Exploring the Homogeneity of Offenders in Crime Hotspots

Tongxin Chen[*1], Tao Cheng [†1], Yang Zhang [‡1]

[1] SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental and Geomatic Engineering, University College London, Gower Street WC1E 6BT, London, U.K.

January 13, 2019

Summary

Exploring the homogeneity of offenders within crime hotspots is helpful for not only understanding how the similar offenders commit the crime following the same spatio-temporal pattern, but also for crime prevention or crime investigation work. In this study, we utilise ST-DBSCAN algorithm to detect crime hotspots using historical theft records in central area of Beijing, China. Leveraging demographic information of the associated offenders, we propose three novel entropy-based indices to measure the similarity of offenders. Results show that the crimes concentrated in a narrow space and time span are usually committed by a group of offenders with similar demographics, which is referred to as homogeneous offenders.

KEYWORDS: Spatial-temporal clustering, homogenous offenders, ST-DBSCAN, entropy, unsupervised learning

1. Introduction

The homogeneity of offenders refers to the offenders with similar social-demographics, i.e. belonging to same offender group. Exploring the homogeneity of offenders who committed crimes at similar time and locations (spatio-temporal concentrated crimes) is meaningful for not only crime prevention, but also for the crime investigation. It is useful for understanding not only the characteristic of the offenders but also why such kind of offenders concentrate on the hotspot. For example, Bowers and Johnson (2012) compared crime hotspots generated by "prolific" offenders and "occasional" offenders. Their study, however, does not examine the characteristic similarities among the offenders within the hotspots. In this study, we utilise a spatial-temporal clustering method to extract crime hotspots. Then, we propose several novel indices to examine the homogeneity of offenders in the same crime hotspot.

2. Methodology

2.1. Data

The study area is a central district of Beijing, China (latitude: 39.7574°~40.0287°, longitudes: 116.1937°~116.5519°) and the crime data is the theft data in the study area ranging from 1st Jan 2014 to 30th Dec 2014. The data consist of 7,802 crime offences committed by 6,754 offenders. The demographic information of offenders include name, ID, age, registered home address (RHA).

2.2. Spatial temporal crime hotspots

[*] tongxin.chen.18@ucl.ac.uk
[†] tao.cheng@ucl.ac.uk
[‡] yang.zhang.16@ucl.ac.uk

Spatial-temporal clustering is a process of grouping objects based on their spatial and temporal similarity. It has been widely applied to crime hotspot analysis (Johnson and Bowers, 2004; Johnson, Summers and Pease, 2009). Among the existing spatial-temporal clustering methods, we utilise a density-based method called Spatial-temporal Density Based Spatial Clustering of Applications with Noise (ST-DBSCAN) algorithm proposed by Birant and Kut (2007). Efficiently, the clusters with same spatial-temporal pattern represent crime hotspots could be detected from ST-DBSCAN considering both space and time concentration. ST-DBSCAN has three predefined parameters, spatial maximum reachable distance (SMRD), temporal maximum reachable distance (TMRD), Minimum number of points within SMRD and TMRD (MinPts). In this study, the SMRD iterates over the range [100, 5000] with step 400 (spatial unit: metre) and TMRD iterates over the range [1, 60] with step 4 (temporal unit: day) to determine the optimal parameter settings. Further, MinPts is set to be three as this is the minimum counts in an offender group or a serial crime.

2.3. The homogeneity of offenders

Various indices have been developed to measure the homogeneity or diversity of detected clusters or groups, such as the Simpson Index in ecology (Simpson, 1949) and the Shannon Entropy Index in information theory (Shannon, 1948). In criminology, Bouhana, Johnson and Porter (2014) used the Simpson Index to evaluate the consistency of MOs (modus operandis) burglars. In this study, we define several novel indices based on Shannon Entropy to assess the homogeneity of offenders in a crime hotspot. Shannon Entropy is a metric of measuring the amount of disorders originally and it can be further established to variations from the calculation. The definitions of the proposed indices are given below.

Definition 1. Entropy of an Attribute (EOA): Suppose attribute A is a random variable with possible values $\{a_0, a_1 \cdots a_i \cdots a_n\}$. The entropy H is defined as:

$$H(A) = -\sum_{i=0}^{n-1} P(a_i) \log_2 P(a_i) \quad (1)$$

where, $H(A)$ indicates the homogeneity of the feature A among offenders' demographic variables in the same crime hotspot. A higher H indicates a lower homogeneity. $P(a_i)$ is the probability of $A = a_i$. For example, suppose we categorise the age of offenders into levels, denoted as $A_{age} = \{level_0, level_1, level_2\}$. If a crime hotspot committed by five offenders with age $level_0$, $level_1$, $level_1$, $level_2$, $level_2$, respectively. The homogeneity of age of offenders in this hotspot is then:

$$H(A_{age}) = -\frac{1}{5} \times \log_2 \frac{1}{5} - \frac{2}{5} \times \log_2 \frac{2}{5} - \frac{2}{5} \times \log_2 \frac{2}{5} \approx 1.5219$$

Definition 2. Entropy of a Cluster (EOC): If Cluster C consists of objects with m attributes $\{A_0, A_1 \cdots A_j \cdots A_m\}$, the entropy of a cluster considering all attributes is defined as:

$$H(C) = \sum_{j=0}^{m} H(A_j) \quad (2)$$

where, $H(C)$ indicates the homogeneity of a cluster and A_j is the j-th attribute of offenders (e.g. age) in a cluster m is the number of attributes' categories in a cluster. And higher H indicates a lower homogeneity of offenders in the hotspots.

Definition 3. Entropy of Clusters (ECs): To measure the quality of the clustering results, we propose to use the entropy of all detected clusters. Clusters S is the set of different clusters from a certain pattern and could be a system that also assessed by the overall entropy. ECs is defined as the weighted sum of EOC is denoted as:

$$H(S) = \sum_{k=0}^{K} H(C_k) * \omega_k \ , \ \omega_k = \frac{|C_k|}{|S|}, \quad S = \{C_0, C_1 \cdots C_K\} \quad (3)$$

where, K is the number of obtained clusters; C_k is the k-thcluster in the clusters set S; $|C_k|$ is the number of objects in cluster C_k; $|S|$ is the total number of objects in all clusters; ω_k is the weight of C_k.

3. Case Study

In this paper, we mainly study the homogeneity of offenders' age and RHA as the representation of demographic homogeneity. This is because offenders with similar ages and coming from the same hometown are prone to constructing crime networks for sharing crime opportunity information. The attribute "age" is categorized into 4 levels: under 18, 18~40, 40~60, >60 and RHA (30 main provinces in China) represents where offenders come from.

To determine the optimal number of the crime clusters, we first use the ECs to measure the quality of clustering results obtained by using different combinations of SMRD and TMRD. For the space time unit showing a narrow range significantly based on outcome, the optimal values of SMRD and TMRD are determined as 500m and 7days, respectively. And the theft crimes can be then clustered by using ST-DBSCAN. Based on the clustering results, we calculate the EOC of each cluster to measure the homogeneity of offenders. The distribution of crime clusters and their EOC are visualized in Figure 1.

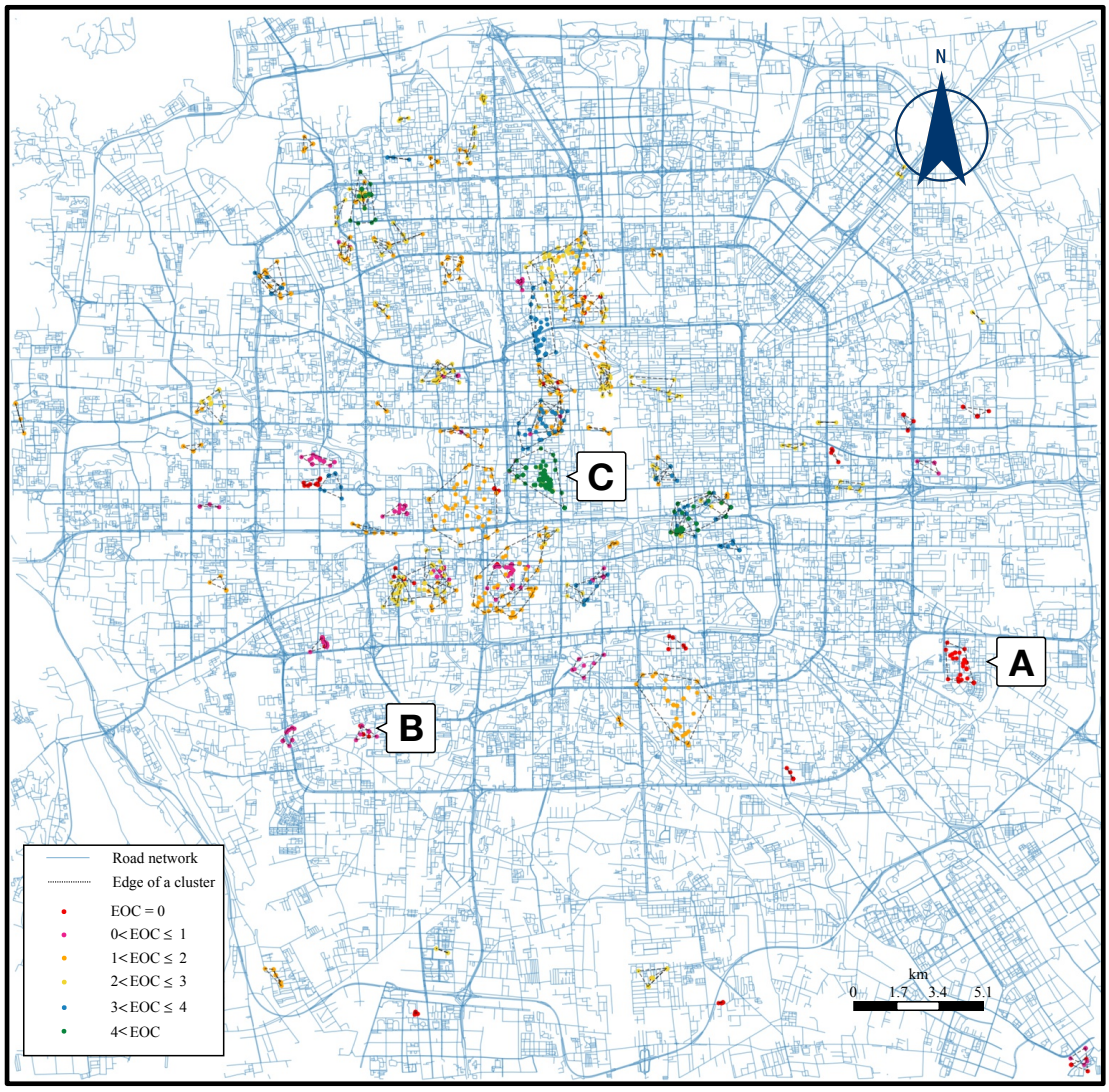

Figure 1 Theft offence clusters in different levels of EOC based on a spatial-temporal unit of 500 m and 7 days

Fig 1 shows crime hotspots (spatial-temporal clusters) are mainly distributed on the west side of the study area. And the hotspots of high-level homogeneity (e.g. red, purple such as Clusters A and B) are further from the city center compared with the hotspots of low-level homogeneity (e.g. orange, green such as Cluster C). In particular, crime hotspot A consists of serial thefts from cars committed by an offender in two weeks since the cars are parking along the roadside lacking guardian. In addition, hotspot B consists of several thefts against pedestrians committed by two co-offenders with the same level of age and from the same RHA. In the hotspot B, we also found another single theft committed by an offender occasionally, whose RHA and age level are different with the co-offenders. Lastly, hotspot C consist of theft offences committed by a diversity of offenders, because the places of offences locate at the largest commercial district of Beijing, which attracts not only lots of consumers but also the potential theft offenders coming from diverse places.

4. Conclusion

The homogeneity of offenders is correlated with a specific pattern of spatial-temporal concertation, which means humongous offenders could be detected in the majority of hotspots through the spatial-temporal clustering algorithm. The proposed Shannon Entropy indices captures the homogeneity of the offenders within the hotspots, which facilitate offences investigation leading to crime prevention.

5. Acknowledgements

This work was partially supported by the UK Economic and Social Research Council Consumer Data Research Centre (CDRC) under Grant ES/L011840/1. Valuable suggestions from anonymous reviewers are gratefully acknowledged.

References

Bowers K and Johnson S (2012). Contrasting hotspots: Did the opportunists make the heat? In: Farrell G and Tilley N (eds) *Crime Prevention Studies*. Monsey NY: Criminal Justice. Lynne Rienner Publishers, pp 212-225.

Bouhana N, Johnson S and Porter M (2014). Consistency and specificity in burglars who commit prolific residential burglary: testing the core assumptions underpinning behavioural crime linkage. *Legal and Criminological Psychology*, 21(1), 77-94.

Derya Birant and Alp Kut (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60,208–221.

Johnson S and Kate B (2004). The stability of space-time clusters of burglary. *British Journal of Criminology*, 44 (1), 55-65.

Johnson S, Summers L and Pease K (2009). Offender as forager? A direct test of the boost account of victimization. *Journal of Quantitative Criminology*, 25, 181–200.

Shannon C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal.* 27, 379–423.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163, 688.

Biographies

Tongxin Chen is a PhD student at UCL SpaceTime Lab. He received his BSc and MSc both in crime investigation and criminology. His research interests include crime mapping, machine learning for crime pattern analysis and geographic profiling.

Tao Cheng is a professor in Geoinformatics at Department of Civil, Environmental and Geomatic Engineering, University College London. She is the Founder and Director of SpaceTimeLab for Big Data Analytics. Her research interests span network complexity, Geocomputation, space-time analytics and Big data mining (modelling, prediction, clustering, visualisation and simulation) with applications in transport, crime, business, health, social media, and natural hazards.

Yang Zhang is a PhD student in department of Civil, Environmental and Geomatic Engineering at University College London. Her research interest includes spatial-temporal data mining, deep learning and complex network, with applications in crime and transportation.