment is regarded as a vertex of $\boldsymbol{G}$;

(ii) $\boldsymbol{E}$ is the edge set, and two consecutive road segments in a path form an edge in $\boldsymbol{G}$, represented as < road segment$_i$, road segment$_{i+1}$>;

(iii) The weight of an edge < road segment$_i$, road segment$_{i+1}$> in $\boldsymbol{G}$ is defined as the number of paths in which road segment$_i$ and road segment$_{i+1}$ are consecutive road segments;

(iv) The strength of a vertex v_i is defined as the sum of the weights of edges linked to v_i.

For a weighted spatially embedded graph $\boldsymbol{G}$= ($\boldsymbol{V}$, $\boldsymbol{E}$), a spatial scan statistic is defined based on the Poisson model. For a sub-graph $\boldsymbol{Z}$, W_Z is the observed sum of the weights of the edges in $\boldsymbol{Z}$, S_Z is the sum of the strengths of the vertexes in $\boldsymbol{Z}$, $\mu(Z)$ is the expected sum of the weights of the edges in $\boldsymbol{Z}$ under the Poisson model, $\mu(Z) = \frac{S_Z^2}{4W_G}$, W_G is the observed sum of the weights of the edges in $\boldsymbol{G}$, S_G is the sum of the strengths of the vertexes in $\boldsymbol{G}$, $\mu(G)$ is the expected sum of the weights of the edges in $\boldsymbol{G}$ under the Poisson model, $\mu(G) = \frac{S_G^2}{4W_G}$.The likelihood ratio statistic of a sub-graph $\boldsymbol{Z}$ can be represented as follows:

$$LR(Z) = \frac{L_Z}{L_0} = \begin{cases} \left(\frac{W_Z}{\mu(Z)}\right)^{W_Z} \left(\frac{W_G - W_Z}{\mu(G) - \mu(Z)}\right)^{W_G - W_Z} & if \ \ \frac{W_Z}{\mu(Z)} > \frac{W_G - W_Z}{\mu(G) - \mu(Z)} \\ 1 & otherwise. \end{cases} \quad (1)$$

3. Ant colony optimization-based spatial scanning method

To detect statistically significant spatial communities efficiently, spatial scan statistic is first used to detect candidate road segments in spatial communities, and then candidate road segments are grouped into spatial communities by using ant colony optimization (Dorigo and Stützle, 2004).

3.1 Detection of candidate road segments in spatial communities based on spatial scan statistic

For each vertex v_i in G, the scanning window (or sub-graph) is defined as the first-order neighbourhood of v_i. For each scanning window, the likelihood ratio statistic in Eq.(1) is calculated, and the significance (p-value) of each scanning window is calculated using a Monte Carlo simulation method. Each simulated dataset is generated by randomly assigning each trajectory on the road network. The p-value of a scanning window Z can be calculated as:

$$p_Z = \frac{\sum_{i=1}^{N_{rep}} I_i}{N_{rep}} \quad (2)$$

Where N_{rep} is the number of simulated datasets (N_{rep} is set to 999 in this study), I is an indicator variable. After the ith simulation, if $LR_{sim}(Z) > LR(Z)$, then $I_i = 1$, otherwise, $I_i = 0$ ($LR_{sim}(Z)$ is the likelihood ratio statistic calculated based on the simulated dataset). To avoid losing any candidate road segments, the significance level is set to 0.1 and the multiple testing problem is not adjusted. Only the candidate road segments need to be grouped by the ant colony optimization method, therefore, the search space of the ant colony optimization method is significantly reduced.

3.2 Discovery of arbitrarily shaped spatial communities based on ant colony optimization

The ant colony optimization method is further employed to identify spatial communities via the random walks of ants in ***G*** based on the walk reachability between candidate road segments (Pei et al., 2011). The parameters of the ant colony optimization method are set according to the suggestion given by Dorigo and Stützle (2004). The number of ants (N_{ant}) is set to the number of candidate road segments, the maximum number of iterations (N_{ite}) is set to 200, the number of elite ants (N_e) is set to 20, the initial pheromone value (M_{ip}) of each candidate road segment is set to the normalized value of the number of paths containing that candidate road segment, the pheromone evaporation value (M_{pe}) is set to 0.1 and the pheromone increment value (M_{pi}) is set to 0.1. The spatial communities are discovered in the following five steps:

Step 1: Randomly select a candidate road segment R_1, and then randomly determine a maximum walking length L of an ant from a Gaussian random function.

Step 2: An ant walks from R_1, and the next road segments R_2 that the ant walks into is determined in proportion to the probability calculated based on the pheromone on R_2. The ant stops walking until the walking length of the ant reaches L, then a sub-graph formed road segments is obtained.

Step 3: Repeat Step 2 N_{ant} times, and obtain N_{ant} sub-graphs. For each sub-graph, calculate the likelihood ratio statistic using Eq.(1). The values of likelihood ratio statistic are sorted in descending order $\boldsymbol{LR}=[LR_1, LR_2,\ldots,LR_{Nant}]$. The first N_e ants in $\boldsymbol{LR}$ are identified as elite ants, and the pheromones of the road segments in sub-graphs generated by elite ants will be increased by M_{pi}. Pheromones on all the road segments volatilize M_{pe}.

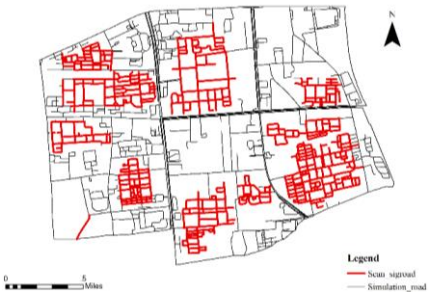
Step 4: Repeat Step1 to Step3 N_{ite} times. All the sub-graphs generated by ants are considered as candidate spatial communities.

Step 5: The p-value of each candidate spatial community is calculated using the method introduced in Section 3.1. The significance level α is set to 0.01. The False Discovery Rate approach (Benjamini and Yekutieli, 2001) is used to control the multiple testing problem, and calculate the adjusted significance level α_{adj}. A spatial community whose p-value is smaller than α_{adj} will be identified as a statistically significant spatial community.

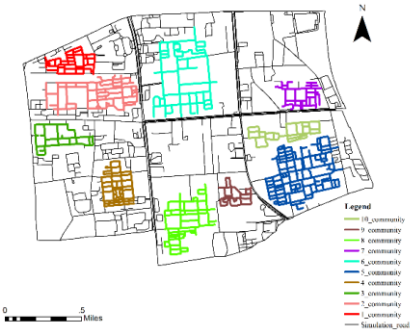
4. Experimental analysis

To evaluate the performance of the proposed method, a synthetic trajectory dataset with ten known spatial communities and 30% random moves was generated on a real road network taken from Beijing, China according to the method introduced by Guo et al. (2018) (shown in Figure 1(a)). The identified candidate road segments are presented in Figure 1(b). In Figure 1(c), one can see that ten predefined spatial communities are well discovered by the proposed method. In Figure 1(d), the communities discovered by the widely-used modularity-based hierarchical clustering (Clauset et al., 2004) are shown. It can be found that only three spatial communities can be identified roughly. The example with GPS trajectories in Beijing will be discussed in the conference presentation. And the statistical significance of the identified networks will show in the figure of the Beijing experiment result.

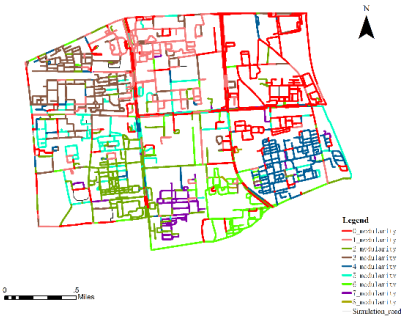
(a) Ten simulated spatial communities



(b) Candidate road segments



(c) Communities detected by the proposed method



(d) Communities detected by modularity

Figure 1. Experimental results on simulated dataset

The proposed method was also applied to detect spatial communities from the taxi GPS trajectory data for 1 day in Beijing, China. The discovered spatial communities clearly reveal the polycentric structure of the city, and will be useful for planning an efficient spatial configuration for the city.

5. Conclusion

In this study, a spatial scanning method based on ant colony optimization is developed for detecting statistically significant spatial communities with arbitrary shapes. By using the road segments as the basic units to represent the paths of vehicles, the network-constraint of the vehicles can be considered and the aggregation problem can be minified. The constructed spatial scan statistic for weighted spatially embedded graph can assess the spatial communities quantitatively and meaningfully. The ant colony optimization-based spatial scanning method can not only detect arbitrarily shaped spatial communities, but also evaluate the statistical significance of each discovered spatial community. Experimental results show that the proposed method is more effective for detecting spatial communities in vehicle movements.

6. 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.

Clauset, A., Newman, M. E. and Moore, C. 2004. Finding community structure in very large networks. *Physical Review E*, 70, 066111.

Dorigo, M. and Stützle, T. 2004. *Ant Colony Optimization*. Cambridge, MA: MIT Press.

Expert, P., Evans, T.S., Blondel, V.D., *et al*. 2011. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences of the United States of America*, 108(9): 7663-7668.

Gao, S., Liu, Y., Wang, Y.L., *et al*. 2013. Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, 17(3): 463-481.

Guo, D.S., Jin, H., Gao, P., *et al*. 2018. Detecting spatial community structure in movements. *International Journal of Geographical Information Science*, 32(7): 1326-1347.

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

Liu, K., Gao, S., Lu, F. 2019. Identifying spatial interaction patterns of vehicle movements on urban road networks by topic modelling. *Computers, Environment and Urban Systems*, 74: 50-61.

Newman, M.E.J. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23): 8577-8582.

Pei, T., Wan, Y., Jiang, Y., *et al*. 2011. Detecting arbitrarily shaped clusters using ant colony optimization. *International Journal of Geographical Information Science*, 25(10):1575-1595.

Rosvall, M., Bergstrom,C.T. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4): 1118-1123.

Wang, B., Phillips, J. M., Schreiber, R., *et al*. 2008. Spatial Scan Statistics for Graph Clustering. In: *Proceedings of the 2008 SIAM International Conference on Data Mining*,727-738.

You, W. and Yaolin, L. 2018. Dasscan: a density and adjacency expansion-based spatial structural community detection algorithm for networks. *ISPRS International Journal of Geo-Information*, 7(4), 159.

Zhang, P. and Moore, C. 2014. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proceedings of the National Academy of Sciences of the United States of America*, 111(51): 18144-18149.

Zhu D, Wang, N. H, Wu, L., *et al*. 2017. Street as a big geo-data assembly and analysis unit in urban studies: a case study using Beijing taxi data. *Applied Geography*, 86: 152-164.