Euro. Jnl of Applied Mathematics (2022), *vol.* 33, *pp.* 369–391

doi:10.1017/S0956792521000048

Modelling Burglary in Chicago using a self-exciting point process with isotropic triggering

CRAIG GILMOUR[1]* and DESMOND J. HIGHAM[2]†

[1]*Department of Mathematics and Statistics, University of Strathclyde, Glasgow, G1 1XH, UK*
email: cgilmour23@gmail.com
[2]*School of Mathematics, University of Edinburgh, Edinburgh, EH9 3FD, UK*
email: d.j.higham@ed.ac.uk

(*Received 3 March 2020*; *revised 14 January 2021*; *accepted 5 March 2021*;
first published online 8 April 2021)

Self-exciting point processes have been proposed as models for the location of criminal events in space and time. Here we consider the case where the triggering function is isotropic and takes a non-parametric form that is determined from data. We pay special attention to normalisation issues and to the choice of spatial distance measure, thereby extending the current methodology. After validating these ideas on synthetic data, we perform inference and prediction tests on public domain burglary data from Chicago. We show that the algorithmic advances that we propose lead to improved predictive accuracy.

Key words: Hawkes process, criminology, non-parametric, hotspots

2020 Mathematics Subject Classification: Primary: 65C20; Secondary: 60J20, 91D99

1 Introduction

Recently, quantitative research in criminal behaviour has exploited the idea that crime does not occur uniformly in space, but rather is more likely to take place in a limited number of high risk 'hot spots' [7, 20]. In particular, it has been proposed that certain types of crime, including burglary and gang violence, arise in highly clustered sequences, and therefore can be modelled in much the same way as seismic events, where there is increased risk of aftershocks in close proximity to an earthquake [15]. For example, a gang shooting may lead to retaliatory acts of violence against rival gangs, and burglars often target houses which have recently been burgled, along with nearby properties [15]. This has motivated the use of self-exciting point process models in a number of related contexts, including burglary and other crimes in Los Angeles [15] and Kent [16], gang rivalries in Los Angeles [9], gun crime and homicides [13, 14], the use of improvised explosive devices during 'The Troubles' [22] and civilian deaths in Iraq [12].

*Supported by EPSRC Programme Grant EP/P020720/1.
†Supported by grant EP/M00158X/1 from the EPSRC/RCUK Digital Economy Programme and by EPSRC Programme Grant EP/P020720/1.

From the perspective of modelling and algorithmics, predictive crime is a relatively new field that combines ideas from applied mathematics, statistics and data science [5, 6]. Our work focuses on the development and evaluation of high-level mathematical models that allow us to incorporate and test possible 'laws of motion' concerning criminal events. In particular, we aim to investigate and improve on current understanding by conducting experiments on real, anonymised, public domain crime data. We are motivated by the seminal work of Mohler *et al*. [15], where non-parametric triggering was suggested, and also by the ideas of Rosser and Cheng [18], who argued for an isotropic formulation.

The main contributions of our work are

- to suggest a more intuitive and effective normalisation in the construction of an isotropic triggering function,
- to consider L_p norms other than Euclidean distance,
- to define a more effective strategy for estimating background rate,
- to compare these formulations on a large-scale data set in order to (a) show how each model interprets the triggering effect and (b) test their predictive power.

While acknowledging that the use of mathematical models to quantify ideas from criminology has undergone a recent dramatic growth, we wish to emphasise that the issue of whether, and if so, how, law enforcement agencies should exploit such information is a separate issue of crucial importance. Moreover, rising above the question of how police forces could exploit model predictions to implement interventions that reduce future crime risk, there are more fundamental issues that must first be addressed around ethics, privacy and fairness [4, 10].

The manuscript is organised as follows. In Section 2, we set up the self-exciting process and discuss techniques that have been proposed to infer a trigger function. Sections 3 and 4 give further details about the inference task and the use of reflection in the kernel density estimation sub-problem. In Section 5 we look at imposing isotropy on the trigger function and suggest a new normalisation procedure. Different distance norms are introduced in Section 6. In Section 7 we test the inference approach on synthetically generated data in order to verify that it is feasible on the size of real data set that we plan to use. Such real data, relating to burglaries in the city of Chicago, is used in Section 8, where we compare the predictive power of the algorithms considered and illustrate the effectiveness of our refinements. In Section 9, we show that further improvements arise if the background rate is estimated differently. Finally, in Section 10 we conclude with a brief discussion.

2 Self-exciting point processes in crime modelling

A spatial-temporal point process, $N(\cdot,\cdot,\cdot)$, is a random measure on a region of $\mathbb{R}\times\mathbb{R}\times\mathbb{R}$ that takes non-negative integer values [17]. Letting x and y denote spatial coordinates in two dimensions, with t denoting time, the process can be characterised by its conditional intensity function

$$\lambda(x,y,t)=\lim_{\Delta x,\Delta y,\Delta t\to 0}\frac{\mathbb{E}(N(x+\Delta x,y+\Delta y,t+\Delta t)-N(x,y,t))}{\Delta x\Delta y\Delta t}. \tag{2.1}$$

Here, $N(x+\Delta x,y+\Delta y,t+\Delta t)-N(x,y,t)$ records the number of events that have taken place with the first spatial coordinate in $(x,x+\Delta x)$, the second in $(y,y+\Delta y)$ and time in $(t,t+\Delta t)$.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

In our context, events are recorded acts of crime. Here, it has been proposed that the conditional intensity function should take the Epidemic-Type Aftershock Sequences (ETAS) form [15]

$$\lambda(x, y, t) = \mu(x, y, t) + \sum_{t > t_i} g(x - x_i, y - y_i, t - t_i). \tag{2.2}$$

In (2.2), μ represents the *background rate*, and g is the *triggering function* that quantifies how an event creates a 'knock-on' effect of increased likelihood of further events nearby in space and future time. Specifying the form of the trigger is a key modelling step. Some authors have proposed specific, parametrised forms; notably, Gaussian in space and exponentially decaying in time [14, 17]. However, we will follow [14] and [18] by adopting a nonparametric approach – here we construct the trigger function from the available data, using an Expectation-Maximization (EM) algorithm.

The EM algorithm attempts to solve an incomplete data problem, where it is not known whether each data point is a background event or a triggered event, but associated probabilities may be calculated [23]. The E-step consists of building a non-negative lower triangular matrix $P \in \mathbb{R}^{N \times N}$, where N is the number of events which have occurred. The entries p_{ii} represent the probability that event i was a background event. Similarly, for $i > j$ the entries p_{ij} represent the probability that event i was triggered by the earlier event j. For the model (2.2), these probabilities are given as

$$p_{ii} = \frac{\mu(x_i, y_i, t_i)}{\lambda(x_i, y_i, t_i)}, \tag{2.3}$$

$$p_{ij} = \frac{g(x_i - x_j, y_i - y_j, t_i - t_j)}{\lambda(x_i, y_i, t_i)}, \quad \text{for } i > j, \tag{2.4}$$

with $p_{ij} = 0$ for $i < j$. In the next section, we describe how these expressions are used within an iterative algorithm.

3 Mohler *et al.* Monte–Carlo iterative procedure

In this work, we assume that the background rate in (2.2) does not depend on t. Following [15] we use variable bandwidth kernel density estimation (KDE), where the bandwidth around each data point is selected according to nearest neighbour distances [15]. The general principle is that data points occurring in close proximity use a small bandwidth.

To describe the procedure in greater detail, let the non-negative lower triangular matrix $P^{(0)}$, with $\sum_{j=1}^{N} p_{ij}^{(0)} = 1$, be an initial guess for P. To control the computational expense, we sample N_b background events and N_o triggered events from P, reducing the number of data points from $N(N+1)/2$ to $N_b + N_o = N$. In effect, based on the probabilities in P, we 'randomly' categorise each event as either a background event or an event triggered by a specific previous one. (In reality, we cannot say with certainty whether an event is a background event or has been triggered.)

Variable bandwidth KDE is then used on these samples in the following way. First, the offspring/parent interpoint distances $\{x_i^o, y_i^o, t_i^o\}_{i=1}^{N_o}$ are scaled to have unit variance in each coordinate. Based on this data, the kth nearest neighbour D_i in three-dimensional Euclidean distance is calculated from each data point i. After transforming back to the original scale and letting

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

$\sigma_x, \sigma_y, \sigma_t$ be the sample standard deviation of each coordinate, the triggering function is then estimated as

$$g(x,y,t) = \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_x\sigma_y\sigma_t(2\pi)^{(3/2)}D_i^3K_i} \times \exp\left(-\frac{(x-x_i^o)^2}{2\sigma_x^2D_i^2} - \frac{(y-y_i^o)^2}{2\sigma_y^2D_i^2} - \frac{(t-t_i^o)^2}{2\sigma_t^2D_i^2}\right). \tag{3.1}$$

Here, K_i is a normalisation constant for each data point i:

$$K_i = \int_0^\infty \frac{1}{\sigma_t\sqrt{2\pi}D_i} \exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2D_i^2}\right) \mathrm{d}t. \tag{3.2}$$

This is required because we 'lose' the density in the negative time region. Such a normalisation does not fully address the issue associated with there being a jump discontinuity for the time element of the triggering function (due to the fact that events cannot trigger events to happen in the past). We return to this topic in Section 4.

It has been reported that, for prediction purposes, variable bandwidth KDE on the background μ is less accurate than fixed bandwidth KDE [15]. The background rate is therefore estimated with a two-dimensional Gaussian kernel used on the samples of background events at each iteration. We found, as in [15], that prediction accuracy was improved by selecting a bandwidth smaller than that suggested by k-fold cross validation [21].

We summarise the full procedure as follows:

Mohler *et al*. Model

1. Set $k = 0$, and select nn_{trig}, the number of nearest neighbours to be used for the bandwidth of the spatial triggering kernel density estimation.
2. Make an initial guess $P^{(0)}$ for P. We may also wish to add the constraint $p_{ij}^{(0)} = 0$ if $t_i - t_j$, $|x_i - x_j|$ or $|y_i - y_j|$ is above a certain threshold.
3. Sample background events $\{x_i^b, y_i^b, t_i^b\}_{i=1}^{N_b}$ and offspring/parent interpoint distances $\{x_i^o, y_i^o, t_i^o\}_{i=1}^{N_o}$, based on the probabilities in each row of $P^{(k)}$, with $N_b + N_o = N$.
4. The background rate is then given by

$$\mu(x,y) = \frac{1}{T}\sum_{i=1}^{N_b} \frac{1}{\sigma_x^b\sigma_y^b 2\pi} \exp\left(-\frac{\left(x-x_i^b\right)^2}{2\sigma_x^{b\,2}} - \frac{\left(y-y_i^b\right)^2}{2\sigma_y^{b\,2}}\right). \tag{3.3}$$

5. Using the sampled triggered events (x_i^o, y_i^o, t_i^o), scale this data to have unit variance in the x, y and t components. Then calculate $D_{i,\text{trig}}$, the nn_{trig}th nearest neighbour, in three-dimensional Euclidean distance, to each data point i. Transforming the data back to its original scale, and with σ_x, σ_y and σ_t the standard deviations in the x, y and t directions, calculate the triggering function using (3.1) and (3.2).
6. Update $P^{(k)}$ with (2.3) and (2.4) using the new μ and g from steps 4 and 5 and set $k = k + 1$.
7. If $k <$ itermax, go to step 3.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

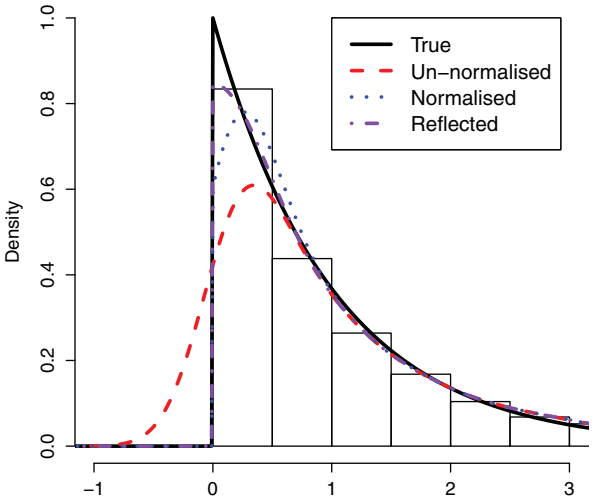

FIGURE 1. An example of KDE on samples taken from an exponential distribution, with the true distribution shown in solid black linetype. The result of standard kernel density estimation is shown in a red dashed linetype. This is undesirable for our purposes because it allows nonzero density on the negative axis. Shown in dotted blue linetype is an alternative where the density is set to zero on the negative axis before normalisation. The reflected kernel density estimator (4.1), shown in purple dash-dot linetype, gives a result closer to the true distribution.

4 Mohler model with reflected time

Rosser and Cheng [18] proposed an alteration to the Mohler model (3.1) to address the jump discontinuity at $t = 0$, namely reflecting the temporal component of the triggering function about $t = 0$. Given a density estimator K, this reflection can be defined as

$$K^*(t) = \begin{cases} K(t) + K(-t), & \text{if } t \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

This reflection approach is illustrated in Figure 1. As no density is 'lost', the normalisation constant is not needed, and the triggering function is estimated as

$$g(x,y,t) = \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_x\sigma_y\sigma_t(2\pi)^{(3/2)}D_i^3} \exp\left(-\frac{(x-x_i^o)^2}{2\sigma_x^2 D_i^2} - \frac{(y-y_i^o)^2}{2\sigma_y^2 D_i^2}\right) \times \left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right). \tag{4.2}$$

5 Isotropic function

Following [18], we consider an *isotropic* form of the triggering function,

$$\lambda(x,y,t|\mathcal{H}_t) = \mu(x,y) + \sum_{t>t_i} f(\sqrt{(x-x_i)^2 + (y-y_i)^2}, t-t_i). \tag{5.1}$$

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

Here, the spatial dependence of f arises only through the Euclidean distance $r = \sqrt{(x - x_i)^2 + (y - y_i)^2}$ from which a preceding event has taken place. In this case, a crime is just as likely to trigger a subsequent crime in any direction, only the distance affects the triggering strength. Justification for the use of an isotropic function is given in [18]; while the specific nature of triggering clearly depends on local features, such as the road network, and urban boundaries, such as rivers and parks, at a global level these directional dependencies will largely 'cancel out', and an isotropic function becomes a more appropriate predictor on average.

Writing $r = \sqrt{x^2 + y^2}$, Rosser and Cheng proposed the specific form

$$f(r,t) = \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t D_i \sqrt{2\pi} K_i} \exp\left(-\frac{(r - r_i^o)^2}{2\sigma_r^2 D_i^2}\right) \times \left(\exp\left(-\frac{(t - t_i^o)^2}{2\sigma_t^2 D_i^2}\right) + \exp\left(-\frac{(-t - t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right). \tag{5.2}$$

Noting that $r \geq 0$, normalising across the whole plane leads to the normalisation constant

$$\begin{aligned} K_i &= \int_{-\pi}^{\pi}\int_0^{\infty} \exp\left(-\frac{(r - r_i^o)^2}{2\sigma_r^2 D_i^2}\right) r \mathrm{d}r\mathrm{d}\theta \\ &= \sigma_r D_i r_i^o \sqrt{2\pi^3}\left(1 + \mathrm{erf}\left(\frac{r_i^o}{\sqrt{2}\sigma_r D_i}\right)\right) + 2\sigma_d^2 D_i^2 \pi \exp\left(-\frac{r_i^{o2}}{2\sigma_r^2 D_i^2}\right), \end{aligned} \tag{5.3}$$

where in this expression we have fixed a small error in [18].

However, there is an inherent issue with the approach to normalisation that leads to (5.3). The kernel density estimator applies its kernel to the distance r at which an event is said to have taken place, and this is swept around the whole plane and normalised. However, the area in the plane associated with the strip $[r, r + \Delta r]$ grows linearly with r. Consequently, if we look at the corresponding marginal probability density function in the r direction under the normalisation (5.3), we see an inherent bias – the expected distance at which the kernel will predict a future event to occur at is greater than the sample it has been given. This effect is illustrated in Figures 2 and 3. Essentially, it is because a 'strip' has a greater area as you move further away from the origin; if there are two distances from the origin at which the probability distribution function has the same height, then overall there will be a greater probability of an event occurring at the distance which is furthest away from the origin.

We now define what we believe to be the correct normalisation. Whereas the Rosser and Cheng isotropic model (5.2) and (5.3) uses a Gaussian kernel in the r-direction and normalises over the whole plane, we first normalise so that the function is a probability density function in the r-direction, then divide by $2\pi r$ when applying the function over the whole plane. In more detail, as in the Mohler model (4.2), we scale the data $\{r_i^o, t_i^o\}_{i=1}^{N_o}$ to have unit variance in each coordinate, and based on this data the k'th nearest neighbour D_i in two-dimensional Euclidean distance is calculated from each data point i. We find the normalisation constant

$$K_i = \int_0^{\infty} \frac{1}{\sigma_r D_i \sqrt{2\pi}} \exp\left(-\frac{(r - r_i^o)^2}{2\sigma_r^2 D_i^2}\right) \mathrm{d}r = 0.5\left(1 + \mathrm{erf}\left(\frac{r_i^o}{\sqrt{2}\sigma_r D_i}\right)\right), \tag{5.4}$$

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

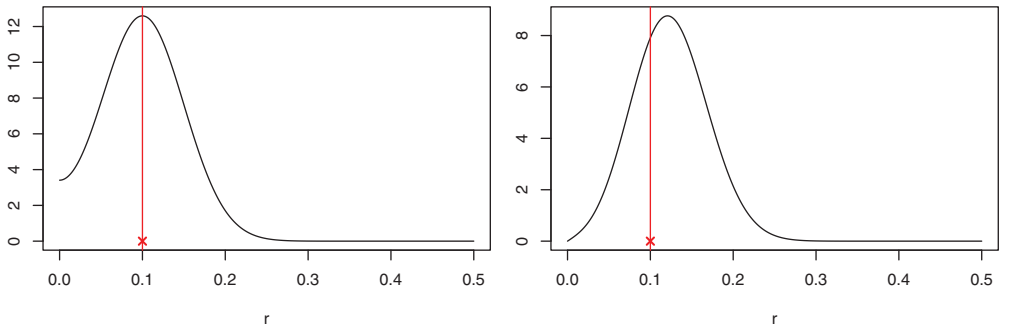

FIGURE 2. On the left is shown the height of the Rosser and Cheng isotropic function (5.2) based on one data point at a distance $r = 0.1$ from the centre. On the right is the corresponding marginal probability density function for the distance r at which we can expect a triggered event to happen, under the normalisation (5.3).

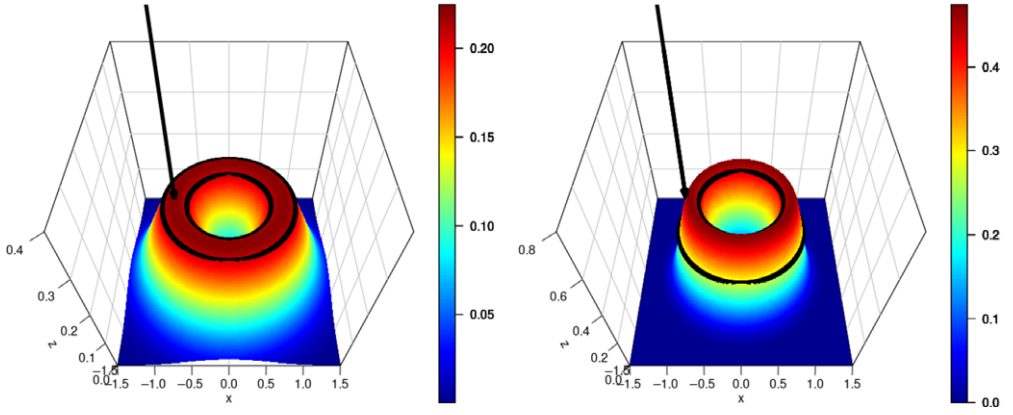

FIGURE 3. On the left is an example of the Rosser and Cheng isotropic model (5.2) and (5.3), and on the right is an example of the new isotropic model (5.5) and (5.6) applied to one data point indicated by the arrow, with no time dependence. The black lines indicate the value of the function at an equal distance from the event in an isotropic sense.

so that the overall density of each kernel is 1. We then have the following estimator of the probability distribution function

$$\hat{h}_k(r,t) = \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t \sigma_r D_i^2 2\pi K_i} \left(\exp\left(-\frac{(r-r_i^o)^2}{2\sigma_r^2 D_i^2}\right)\right) \times \left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right), \tag{5.5}$$

with $r = \sqrt{x^2+y^2}$, and hence

$$f(r,t) = \frac{\hat{h}_k(r,t)}{2\pi r}. \tag{5.6}$$

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

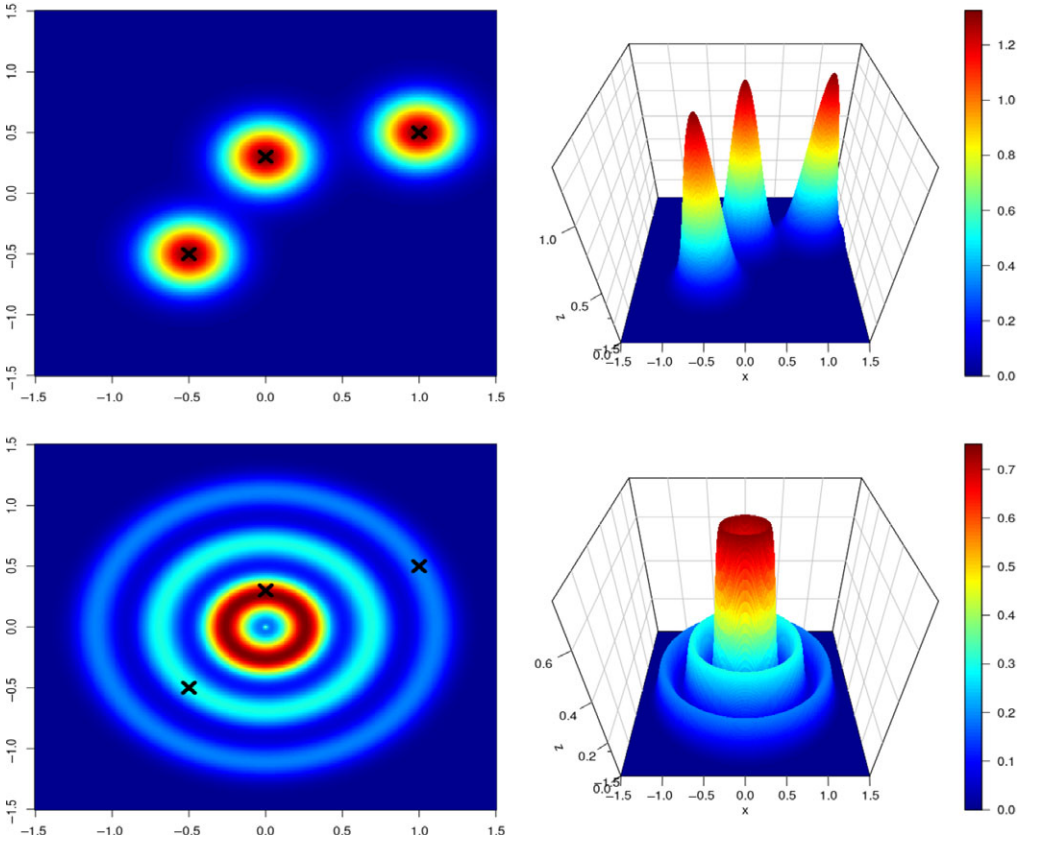

FIGURE 4. The upper figures correspond to a non-isotropic model and the lower figures to an isotropic version, with same data denoted by crosses. These figures show the triggering function in the x-y directions without time dependence. The origin represents the location of the ancestor event for each triggered event.

Note that, because we are estimating the probability density function for the distance at which an event will occur from the origin, we typically have $\hat{h}_k(r,t) \geq 0$ for r very close to 0, leading to large values of $f(r,t)$. Hence, to alleviate possible numerical overflow problems, in practice we set $f(0,t) = 0$, so events recorded at the exact same location are said to have not been triggered by each other.

Using this new normalisation, Figure 4 illustrates the difference between isotropic and non-isotropic triggering.

6 Alternative distance measures

6.1 Manhattan distance

The road network in the city of Chicago is laid out on a grid plan (see Figure 5). Based on both pedestrian and vehicular movement, this suggests that spatial distance may be better measured using a Manhattan rather than a Euclidean norm.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

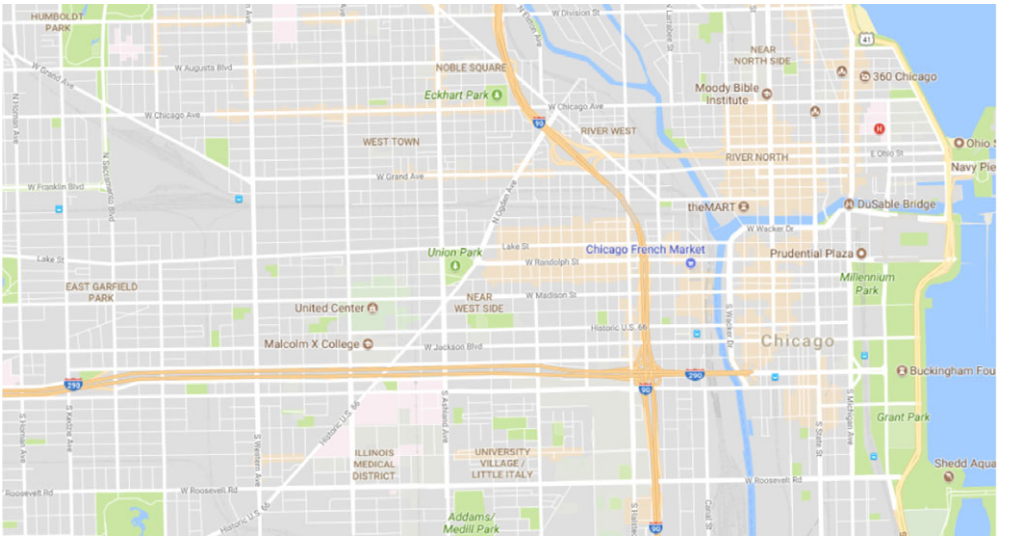

FIGURE 5. Example of the block structure of the road network in Chicago [2].

Using Manhattan distance $d = |x| + |y|$ instead of distance measure r from the previous section, we have the alternative estimator $\bar{h}_k$

$$\bar{h}_k(d,t) = \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t \sigma_d D_i^2 2\pi K_i} \left(\exp\left(-\frac{(d-d_i^o)^2}{2\sigma_d^2 D_i^2}\right)\right) \times \left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right). \tag{6.1}$$

Then, f in (5.1) becomes

$$\bar{f}(d,t) = \frac{\bar{h}_k(d,t)}{4d}. \tag{6.2}$$

In Figure 6, we show how the example in Figure 4 changes in this case. Again we face the same potential issues when looking at point estimates, namely that $\bar{h}_k(d,t) > 0$ for d close to zero can lead to large values of $\bar{g}(d,t)$, and we deal with them in the manner described in Section 5.

6.2 Chebyshev distance

The Manhattan and Euclidean distance measures correspond to the cases $p = 1$ and $p = 2$, respectively, in the general Minkowski, or L_p, setting. It has been argued that other choices of Minkowksi parameter may give a more effective representation of distance in an urban environment [8]. Hence, we will also consider the $p = \infty$, or Chebyshev, case, $m = \max(|x|, |y|)$. Intuitively, this measure is relevant if our perception of the distance between two points is dominated by the maximum of the two x-y distances. The resulting model could potentially flag streets which are of increased risk after an initial event. This leads to the estimator

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

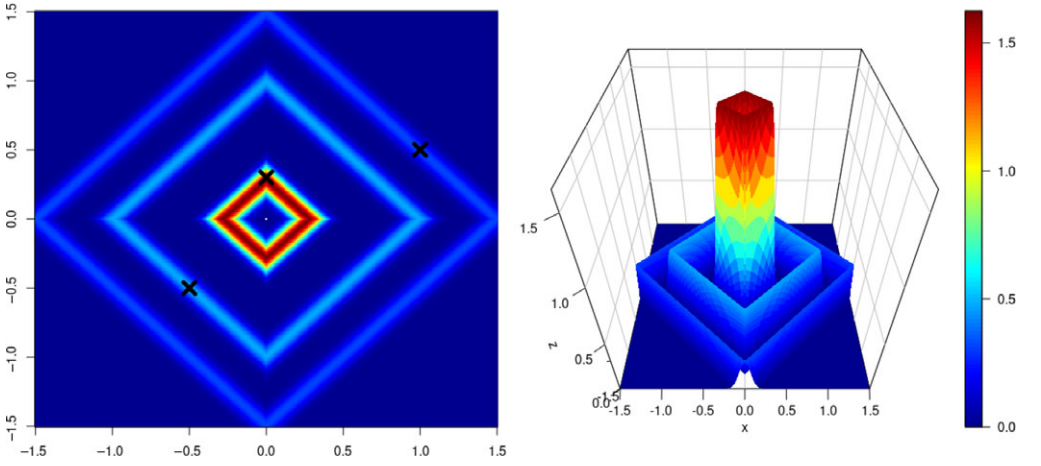

FIGURE 6. Manhattan distance analogue of Figure 4.

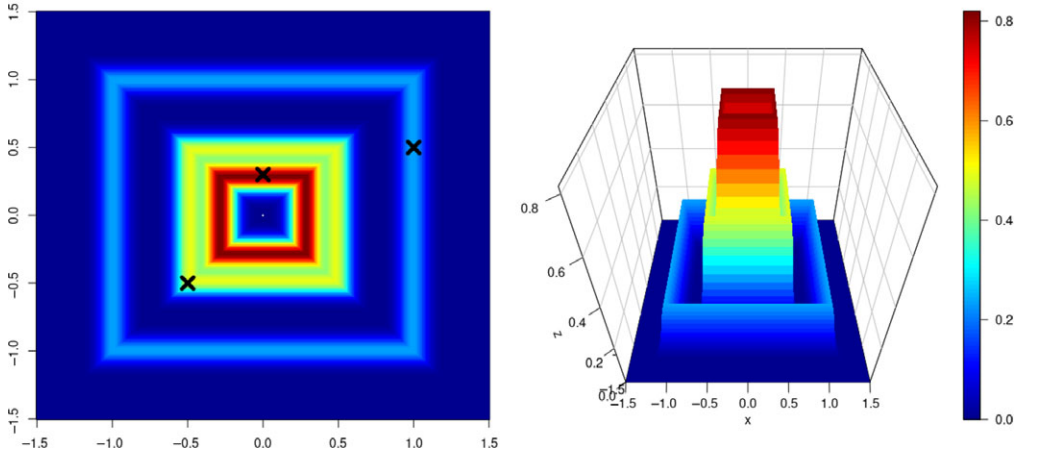

FIGURE 7. Chebyshev distance analogue of Figure 4.

$$\tilde{h}_k(m,t) = \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t \sigma_m 2D_i^2 \pi K_i} \left(\exp\left(-\frac{(m-m_i^o)^2}{2\sigma_m^2 D_i^2}\right)\right) \times \left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right), \tag{6.3}$$

and isotropic trigger

$$\tilde{f}(m,t) = \frac{\tilde{h}_k(m,t)}{8m}, \tag{6.4}$$

with the same fix as seen in Section 5 used to deal with small m values.

In Figure 7, we show the Chebsyhev analogue of Figure 4.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

7 Simulations

In this section, we test the EM algorithm on data simulated from a known isotropic spatial-temporal point process of the type (2.2). Our main aims are (a) to illustrate the normalisation issue and (b) to show that the overall algorithmic approach is capable of useful inference when the data comes from an appropriate distribution. We generate the data from an algorithm proposed by Zhuang, Ogata, and Vere-Jones [24]. This begins by simulating events from the background rate, the so-called first generation. We then simulate events that are triggered by each first-generation event, and then simulate the events generated by this generation of events, and so on, until the process terminates.

We chose a triggering function with exponential decay in time and a Gaussian profile in space, that is,

$$g(x, y, t) = \alpha\omega e^{-\omega t} \cdot \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}, \tag{7.1}$$

over an area resembling the city of Chicago, with parameters $\alpha = 0.3$, $\omega = 0.5$, $\sigma = 0.1$ and constant background rate $\mu = 0.02$, over time period $T = [0, 365]$. We then use the EM algorithm on these simulations to recover features of the triggering function. We repeated for 10 independent sets of data. Each contained around 6000 events, in line with the real data set that we study in Section 8. Following [18], the matrix P was initialised with the values

$$p_{ii} = 1, \tag{7.2}$$

$$p_{ij} = \exp(-\omega_{\text{init}}(t_i - t_j)) \exp\left(-\frac{(x_i - x_j)^2 + (y_i - y_j)^2}{2{\sigma_{\text{init}}}^2}\right), \quad \text{for } i > j, \tag{7.3}$$

with $\omega_{\text{init}} = 0.1$, $\sigma_{\text{init}} = 0.3$, normalised so that rows sum to 1. We found the results to be insensitive to the choice of starting matrix. In all tests that we report, the algorithm was run for 100 iterations. We found this to be a suitable value beyond which results were stable, and it is consistent with previously reported experiments [11, 15].

We define the marginal in the t-direction as

$$f_{\text{marg}}(t) = \int_0^r h(r, t)dr,$$

and the marginal in the r-direction as

$$f_{\text{marg}}(r) = \int_0^t h(r, t)dt,$$

where h is the probability density function.

In Figure 8, we show the marginals in the r-direction when we fitted the Rosser and Cheng isotropic model (5.2) and (5.3), using constant background bandwidth $\sigma_x^b = \sigma_y^b = 0.3$. The left and right pictures show the case where D_i corresponds to the 15'th and 50'th nearest neighbour, respectively. We observe from the figure that for the bandwidth with 15 nearest neighbours, the marginal results are more volatile. With 50 nearest neighbours the volatility decreases, but there is a noticeable shift to the right compared with the true parameters of the triggering function. This shift may be explained by the effect illustrated in Figure 2.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

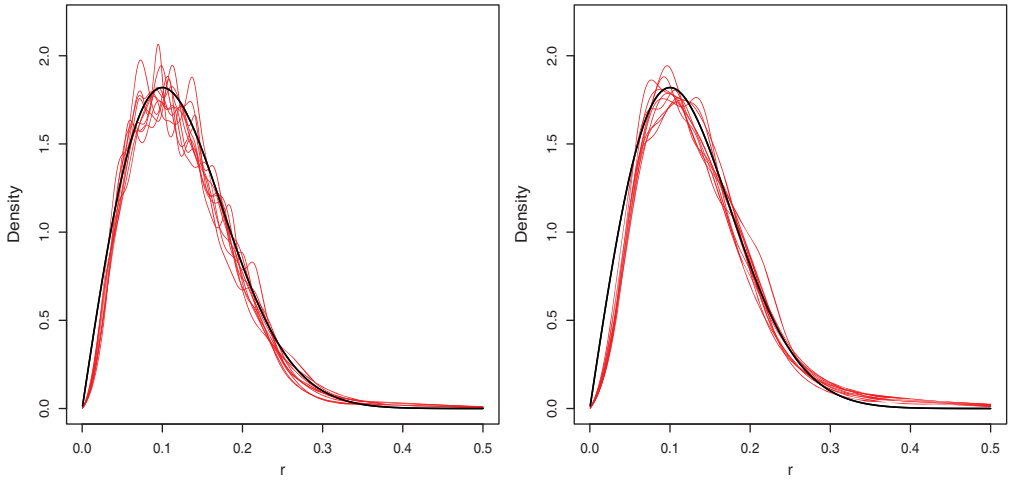

FIGURE 8. Inferred marginals for the triggering function in the r-direction for the Rosser and Cheng isotropic model (5.2) and (5.3), are shown in red/light line type for 10 independent simulated time series. True marginal for the triggering function is shown in black/darker line type. Left: variable bandwidth according to the 15'th nearest neighbour. Right: 50'th nearest neighbour.

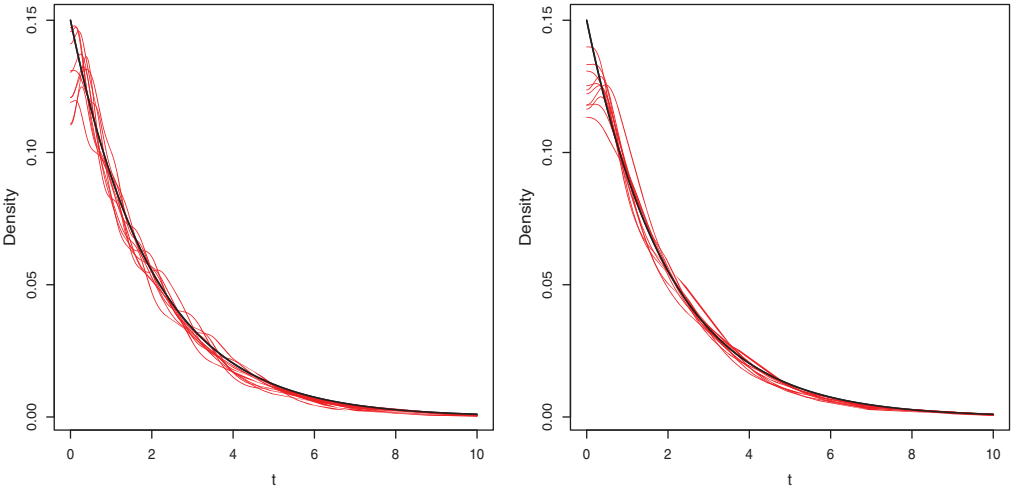

FIGURE 9. Analogue of Figure 8 for marginals in the t-direction.

In Figure 9, we show the analogue of Figure 8 where marginals are computed in the t-direction. We can see that in both cases we approximate the time component of the triggering function to similar accuracy, with the greatest discrepancy appearing to occur around $t = 0$.

In Figure 10, we show the marginals in the r-direction when we use the isotropic model (5.5) with new normalisation (5.6) with constant background bandwidth $\sigma_x^b = \sigma_y^b = 0.3$, and with D_i relating to the 50'th nearest neighbour. Comparing with Figures 8 and 9, we see that the r marginals more closely resemble those of the exact triggering function, whereas both versions

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

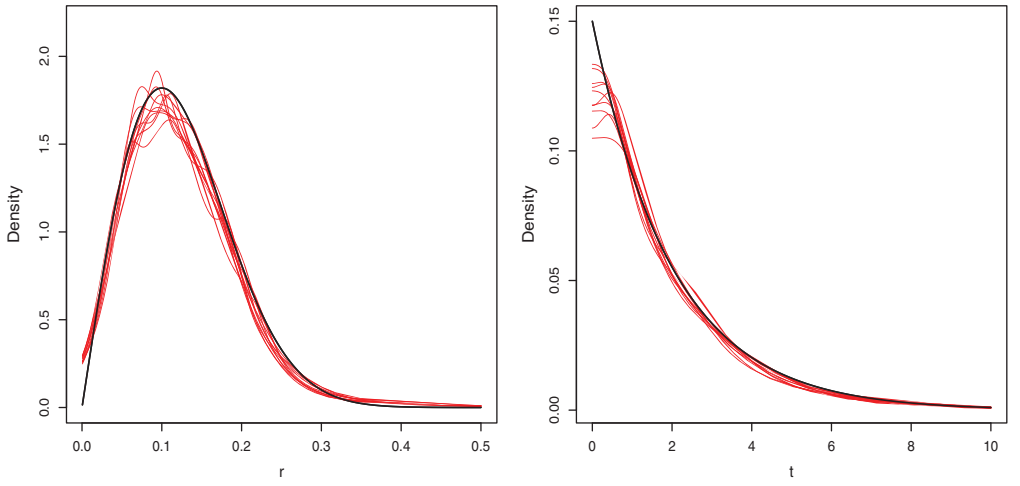

FIGURE 10. The figure on the left displays the marginals in the r-direction for the isotropic model (5.5) with new normalisation (5.6) in red for the 10 simulated processes, compared with the true parameters for the triggering function in black, with variable bandwidth according to the 50'th nearest neighbour. The figure on the right displays the respective marginals in the t-direction.

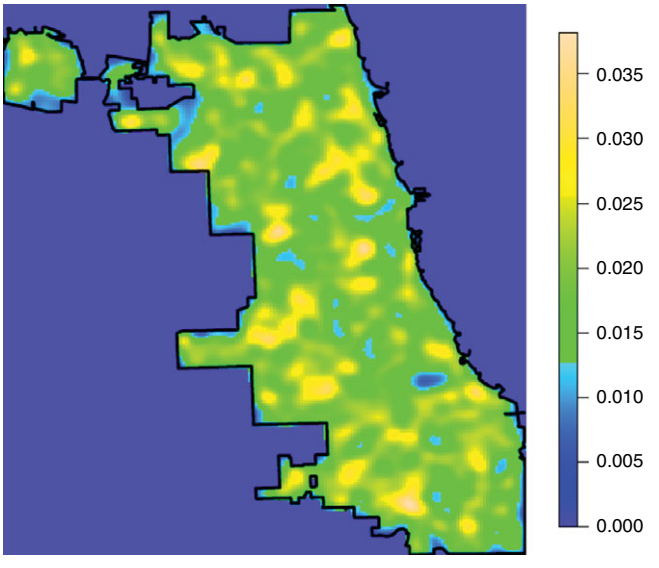

FIGURE 11. Inferred background intensity. Correct value is 0.02

give comparable results for the t marginals. These computations suggest that the normalisation (5.6) is more appropriate in this context.

An example of the background intensity estimated by the isotropic model (5.5) with new normalisation (5.6) for one run is shown in Figure 11. We observe a good approximation to the true constant background rate of $\mu = 0.02$.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

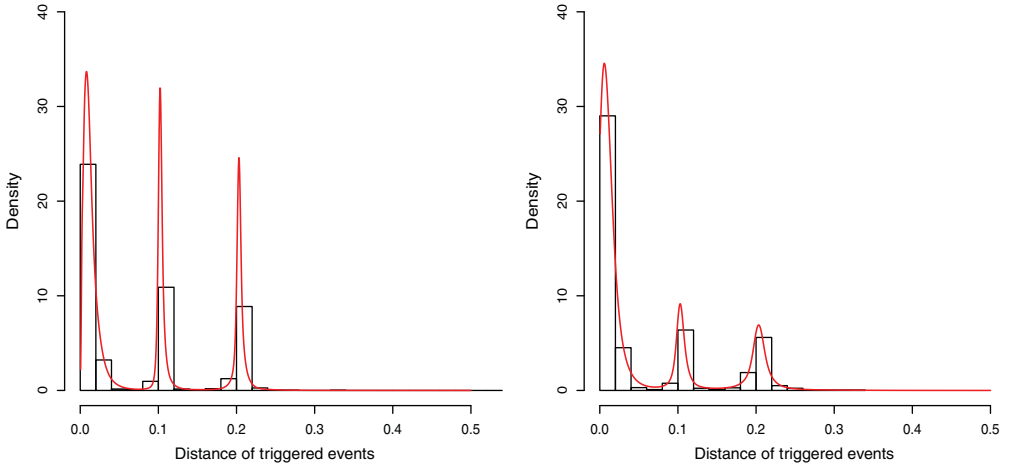

FIGURE 12. Histogram of the distance of the presumed triggered events for the Rosser and Cheng isotropic model (5.2) and (5.3) on the left and for the isotropic model (5.5) with new normalisation (5.6) on the right.

8 Prediction results on real data

Having confirmed that the EM algorithm can be used to recover useful information when data are produced from a spatial-temporal point process, we now use the algorithm to infer background rates and triggering functions for a real data set. After calibrating different isotropic models in this way, we compare their predictive performance.

We fit the models to publicly available burglary data from the city of Chicago. We use 13,044 reported burglaries in the city from 2015. Information about where to find the data, and how to download the codes used in these experiments, can be found in the Acknowledgements section. To be concrete, we measure distance in kilometres and time in days. As with the simulations in the previous section, we initialised the matrix P with (7.2) and (7.3) with $\sigma_{\text{init}} = 0.5$, $\omega_{\text{init}} = 0.2$ and used 100 iterations of the EM algorithm, as we will do with all subsequent model fits. We selected constant background rate bandwidths of $\sigma_x^b = \sigma_y^b = 0.15$, which were found to give good predictive results. To improve computational efficiency we set the triggering function to zero for distances greater than 1km, or time periods greater than 90 days.

Figure 12 displays the disparity between the triggered events detected by the Rosser and Cheng isotropic model (5.2) and (5.3) and the isotropic model (5.5) with new normalisation (5.6), along with the estimated marginals in the r-direction. The distance is given in km, as it is throughout this paper. As we observed with simulated data in Figure 8, we can observe a slight shift to the right for the normalisation (5.3) compared with (5.6). We can see that we observe spikes occurring at distances of around 100 and 200 m from the original event, which corresponds to the grid network in Chicago where a street block is approximately 200 by 100 m [1].

In Figure 13, we show the locations of the events that are assumed to be triggered for

- the original Mohler model with reflected time (4.2) with 50 nearest neighbours,
- the isotropic model (5.5) with new normalisation (5.6) with 50 nearest neighbours,
- the Manhattan model (6.1) and (6.2) with 50 nearest neighbours, and
- the Chebyshev model (6.3) and (6.4) with 50 nearest neighbours.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

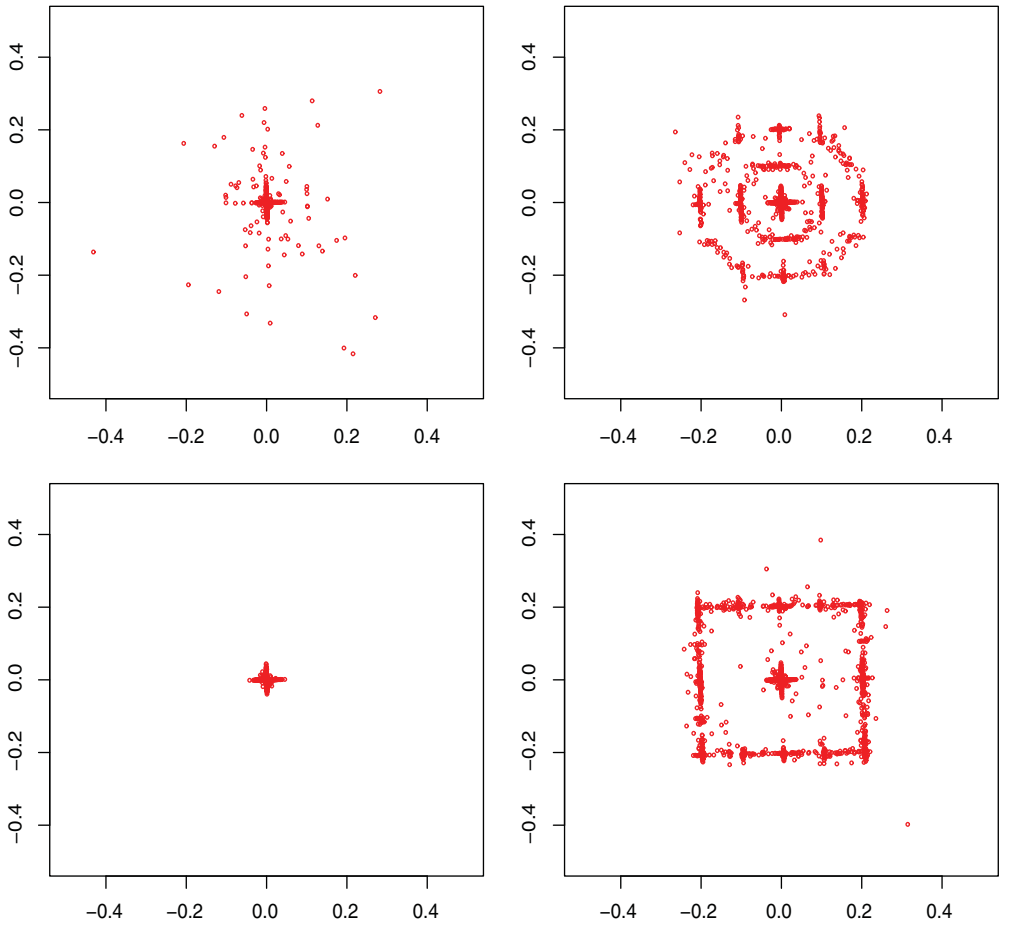

FIGURE 13. These figures display the locations of events which are assumed to be triggered. Top left: original Mohler model (4.2); Top right: new isotropic model (5.5) and (5.6). Lower left: isotropic with Manhattan norm distance (6.1) and (6.2). Lower right: isotropic with Chebyshev distance (6.3) and (6.4).

The origin denotes the preceding event which is said to have triggered the event, and so each figure can give us a sense of the locations that events are likely to be triggered in. This type of figure was shown previously in [15]. The x and y components are given in kilometres.

These figures show how the inferred triggering effect differs between the models. The top left picture is for the original Mohler model (4.2) with D_i selected according to the 50'th nearest neighbour. The top right is for the new isotropic model (5.5) and (5.6) with D_i selected according to the 50'th nearest neighbour. The bottom left is for the model using Manhattan distance (6.1) and (6.2) with D_i selected according to the 50'th nearest neighbour. The bottom right is for the Chebyshev model (6.3) and (6.4) with D_i selected according to the 50'th nearest neighbour.

We see that the non-isotropic Mohler model (4.2) suggests that triggering takes place mostly in the immediate vicinity of the original event. Results indicating that triggering took place almost exclusively in the North-South direction were obtained when D_i was selected according to the 15'th nearest neighbour as in [18] (which found similar results for certain areas of Chicago);

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

Table 1. *Estimated branching ratio for the four models when fitted with burglary data from 2015, along with the number of events which are estimated to have been background or triggered events.*

Model	Branching Ratio	Est. bg events	Est. trig events
Mohler	0.069	12,150	894
Isotropic	0.166	10,874	2170
Manhattan	0.118	11,508	1536
Chebyshev	0.195	10,494	2550

however, this led to far inferior predictive results. The new isotropic model (5.5) and (5.6) suggests crime spreading in all directions, with peaks at around one and two hundred metres from the original event. The Manhattan model (6.1) and (6.2) shows triggered events concentrated in close proximity to the original event, with a small amount of spread in the four coordinate directions. The Chebyshev model (6.3) and (6.4) indicates burglaries being triggered both in very close neighbours and also in a 200 by 200 m block centred at the parent event. In Table 1, we show the branching ratio that is estimated by the four models, along with the number of events in 2015 which were estimated by the models to have been background or triggered events.

We now judge the predictive performance of the competing models. Following [14, 15], we record how much actual crime takes place in locations that are predicted to have the highest intensity. In more detail, starting from the beginning of 2016, we split Chicago into 75 m $\times$ 75 m grids. After fitting our models to burglary data from 2015, we order the grid cells in terms of the value of the predicted intensity for each grid cell using the events that have happened up until that point. We then look at the proportion of overall activity over the next hour that was correctly 'predicted' by the model at each point of the ranking. For example, at rank k, using the top k grid cells in terms of intensity, we record the proportion of the recorded crime in Chicago occurring in these cells over the next hour. We then recalculate the predicted intensity an hour later for each grid cell and rank the grid cells again and look at how well we have predicted the next hour's crime and so on. The resulting information, which is comparable to that in a receiver operating characteristic (ROC) curve, shows what proportion of the total activity we can predict by focusing on the regions that are regarded as the most susceptible. In practical terms, this shows how effectively the model would allocate the limited law enforcement resources to locations where crime is going to take place in the near future (making the assumption that such allocation will be productive in either stopping the crime taking place, or apprehending the perpetrators).

In Figure 14, we show results for the occurrence of burglaries on an hourly basis for the first three months in 2016 in Chicago, having fitted the models to burglary data from the 12 months in 2015. These results are a sum of the hourly results over the three months. We fitted the Rosser and Cheng isotropic model (5.2) and (5.3) while finding D_i corresponding to the 15'th nearest neighbour as was done in their paper [18], and for the new isotropic model (5.5) and (5.6) using D_i corresponding to the 50'th nearest neighbour. We find the new isotropic model benefits from taking a larger number of nearest neighbours to assign the bandwidths. We can see that our isotropic model offers predictive improvements in relation to the Rosser model when fitted to this real burglary data.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

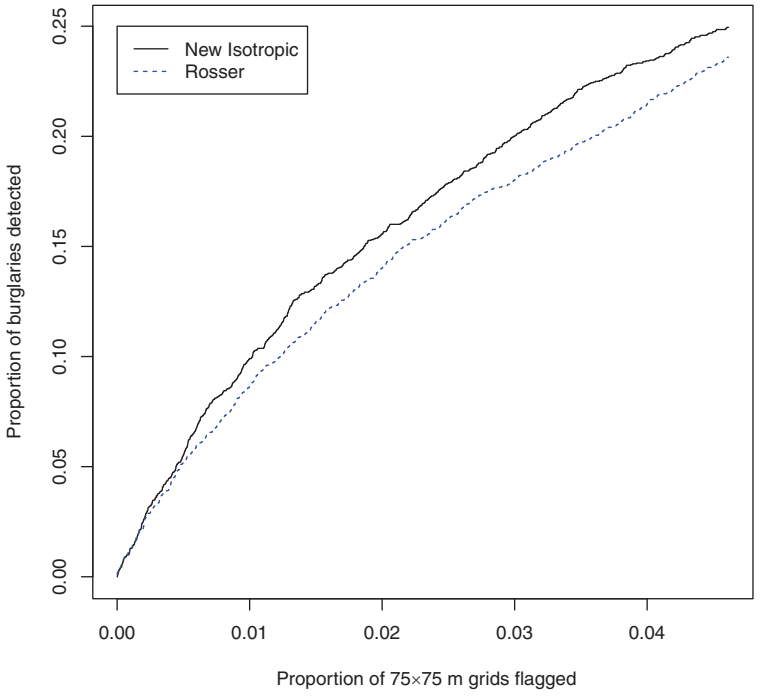

FIGURE 14. Proportion of burglaries predicted in 2016 for the Rosser and Cheng isotropic model (5.2) and (5.3) and the new isotropic model (5.5) and (5.6).

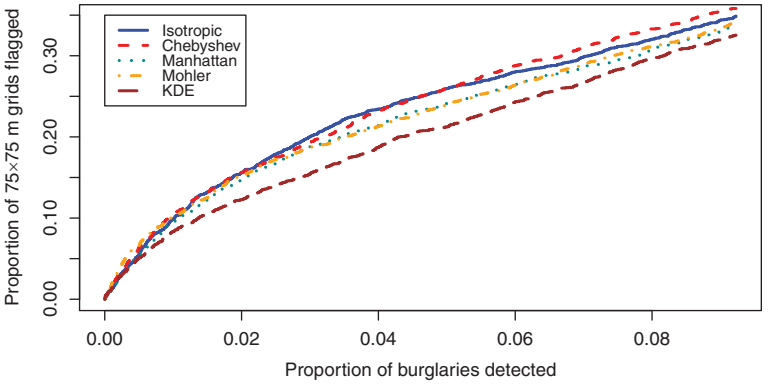

FIGURE 15. Proportion of burglaries predicted in 2016 for various models.

In Figure 15, we perform the same comparison between the Mohler model (4.2) with D_i corresponding to the 50'th nearest neighbour, as well as the new isotropic model (5.5) and (5.6), the Manhattan model (6.1) and (6.2) and the Chebyshev model (6.3) and (6.4) using D_i corresponding to the 50'th nearest neighbour. From this point on all references to isotropic models will be using the new normalisation. We find that that there is little difference between the performance of the models; the Chebyshev and isotropic model appearing to perform slightly better than the

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

other two models on the high-activity cells. We also show the prediction performance of a simple kernel density estimation of events from the previous 6 months without a triggering component.

9 Omitting observations from estimation of the background rate

Motivated by an approach used in a slightly different context [14], we now consider a variation of the inference algorithm. Here, we calculate the background rate for each event while omitting events which have occurred at that location in the kernel density estimation. Such an implementation requires us only to sample triggered events and allows us to use all the original data to estimate the background rate. We spell out the details below.

Isotropic Function Algorithm

1. Set $k = 0$, and select the number of nearest neighbours to give the bandwidth for the spatial triggering kernel density estimation nn_{trig}.
2. Make an initial guess $P^{(0)}$ for the matrix P, with $p_{ij} = 0$ for $j > i$, and $\sum_{j=1}^{n} p_{ij} = 1$ for all i. We may also wish to initiate the matrix with $p_{ij} = 0$ if $t_i - t_j$ or $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ are above a certain threshold.
3. Sample background events $\{x_i^b, y_i^b, t_i^b\}_{i=1}^{N_b}$ and offspring/parent interpoint distances $\{x_i^o, y_i^o, t_i^o\}_{i=1}^{N_o}$, based on the probabilities in each row of P_k, with $N_b + N_o = N$.
4. The background rate is updated with

$$\mu(x,y) = \frac{1}{T} \sum_{i=1,(x,y)\neq(x_i,y_i)}^{N} p_{ii} \frac{1}{\sigma^2 2\pi} \exp\left(-\frac{((x-x_i)^2 + (y-y_i)^2)}{2\sigma^2}\right). \tag{9.1}$$

5. Using the sampled triggered events (r_i^o, t_i^o), we scale this data to have unit variance in the r and t components. We then calculate $D_{i,\text{trig}}$, the nn_{trig}th nearest neighbour two-dimensional Euclidean distance to each data point i. Then, transforming the data back to its original scale, and with σ_r^b and σ_t^b denoting the standard deviations in the r and t directions, we calculate the triggering function as

$$\hat{f}(r,t) = \frac{\alpha}{N_o} \sum_{i=1}^{N_o} \frac{1}{\sigma_t \sigma_r D_i^2 4\pi^2 r K_i} \left(\exp\left(-\frac{(r-r_i^o)^2}{2\sigma_r^2 D_i^2}\right)\right) \times \left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right), \tag{9.2}$$

   where α is the branching ratio

$$\alpha = \frac{\sum_{i\neq j} p_{ij}}{N}. \tag{9.3}$$

6. With the new μ and g from steps 4 and 5, we update from $P^{(k)}$ to $P^{(k+1)}$ and set $k = k + 1$.
7. If $k <$ itermax, go to step 3.

9.1 Prediction results with adapted background

As we might expect, changing the algorithm in this way was found to increase the inferred triggering rate. Figure 16, which may be compared with Figure 13, shows the triggered events after

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

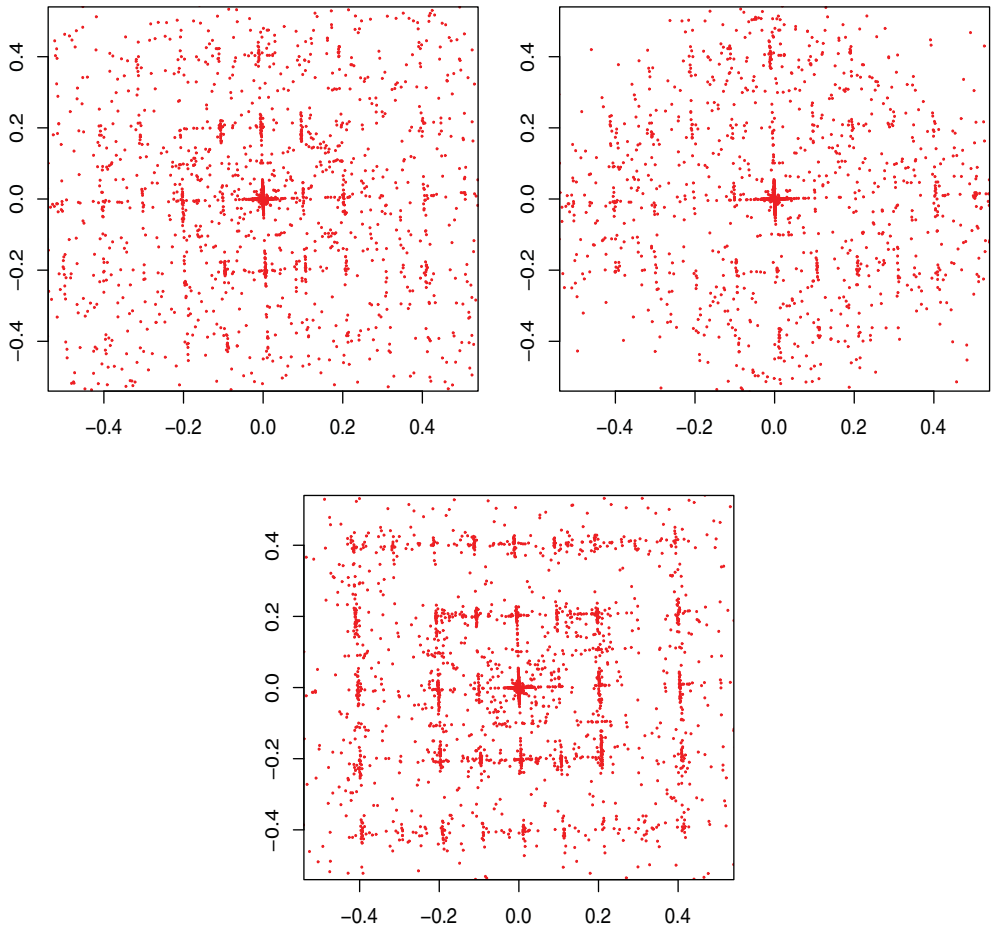

FIGURE 16. Locations of triggered events with adapted background algorithm. Top left: Euclidean distance. Top Right: Manhattan distance. Bottom: Chebyshev distance.

100 iterations of the EM-algorithm for the new isotropic model (5.5) and (5.6) and Chebyshev models (6.3) and (6.4). More events are assumed to have been triggered, and, moreover, the triggering effect is inferred to have spread further than with the version that does not use the adapted background.

Figure 17 displays the marginals in the r and m components for the isotropic (5.5) and (5.6) and Chebyshev models (6.3) and (6.4), respectively, with the adapted background rate. For the Chebyshev model, we can see clear spikes occurring roughly every 200 m.

Figure 18 shows the time marginal $f_{\text{marg}}(t)$ with the isotropic model (5.5) and (5.6) when observations are omitted from the background rate, along with the estimate of the background intensity across Chicago. In this case, the inferred triggering effect is at its greatest in the immediate aftermath of a crime, dropping steadily over the next few weeks. The estimation of the background rate indicates that burglary is highly localised, as we would expect. In Table 2, we show the branching ratio that is estimated by the four models, along with the number of events in 2015 which were estimated by the models to have been background or triggered events.

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

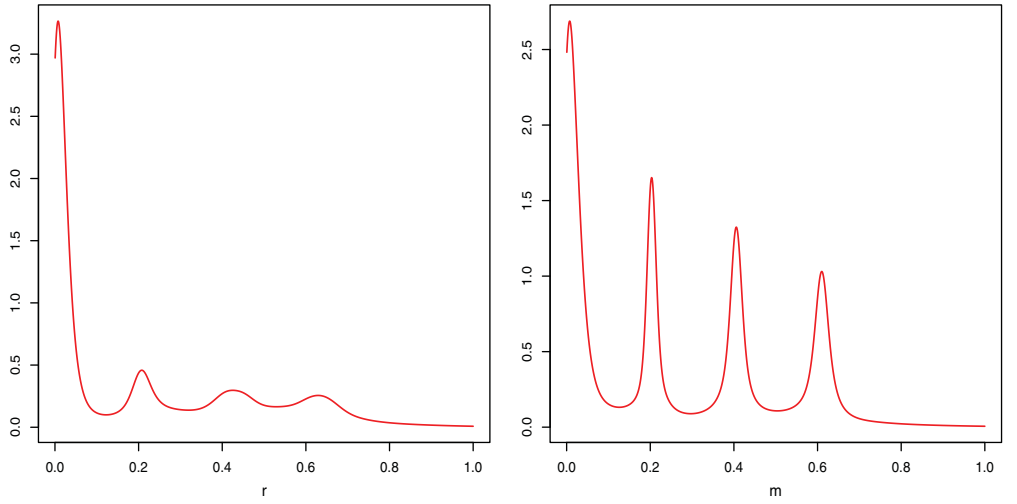

FIGURE 17. These figures display the marginal in the r component in kilometres on the real burglary data for the isotropic model (5.5) and (5.6) when observations are omitted from the background rate (left), and the marginal in the m-direction for the Chebyshev model (6.3) and (6.4) when observations are omitted from the background rate (right).

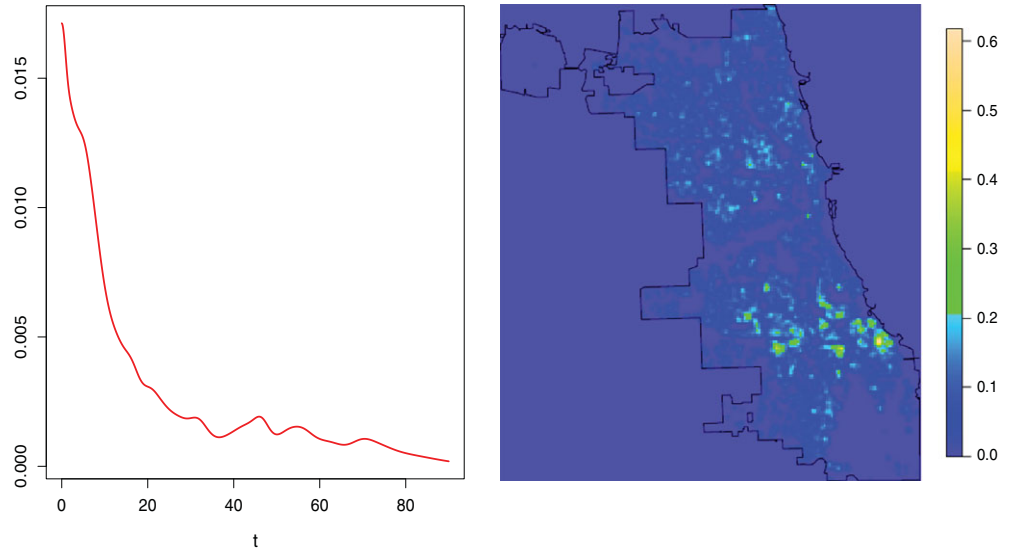

FIGURE 18. These figures display the marginal $f_{\text{marg}}(t)$ in days when the isotropic model (5.5) and (5.6) is fitted (left) and the estimate of the background rate with this model (right).

In Figure 19, which can be compared with Figures 14 and 15, we compare the prediction results of the new isotropic model (5.5) and (5.6), the Manhattan model (6.1) and (6.2) and the Chebyshev model (6.3) and (6.4) using this adapted background approach and the isotropic model with the original background from the previous section. We can see a clear improvement in this measure of predictive power when this change is made. All the models with the adapted

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

Table 2. *Estimated branching ratio for the four models with adapted background when fitted with burglary data from 2015, along with the number of events which are estimated to have been background or triggered events*

Model	Branching Ratio	Est. bg events	Est. trig events
Mohler	0.084	11,950	1094
Isotropic	0.256	9702	3342
Manhattan	0.232	10,024	3020
Chebyshev	0.317	8904	4140

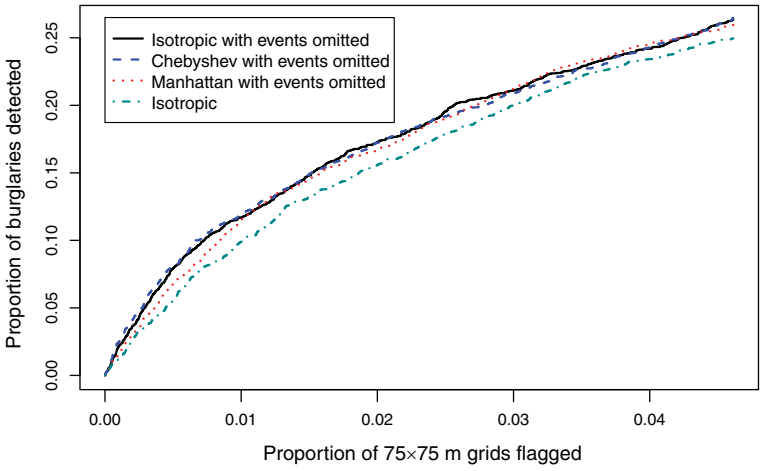

FIGURE 19. Proportion of burglaries predicted in 2016 for the isotropic (5.5) and (5.6), Manhattan (6.1) and (6.2) and Chebyshev models (6.3) and (6.4) with observations omitted from the background, along with the isotropic model.

background approach perform similarly, with the Manhattan model giving slightly worse results than the other two.

10 Discussion

Our aim in this work was to develop and fit self-exciting point processes to publicly available crime data from Chicago, with the intention of building on existing models in order to better understand how crime events arise and more accurately predict their future occurrence.

For this data set, the best-performing models when using the original measurement of the background rate were the models using Chebyshev (6.3) and (6.4) and Euclidean (5.5) and (5.6) distance, which marginally outperformed the model using Manhattan (6.1) and (6.2) distance and the Mohler model (4.2). Further improvement in predictive performance was achieved when the adaptation to the measurement of the background rate was made in Section 9. With this change to the way the background rate was found, the prediction of future burglaries of burglaries improved similarly for all the distance measures used, with the Chebyshev (6.3) and (6.4) and Euclidean (5.5) and (5.6) models slightly outperforming the Manhattan model (6.1) and (6.2).

https://doi.org/10.1017/S0956792521000048 Published online by Cambridge University Press

There are many interesting directions in which this work could be extended. For example, we could allow these models to include a hierarchy of crime types, as in [14], where minor crimes are assumed to trigger more serious offences. Further, by incorporating extra spatial information to the model, such as locations where crime cannot possibly take place, we could further improve predictive power. It may also be beneficial to allow the triggering function to vary with the location of the original crime. Of course, in all such extensions, a balance must be struck between incorporating useful information and overfitting a heavily parametrised model. Other possible areas of interest include using recent ideas from network science to capture key features of the street network [19].

We also emphasise that this work has focused on understanding and improving the performance of mathematical modelling and inference techniques on a public domain, anonymised data set. As pointed out in Section 1, a separate, and more fundamental, question is whether law enforcement agencies can and should make operational use of such techniques to improve resource allocation. To address such a question requires consideration of a diverse range of issues, including accountability, transparency, privacy and bias.

Acknowledgements

The City of Chicago has a publicly available data set consisting of reported crimes in Chicago [3]. This data contain all reported incidents of crime from 2001 to the present day, along with other information including the date and time each crime took place and the location that it occurred. We fitted all the models on recorded burglaries from the year 2015. We removed the entries containing NA's for the location data, along with duplicate entries for burglaries recorded at the exact same time and location as another. It should be noted that the location data are partially redacted but lie on the same block as the actual incident, while the information on the time the incident took place is often a best estimate of when the crime occurred. R code for the experiments conducted in this work is available from https://www.maths.ed.ac.uk/∼dhigham/algfiles.html.

Conflict of interest

None.

References

[1] Street and Site Plan Design Standards, City of Chicago (2007) https://www.cityofchicago.org/dam/city/depts/cdot/StreetandSitePlanDesignStandards407.pdf [Online; accessed 13-November-2018].

[2] Google Maps (2018) https://www.google.com/maps/@41.8787372,-87.6812593,13.5z [Online; accessed 1-November-2018].

[3] City of Chicago Data Portal (2020) https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data [Online; accessed 13-January-2020].

[4] Bennett Moses, L. & Chan, J. (2018) Algorithmic prediction in policing: assumptions, evaluation, and accountability. *Policing Soc.* **28**(7), 806–822.

[5] Berestycki, H., Johnson, S., Ockendon, J. & Primicerio, M. (2010) Criminality. *Eur. J. Appl. Math.* **21**(4–5).

[6] Bertozzi, A., Johnson, S. & Ward, M. (2016) Mathematical modelling of crime and security: special issue of EJAM. *Eur. J. Appl. Math.* **27**(3), 311–316.

[7] BRAGA, A. A. (2001) The effects of hot spots policing on crime. *ANNALS Am. Acad. Polit. Soc. Sci.* **578**(1), 104–125.
[8] CROSBY, H., DAMOULAS, T., CATON, A., DAVIS, P., DE ALBUQUERQUE, J. P. & JARVIS, S. A. (2018) Road distance and travel time for an improved house price Kriging predictor. *Geo-Spatial Inf. Sci.* **21**, 185–194.
[9] EGESDAL, M., FATHAUER, C., LOUIE, K., NEUMAN, J., MOHLER, G. & LEWIS, E. (2010) Statistical and stochastic modeling of gang rivalries in Los Angeles. *SIAM Undergraduate Res. Online* **3**, 72–94.
[10] KARPPI, T. (May 2018) 'The computer said so': On the ethics, effectiveness, and cultural techniques of predictive policing. Social Media+ Society. https://doi.org/10.1177.2056305118768296
[11] LEWIS, E. & MOHLER, G. (2011) A nonparametric EM algorithm for multiscale Hawkes processes. *J. Nonparam. Stat.* **1**(1), 1–20.
[12] LEWIS, E., MOHLER, G., BRANTINGHAM, P. J. & BERTOZZI, A. L. (2012) Self-exciting point process models of civilian deaths in Iraq. *Secur. J.* **25**(3), 244–264.
[13] LOEFFLER, C. & FLAXMAN, S. (2018) Is gun violence contagious? A spatiotemporal test. *J. Quant. Criminol.* **34**, 999–1017.
[14] MOHLER, G. (2014) Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecasting* **30**(3), 491–497.
[15] MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. & TITA, G. E. (2011) Self-exciting point process modeling of crime. *J. Am. Stat. Assoc.* **106**(493), 100–108.
[16] MOHLER, G. O., SHORT, M. B., MALINOWSKI, S., JOHNSON, M., TITA, G. E., BERTOZZI, A. L. AND BRANTINGHAM, P. J. (2015) Randomized controlled field trials of predictive policing. *J. Am. Statist. Assoc.* **110**(512), 1399–1411.
[17] REINHART, A. (2018) A review of self-exciting spatio-temporal point processes and their applications. *Statist. Sci.* **33**(3), 299–318.
[18] ROSSER, G. & CHENG, T. (2016) Improving the robustness and accuracy of crime prediction with the self-exciting point process through isotropic triggering. *Appl. Spat. Anal. Pol.*, 1–21.
[19] ROSSER, G., DAVIES, T., BOWERS, K. J., JOHNSON, S. D. & CHENG, T. (2017) Predictive crime mapping: Arbitrary grids or street networks? *J. Quant. Criminol.* **33**(3), 569–594.
[20] SHERMAN, L. W., GARTIN, P. R. & BUERGER, M. E. (1989) Hot spots of predatory crime: routine activities and the criminology of place. *Criminology* **27**(1), 27–56.
[21] SILVERMAN, B. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.
[22] TENCH, S., FRY, H. AND GILL, P. (2016) Spatio-temporal patterns of IED usage by the provisional Irish republican army. *Eur. J. Appl. Math.* **27**(3), 377–402.
[23] VEEN, A. & SCHOENBERG, F. P. (2008) Estimation of space–time branching process models in seismology using an EM–type algorithm. *J. Amer. Stat. Assoc.* **103**(482), 614–624.
[24] ZHUANG, J., OGATA, Y. & VERE-JONES, D. (2004) Analyzing earthquake clustering features by using stochastic reconstruction. *J. Geophys. Res. Solid Earth* **109**(B5).