Investigating the impacts of training data set length (T) and the aggregation unit size (M) on the accuracy of the self-exciting point process (SEPP) hotspot method

Monsuru Adepeju[1] and Andy Evans[2]

[1, 2] School of Geography, University of Leeds, LS21 1HB
[1]M.O.Adepeju@leeds.ac.uk

KEYWORDS: self-exciting, point process, crime prediction, temporal, aggregation

Abstract

This study examines the impacts of two variables; the training data lengths (*T*) and the aggregation unit sizes (M); on the accuracy of the self-exciting point process (SEPP) model during crime prediction. A case study of three crime types in the South Chicago area is presented, in which different combinations of values of *T* and *M* are used for 100 daily consecutive crime predictions. The results showed two important points regarding the SEPP model: first is that large values of *T* are likely to improve the accuracy of the SEPP model and second is that, a small aggregation unit, such as a 50m x 50m grid, is better in terms of capturing local repeat and near-repeat patterns of crimes.

1. Introduction

The self-exciting point process (SEPP) model is currently considered the most 'state-of-the-art' technique for generating hotspots of crime on a 2D Euclidean space, discretised into a system of square grids (Mohler et al., 2011; Adepeju et al., 2016). The length (*T*) of the training data set as well as the size (*M*) of the grid unit that is used to aggregate the SEPP's kernel estimations, are two important aspects which hugely impact on the level of accuracy that is achievable via SEPP (Mohler, 2014). In Mohler's work, a brief demonstration of the impact of changes to *M* was presented through using two values for the grid system (i.e. 75m x 75m and 150m x 150m). However, it is argued that this demonstration is not robust enough to draw any solid conclusions with regard to the real impact of changes in grid values on the performance of the SEPP. Mohler's work also claimed that a large value of T is likely to produce better accuracy than a smaller value of *T*. Yet, no conclusive empirical support for this claim exists in his article.

The aim of this study therefore, is to investigate the impacts of these two variables; T and M; more comprehensively, which will be achieved through creating a list of values for each variable and conducting crime predictions with all the possible pairs of values from the two lists. Furthermore, to ensure a more robust analysis, three different crime types have been chosen from a publicly available crime database of the South Chicago area. Due to these methods being utilised, it is expected that a better insight will be gained in relation to the best values of T and M that will generate the best predictions for the SEPP. Theoretically, the investigation of T and M in any spatiotemporal data analysis is synonymous with the temporal boundary problem (Cheng and Adepeju, 2014), and the modifiable areal unit problem - MAUP (Openshaw, 1981), in the time and space dimensions respectively.

The structure of this paper is as follows: a brief description of the SEPP model will first be provided, along with an explanation of how the two variables; T and M; factor into the model's results. This is followed by the description of the crime data sets, the values of T, the values of M, and lastly, the accuracy measurement. Moreover, the results and their discussion are presented. Finally, the conclusion and recommendations for future work are provided.

2. The self-exciting point process (SEPP)

The self-exciting point process (SEPP) is a modelling framework that has been used for several decades, particularly in the field of seismology so as to predict patterns of earthquake aftershocks. The model describes a dynamic point process in space and time in which a set of events may trigger further events within their spatial and temporal neighbourhood (Mohler et al., 2011). This idea was subsequently likened to the criminological theory of near repeat victimisation, and the model has thus been applied as a predictive framework for crime data (Mohler et al., 2011; Adepeju et al., 2016).

At the core of the SEPP model is the conditional intensity, $\lambda(t, x, y)$, which estimates the density of the expectation of crime within a small neighbourhood of a region (x, y) at time t. It is all conditional upon the history of events bounded by time $[t_0, t_i]$, where t_i, represents the present time. The range $T = t_i - t_0$ is the length of the dataset (training).

Theoretically, the SEPP takes the form,

$$\lambda(x, y, t) = \mu(x, y) + \sum_{t>t_i} g(x - x_i, y - y_i, t - t_i) \qquad \text{Equation 1}$$

Where μ represents the background (stationary) intensity and g represents the triggering function. Given Equation 1, it can be observed that all crimes that have occurred prior to t_i may theoretically contribute to an additional expectation being attributed to the current criminal activities taking place. In order to estimate (de-cluster) μ and g, different approaches have been proposed in the previous literature, including stochastic de-clustering (Zhuang et al., 2002), the maximum likelihood estimate (Daley and Vere-Jones, 2003) and the kernel-based approach (Mohler et al., 2011). For the purposes of this study, the kernel-based approach is employed. Thus, the resulting output is synonymous to adding two KDEs; one, for the background function, and the other, for the triggering function. The selection of the bandwidth is carried out using Silverman's rule approach, which computes optimal kernel bandwidths that best fit the spatial and temporal extent of a given dataset.

Given the two computed kernels, the final risk (hotspot) value inside each grid unit covering the study area is obtained by adding up all of the kernel estimates in each unit (*M*). Thus, the kernels are defined irrespective of a discretisation of space (i.e. the spatial grid unit size), and the latter is primarily for the purposes of visualisation and easy evaluation. While different sizes of grid unit have been used in previous studies, there has not been a comparative study of the impacts of the different grid unit sizes on the resulting accuracy of the final hotspot which is generated.

3. Study area and variables

3.1 Data sets

In this study, an open source data set of three different crime types of South Chicago, between the period of 1st March 2011 and 8th January 2012, is used. These crimes are; *theft-of-motor vehicle* (2,202 records), *residential burglary* (3,405 records), and *assault* (2,978 records). Each data set is divided into a *training set* (1st March 2011 to 30th September 2011) and an *evaluation set* (1st October 2011 to 8th January 2012 [100 days]). Further subsetting of the data sets to derive *T* is described in Subsection 3.2. The data set is downloaded from the official website of City of Chicago: https://data.cityofchicago.org/.

3.2 Creating different training data length (*T*)

In order to create different training data lengths (T), the common end dates of the trainings must first be identified; t_i = 30th Sept. 2011. From t_i measuring backwards in time, 90 days', 150 days' and 210 days' worth of data sets were collected. These are called $T_{initials}$, representing the three different training data sets (*T*) created. Figure 1 describes how each of these values are used separately in the training of SEPP. For example, given $T_{initial}$, the SEPP model is trained and evaluated (i.e. the accuracy calculated) in relation to the next day's crime $[t_{i+1} - t_i]$. Furthermore, instead of using a rolling length of training (i.e. $[t - T, t]$, with t being the present day), the T is allowed to vary incrementally, as the training – evaluation cycles continue (Figure 1).

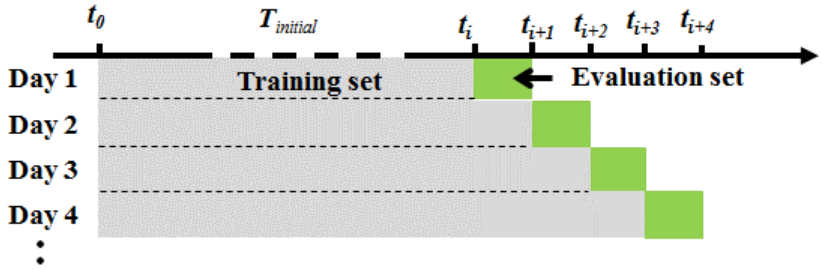

Figure 1: Details of the training and evaluation of the SEPP model, given a value of T (here denoted as $T_{initial}$). On the first day of training, the data length $T = T_{initial}$, and was evaluated in terms of $[t_{i+1} - t_i]$. On the second day of training, the data length $T = T_{initial} + 1\ day$, and was evaluated in terms of $[t_{i+2} - t_{i+1}]$. On the third day of training, the data length $T = T_{initial} + 2days$, and was evaluated in terms of $[t_{i+3} - t_{i+2}]$. This was completed for 100 steps.

In the above figure, the application considered is that of predictive policing, which involves producing hotspot maps consisting of ranked evaluated grid units, on a daily basis. Thus, the increment of *T* and the evaluation data size are based on a one day step.

3.3 Creating different aggregation grid sizes (*M*)

The study area is discretised using four different grid unit systems, as shown in Figure 2. The following is the list of grid systems created, along with the total number of units covering the entire area:

- 50m x 50m grid system (a total of 21,828 units)
- 100m x 100m grid system (a total of 5,593 units)
- 150m x 150m grid system (a total of 2,539 units)
- 250m x 250m grid system (a total of 951 units)

The boundary of the area is then overlaid on the grid to create a new grid system with cropped marginal. During the training, the un-cropped grid system is used, while the cropped grid system is utilised during the evaluation process.

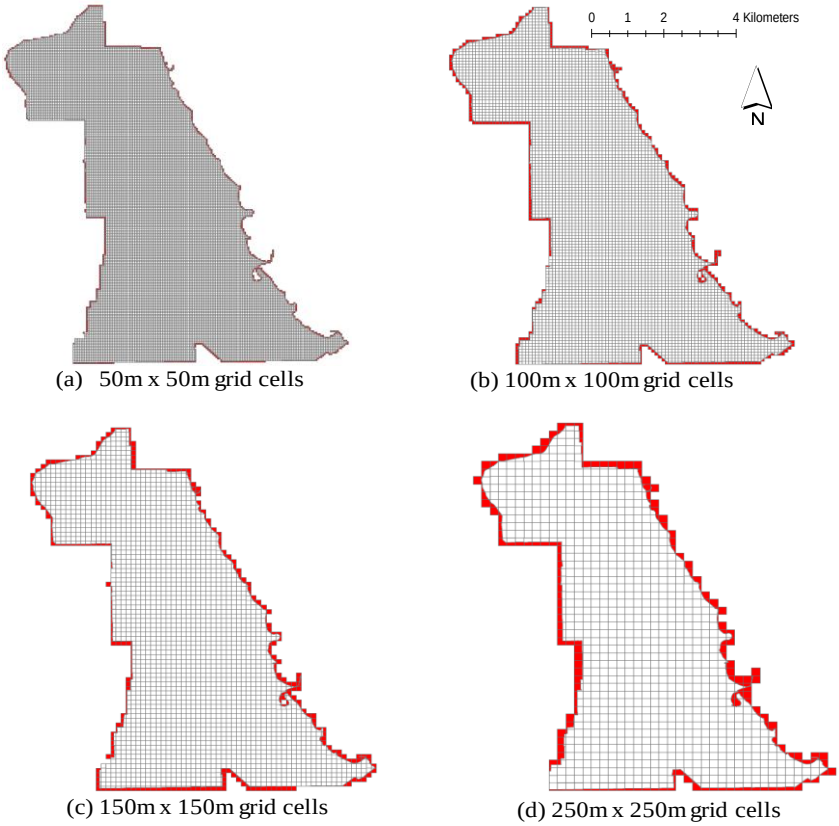

Figure 2: Different grid unit sizes (*M*) created for hotspot generation (study area: South Chicago). The grids outside the boundary are cropped during the evaluation to ensure uniform area coverage.

No grid systems smaller than 50m x 50m were created here, as it would have generated too many cells to utilise in practice; the study area being so large. Furthermore, no grid size beyond 250m x 250m was created since it would be considered too large for capturing local hotspots within an urban space (Adepeju et al., 2016).

3.4 Measuring the accuracy (evaluation)

Having generated a hotspot surface, an empirical accuracy measure can be obtained by first ranking the grid units in descending order of their risk values (summed kernel estimations per unit), and calculating the proportion of crimes (of the next one day) that fall within a chosen hotspot coverage area (Rosser et al., 2016). This accuracy measure is often referred to as 'the hit rate', and is usually expressed in percentages (Bowers et al., 2004). The accuracy measurement may also be conducted at increasing hotspot coverages. The results of this are an accuracy profile; an approach which is used in this study. This approach allows for flexible evaluation in terms of the allocation of police resources.

4. Results and Discussion

4.1 Impacts of training data lengths (*T*) at different aggregation unit sizes (*M*)

Figure 3 shows the mean accuracy of 100 daily predictions by incrementing the hotspot coverage to different amounts. We set 20% as a cut-off point for each plot. In practice, hotspot patrols are generally below 20% coverages for an area as large as South Chicago.

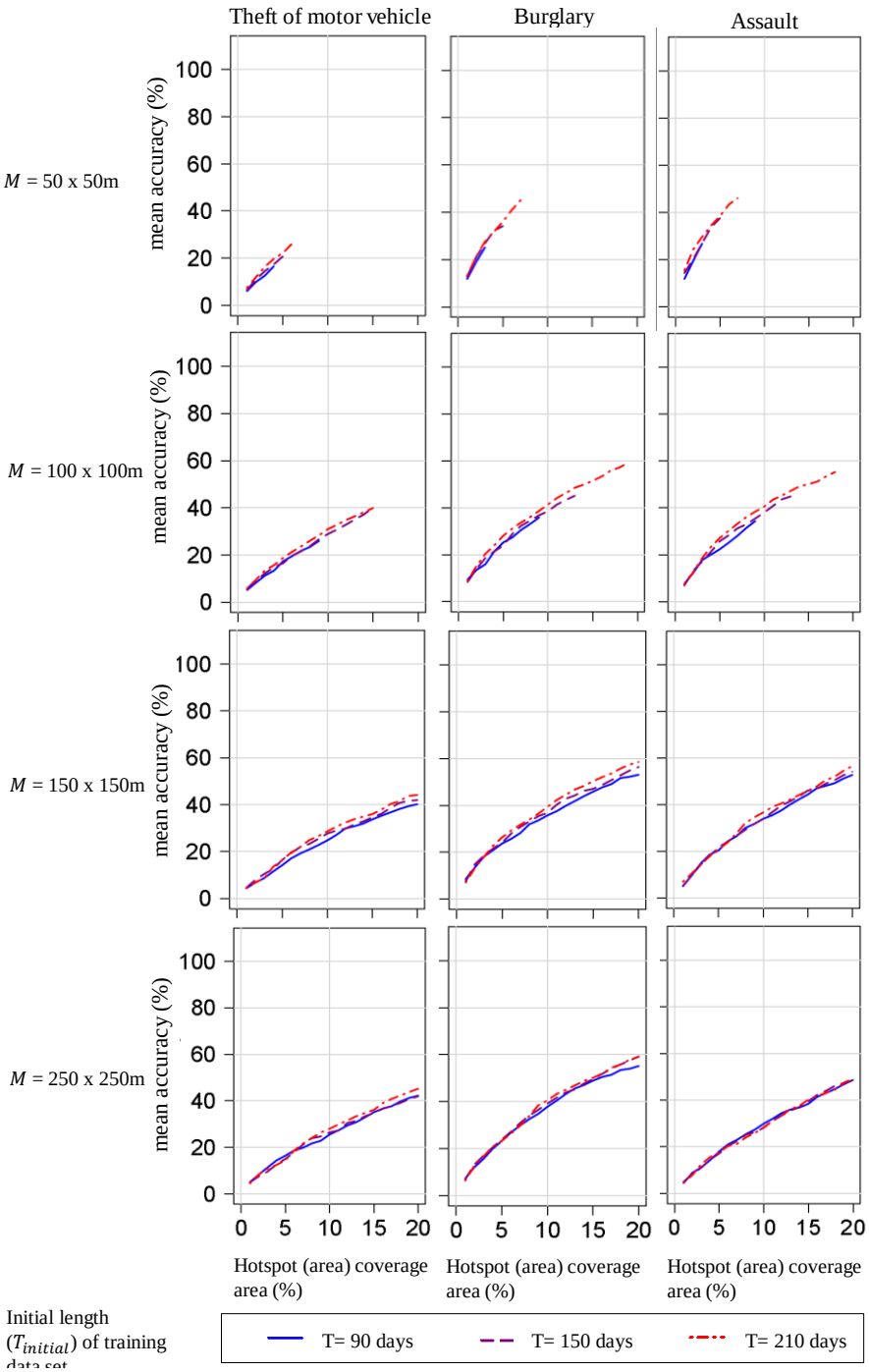

Figure 3: Accuracy profiles of SEPP for three training data set lengths (*T*) evaluated at various levels of aggregation unit sizes (*M*).

Figure 3 demonstrates that the SEPP model produced better accuracy with larger training data sets. This can be seen in more than 80% of cases across all coverage areas, grid sizes and crime types. For example, at T = 210, SEPP marginally outperforms both T = 150 days and T = 90 days in the majority of cases. Furthermore, increased training data sets ensure larger hotspot coverage. This is apparent in the plots at grid sizes of 50m x 50m and 100m x 100m.

4.2 The impact of aggregation unit sizes (*M*) on different training data length (*T*)

Figure 4 shows that as the aggregation size decreases, the accuracy of predictions increases. This implies that small aggregation sizes are able to capture local patterns of repeat and near repeat events more accurately. The highest accuracy level is achieved by the smallest aggregation size (i.e. 50m x 50m) across all the three training data lengths (T) and crime types. However, the use of such small aggregation sizes may not guarantee sufficient coverage for the purposes of practical crime intervention. For example, none of the accuracy profiles generated by 50m x 50m grid sizes cover up to 10%, a coverage level that may be considered reasonable enough for use by most police departments in the majority of urban cities.

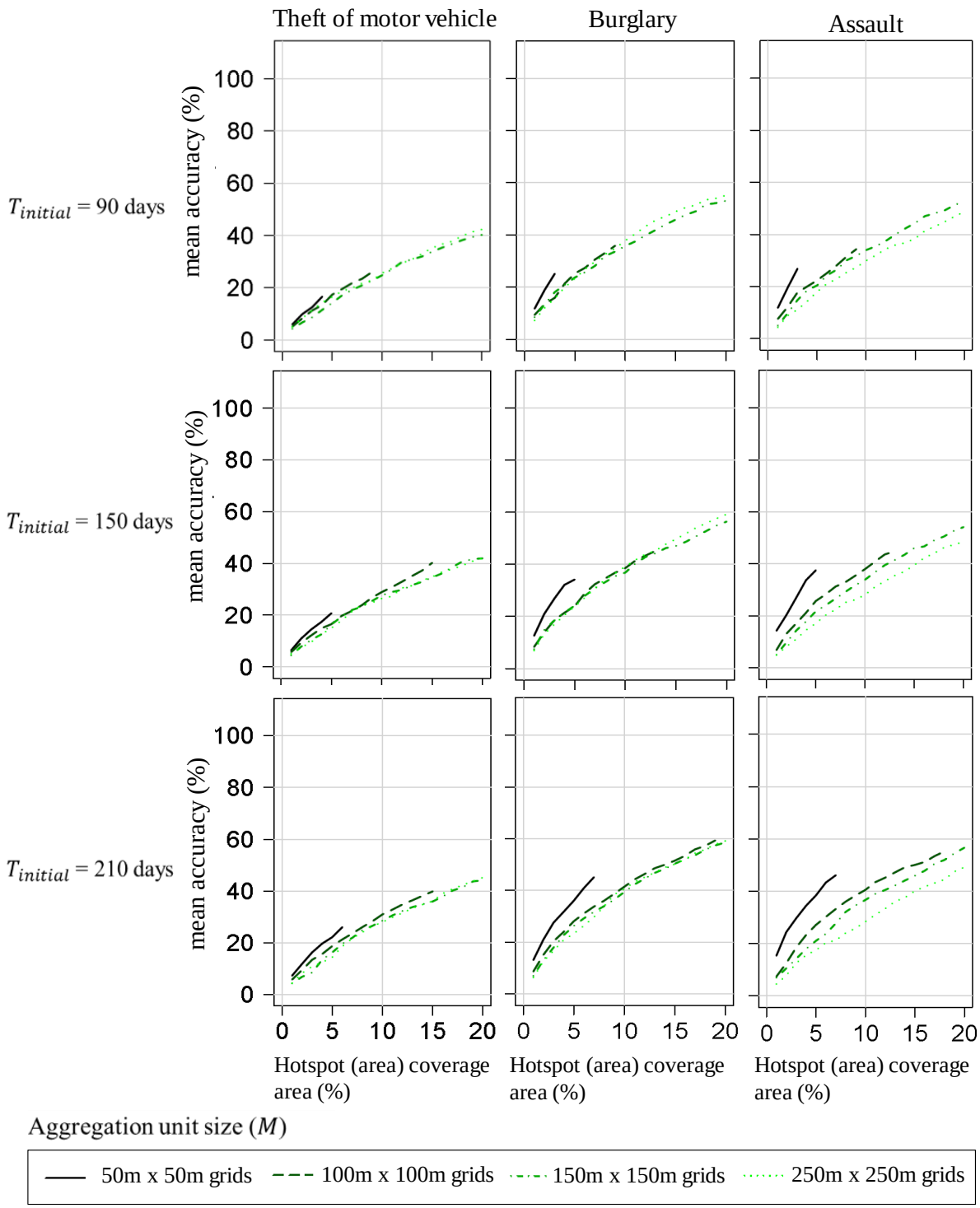

Figure 4: Accuracy profiles of SEPP for different aggregation unit sizes (M), evaluated at various training data set lengths (*T*).

5. Conclusion and Current work

This study has helped with gaining insights into how the training data set lengths (*T*) and aggregation unit sizes (*M*) impact on the accuracy of the SEPP hotspot model. The pattern of accuracy observed using different values of *T* validates earlier arguments by Mohler (2014), who said that a large training data set will allow the background (μ) component of an SEPP algorithm to be modelled more

accurately, thereby ensuring better results. The pattern of accuracy observed using different sizes of aggregation units (*M*) clarifies two important points; first, that MAUP is also a source of bias in the results of the SEPP, and second, that the near-repeat patterns of the specified crimes are highly localised. Therefore, the highly-localised predictions provided by the 50m x 50m grids, despite their lower coverage, are considered the best for crime interventions.

Although, this study only focuses on two variables; the training data length (*T*) and aggregation unit size (*M*); there are several other variables (or factors) that can impact on the performance of the SEPP models, which have not yet been investigated in any of the previous studies. Examples of these variables include the declustering function, which can be used in order to separate the μ and g components of the SEPP model (Equation 1), the spatial and temporal bandwidths, and so on. Future work could thus, investigate these variables to enable optimal model selection.

6. Acknowledgements

This work was funded by the UK Home Office Police Innovation Fund, through the project "More with Less: Authentic Implementation of Evidence-Based Predictive Patrol Plans".

The researchers also wish to acknowledge the contribution of the government of the City of Chicago for making the crime incident data set publicly available.

7. 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, *30*(11), pp.2133-2154. At https://doi.org/10.1080/13658816.2016.1159684.

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

Cheng, T. and Adepeju, M., 2014. *Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection*. PloSOne, *9*(6), p.e100465.

Daley, D., and Vere-Jones, D., 2003. *An Introduction to the Theory of Point Processes*. (2nd ed.), *New York: Springer*.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. and Tita, G. E., 2011. *Self-exciting point process modelling of crime*. *Journal of the American Statistical Association*, *106*(493), pp.100-108.

Mohler, G., 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, *30*(3), pp.491-497.

Openshaw, S., and Taylor, P.J., 1981. *The modifiable areal unit problem*. In: Wrigley N and Bennett RJ (eds.) Quantitative Geography: A British View. London: Routledge and Kegan Paul: pp. 60–70.

Rosser, G., Davies, T., Bowers, K. J., Johnson, S.D. and Cheng, T., 2016. *Predictive Crime Mapping: Arbitrary Grids or Street Networks?.* Journal of Quantitative Criminology, pp.1-26.

Zhuang, J., Ogata, Y. and Vere-Jones, D., 2002. *Stochastic declustering of space-time earthquake occurrences*. Journal of the American Statistical Association, *97*(458), pp.369-380.

8. Biography

Monsuru Adepeju is a Research Fellow at the School of Geography, University of Leeds. His current research interest is in the development and provision of cheap and accessible predictive policing services via web and mobile platforms for various law enforcement bodies worldwide.

Dr. Andy Evans is a Senior Lecture in GIS and Computational Geography. His current research interest is in developing predictive policing methodologies, in particular utilising artificial intelligence.