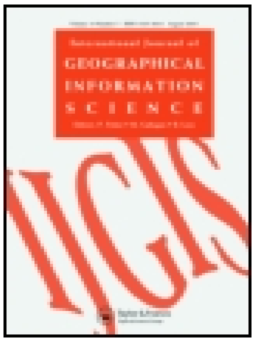
Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study

Article *in* International Journal of Geographical Information Science · April 2016

Impact Factor: 1.66 · DOI: 10.1080/13658816.2016.1159684

Monsuru Adepeju
University College London
7 PUBLICATIONS **6** CITATIONS
SEE PROFILE

T. Cheng
University College London
85 PUBLICATIONS **964** CITATIONS
SEE PROFILE



Taylor & Francis
Taylor & Francis Group

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study

Monsuru Adepeju, Gabriel Rosser & Tao Cheng

To cite this article: Monsuru Adepeju, Gabriel Rosser & Tao Cheng (2016): Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study, International Journal of Geographical Information Science

To link to this article: http://dx.doi.org/10.1080/13658816.2016.1159684

Published online: 07 Apr 2016.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
http://dx.doi.org/10.1080/13658816.2016.1159684

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study

Monsuru Adepeju, Gabriel Rosser and Tao Cheng

SpaceTimeLab, Department of Civil, Environmental & Geomatic Engineering, University College London, London, United Kingdom

ABSTRACT

Many physical and sociological processes are represented as discrete events in time and space. These spatio-temporal point processes are often sparse, meaning that they cannot be aggregated and treated with conventional regression models. Models based on the point process framework may be employed instead for prediction purposes. Evaluating the predictive performance of these models poses a unique challenge, as the same sparseness prevents the use of popular measures such as the root mean squared error. Statistical likelihood is a valid alternative, but this does not measure absolute performance and is therefore difficult for practitioners and researchers to interpret. Motivated by this limitation, we develop a practical toolkit of evaluation metrics for spatio-temporal point process predictions. The metrics are based around the concept of hotspots, which represent areas of high point density. In addition to measuring predictive accuracy, our evaluation toolkit considers broader aspects of predictive performance, including a characterisation of the spatial and temporal distributions of predicted hotspots and a comparison of the complementarity of different prediction methods. We demonstrate the application of our evaluation metrics using a case study of crime prediction, comparing four varied prediction methods using crime data from two different locations and multiple crime types. The results highlight a previously unseen interplay between predictive accuracy and spatio-temporal dispersion of predicted hotspots. The new evaluation framework may be applied to compare multiple prediction methods in a variety of scenarios, yielding valuable new insight into the predictive performance of point process-based prediction.

ARTICLE HISTORY
Received 22 September 2015
Accepted 11 February 2016

KEYWORDS
Point process; hotspot; prediction; space-time; crime

1. Introduction

1.1. *Spatio-temporal point processes and hotspot prediction*

Many physical and sociological processes of interest take the form of events that occur at discrete points in space and time, including crime (Mohler *et al*. 2011), earthquakes (Zhuang *et al*. 2002) and infrastructure failures (Ertekin *et al*. 2015). These are referred to as spatio-temporal point processes (henceforth denoted STPPs). Unlike processes that

CONTACT Tao Cheng ✉ tao.cheng@ucl.ac.uk

can be measured at regular time intervals and fixed spatial locations such as road traffic (Haworth *et al*. 2014) and ocean surface temperature (Rayner 2003), STPPs are only observed through the monitoring and reporting of incidents. Modelling STPPs is therefore challenging; whereas many statistical models (such as ARIMA or STARIMA) have been developed to analyse discrete space-time series (Cheng *et al*. 2014), such approaches are only applicable to point data if they are first aggregated into similar discrete space-time cubes to give event counts. However, many STPPs of interest are sparse, meaning that counting events over space and time intervals of interest results in mainly zero counts. For example, in the case of crimes, the majority of urban streets witness no crime at all on any given day. Conventional statistical modelling approaches are therefore unsuitable for these sparse STPPs.

Fortunately, a well-developed mathematical framework exists to describe STPPs (Daley and Vere-Jones 2006). According to this, events (points) arise from an underlying but unobservable intensity function whose value varies in time and space. The result is a point cloud, with each point representing a single event. Models based on the STPP generally attempt to describe the underlying intensity function (or a statistical summary thereof) based on the observed point data. When the events have other attributes ascribed to them, for example the Richter scale severity of an earthquake, the process becomes a marked STPP (Mohler 2014). In this study, we are concerned with the unmarked STPP.

The STPP framework provides a unifying model of processes as diverse as the distribution of cells within a retina (Diggle 1986), sightings of rats in cities (Tamayo-Uria *et al*. 2014) and the outbreak of violence (O'Loughlin *et al*. 2010). Models based on the STPP may be used for both retrospective analysis and forecasting. For example, retrospective analysis of seismic data based on an STPP model has been used to analyse the emergence of correlated clusters of earthquake activity in historic datasets (Zhuang *et al*. 2002). Predictions are generated by evaluating the intensity function of STPP-based models at future times. A common aim of a predictive method is to identify regions that are expected to have a high associated point intensity in the future, such as an area with a high crime count or a large population of infected individuals. We use the general term 'hotspots' to describe such regions. Predicted hotspots are useful in practice for planning proactive interventions or responses to future events, as their interpretation is generally straightforward in all application areas. Key examples include anticipating the occurrence of earthquakes (Marzocchi *et al*. 2012) and the focus of this study, namely the risk of crime (Mohler *et al*. 2011, Perry *et al*. 2013).

1.2. *Challenges and the state-of-the-art in evaluating STPP-based predictions*

As with any other prediction method, forecasts generated using STPP-based models must be evaluated with reference to a ground truth to ensure that they are reliable. This is typically achieved with historic datasets, by making predictions for periods of time for which real data have been collected and comparing the predicted outcome with experimental observations. This conceptually simple evaluation process is, however, challenging in practice. On one hand, likelihood-based approaches have been applied to compare earthquake prediction models (Marzocchi *et al*. 2012). This is an important statistical tool for model selection, however likelihood values cannot be used to measure absolute predictive performance in any interpretable fashion. On the other hand, just as conventional statistical prediction approaches cannot be used to model sparse STPP data, the methods typically used to evaluate their performance, such as the root mean square error (RMSE), cannot be applied to STPP-based predictions for the same reason. Measures such as RMSE are designed to evaluate the mean performance over all space and time and are therefore poorly suited to sparse STPP data. The large number of zero counts that occur when a sparse STPP is aggregated would strongly influence the RMSE and reduce its sensitivity to predictive performance in the more relevant hotspot regions.

Perhaps as a result of the challenges involved with evaluating predictions made against a STPP, the process of evaluating them has received little attention relative to the large number of predictive methods developed. In the case of crime prediction, there is little consensus in academic circles on how best to assess and compare a new method (Bowers *et al.* 2004) and systematic evaluation is virtually absent in operational environments (Perry *et al.* 2013). A recent review of prediction methods for the spread of infectious disease also ignores the issue of evaluation (Woolhouse 2011). Where STPP-based prediction methods are evaluated, the majority of studies in criminology (Mohler *et al.* 2011), epidemiology (Kulldorff *et al.* 2005) and geophysics (Vere-Jones 1995) focus on measures related to accuracy, including concepts such as true positive and false discovery rates.

When computing predictive accuracy scores, the observed performance may vary a great deal from one time window to the next due to the inherent stochasticity of the data. It is therefore common practice to use the mean value of the aforementioned measures, obtained by aggregating multiple consecutive prediction time windows (Mohler *et al.* 2011). However, comparing mean values amongst different predictive methods is insufficient to infer that one method outperforms another, as the process of taking an average hides the inherent variability in the results.

1.3. *Aims and objectives*

The difficulties associated with evaluating predictions against STPP data hamper the development and operational uptake of new methods in many fields. Whilst the prediction accuracy is an important measure of predictive performance, several other aspects should also be considered to gain fuller insight of prediction methods. We address this shortcoming by developing a more comprehensive set of evaluation metrics, which are described in detail below. This is the primary motivation and contribution of this study. The second motivation of this study is the development of a novel method that compares the outcomes of different STPP methods over all time windows, permitting a rigorous statistical comparison of prediction accuracy. This is an important advance beyond the comparison of mean predictive accuracy values, which cannot indicate statistical significance levels.

We focus throughout on the comparison of STPP models for hotspot prediction. In addition to their ease of interpretation, hotspots are also useful as a methodological tool, as they represent spatial subregions on which we may focus our evaluation efforts. By predicting the location of high-intensity regions, hotspots represent a natural choice of spatial domain that can be used to assess predictive performance.

In the following sections, we will present a practical toolkit of robust and interpretable evaluation metrics, which allows the full and systematic assessment of STPP predictive methods. We then proceed to apply this toolkit, using crime prediction as our case study. Four varied crime prediction methods are considered using crime datasets from the London Borough of Camden, UK and the City of Chicago, Illinois, USA. This case study provides a key example of how our metrics should be applied routinely to new predictive methods, providing a comprehensive assessment of their performance in line with the strongest existing techniques. Finally, we discuss the implications of our findings in terms of predictive hotspot policing and in the wider context of STPP prediction.

2. Hotspot predictive methods

In this study, we consider four methods for generating a predictive hotspot map, three of which are successful existing approaches: (1) the prospective hotspot (PHotspot) method (Bowers *et al.* 2004), (2) the self-exciting point process (SEPP) model (Mohler *et al.* 2011) and (3) a prospective kernel density estimate (PKDE) (Chainey *et al.* 2008). We provide brief details of the implementation of these methods in the following section. These and many other contemporary methods are discussed by Perry et al. in their report (Perry *et al.* 2013). For the fourth method, the space-time scan statistic (STSS) (Kulldorff *et al.* 2005) is innovatively modified to generate hotspots for crime prediction. We also consider a method for randomly generating hotspot maps that is useful as a null hypothesis for comparison with the other four methods. We now briefly explain how these methods are implemented.

2.1. *Prospective hotspot (PHotspot)*

The PHotspot method is based on the widespread observation of the near-repeat victimisation phenomenon in burglary data (Bowers *et al.* 2004), in which properties nearby the site of a recent burglary are at heightened risk of themselves being targeted for some ensuing time period. Bowers et al. quantify the spatial and temporal extent of the communicable risk, using this to inform their method (Bowers *et al.* 2004). PHotspot is similar in philosophy to a spatio-temporal KDE (STKDE), in which overlapping kernels are combined additively to compute the risk at a given time and location. However, unlike the STKDE, the risk function is an atypical form and the bandwidths are set a priori using empirical data analysis. This leads to spatial and temporal bandwidths of 400 metres and 2 months, respectively (Bowers *et al.* 2004, Johnson and Bowers 2004).

The PHotspot method has previously only been applied to burglary crime data. One of the contributions of this work is determining how effective the method is in the context of different crime types. We use the same bandwidths in all cases, since there have been no studies suggesting alternative values. It is possible to derive optimal bandwidths directly from the crime data using maximum likelihood methods, however this is beyond the scope of the current study and the subject of ongoing work.

2.2. *Self-exciting point process (SEPP)*

The SEPP is a modelling framework that originated in the field of seismology where it was used to describe the pattern of aftershocks in earthquakes. Mohler et al. applied this model to crime prediction, the essence of which is that crimes occur due to a spatio-temporally heterogeneous risk background and that crime events may trigger further crimes in their spatial and temporal neighbourhood. The theoretical and computational details of the SEPP are all found in Mohler *et al.* (2011), including a discussion of the criminological basis for the method. Following the authors' suggestion, we apply an upper limit on the temporal and spatial extent of the triggering process: crimes may trigger subsequent events for a time window of up to 60 days and over a distance of 300 m. The algorithm proceeds by repeatedly performing density estimates using a variable bandwidth KDE. We found that it was necessary to impose minimum bandwidths, since overly narrow values are otherwise selected, preventing the algorithm from converging. We selected a value of 30 m and 0.5 days for this purpose. This requirement is probably related to the geocoding process; a large proportion of the crimes are snapped to either a grid square centroid or the midpoint of a street segment.

2.3. *Prospective kernel density estimate (PKDE)*

Spatial KDEs (SKDEs) are ubiquitous in the analysis of crime data (Chainey *et al.* 2008, Malik *et al.* 2014) and offer a generic means of estimating an underlying spatial distribution from a data sample. In the context of crime mapping, the popularity of the SKDE is in part due to the smoothly varying heat maps that it generates (Eck *et al.* 2005, Chainey *et al.* 2008). The SKDE requires that data be aggregated over some time window prior to the latest datum; the temporal element within this window is then discarded and all data points are treated equivalently. There are, however, no general guidelines for selecting the size of the time window.

We define the PKDE method as follows. Crime data are aggregated over a rolling 60 day time window and the SKDE is computed using 'sum of asymptotic mean squared error' (SAMSE) plug-in bandwidth selection, implemented in the R package *ks* (Duong and Hazelton 2005). Predictions are then made for a one day period as described. The PKDE is the only purely spatial method of the four considered, although the use of a rolling time window means that the generated hotspots nonetheless change gradually over time. This method is an important inclusion as it is expected to generate more stable hotspot predictions compared with the other spatio-temporal methods.

2.4. *Prospective space-time scan statistic (PSTSS)*

The STSS is a surveillance method primarily used in the field of epidemiology for the detection of clusters of diseases (Kulldorff *et al.* 2005). It has been widely applied in several other fields such as forestry (Tuia *et al.* 2008), ecology (Tonini *et al.* 2009) and criminology (Nakaya and Yano 2010, Cheng and Adepeju 2014). Most applications of the STSS in criminology have been retrospective in nature, the goal being to detect significant historical crime hotspots. The exception is a recently developed prospective scanning method that operates on a road network (Shiode and Shiode 2014). Crime prediction on networks is a relatively new development and beyond the scope of this study.

To our knowledge, the STSS framework has not previously been applied to the grid-based predictive hotspots considered here. This has prevented any comparisons

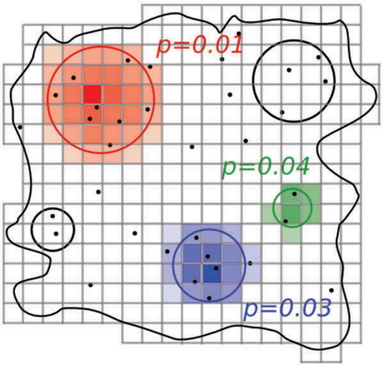


Figure 1. Producing ranked grid predictions from the PSTSS. Circular clusters are selected in ascending *p*-value order (red, blue, green). In each cluster, grid squares near the centre (darker shades) are ranked above those nearer the perimeter (lighter shades). The process is halted partway through filling the green cluster as the desired coverage has been achieved.

between the predictive hotspot methods discussed above and STSS. We now show how the STSS may be modified to make it suitable for this purpose, providing predictions for future events rather than retrospective classification. We name this new approach the PSTSS.

Clusters identified in historical datasets by the STSS may cover any past time period and may disappear before the most recent time point in the dataset. Long-departed hotspots are of little use in short-term proactive policing, therefore we assume that only clusters that survive up until the end time of the dataset are relevant. The purely spatial representation of surviving clusters at time t_n is used as the basis of a prediction relating to the time period starting at t_n and lasting 24 hours.

The circular surfaces of surviving clusters represent hotspots where crimes are likely to occur in future, quantified by the significance level (*p*-value). When ranked by decreasing significance, they have most of the characteristics of the other hotspot methods described above, with the exception of the shape of the hotspots, which are circular and therefore not directly comparable. We therefore develop a method of converting the circular regions of the PSTSS into grid-based hotspots based on the assumption that the risk intensity in a region decreases upon moving from the centre towards the perimeter. Overlaying a regular grid on the circularly clustered surface, we rank all grid squares that intersect a cluster, first by the *p*-value of the overlapping cluster, then by distance from the cluster centre. Figure 1 illustrates this procedure.

In the original implementation of STSS (Kulldorff *et al.* 2005), a minimum *p*-value threshold is applied to the reported clusters, to render the output more manageable. As PSTSS returns only those clusters that persist until the current day, no such threshold is necessary and we simply consider all clusters.

2.5. *Predictive selective random algorithm (PSRA)*

It is often valuable to compare predictive methods with a method that generates hotspots purely at random. This baseline method acts as a null hypothesis, allowing us to test whether the other methods show significantly improvements over a random algorithm. In specifying this algorithm, we wish to emulate a naïve hotspot generation process. The simplest possible choice is to pick grid squares with equal probability until the desired coverage has been reached. However, this is unfairly simplistic as many of the grid squares cover areas that cannot possibly contain certain crime events. For example, regions containing no retail outlets cannot experience shoplifting, and residential burglaries cannot happen from within a park.

We therefore propose the PSRA, in which hotspots are drawn with equal probability from the pool of grid squares that have ever received at least one crime of the type being considered. Land use data might provide a more accurate measure of whether crimes can occur in different regions, however for this study we use the presence or absence of crime records as a straightforward proxy for land use to avoid additional complications. This may result in the removal of some grid squares where the absence of crime is due to low victimisation rates rather than land use. The net effect is to make the PRSA baseline slightly more conservative and has no significant effect on our results.

3. A framework for assessing hotspot predictive methods

Here we will develop a new set of metrics to evaluate four varied aspects of STPP-based hotspot prediction. The first, accuracy, measures the number of crimes that fall within predicted hotspots during the time window of interest. The second, compactness, describes the spatial arrangement of the hotspots in terms of the level of dispersion. This is relevant to planning interventions or proactive responses. Thirdly, the dynamic variability of a predictive method describes the extent to which the predicted hotspots change on consecutive time windows. Finally, the complementarity between multiple methods indicates the numbers of crimes uniquely detected by a single method.

Throughout this section, we denote the total area of the region of interest by A, and the area covered by predicted hotspots by a. These hotspots represent a prediction for the locations of crimes in some future time window, w_t. We use $w_t = 1$ day in our case study. Other studies consider different time windows of one month (Gorr *et al.* 2003), one week and two days (Bowers *et al.* 2004). In this study, we are most concerned with the short timescale required by patrolling police officers, who are, in the case of Camden, assigned tasks on a daily basis. We expect that this practice is unlikely to differ significantly between modern urban police forces. However, all of the results presented here are theoretically applicable to any forecasting time window.

Evaluation metrics are computed by specifying the area coverage a/A, and assessing the resulting hotspot map (Bowers *et al.* 2004). Alternatively, some studies vary the coverage as the independent variable, plotting the value of a given metric against coverage to determine the relationship (Mohler *et al.* 2011). In this sense, the coverage is treated as a discrimination threshold in a binary classification task and the output resembles a receiver operating characteristic (ROC) curve (Brown and Davis 2006), though the interpretation is not identical. We use a mixture of these approaches in the current study. A variable coverage level provides a more detailed evaluation, but it is more practical to consider a fixed coverage in order to simplify the comparison process between methods. Where a fixed coverage is required, we choose 20%, following

Bowers et al. (Bowers *et al.* 2004). We later show evidence that suggests that this value has little effect on our conclusions when varied by around 10% in either direction.

For operational and comparison purposes, we define a hotspot as one or more subregions on a regular spatial grid defined on the domain of interest. The length of the grid square side may be defined by the user according to the level of spatial granularity desired in the predictions. We use a side length of 250 m throughout this study, but other authors have considered predictions at finer scales (Bowers *et al.* 2004, Johnson *et al.* 2007). The grid size is chosen to reflect a reasonable patrolling area for police, in addition to managing computational demands. In various conversations with police officers in the London Borough of Camden, this cell size was deemed sufficient. Further experiments (data not shown) revealed little variation in our results when a grid with length 150 m is used.

For a given time window, the grid squares are ranked by predicted intensity, following several preceding studies (Bowers *et al.* 2004, Mohler *et al.* 2011). To obtain hotspots, grid squares are selected in decreasing intensity order until a desired coverage area is obtained. By defining hotspots using a fixed coverage, we achieve a standardised scale on which to compare different methods, assuming a constant level of available police resources. Alternative methods include defining a threshold intensity value and labelling all regions with intensity above the threshold as hotspots (Roerdink and Meijster 2000), using a scanning algorithm (Patil *et al.* 2010), or selecting regions based on a statistical test for significance (Brimicombe 2012). These approaches emphasise the varying level of demand for police presence over space and time. Whilst it would be interesting to assess prediction methods using a variable resource level, it is beyond the scope of the present study.

3.1. *Accuracy and statistical significance*

In the case of accuracy metrics, we denote the total number of crimes in the prediction time window by N, and the number of crimes captured in the hotspots by n. Since N may be small for any given prediction time window, all performance metrics are subject to significant random variation between different time periods. We therefore evaluate prediction methods using many consecutive time windows to reduce the variance in our estimate of the metrics (Mohler *et al.* 2011).

We consider two different measures of predictive accuracy. The hit rate is the proportion of crimes accurately captured by the defined hotspot area, n/N (Bowers *et al.* 2004). This quantity varies with a and is only meaningful at attainable coverage levels. A large hotspot may yield a high hit rate, but the associated coverage could be so large that it is useless for effective police deployment. To mitigate this drawback, Chainey et al. (Chainey *et al.* 2008) propose a measure called the predictive accuracy index (PAI), given by the ratio of hit rate to area coverage,

$$\mathrm{PAI} = \left(\frac{n}{N}\right) / \left(\frac{a}{A}\right). \tag{1}$$

The PAI is interpreted as the ratio of crime density in predicted hotspots to the crime density over the whole region. There is no general consensus as to the most appropriate measure of predictive accuracy in crime prediction.

Having computed measures of predictive accuracy as described above, we wish to compare different prediction methods to determine which has the strongest performance. A straightforward option is to compare mean values directly, however the statistical significance of any such result is unknown. Rather than pooling all prediction accuracy measures to compute a mean value, we therefore consider the values as a time series, with a regular time interval $w_t = 1$ between observations. Comparing the predictive accuracy of two different predictive methods for a given coverage level is then equivalent to comparing two paired time series. This is complicated by the possibility of temporal autocorrelation in the values, due to the fact that crime levels are themselves correlated in time (Johnson and Bowers 2004). We therefore make the simplifying assumption that the difference in predictive accuracy between two methods is independent of the underlying crime rate.

Under this assumption, the time series of differences comprises independent and identically distributed (iid) random variables. The comparison of the two methods may then be achieved through standard statistical tests for paired observations. In this study we use Wilcoxon's signed-rank test (WSR), as it is well established and demonstrates good statistical power (Diebold and Mariano 1995). WSR is a distribution-free test that is used to compare two related samples by assessing whether their mean population ranks differ. The test statistic is given by

$$W = \sum_{i=1}^{N} \left[\mathrm{sgn}\left(y_{2,i} - y_{1,i} \right) \cdot R_i \right], \qquad (2)$$

where N is the sample size, $y_{1,i}$ denotes sample i from prediction method 1, R_i is the rank of the difference and sgn is the sign function. By convention, $\mathrm{sgn}(0) = 0$, meaning that exact matches are not included. Having computed W, the statistical significance value can be obtained using a lookup table. We use a single-tailed lookup, which effectively tests whether one method is significantly better than another. The test is implemented in the R package *stats*. In this study, we consider four prediction methods, so six pairwise comparisons are needed to test all combinations. We apply a Bonferroni correction to avoid false positives.

3.2. *Compactness*

Measures of prediction accuracy do not reflect the spatial distribution of hotspots. Besides accurate predictions, police establishments and medical services are interested in whether the identified hotspots can easily be addressed with the available resources. In the context of crime, officers must be capable of physically patrolling hotspots (Bowers *et al.* 2004), while in the field of epidemiology quarantine regions must be arranged optimally to slow or prevent disease transmission (Riley *et al.* 2014). We refer to this property as hotspot compactness. The notion of compactness reflects the intuitive idea that more compact and connected hotspots are easier to patrol and therefore considered more effective from a policing standpoint. We note that compactness is not the primary objective of the prediction methods considered here as it does not feature explicitly in their implementation. Nonetheless, its inclusion is warranted because it is an important operational consideration and a useful characterisation metric.

The area:perimeter (AP) ratio is a well-established means of measuring compactness. A significant drawback of this metric is that it has units of length, making it dependent

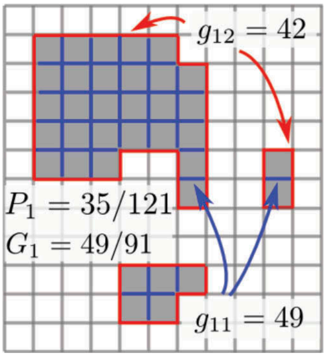


Figure 2. Computing the clumpiness index (CI). Grey regions indicate hotspots, red lines denote cross-class edges, blue lines denote within-class edges. The values shown here give a CI of 0.35, corresponding to mild levels of aggregation.

on the scale of the region and grid squares and preventing comparisons across different geographic regions. Furthermore, the value of the AP ratio is theoretically unbounded, meaning its absolute value is hard to interpret.

In light of these drawbacks, we adapt an alternative measure of compactness called the clumpiness index (CI) (Turner 1989). This metric is included in the spatial analysis package FRAGSTATS (McGarigal *et al.* 2012). The CI gives the deviation of the proportion of same-class adjacencies (i.e. grid square edges shared between like classes) from a complete spatial randomness model. Intuitively, hotspots that are aggregated into a few large clusters of grid cells will have many same-class adjacencies, resulting in a large value of the CI. To compute the CI, we classify grid squares into two classes, (1) hotspots and (2) cold spots. We compute the total number of shared edges between similar and different classes, denoted $g_{i,j}$, where $i, j \in \{1, 2\}$ are class label indicators. The CI of the hotspot class is then given by

$$\mathrm{CI} = \begin{cases} \frac{G_1 - P_1}{1 - P_1}, G_1 < P_1; P_1 \geq 0.5, \mathrm{or} G_1 \geq P_1 \\ \frac{G_1 - P_1}{P_1}, \mathrm{otherwise}, \end{cases} \tag{3}$$

where $G_i = g_{1,1}/(g_{1,1} + g_{1,2})$ and P_1 is the proportion of the total area occupied by the hotspot class. An example is shown in Figure 2. The CI is bounded between −1 (maximal disaggregation corresponding to a checkerboard arrangement) and 1 (maximal aggregation, corresponding to complete coverage by one class).

Use of the CI avoids the drawbacks of the AP ratio. It is independent of the spatial scale of the domain, and has well-defined baseline values for randomly distributed, maximally aggregated and maximally disaggregated hotspots. This makes it easier not only to compare different hotspot maps to one another but also to interpret the results relative to a common baseline.

3.3. *Dynamic variability*

Another major distinguishing feature between predictive methods is the level of variation in the spatial distribution of hotspots between consecutive forecasting time

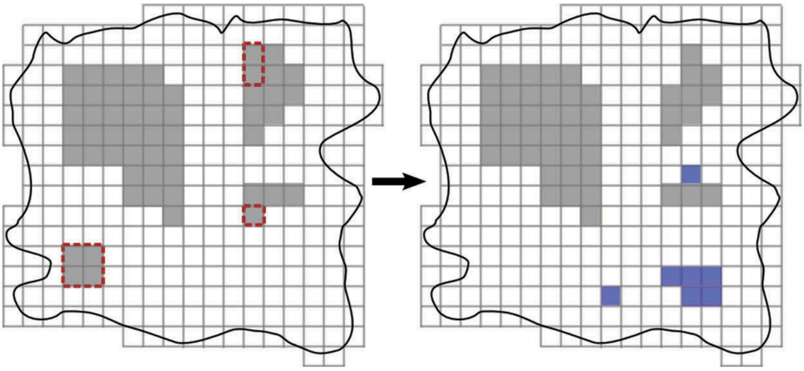


Figure 3. The components of the dynamic variability index for consecutive days t_n and t_{n+1}. Grey indicates repeated hotspots, blue indicates emerging hotspots and red indicates disappearing hotspots.

windows. Intuitively, we expect that some methods will produce very different hotspot maps between any two days, while others will be relatively unchanging in their predictions. We introduce a measure termed the dynamic variability index (DVI) to describe this phenomenon. Figure 3 illustrates the process of computing the DVI. Considering the predictive hotspot maps generated on two consecutive days, t_n and t_{n+1}, we define three classes of hotspot areas: repeating hotspots are those that appear in both hotspot maps, emerging hotspots are those that are present on day t_{n+1} but not t_n, and disappearing hotspots are present on day t_n but not t_{n+1}. The total number of hotspots in each category is denoted, R, E and D, respectively. Note that since the coverage proportion is fixed for the hotspots, $D = E$ by symmetry, and $E + R = 20\%$ of the total area. We now define the DVI as the proportion of all hotspot regions that are emerging,

$$\mathrm{DVI} = E/(E + R). \tag{4}$$

As with the previously described metrics, we repeatedly compute the DVI for every consecutive pair of days in the case study.

3.4. *Complementarity*

Whilst metrics are highly informative summary statistics, they give little or no insight into the underlying differences between predictive methods. In reality, whilst one method may have a higher accuracy (or compactness) score than another, they may provide complementary information by identifying spatially distinct hotspots and therefore detecting different crimes. In the field of machine learning, ensemble prediction models – in which the outputs of multiple prediction models are combined – have repeatedly shown better performance than their constituent elements (Opitz and Maclin 1999). We expect that in future a similar approach will lead to improved STPP-based ensemble predictions, however this requires detailed investigation into the degree of overlap between the constituent methods. We denote this property predictive complementarity, which quantifies the extent to which different methods detect the same crimes in a dataset. To our knowledge, no studies to date have considered this property for STPP data.

We visualise the complementarity between members of a group of STPP-based predictive methods using a Venn diagram. This provides a graphical tool that is easy to interpret. Key phenomena of interest are the number of unique detections, that is where a single method in the group detects a crime, and significant overlap between all methods.

4. Case study

We now consider crime datasets from two case studies, namely the London Borough of Camden, London, UK, and South-side, Chicago, IL (henceforth referred to as Camden and South Chicago, respectively). We consider three crime types that together make up a large proportion of all records in each dataset. In both cases, we include burglary and violent assault crimes in our analysis. The similarity between matching crime types is approximate, in the sense that different police departments may apply different definitions when classifying crime. We elected to use matching crimes for consistency; a precise equivalence is not required for our conclusions to hold. We also consider shoplifting crimes in Camden and theft of vehicle crimes in South Chicago. The date range is the same in both cases: 1st March 2011 to 6th January 2012, which is the full extent of the Camden dataset. As discussed, crimes are aggregated daily and predictions are made one day ahead ($w_t = 1$). We evaluate each prediction method on the final 100 days of the dataset (28 September 2011–6 January 2012).

Since the focus of this study is methodological, we have avoided consideration of how the accuracy and success rate of geocoding each crime may affect our results. This is an important issue that may have significant effects on the inferred hotspots (Brimicombe *et al*. 2007), however a detailed investigation lies outside of the scope of this study.

Figure 4 shows the spatial distribution of crimes in the two case study regions. The length of each grid square side is 250 m; this is the grid used to define hotspots. Shoplifting crime in Camden is highly concentrated in a few regions near commercial areas. In comparison, South Chicago has fewer areas with no crime; the majority of crime-free squares lie over parks or transport infrastructure (not shown). Burglaries are dispersed in Camden and more confined in South Chicago, while the opposite trend is apparent for violent crimes.

A full summary of the evaluation metrics, recorded at a fixed hotspot coverage of 20%, is given in Tables 1 and 2 for Camden and South Chicago, respectively. The mean and standard deviations are calculated by aggregating over the 100 prediction results.

4.1. *Predictive accuracy*

Figure 5 shows the mean hit rate of each method against coverage, averaged over the 100 days of predictions. The results differ significantly by region; the observed accuracy is more variable in Camden, both by crime type and prediction method. The mean hit rates are more similar in South Chicago, though PSTSS has relatively lower accuracy compared with the other methods. The performance of PSTSS relies on providing sufficient historical data to estimate the background parameters. South Chicago has an area approximately 2.4 times larger than Camden, but a similar number of violent

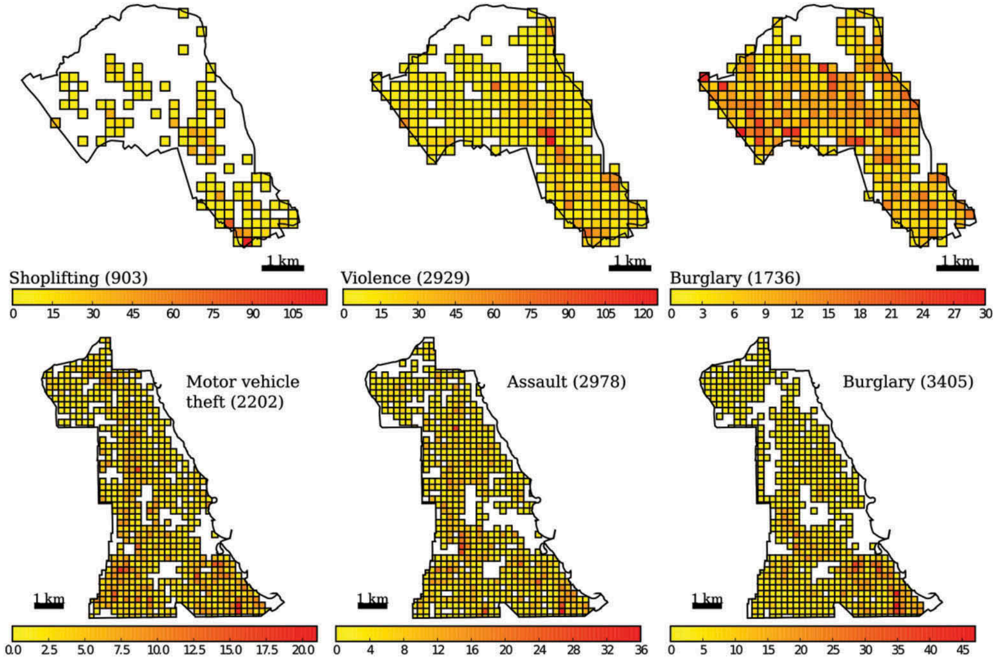


Figure 4. Spatial distribution of crimes in Camden (top) and South Chicago (bottom), aggregated over the study period. Numbers in brackets show the total number of crime records in the dataset. White regions indicate that no crimes occurred in that area.

Table 1. Evaluation metrics for Camden at 20% coverage level.

		Accuracy				Hotspot compactness				Variability	
		Hit rate		PAI		AP ratio		CI		DVI	
Crime Type	Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Shoplift	PSTSS	81.3	27.6	4.07	1.38	197.5	15.9	0.42	0.04	14.9	11.0
	PKDE	74.3	29.8	3.72	1.49	**264.0**	26.8	**0.55**	0.04	2.7	1.7
	SEPP	**91.5**	20.1	**4.58**	1.01	160.8	9.6	0.31	0.03	6.0	2.1
	PHotspot	85.1	27.0	4.26	1.35	177.0	13.4	0.37	0.04	**19.2**	9.2
	PSRA	81.6	27.3	4.08	1.37	95.5	3.6	0.02	0.04	23.4	1.9
Violence	PSTSS	46.5	20.0	2.33	1.00	220.2	20.5	0.46	0.04	10.8	8.0
	PKDE	51.7	19.5	2.59	0.98	**274.9**	25.2	**0.54**	0.04	2.6	3.9
	SEPP	**59.7**	19.8	**2.99**	0.99	113.0	6.1	0.12	0.03	4.5	1.6
	PHotspot	52.2	19.9	2.61	1.00	163.6	15.1	0.32	0.05	**21.1**	6.7
	PSRA	26.3	18.5	1.32	0.93	82.3	3.2	−0.30	0.10	71.0	4.1
Burglary	PSTSS	34.4	22.0	1.72	1.10	243.5	30.5	**0.51**	0.05	3.7	5.6
	PKDE	38.8	24.2	1.94	1.21	**252.5**	41.7	0.50	0.07	2.3	1.9
	SEPP	**47.4**	26.3	**2.37**	1.32	97.9	5.3	0.02	0.05	1.4	1.2
	PHotspot	34.9	23.0	1.75	1.15	160.4	19.2	0.30	0.06	**5.3**	4.4
	PSRA	23.9	21.8	1.20	1.09	81.0	3.1	−0.34	0.10	72.2	4.2

Bold indicates the greatest mean value, excluding the PSRA. SD denotes standard deviation.

crime records. The density of burglaries is similar between the regions, but many of the crimes are concentrated on a few small areas in South Chicago. These two factors may lead to a poor background estimate and low accuracy in PSTSS.

Table 2. Evaluation metrics for South Chicago at 20% coverage level.

		Accuracy				Hotspot compactness				Variability	
		Hit rate		PAI		AP ratio		CI		DVI	
Crime type	Method	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Motor vehicle theft	PSTSS	32.0	21.1	1.59	1.05	235.3	22.3	0.52	0.04	14.0	7.70
	PKDE	39.1	20.5	1.96	1.03	**409.8**	22.4	**0.76**	0.02	4.34	2.00
	SEPP	**39.8**	20.0	**1.99**	1.00	240.0	0	0.54	0	0	0
	PHotspot	37.8	21.6	1.89	1.08	158.8	9.90	0.33	0.03	**14.0**	5.22
	PSRA	29.8	19.6	1.49	0.98	81.72	1.96	−0.31	0.06	70.4	2.88
Assault	PSTSS	29.5	17.6	1.48	0.88	238.2	17.7	0.53	0.04	14.2	7.11
	PKDE	37.0	17.7	1.85	0.88	**357.6**	16.8	**0.72**	0.02	3.87	1.42
	SEPP	**42.2**	19.7	**2.11**	0.99	136.0	2.71	0.24	0.01	0.84	0.57
	PHotspot	39.7	20.3	1.99	1.02	156.5	7.90	0.31	0.03	**18.4**	5.30
	PSRA	30.5	16.8	1.52	0.84	81.9	1.85	−0.30	0.06	69.4	2.79
Burglary	PSTSS	36.4	18.2	1.82	0.91	226.0	14.3	0.51	0.03	12.60	7.80
	PKDE	48.8	17.9	2.44	0.89	**402.0**	25.8	**0.75**	0.02	2.57	1.16
	SEPP	**51.5**	18.5	**2.57**	0.93	156.0	3.62	0.31	0.01	0.77	0.48
	PHotspot	48.4	16.9	2.42	0.85	171.5	9.67	0.37	0.03	**18.3**	6.66
	PSRA	30.1	16.5	1.51	0.83	83.7	2.21	−0.24	0.07	68.6	2.63

Bold indicates the greatest mean value, excluding the PSRA. SD denotes standard deviation.

With the exception of lower coverage levels in South Chicago burglaries, SEPP is the most accurate method in both case studies irrespective of crime type. The SEPP correctly identifies around 50% more shoplifting crimes than PKDE at coverage levels of around 6%. In South Chicago the difference is considerably more minor.

As the coverage approaches 30%, the mean hit rate of all of the methods for shoplifting converges due to the spatial confinement of the data. A similar effect is expected for the other crime types, but at coverage levels greater than those shown here. This is also the reason why the PSRA performs so well on the shoplifting data. Figure 5 also suggests that, with the exception of shoplifting, our conclusions about relative accuracy would vary little were the fixed coverage level varied in the range of 10–30%.

The WSR test results shown in Figure 6 provide additional insight that cannot be gained from the mean predictive accuracy results. In the case of Camden burglary, SEPP outperforms PKDE with a lower statistical significance ($p \leq 0.05$) than the other methods ($p \leq 0.001$ in all cases), despite exhibiting similar mean values. Conversely, PHotspot is significantly more accurate than PSTSS for Camden violence data ($p \leq 0.05$), while PKDE is not. This would be impossible to determine simply using the mean and standard deviations listed in Table 1. In the South Chicago assault data the SEPP outperforms PKDE, despite it performing very similarly to the PHotspot method. These results illustrate the value of treating prediction accuracy results as paired time series.

Table 2 indicates that all methods are approximately 25% more accurate in South Chicago when applied to burglary data, relative to the other crime types. However, the PRSA does not change. This suggests that the improvement in accuracy is due to the increased spatiotemporal clustering of the burglary crimes, making them generally easier to predict. The PAI of the PRSA may be interpreted as the improvement gained simply by restricting the spatial domain in which hotspots may be defined. In South Chicago, this improvement is around 50% for all crime types. In Camden it varies considerably, from 300% in the case of shoplifting to only 20% for burglary.

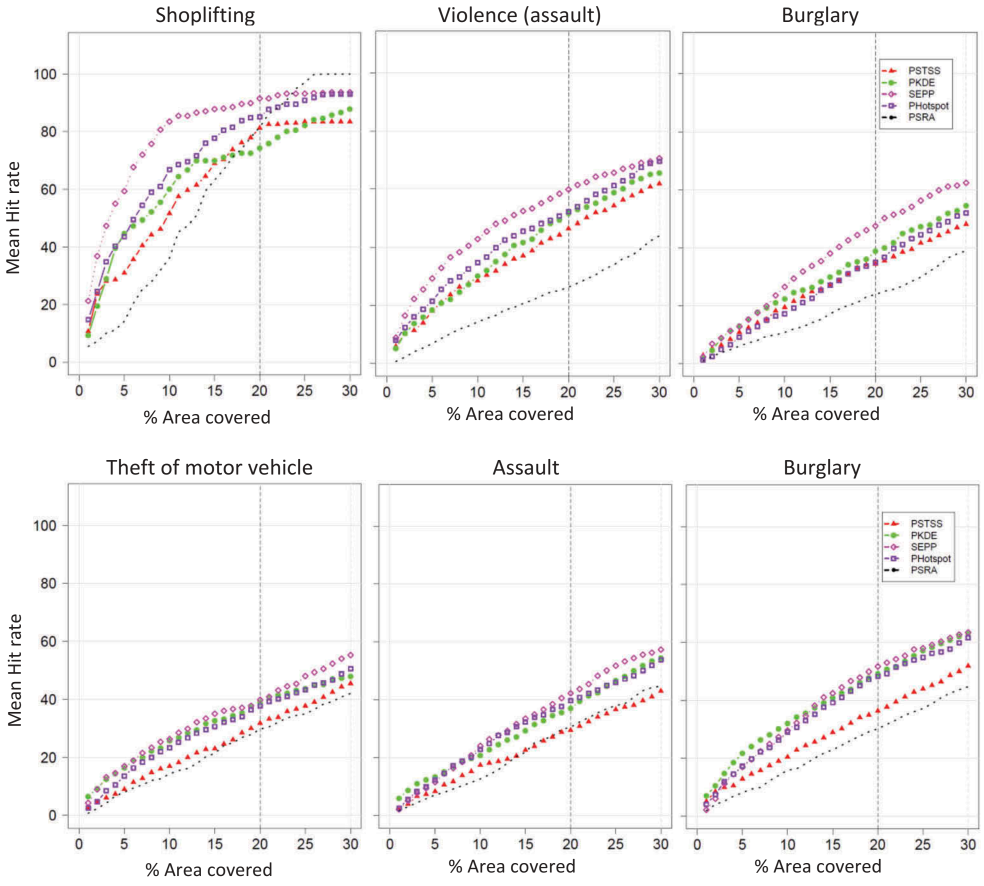


Figure 5. Variation of mean hit rate with area coverage for the three crime types and five predictive methods in Camden (top) and South Chicago (bottom). The dashed line shows the fixed coverage level used throughout.

4.2. *Compactness*

Compactness is interpreted as the ease with which the hotspots generated by each predictive method can be patrolled and is summarised in Tables 1 and 2. The AP ratio and CI are ranked in the same order throughout, except for a very minor discrepancy in the Camden burglary dataset. However, the AP ratio cannot be compared between the two regions. It is generally higher in the larger South Chicago region, which is expected since the AP ratio has units of length. In contrast, the CI is normalised and consistent across case studies. The consistently greater CI in South Chicago relative to Camden is therefore reflective of the underlying crime data.

The compactness results differ dramatically from the accuracy, with the less accurate PKDE and STSS methods achieving the highest compactness values for all crime types. The most accurate method, SEPP, achieves consistently poor compactness scores across all crime types. Whilst accurate, the predicted hotspots of the SEPP may therefore be too

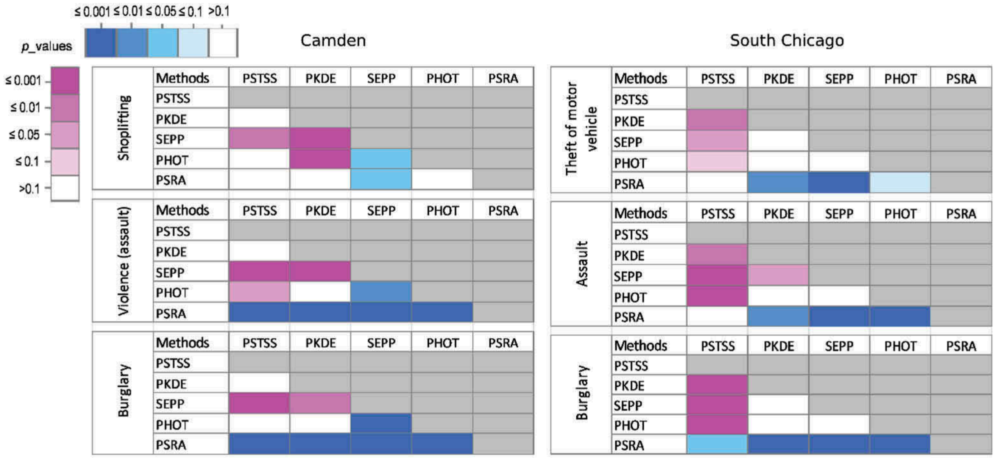


Figure 6. Results of the Wilcoxon Signed Rank (WSR) test at a fixed 20% coverage, showing the significance of any differences between predictive accuracy of the methods. Pink shades indicate that the row method is more accurate, blue shades indicate that the column method is more accurate.

diffuse to cover by police patrol. The sole exception, motor vehicle theft, is discussed below, in the section on DVI.

The variation between crime types in CI is highly dependent on the predictive method. PKDE, PSTSS and PHotspot show similar values across all crime types for a given region. In contrast, the CI of SEPP varies significantly by crime type. In the case of shoplifting, this is likely to be the result of the high concentration of shoplifting crimes within small regions. In this situation, the background and triggering components of the SEPP coincide very closely, producing several highly compact hotspot regions. In contrast, for burglary and assaults, there are more opportunities for the components to be spatially disparate, resulting in lower compactness.

4.3. *Dynamic variability*

The DVI is interpreted as the percentage of grid squares identified as hotspots that change between consecutive days. Tables 1 and 2 demonstrate that the PHotspot and PSTSS methods are highly variable, while the SEPP and PKDE methods remain almost constant over consecutive predictions. For example, the DVI of PHotspot when applied to Camden violent crimes is almost double that of the next highest method and results in an average of 16 hotspots changing between consecutive predictions. Conversely, the PKDE DVI equates to an average change of only two to four hotspot grid squares for all crime types and regions.

The DVI reveals that SEPP is completely invariant with respect to time when applied to TMV crimes in South Chicago. In this case, the SEPP has no triggering component; the background is purely spatial and computed using all data from before 28th September 2011. It is remarkable, therefore, that the predictive accuracy of the SEPP remains the highest over a period of 100 day-long predictions by a narrow margin. This suggests that

TMV crimes arise primarily due to purely spatial heterogeneity, and exhibit little spatio-temporal clustering.

The pattern of DVI scores observed in both case studies encapsulates the primary differences between the methods. In the case of the PKDE, roughly 1/60th of the data in the time aggregation window change with each prediction day, equivalent to around 8 crimes in a pool of 500 in the case of assaults. This has a very limited effect on the PKDE, so predictions are highly consistent between consecutive days. In contrast, the PHotspot method is highly sensitive to newly added crime data, since it weights recent crimes most highly when computing risk values. The SEPP lies somewhere between these extremes; the triggering component is equivalent to the PHotspot method while the background process is similar to the PKDE. The DVI of STSS is consistently high, indicating that clusters change significantly between consecutive days. Considering this information along with the CI, we see that the PSTSS method varies a great deal, but generally forms cohesive hotspot clusters in the process. This is as expected based on the selection of hotspot squares from contiguous circular regions (see Figure 1).

4.4. *Complementarity*

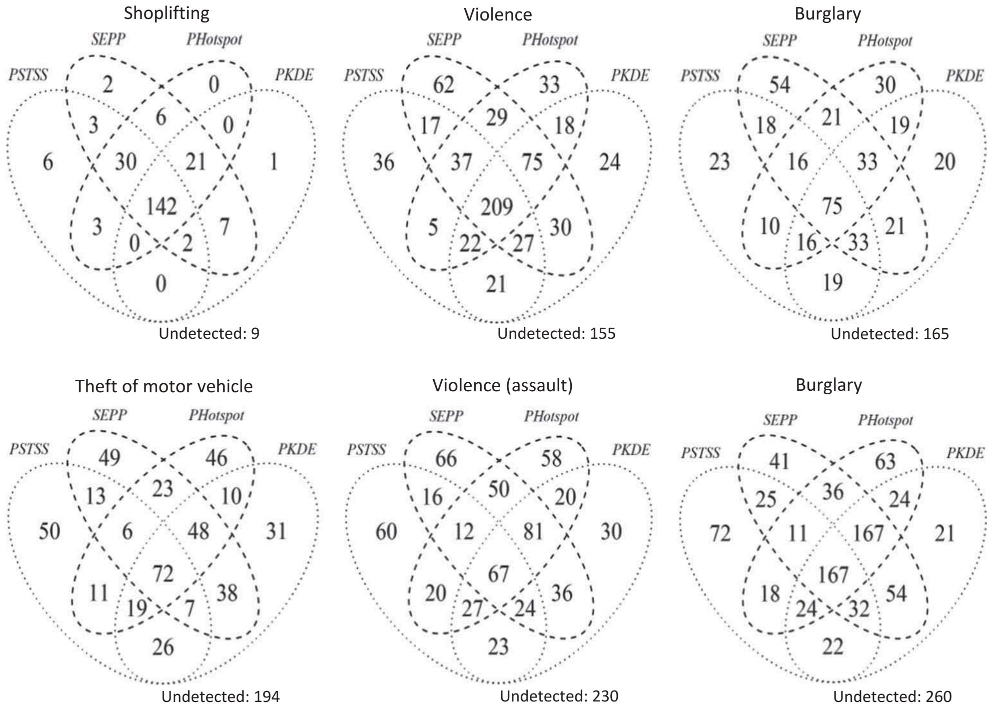
Analysing complementarity involves quantifying the degree of overlap between the sets of crimes identified by each method. As before, we use a fixed coverage of 20% for our analyses. The results are summarised in Figure 7.

In the case of shoplifting, the SEPP places hotspots over all but 10 of the 223 crimes identified by all methods combined. Here, ensemble methods are unlikely to lead to more accurate predictions. In contrast, for violent and burglary crimes in Camden the spread across all methods is much greater. Each method is successful at identifying, exclusively, a substantial number of crimes outside the ones jointly captured by the other methods. This demonstrates the utility of the complementarity: the WSR test revealed that the SEPP is significantly more accurate than the other three methods at 99.9% significance level in five of the six cases, yet all the methods identify a unique subset of crimes. In this case, an ensemble of methods might be expected to have significantly higher accuracy than its constituent parts.

In South Chicago, crimes are distributed more equally among the methods. In the case of burglary, of the 1037 crimes in total the SEPP, PHotspot and PKDE methods jointly identify 16% that the PSTSS method failed to capture. However, the PSTSS identified a further 7% of crimes that no other method captured. Despite being the least accurate method overall, around 20% of the crimes identified by PSTSS are detected exclusively by that method. Considering that PSTSS has some of the highest DVIs after PHotspot, we conclude that PSTSS is able to capture more emerging hotspots than any other method in South Chicago, though it performs more poorly with established hotspots.

5. Summary and conclusions

The primary motivation for this study is the need for greater insight into predictive methods for sparsely observed STPPs in terms of their evaluation and comparison. This addresses an increasingly pressing need in certain domains such as criminology and



Shoplifting

Violence

Burglary

Theft of motor vehicle

Violence (assault)

Burglary

Figure 7. Venn diagrams showing the total number of crimes correctly identified by each method at a fixed coverage of 20% in Camden (top) and South Chicago (bottom).

epidemiology in which the absence of systematic methodological assessments of predictive hotspot methods hinders analysts from making the best selection from the available methods and prevents researchers from demonstrating conclusively the contribution of their novel methods. Furthermore, the relative paucity of approaches to evaluation, primarily limited to the assessment of accuracy, limits the insight researchers may gain into the nature of prediction methods.

We tackled this problem by developing a toolkit of assessment metrics that can be applied routinely to any predictive method that generates forecasts based upon sparse STPP observations. We used crime prediction as our case study, however the methods developed here are applicable to any similar STPP prediction problem. In formalising the evaluation framework, we have combined four assessment measures, namely: accuracy, compactness, dynamic variability and complementarity. To demonstrate our evaluation toolkit thoroughly, we tested it on four different prediction methods, SEPP, PHotspot, PKDE and PSTSS. The PSTSS method has not previously been applied to predict crime hotspots and one important contribution of this study is the detailed assessment of its performance. In order to test the robustness of this evaluation framework, we applied it to crime data from two very different regions, one in London, UK and one in Chicago, IL, USA.

As the mean value of predictive accuracy does not permit statistically rigorous inference, we developed a novel hypothesis-based approach to identify methods with significantly higher accuracy. Treating predictive accuracy results as paired time series, our approach is based on the established WSR test. This added much-needed detail about the confidence we may place in the results, highlighting two instances in the Camden region in which the apparently obvious superiority of the SEPP is less significant than might be assumed. In the opinion of the authors, a statistically derived test such as this should be carried out in all future evaluations of STPP prediction methods to avoid spurious inferences.

Regarding measures of compactness, we assume that layouts with a high CI are more easily to patrol by police (Bowers *et al*. 2004). In reality, further work is required to determine the complex relationship between measures such as the CI and the ease of patrolling a region. Furthermore, it is likely that methods in which predictions are generated on road networks (Shiode and Shiode 2014) will resolve this open question, as police patrols generally follow these networks. Such approaches require similar metrics to those used here, adapted for application to a network. Network prediction and assessment is the subject of ongoing work by the authors. A further insightful characterisation of predictive performance is the DVI, which measures the extent to which the predicted hotspots change from day to day. Whilst this value is clearly linked to the variability of the underlying crime data, comparing the DVI between methods reveals hitherto unseen properties. Notably, the DVI indicates that the hotspot predictions of PHotspot and PSTSS are significantly more changeable from day to day than SEPP and PKDE. This is an important operational consideration; for example, in some situations it may not be possible to respond to these substantial changes sufficiently rapidly.

Moving beyond single score metrics, we quantified the predictive complementarity of the four predictive methods. This novel approach demonstrates that the SEPP has the potential of achieving a high level of accuracy especially for crimes that are tightly clustered in space. This is demonstrated in the Camden shoplifting dataset, where the SEPP correctly predicted all but 10 of the 223 crimes identified by all of the methods. In all other cases, however, the predictive complementarity was much smaller, with all methods uniquely identifying a substantial number of crimes. An ensemble predictive method might give accuracy gains in those cases. In particular, the complementarity results of all crime types in South Chicago suggest that PSTSS – despite fairly disappointing predictive accuracy – might complement one of the other methods, as it uniquely identified a large percentage of crimes. To our knowledge, ensemble methods – in which multiple prediction hotspots are combined to give a new hotspot map – have not previously been attempted. Further work is required to determine how best to combine multiple hotspot maps.

The metrics considered in this study are not exhaustive. For example, the software package FRAGSTATS (McGarigal *et al*. 2012) offers multiple alternatives to the CI. Further work is required to determine which measures are most effective for characterising predictive methods. However, it is likely that the findings will be domain-specific. A further metric not considered here is the recapture rate index (RRI), which was applied to crime prediction by Levine (Levine 2008). This measures predictive precision, in terms of the relative variability of the PAI over successive prediction windows, standardised to take temporal variation in crime density into account. In Levine's study, the time window is a full year. It is not immediately apparent how to adapt this measure for our purposes, where the crime density varies a great deal more from day to day, though this is the focus of future work. The WSR test and reported standard deviations of the PAI (Tables 1 and 2) partially address the issue of precision, albeit whilst ignoring the question of standardisation.

The evaluation framework developed here is widely applicable to the prediction of hotspots with spatio-temporal point data. Whilst we opted to use crime data to demonstrate its utility, the framework can be used to assess predictions in fields such as ecology (Tuia *et al*. 2008), geophysics (Marzocchi *et al*. 2012) and civil engineering (Ertekin *et al*. 2015). Predictive accuracy measures such as hit rate or PAI have a similar meaning and interpretation irrespective of domain and can therefore be applied directly. The concept of compactness is also readily interpretable when viewed in the context of other STPP datasets. For example, in epidemiology the concept of compactness will also enable health organisations to understand how a proactive intervention against an infectious disease might be realised. The notion of dynamic variability can be used to study how hotspot surfaces change over time, providing a novel and powerful means of distinguishing the types of forecasts generated by different methods and useful insight into the underlying mechanism responsible for temporal variations in predictions. Finally, the ability to measure predictive complementarity paves the way to future research into combining hotspot methods for improved prediction of STPPs.

Acknowledgements

The authors would like to acknowledge the Metropolitan Police Service (MPS) for provision of the crime data in the London Borough of Camden. They are also grateful to Trevor Adams for many valuable discussions about the manuscript and related work. The results presented and views expressed in this manuscript are the responsibility of the authors alone and not Trevor Adams or the MPS.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is part of the project Crime, Policing and Citizenship (CPC): Space-Time Interactions of Dynamic Networks (www.ucl.ac.uk/cpc), supported by the UK Engineering and Physical Sciences Research Council [EP/J004197/1].

ORCID

Monsuru Adepeju http://orcid.org/0000-0002-9006-4934
Gabriel Rosser http://orcid.org/0000-0001-9482-573X
Tao Cheng http://orcid.org/0000-0002-5503-9813

References

Bowers, K.J., Johnson, S.D., and Pease, K., 2004. Prospective hot-spotting: the future of crime mapping? *British Journal of Criminology*, 44 (5), 641–658. doi:10.1093/bjc/azh036

Brimicombe, A., 2012. Did GIS start a crime wave? SatNav theft and its implications for geo-information engineering. *The Professional Geographer*, 64, 3. 430–445. 10.1080/00330124.2011.609778

Brimicombe, A.J., Brimicombe, L.C., and Li, Y., 2007. Improving geocoding rates in preparation for crime data analysis. *International Journal of Police Science & Management*, 9 (1), 80–92. doi:10.1350/ijps.2007.9.1.80

Brown, C.D. and Davis, H.T., 2006. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80 (1), 24–38. doi:10.1016/j.chemolab.2005.05.004

Chainey, S., Tompson, L., and Uhlig, S., 2008. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21 (1–2), 4–28. doi:10.1057/palgrave.sj.8350066

Cheng, T. and Adepeju, M., 2014. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *Plos One*, 9 (6), e100465. doi:10.1371/journal.pone.0100465

Cheng, T., *et al*., 2014. A dynamic spatial weight matrix and localized space-time autoregressive integrated moving average for network modeling. *Geographical Analysis*, 46 (1), 75–97. doi:10.1111/gean.12026

Daley, D.J. and Vere-Jones, D., 2006. *An introduction to the theory of point processes: Volume I: Elementary theory and methods*, Springer Science & Business Media. Available from: https://books.google.com/books?id=6Sv4BwAAQBAJ&pgis=1 [Accessed 3 August 2015].

Diebold, F.X. and Mariano, R.S., 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 253–265.

Diggle, P.J., 1986. Displaced amacrine cells in the retina of a rabbit: analysis of a bivariate spatial point pattern. *Journal of Neuroscience Methods*, 18 (1–2), 115–125. doi:10.1016/0165-0270(86)90115-9

Duong, T. and Hazelton, M.L., 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32 (1), 485–506. doi:10.1111/j.1467-9469.2005.00445.x

Eck, J.E., et al., 2005. *Mapping crime : understanding hot spots*. Washington, DC: National Institute of Justice. Available from:: http://www.nij.gov/topics/technology/maps/pages/ncj209393.aspx

Ertekin, Ş., Rudin, C., and McCormick, T.H., 2015. Reactive point processes: A new approach to predicting power failures in underground electrical systems. *The Annals of Applied Statistics*, 9 (1), 122–144. doi:10.1214/14-AOAS789

Gorr, W., Olligschlaeger, A., and Thompson, Y., 2003. Short-term forecasting of crime. *International Journal of Forecasting*, 19, 579–594. doi:10.1016/S0169-2070(03)00092-X

Haworth, J., et al., 2014. Local online kernel ridge regression for forecasting of urban travel times. *Transportation Research Part C: Emerging Technologies*, 46, 151–178. doi:10.1016/j.trc.2014.05.015

Johnson, S.D., et al., 2007. *Prospective crime mapping in operational context*. London, UK: Home Office.

Johnson, S.D. and Bowers, K.J., 2004. The burglary as clue to the future: the beginnings of prospective hot-spotting. *European Journal of Criminology*, 1 (2), 237–255. doi:10.1177/1477370804041252

Kulldorff, M., et al., 2005. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*, 2 (3), 0216–0224. doi:10.1371/journal.pmed.0020059

Levine, N., 2008. The “Hottest” part of a hotspot: comments on “the utility of hotspot mapping for predicting spatial patterns of crime.”. *Security Journal*, 21 (4), 295–302. doi:10.1057/sj.2008.5

Malik, A., et al., 2014. Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement. *IEEE Transactions on Visualization and Computer Graphics*, 20 (12), 1863–1872. doi:10.1109/TVCG.2014.2346926

Marzocchi, W., Zechar, J.D., and Jordan, T.H., 2012. Bayesian forecast evaluation and ensemble earthquake forecasting. *bulletin of the Seismological Society of America*, 102 (6), 2574–2584. doi:10.1785/0120110327

McGarigal, K., Cushman, S.A., and Ene, E., 2012. FRAGSTATS v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps. Available from: http://www.umass.edu/landeco/research/fragstats/fragstats.html

Mohler, G., 2014. Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, 30 (3), 491–497. doi:10.1016/j.ijforecast.2014.01.004

Mohler, G.O., et al., 2011. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106 (493), 100–108. doi:10.1198/jasa.2011.ap09546

Nakaya, T. and Yano, K., 2010. Visualising crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14 (3), 223–239. doi:10.1111/j.1467-9671.2010.01194.x

O'Loughlin, J., Witmer, F.D.W., and Linke, A.M., 2010. The Afghanistan-Pakistan wars, 2008-2009: micro-geographies, conflict diffusion, and clusters of violence. *Eurasian Geography and Economics*, 51, 4. 437–471. 10.2747/1539-7216.51.4.437#.VemYo5P6nl8

Opitz, D. and Maclin, R., 1999. Popular ensemble methods: an empirical study. *Journal of Artificial Intelligence Research*, 11, 169–198. Available from:: http://jair.org/papers/paper614.html [Accessed 21 January 2015].

Patil, G.P., Joshi, S.W., and Koli, R.E., 2010. PULSE, progressive upper level set scan statistic for geospatial hotspot detection. *Environmental and Ecological Statistics*, 17 (2), 149–182. doi:10.1007/s10651-010-0140-1

Perry, W.L., et al., 2013. *Predictive policing: the role of crime forecasting in law enforcement operations*. Santa Monica, CA: RAND Corporation. Available from:: http://www.rand.org/pubs/research_reports/RR233.html

Rayner, N.A., 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, 108 (D14), 4407. doi:10.1029/2002JD002670

Riley, S., et al., 2014. Five challenges for spatial epidemic models. *Epidemics*, 10, 68–71. doi:10.1016/j.epidem.2014.07.001

Roerdink, J.B.T.M. and Meijster, A., 2000. The watershed transform: definitions, algorithms and parallelization strategies. *Fundamenta Informaticae*, 41 (1,2), 187–228. Available from:: http://dl.acm.org/citation.cfm?id=2372488.2372495 [Accessed 10 September 2015].

Shiode, S. and Shiode, N., 2014. Microscale prediction of near-future crime concentrations with street-level geosurveillance. *Geographical Analysis*, 46 (4), 435–455. doi:10.1111/gean.2014.46.issue-4

Tamayo-Uria, I., Mateu, J., and Diggle, P.J., 2014. Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spatial Statistics*, 9, 192–206. doi:10.1016/j.spasta.2014.03.005

Tonini, M., Tuia, D., and Ratle, F., 2009. Detection of clusters using space–time scan statistics. *International Journal of Wildland Fire*, 18 (7), 830. doi:10.1071/WF07167

Tuia, D., et al., 2008. Scan statistics analysis of forest fire clusters. *Communications in Nonlinear Science and Numerical Simulation*, 13 (8), 1689–1694. doi:10.1016/j.cnsns.2007.03.004

Turner, M.G., 1989. Landscape ecology: the effect of pattern on process. *Annual Review of Ecology and Systematics*, 20 (1), 171–197. doi:10.1146/annurev.es.20.110189.001131

Vere-Jones, D., 1995. Forecasting earthquakes and earthquake risk. *International Journal of Forecasting*, 11 (4), 503–538. doi:10.1016/0169-2070(95)00621-4

Woolhouse, M., 2011. How to make predictions about future infectious disease risks. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 366 (1573), 2045–2054. doi:10.1098/rstb.2010.0387

Zhuang, J., Ogata, Y., and Vere-Jones, D., 2002. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97 (458), 369–380. doi:10.1198/016214502760046925