Determining the optimal spatial scan extent (*K*) of a Prospective space-time scan statistics (PSTSS) that maximises the predictive accuracy of crime prediction

Monsuru Adepeju[*1], Tao Cheng[†1]

[1, 2]Civil, Environmental and Geomatic Engineering, University College London
WC1E 6BT, Gower Street, London

April 18, 2017

Summary

The input value for the maximum spatial scan extent (*K*) of a prospective space-time scan statistical (PSTSS) technique is one important parameter that ensures accurate detection of the predictive hotspot of geographical point events. Currently, there is no general consensus on how to determine the optimal value of *K* that maximises the predictive accuracy, especially in crime hotspot prediction. To address this issue, this study proposes a strategy in which different values of *K* are used to generate accuracy profiles that can be compared to determine the optimal value of *K*. A case study presented shows that the best accuracy is obtained at *K*=250m, which is equivalent to the spatial aggregation of the crime dataset used. It is contended that this strategy could be extended to other geographical surveillance tasks.

KEYWORDS: Space-time, Crime, Hotspot prediction, Spatial scan extent, predictive accuracy.

1. Introduction

In the use of Prospective space-time scan statistical (PSTSS) technique for crime prediction (Adepeju et al. 2016), the maximum spatial scan extent (*K*) is primarily set to ensure that all possible sizes of geographical hotspots are identified to achieve accurate results. The *K* represents the maximum size that any hotspot identified can assume during a prediction task. However, how to determine the best value of *K* that maximises the predictive accuracy has remained an open question. Previous studies (Kulldorf et al. 2005) have suggested setting the value of *K* as half the spatial size of the study area. While this suggestion has a number of merits (Neill, 2006), it usually results in an extremely large computational time, especially when the number of unique spatial locations to be scanned is very large. A different approach is employed in Adepeju et al. (2016) in which a small value of *K* (i.e. 750m) was set based on the observed spatial aggregation of the crime data used. While this approach allowed faster computation, the predictive accuracy of the PSTSS was observed to be very low as compared to other predictive methods. In this study, we suggest that this predictive accuracy can be improved by determining the optimal value of *K*, rather than using any of the aforementioned suggestions.

This study proposes a new strategy by which the optimal value of *K*, that maximises the predictive accuracy of a PSTSS technique, can be determined. The strategy entails using different values of *K* to carry out predictions, while the profiles of the predictive accuracy (i.e. *mean hit rate*) are generated and continuously monitored. The profile of *mean hit rate* will be examined at varying spatial coverages of the hotspot, for a more robust analysis.

The structure of this abstract is as follows: the description of PSTSS technique and how the technique is used to generate predictive hotspots of crime. This is followed by the description of the objective function, i.e. the *mean hit rate* to be used. We then present the results of the three study areas. Lastly, the discussion and conclusion are presented.

[*] monsuru.adepeju.11@ucl.ac.uk
[†] tao.cheng@ucl.ac.uk

2. Prospective space-time scan statistical (PSTSS), and the value of *K*

The PSTSS is a surveillance technique that identifies regions (cylinders) of emerging risks of a geographical point datasets *N*, distributed in space and time (Kulldorff et al. 2005). The technique works by placing, on every unique point location (x, y), a cylinder whose widths (spatial extents) and lengths (temporal durations) are varied continuously, until certain maximum values are reached. These maximum values, denoted as *K* and *T*, are the *maximum spatial scan extent* and the *maximum temporal scan duration* that the varying cylinders cannot exceed. For the purpose of this study, we fixed *T* as half the temporal length of a *prediction set* (See Figure 3c), while we focus on how to determine the optimal value for *K*. A prediction set is the dataset used to generate the cylinders, from which the predictive hotspots are generated. The prediction set is bounded by a start date (t_o) and an end date (t_p), with all cylinders having the same end date as t_p (See Figure 1).

Given a cylinder containing *n* data points, the data points are approximated as Poisson distribution with the mean μ (Kulldorff et al. 2005). The risk value *S* of the cylinder is estimated by the log value of:

$$S = \left(\frac{n}{\mu}\right)^n \left(\frac{N-n}{N-\mu}\right)^{N-n} \quad (1)$$

The cylinders required for generating the predictive hotspot are extracted such that no overlap exists between the cylinders (e.g. Figure 1a). The final map representing the predictive hotspot is then generated by intersecting the top circular area of the cylinders with a system of square grids (Figure 1b). The risk level of the grid units is assigned in the order of the proximity of the grid units to the centroids of top circular area of their respective intersecting cylinder, where the cylinders are selected in the order of magnitude of their *S* values. The grid units that are ranked first and last are the most 'risky' and 'least' risky locations, respectively. The rest of the grid units that are not intersected by any cylinder are simply referred to as the cold spots (i.e. regions of no risk at all).

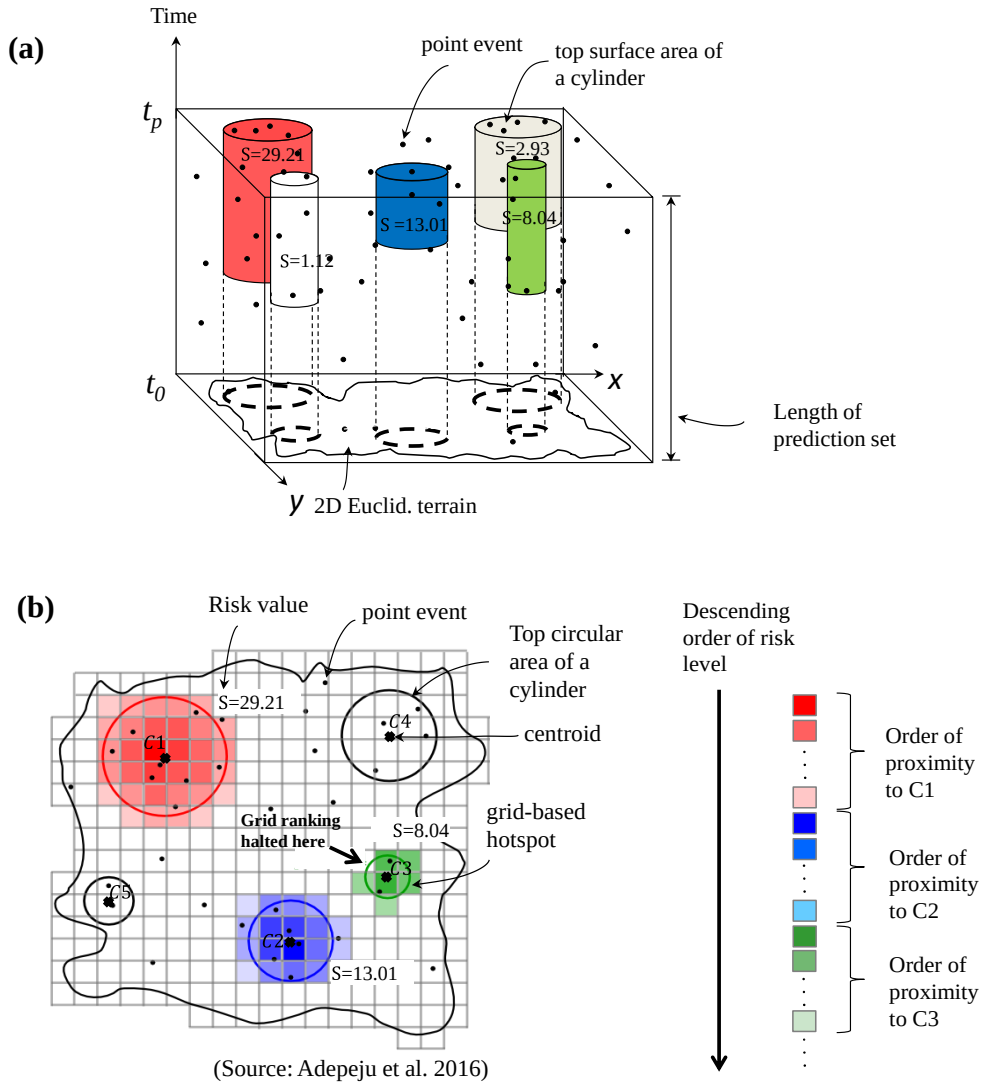

Figure 1. Producing predictive hotspot of point events using PSTSS technique. (a) shows cylinders, identified by PSTSS, representing regions of emerging risk of the point events (b) generated grid-based predictive hotspot based on the cylinders in (a). *C*1, *C*2, ….*C*5 are the centroids of the top circular area of the cylinders. Note: the *S* values shown are assumed.

If for example, the point events in Figure 1 is a crime dataset, the shaded grid units will represent regions at risk of being victimised, imminently. Starting from the most 'risky' grid unit (i.e. the darkest red shade), one can delineate just an amount of hotspot coverage of interest. For example, in Figure 1b, the ranking process is halted partway through the filling of the green circle, giving a hotspot coverage of 11.9% (i.e. $\frac{\text{no. of selected grid units}}{\text{total no. of grid units}} = \frac{35}{294}$). This flexibility is very important in a real practical environment, where the police may be interested in a certain hotspot coverage based on the capacity of their resources.

3. The proposed strategy to determine optimal value of *K*

The proposed strategy to determine the optimal value of *K* involves setting a particular value of *K* to generate predictive hotspot of crime and evaluate the accuracy of the hotspot at an hotspot coverages, using the *hit rate* measure (Bowers et al. 2004). The *hit rate* is defined as the proportion of crimes accurately captured by the predicted hotspot:

$$\text{hit rate} = \frac{a_c}{A} \times 100\% \qquad (1)$$

Where a_c is the actual number of future crimes captured by the hotspot at a coverage c, and A, the total number of future crimes that can be captured. When this process is repeated over a total number of time steps j, the *mean hit rate* at c can be calculated as

$$Mean\ hit\ rate_c = \left(\frac{\sum_{i=1}^{j} \left(\frac{a_{c,i}}{A_i} \times 100 \right)}{j} \right) \% \quad (2)$$

Figure 2 is an illustration of a *mean hit rate* profile generated by calculating the *mean hit rate* at varying hotspot coverages.

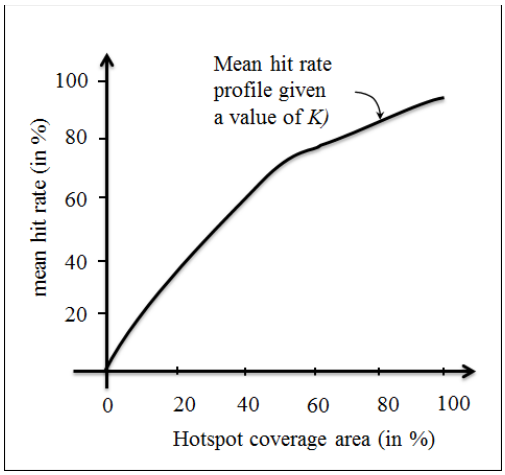

Figure 2. Mean hit rate profile of a number of consecutive predictions, based on a value of *K*

The descriptions above constitute a predictive analysis for one value of *K*. We suggest that by running the same analysis, but use different values of *K*, the resulting *mean hit rate* profiles can be compared to determine the profile with the best (highest) *mean hit rate* at the increasing spatial coverages. This value of *K* that produces the best *mean hit rate* will be considered optimal for the dataset in question.

4. Study areas, datasets, and details of analysis

For this study, crime case data of Camden and South Chicago used in Adepeju et al. (2016) is also used here. In addition, a crime case dataset of new study area - District 10 of San Francisco, is added (Figure 3a), in which two new crime types, 'larceny/theft' and 'drug/narcotic', which are not present in either of Camden or South Chicago are included. In all, we have six (6) different crime types across the three study areas (Figure 3b). A system of 250m by 250m grid is used for aggregating the dataset and will also be used as the unit of the predictive hotspot.

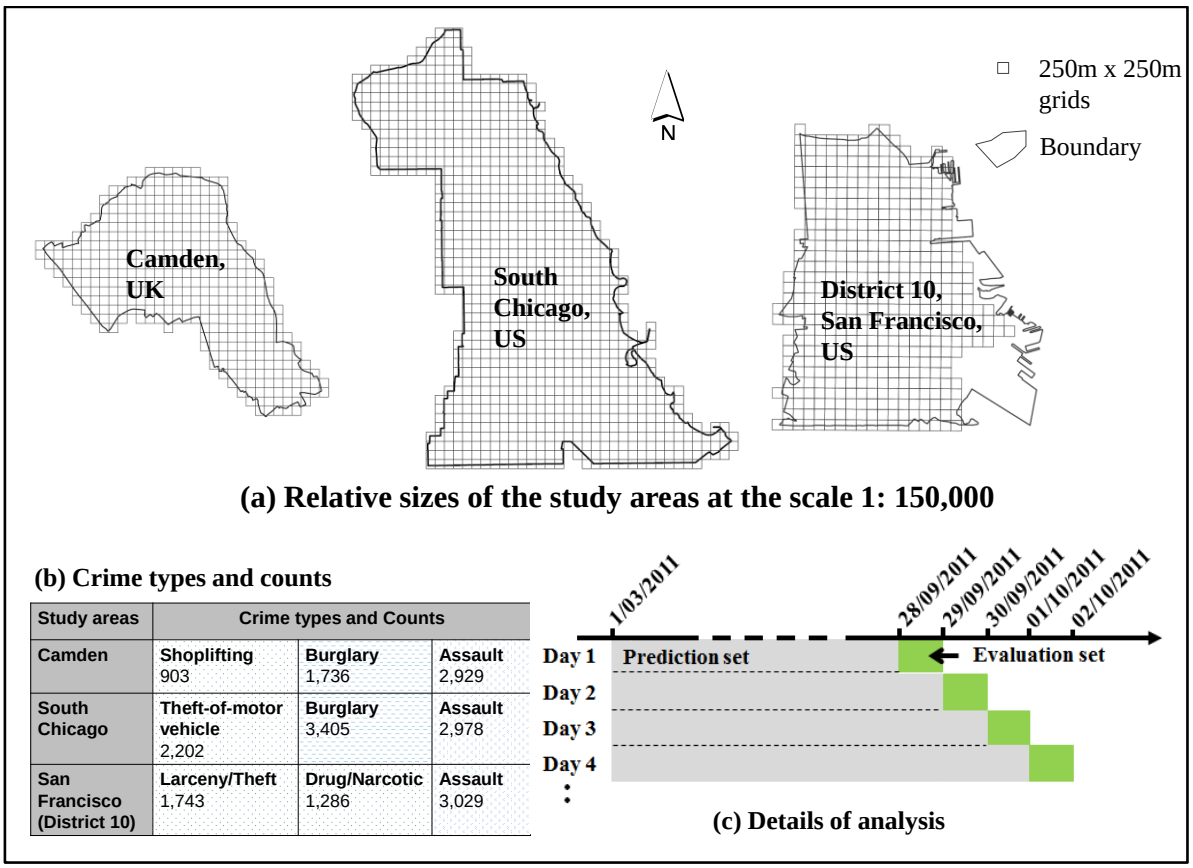

Figure 3. Study areas, dataset and details of analysis.

The temporal range of each crime type is from 1st March 2011 to 6th January 2012. For this analysis, the first *prediction set* is the data range from 1st March 2011 to 28th September 2011 (7 months), while the *hit rate* will be evaluated based on the next one-day dataset (i.e. dataset of 29th September 2011). While the start date of the *prediction set* will be fixed as 1st March 2011, the end dates will be made to increase by 1 day as the analysis progresses, for a given value of *K* being examined (see Figure 3c).

A 100 daily consecutive prediction will be carried out. The list of values of *K* to be used is [250m, 500m, 750m and 1000m].

5. Results and discussion

Figure 4 shows the plot of mean hit rate profiles for different values of *K*, at varying hotspot coverage, for different crime types.

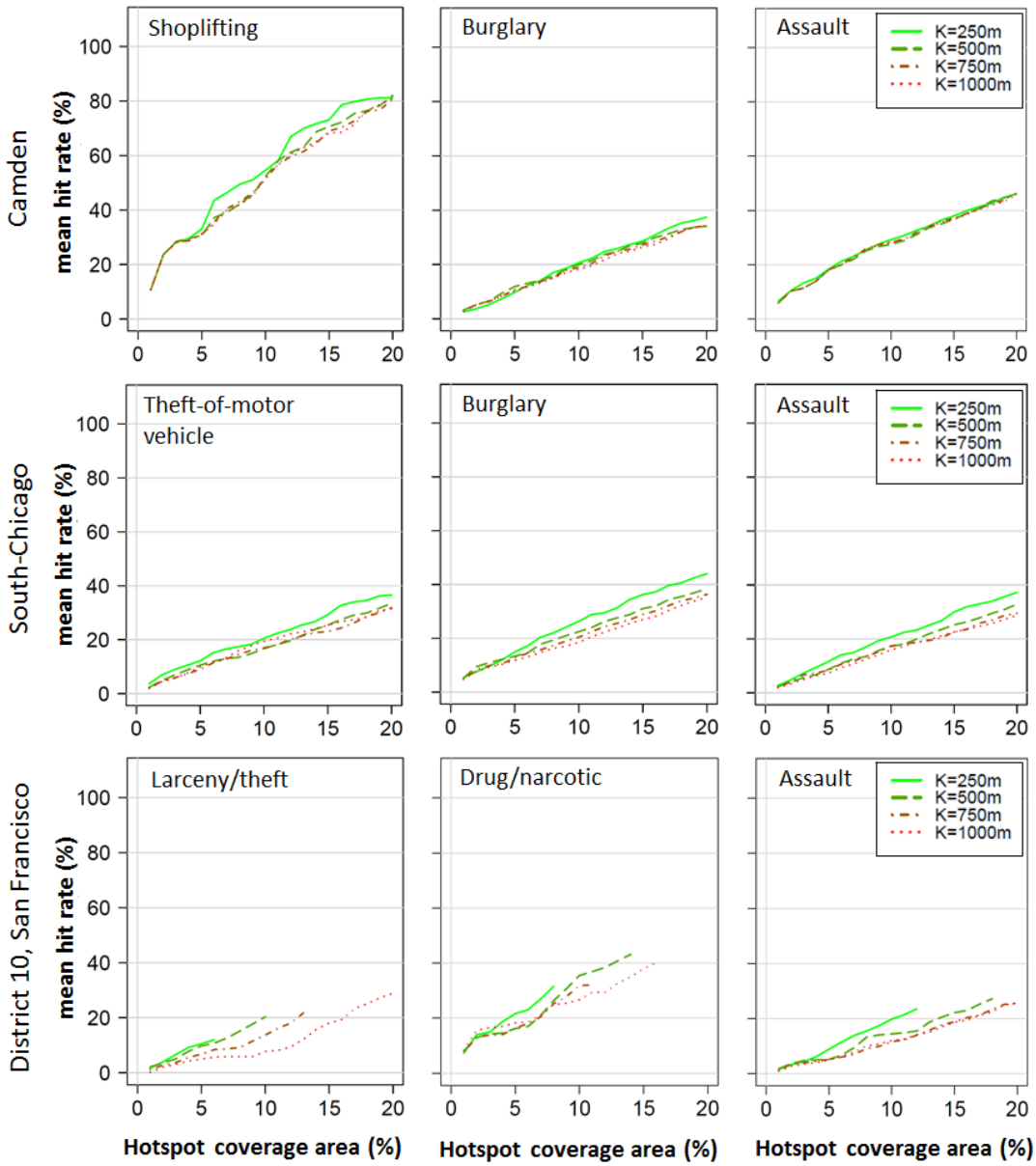

Figure 4. Profiles of m*ean hit rate* for different values of *K*, at varying hotspot coverage, for different crime types.

An increase in predictive accuracy (mean hit rate) is observed across all the profiles as the value of *K* decreases. This pattern is observed at all possible spatial coverages attainable by the hotspots generated based on each value of *K*. Comparing the profiles obtained at K=250m to the profiles at K=750m of the earlier study (Adepeju et al. 2016), a significant improvement in the accuracy can be seen. For example, not less than 15% improvement in accuracy is gained at exactly 20% coverage level of all predictions in South Chicago. It can be deduced based on this result that the PSTSS technique is able to capture the emerging risk of crime more effectively when its scan extent is restricted to a much smaller aggregation level of the dataset.

6. Conclusion

The goal of this study is to determine the optimal value of spatial scan extent *K* of PSTSS technique that maximises the accuracy of a crime hotspot prediction. The outcomes of this study reveal that the accuracy of PSTSS is maximised when *K*=250 which is equivalent to the spatial aggregation of the dataset. Additionally, setting the value of *K*=250 ensures that the hotspots are detected much faster than any other bigger value for K.

Based on this study, we suggest that the optimal scan extent of other surveillance techniques can be determined by first identify the evaluation measure to be maximised and then employ a similar strategy of varying K value while the evaluation measure is being monitored.

7. Future work

Our ongoing work on the PSTSS technique includes the development of the network-based version of the technique for the purpose of crime prediction. This work will also include the experimentation with *K* values below the spatial aggregation level of the dataset used.

8. Acknowledgment

The authors acknowledge the support of the CPC (Crime, Policing and Citizenship) project supported by UK EPSRC (EP/J004197/1), in collaboration with the London Metropolitan Police, who provided datasets used in this study.

9. Biography

Monsuru Adepeju is a completing PhD student at the SpaceTimeLab for Big Data Analytics, at University College London. His PhD focuses on modelling sparse spatio-temporal point process (STPP), with a special application in predictive policing. A significant part of his PhD outputs included a novel predictive framework of STPP datasets, which has been demonstrated to have a great potential for predicting crime very accurately.

Tao Cheng is a Professor in GeoInformatics, and Director of SpaceTimeLab for Big Data Analytics, at University College London. Her research interest span network complexity, Geocomputation, integrated spatio-temporal analytics and big data mining, with applications in transport, crime, health, social media, and environmental monitoring.

References

Adepeju, M., Rosser, G. and Cheng, T., 2016. Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions-a crime case study. International Journal of Geographical Information Science, pp.1-22.

Bowers, K.J., Johnson, S.D., and Pease, K., 2004. Prospective hot-spotting: the future of crime mapping? British Journal of Criminology, Vol. 44(5), pp.641–658

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., Mostashari, F., 2005. A space–time permutation scan statistic for disease outbreak detection. PLoS Med, Vol. 2(3), .e59. doi:10.1371/journal.pmed.0020059

Neill, D.B., 2006. Detection of spatial and spatio-temporal clusters (Doctoral dissertation, University of South Carolina).