Self-Exciting Point Processes and their Applications to Crime Data

Craig Gilmour
Department of Mathematics and Statistics
University of Strathclyde
Glasgow, Scotland, UK

April 2019

This thesis is submitted to the University of Strathclyde for the degree of Doctor of Philosophy in the Faculty of Science.

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed: Date:

Acknowledgements

I would like to thank the Engineering and Physical Sciences Research Council (EPSRC) for providing funding for my research training, and conference attendance. I would also like to thank Banff International Research Station for Mathematical Innovation and Discovery (BIRS) for their invitation to the "Mathematical Criminology and Security" workshop between 18-22 March 2019, and for covering the accommodation costs.

I would like to offer a special thanks to Professor Des Higham for his expert supervision and guidance throughout. I also wish to express my gratitude to Professor Peter Grindrod and Dr John MacKenzie for their careful consideration and comments on the thesis, and to Dr Wei Yang for granting me the opportunity to carry out this research.

Finally I'd like to thank all my friends and family for their help and support throughout. A particular mention should go to my parents for their constant support; to my sister Fiona for her encouragement; to Romain, Andrew and Grant for being there from the start; to Joseph and Feargus for keeping me sane in the office; and to Tommy, Ben and Jack for their friendship.

Abstract

Self-exciting point processes describe a type of point process where the occurrence of an event leads to an increased probability of a future event taking place. Such subsequent events can be said to have been ‘triggered’ by the previous event. Originally used to model earthquakes, in recent years this class of processes has been utilised in crime modelling to describe how crime events can propagate future crime. In this thesis we look at ways of building on the existing literature which uses self-exciting point processes to model crime, with the intention of improving the ability to predict when and where future crime may occur. We introduce a new parametric form for the triggering function, which can be utilised in situations where an initial event does not immediately lead to an increased risk of further events, but rather the risk increases over a period of time after the event. We also introduce adaptations to existing non-parametric methods which offer us fresh insight into the dynamics of crime. We validate our models using publicly available crime data from the city of Chicago.

Contents

Chapter 1

Introduction

In recent years self-exciting point processes have been increasingly used in order to model crime, based on the idea that crime is more likely to occur in the aftermath of a previous crime. In this thesis we concentrate on fitting existing and novel self-exciting point processes to publicly available crime data from Chicago, with the intention of building on existing models in order to more successfully predict where future crime events may occur.

In Chapter 2 we introduce the definition of a temporal self-exciting point process, along with one of the most common parametric forms which the triggering function takes in the literature, the epidemic type aftershock sequence (ETAS) model. We also explain how we infer estimates for the model parameters of self-exciting point processes throughout the thesis, namely via an EM-algorithm.

We propose a new parametric form for the triggering function in Chapter 3, which we name the delayed criminal response (DCR) model, which allows the risk of a future crime occurring to increase for a period after an event, as an alternative to the ETAS function where the risk decays after the immediate increase when an event occurs. We run a number of tests on the EM-algorithm for the ETAS and DCR model on simulated data to verify how well we can recover the true parameters of a self-exciting point process.

In Chapter 4 we fit temporal self-exciting point processes to publicly available burglary data from the city of Chicago, using both the ETAS and DCR model, and compare the performance of both models in terms of predicting future crime. We introduce spatio-temporal point processes and fit parametric versions of these to the burglary data using both the ETAS and DCR triggering function.

In 2008 Marsan and Lengliné proposed a method which could estimate Hawkes processes for earthquakes without making any prior parametric assumptions [45]. In Chapter 5 we model burglary in Chicago taking advantage of this method to incorporate a non-constant background rate, and a non-parametric form for the triggering function using weighted histograms. We give predictive results for unseen burglary data for this method.

In Chapter 6 we alter a model previously proposed by Mohler et al. [48] to build an isotropic version of this non-parametric triggering function. Using Manhattan distance and the maximum norm, we propose further alternatives to this function which take advantage of the grid structure of the street network in Chicago, to give fresh insight into the dynamics of burglary within Chicago. In Chapter 7 we give a summary of the results obtained within this thesis, along with suggestions for future work.

Research results from this work have been presented by the author at

- "Mathematical Criminology and Security" workshop, Banff International Research Station for Mathematical Innovation and Discovery (BIRS), 18-22 March 2019
- 27th Biennial Numerical Analysis Conference, University of Strathclyde, 27-30 June 2017
- SIAM-UKIE Student Chapter Conference, NUI Galway, 26 May 2017
- British Applied Mathematics Colloquium, University of Surrey, 10-12 April 2017

Chapter 2

Self-Exciting Point Processes

2.0.1 Mathematical Modelling of Crime

In recent years mathematics has increasingly found new applications in the social sciences where modelling and simulation can be used to shed light on some aspect of human behaviour. One such area in which mathematics has been applied is crime modelling, with the European Journal of Mathematics having recent special issues dedicated to this field [10, 12].

Within criminology there are several areas in which mathematical models have been utilised to inform the decision making process for police or policy makers, including the investigation of the effect severity of sentencing can have on the crime rate [20], how limited police resources can be allocated in the most efficient way [21, 65], and the effect social factors can have on the prevalence of crime [36].

Many types of mathematical methods have been used to model different problems within criminology. Reaction-diffusion equations have been used to describe the formation and expansion of 'crime hotspots', and how these can be eradicated or displaced with police suppression [63]. Other research has borrowed ideas from epidemiology to model crime as an infection with the police response similar to an immune response [7]. Lotka-Volterra equations have been used to model gang interactions [14], while ideas from network science have been used to investigate the formation of criminal networks [46]. Agent-based models, involving the interactions of people represented as so called agents who follow a set of simple rules, have been used to model the interaction between street gangs [32] and the behaviour of burglars [44]. Game theory has been used to investigate the behaviour of criminals, and optimal strategies to counter criminal behaviour [24, 64]. Other approaches have sought to utilise data from social media in order to more accurately assess the risk of crime [25, 43]

One area where mathematical methods have developed significant traction within police departments is with models aiming to predict future crime [29], with so called ‘predictive policing’ being used by real police forces to target their resources more effectively [55]. ‘PredPol’ is a commercial company which offers software for police to target their patrols [1], with this software being used across the United States and gaining widespread media attention [3, 11]. Fundamental to the use of ‘predictive policing’ is the idea that as well as predicting future crime, police forces can use these predictions to implement some intervention technique which will reduce the future crime risk [55]. The effectiveness of such methods in terms of reducing future crime has garnered mixed reviews [13, 33], while great debate around the ethics of using these techniques has also been generated [9, 35].

2.0.2 Self-Exciting Point Processes in Crime Modelling

Much of the research into criminal behaviour focuses on so-called ‘crime hot spots’, and the idea that crime doesn’t occur uniformly in space, but rather is more likely to take place in a limited number of high risk areas [13, 62]. It has been proposed that certain types of crime, including burglary and gang violence, take place in highly clustered event sequences, and therefore can be modelled in much the same way as seismic events, where earthquakes increase the risks of aftershocks occurring in close proximity to the original earthquake [48]. For example, a gang shooting may lead to retaliatory acts of violence against rival gangs, or burglars often target houses which have recently been burgled and nearby houses [48]. Such self-exciting point processes have been used to model crime in a number of studies, including to model burglary and other crimes in Los Angeles [48] and Kent [49], to model gang rivalries in Los Angeles [23], to predict gun crime and homicides in Chicago [47], to model the use of improvised explosive devices during “The Troubles” [69], and to model civilian deaths in Iraq [39].

2.0.3 Temporal Point Processes

A temporal point process is essentially a list of times $(t_1, t_2, \ldots, t_n)$ where events have taken place [57]. The times arise from probabilistic modelling arguments. Such processes can be described by a conditional intensity function

$$\lambda^*(t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}(N(t + \Delta t) - N(t)|\mathcal{H}_t)}{\Delta t}, \tag{2.0.1}$$

where $\mathcal{H}_t$ is the history of the process prior to time t, and $N(t)$ is a counting process which describes the number of points which have occurred up to time t. In essence this function gives the expected number of events relative to an infinitesimal future time increment, depending on the history of the process [57]. The $*$ notation is typically used to denote that the function depends on the history of the process to avoid writing this explicitly. The number of events which we can expect to observe by time T in such a process can be written

$$\mathbb{E}(N(T)) = \int_0^T \lambda^*(t)\,\mathrm{d}t. \tag{2.0.2}$$

From a modelling perspective, the key task is to specify the form of the conditional intensity function $\lambda^*(t)$. A simple type of point process is the Poisson point process. It has the memoryless property; the conditional intensity function is constant over time, i.e. $\lambda^*(t) = \lambda$ [57]. In such a process the occurrence of events does not depend on events in the past. The distribution of the *waiting times*, the times between consecutive events, follows an exponential distribution. The memoryless property means that the distribution of how long we can expect to wait for the next event from a certain time t remains the same regardless of how long we have been waiting since the previous event.

The Hawkes process is a self-exciting point process whose conditional intensity λ^* increases in the aftermath of an event [31]. The conditional intensity λ^* is defined as

$$\lambda^*(t) = \mu(t) + \sum_{t_i<t} g(t - t_i). \tag{2.0.3}$$

In the Hawkes process we can consider μ to be the background rate, and g to be a triggering function which increases the intensity of the process in the aftermath of an event. Cases where g can take negative values have been investigated in the context of non-linear Hawkes processes [16], but as $\lambda^* \geq 0$ is a requirement, we will assume throughout that g and μ are non-negative.

A common form for the triggering function g is known as the ETAS (epidemic type aftershock sequence) model

$$g(t) = \alpha\omega e^{-\omega t}, \tag{2.0.4}$$

where t_i are the times of the events in the process, and α and ω are non-negative parameters. Such models were developed in the context of investigating earthquakes, where the probability of an aftershock event happening in the aftermath of an earthquake would be increased [52]. In particular, ω governs the speed at which the intensity function decays after an event, and μ in equation (2.0.3) is the background rate [1]. We show below that α gives the expected number of events which an event triggers.

The Hawkes process can be viewed as a branching process [37], which will inform us when it comes to simulating such processes. Events occur according to the background rate μ. An event which happens at time t_i produces offspring events at the rate $g(t - t_i)$ for times $t > t_i$ independently of each other [37]. The direct offspring of the background events are known as *first-generation offspring*, and these individual events can then trigger further events independently of each other, which are known as *second-generation offspring*, and so on. A simple demonstration of such a process is shown in Figure 2.1.

[1] In some of the literature this triggering function is given in the equivalent form $\alpha e^{-\beta t}$, with the interpretation that each arrival increases the intensity by α, with this influence decaying at rate β [37]. We use the form $\alpha\omega e^{-\omega t}$ because it has the useful property that each event is expected to trigger α events.

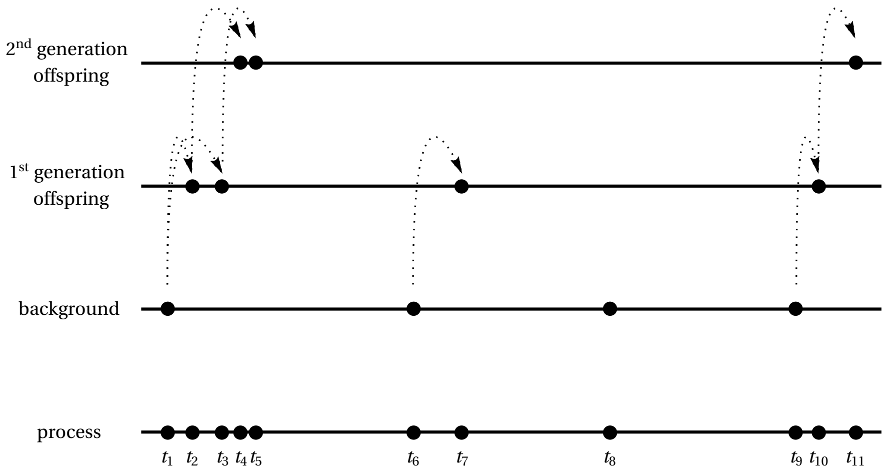
2nd generation offspring
1st generation offspring
background
process
t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11}

Figure 2.1: A simple demonstration of a Hawkes process. Background events may or may not trigger *first-generation offspring* events, which in turn can trigger *second-generation offspring* events, and so on. In Hawkes processes we tend to see a greater clustering of events than we would expect to see in a simple Poisson process.

The number of events a single event is said to trigger in the next generation is a Poisson process with mean n, where n is known as the *branching ratio* [37]. The branching ratio of the ETAS model (2.0.4) is

$$n = \int_0^\infty \alpha\omega e^{-\omega t}\,\mathrm{d}t = \alpha \tag{2.0.5}$$

and can be thought of as the expected number of events in the first-generation offspring of an event, or alternatively as the proportion of events inside the model that are triggered events when $n \leq 1$.

It can be shown that in the defective case, i.e. when $n < 1$ [37], and when the background rate μ is constant,

$$\mathbb{E}[\lambda^*(t)] \to \frac{\mu}{1-n}, \quad \text{as } t \to \infty. \tag{2.0.6}$$

The expected number of descendant events from an initial event is $\sum_{i=1}^{\infty} n^i$, therefore when $n < 1$, the expected number of descendant events from an initial event is $\frac{n}{1-n}$. When $n \geq 1$ the process explodes, as each event is expected to produce an infinite number of descendants.

2.1 Finding Parameters

When faced with a dataset that is believed to have arisen from a point process, our primary concern is typically the inverse problem; inferring estimates for the model parameters and comparing a set of putative models. The standard method for calibration is to estimate the parameters which maximise the likelihood function of the model.

Given point process data $(t_1, t_2, \dots, t_n)$ on an interval $[0, T)$, the likelihood function is given by

$$L = \left(\prod_{i=1}^{n} \lambda^*(t_i) \right) \exp\left(- \int_0^T \lambda^*(t)\, dt \right). \tag{2.1.1}$$

We provide a derivation of this, along the lines of Rasmussen [57], as follows. The conditional intensity function can be defined as

$$\lambda^*(t) = \frac{f^*(t)}{1 - F^*(t)}, \tag{2.1.2}$$

where f^* is the conditional density, and F^* is the corresponding cumulative distribution function. From this we get

$$\lambda^*(t) = \frac{\frac{\mathrm{d}}{\mathrm{d}t} F^*(t)}{1 - F^*(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \log(1 - F^*(t)). \tag{2.1.3}$$

Integrating both sides from t_n to t, where t_n is the last event before t, we get

$$\int_{t_n}^{t} \lambda^*(s)\, \mathrm{d}s = -(\log(1 - F^*(t)) - \log(1 - F^*(t_n))) = -\log(1 - F^*(t)), \tag{2.1.4}$$

as $F^*(t_n) = 0$ (as t_{n+1} occurs at t_n with probability zero). This can be rewritten as

$$F^*(t) = 1 - \exp\left(- \int_{t_n}^{t} \lambda^*(s)\, \mathrm{d}s \right). \tag{2.1.5}$$

Combining this with (2.1.2) gives

$$f^*(t) = \lambda^*(t)(1 - F^*(t)) = \lambda^*(t) \exp\left(- \int_{t_n}^{t} \lambda^*(s)\, \mathrm{d}s \right). \tag{2.1.6}$$

The likelihood function is the joint density function of all the points $(t_1, t_2, \ldots, t_n)$ on $[0, T)$,

$$L = f^*(t_1) f^*(t_2) \ldots f^*(t_n)(1 - F^*(T)). \tag{2.1.7}$$

Using (2.1.6) and (2.1.2) we can obtain the likelihood function (2.1.1).

As shown in [54], we therefore find the log-likelihood of the Hawkes process (2.0.3) with constant background rate μ and ETAS triggering function (2.0.4) to be

$$\begin{aligned} \log L &= \sum_{i=1}^{n} \log\left(\lambda^*(t_i)\right) - \int_0^T \lambda^*(t)\,dt \\ &= \sum_{i=1}^{n} \log\left(\mu + \alpha\omega \sum_{t_j<t_i} e^{-\omega(t_i-t_j)}\right) - \mu T - \sum_{i=1}^{n} \int_{t_i}^{T} \alpha\omega e^{-\omega(T-t_i)}\,dt \quad (2.1.8) \\ &= \sum_{i=1}^{n} \log\left(\mu + \alpha\omega \sum_{t_j<t_i} e^{-\omega(t_i-t_j)}\right) - \mu T + \sum_{i=1}^{n} \alpha(e^{-\omega(T-t_i)} - 1). \quad (2.1.9) \end{aligned}$$

Standard iterative methods may not be ideally suited to finding the parameters which maximise the likelihood of a dataset given the flatness of the likelihood function that is sometimes encountered, as shown by Veen and Schoenberg [70]. Another issue is that the log-likelihood can often be multimodal, meaning numerical methods can often converge on local maxima.

The expectation-maximisation algorithm (EM-algorithm) has been found to alleviate some of these problems. The EM-algorithm is an iterative algorithm which computes maximum likelihood estimators [19]; in the case of the Hawkes process, the branching structure and the fact we don't know whether each event is a background event or a triggered event is regarded as the incomplete data. We will give an explanation for why the EM-algorithm can be used for Hawkes processes as follows.

If we know which events are background events and which have been triggered by other events, we can work with the complete data log-likelihood $L_{\text{cd}}(\mu, g)$, which can be written as

$$\begin{aligned}L_{\mathrm{cd}}(\mu, g) =& \underbrace{\sum_{i=1}^{n} U_{ii} \log(\mu(t_i)) - \int_0^T \mu(t)dt}_{l_{\mathrm{back}}} \\ &+ \underbrace{\sum_{i>j}^{n} U_{ij} \log(g(t_i - t_j)) - \sum_{i=1}^{n} \int_{t_i}^{T} g(t - t_i)dt}_{l_{\mathrm{trig}}}, \qquad (2.1.10)\end{aligned}$$

where

$$U_{ii} = \begin{cases} 1, & \text{if } i \text{ is a background event} \\ 0, & \text{otherwise} \end{cases} \qquad (2.1.11)$$

$$U_{ij} = \begin{cases} 1, & \text{if } i \text{ is caused by } j \\ 0, & \text{otherwise.} \end{cases} \qquad (2.1.12)$$

We see that the complete data log-likelihood L_{cd} can be separated into two parts: the likelihood l_{back} relating to the background events, and the likelihood l_{trig} relating to the triggered events [38]. As we never know with certainty whether an event is a background event or has been triggered by others, we will use the expected value $\mathbb{E}(L_{\mathrm{cd}}(\mu, g))$ of the complete data log-likelihood

$$\begin{aligned}\mathbb{E}(L_{\mathrm{cd}}(\mu, g)) =& \sum_{i=1}^{n} p_{ii} \log(\mu(t_i)) - \int_0^T \mu(t)dt \\ &+ \sum_{i>j}^{n} p_{ij} \log(g(t_i - t_j)) - \sum_{i=1}^{n} \int_{t_i}^{T} g(t - t_i)dt, \qquad (2.1.13)\end{aligned}$$

where

$$p_{ii} = \frac{\mu(t_i)}{\mu(t_i) + \sum_{j=1}^{i-1} g(t_i - t_j)}, \qquad (2.1.14)$$

$$p_{ij} = \frac{g(t_i - t_j)}{\mu(t_i) + \sum_{j=1}^{i-1} g(t_i - t_j)}. \qquad (2.1.15)$$

The probability p_{ii} that an event t_i is a background event is just the ratio of the background rate to the total intensity at time t_i, and similarly the probability p_{ij}

that an event t_i was caused by an event t_j is the ratio of the triggering function contribution of the event t_j to the total intensity at time t_i.

The expected value $\mathbb{E}(L_{\text{cd}}(\mu, g))$ of the complete data log-likelihood of the Hawkes process (2.0.3) with constant background rate μ and the ETAS triggering function (2.0.4) is

$$\begin{aligned}
\mathbb{E}(L_{\text{cd}}(\mu, g)) =& \sum_{i=1}^{n} p_{ii} \log(\mu) - \int_0^T \mu dt \\
&+ \sum_{i>j}^{n} p_{ij} \log(\alpha\omega e^{-\omega(t_i - t_j)}) - \sum_{i=1}^{n} \int_{t_i}^{T} \alpha\omega e^{-\omega(t-t_i)} dt \\
=& \sum_{i=1}^{n} p_{ii} \log(\mu) - \mu T \\
&+ \sum_{i>j}^{n} p_{ij} \log(\alpha\omega e^{-\omega(t_i - t_j)}) - \alpha \sum_{i=1}^{n} (1 - e^{-\omega(T-t_i)}). \qquad (2.1.16)
\end{aligned}$$

Taking partial derivatives of this expression with respect to the parameters $\boldsymbol{\theta} = (\alpha, \mu, \omega)$, we obtain

$$\frac{\partial \mathbb{E}(L_{\text{cd}}(\boldsymbol{\theta}))}{\partial \mu} = \frac{1}{\mu} \sum_{i=1}^{n} p_{ii} - T, \qquad (2.1.17)$$

$$\frac{\partial \mathbb{E}(L_{\text{cd}}(\boldsymbol{\theta}))}{\partial \alpha} = \frac{1}{\alpha} \sum_{i>j}^{n} p_{ij} + \sum_{i=1}^{n} e^{-\omega(T-t_i)} - n, \qquad (2.1.18)$$

$$\frac{\partial \mathbb{E}(L_{\text{cd}}(\boldsymbol{\theta}))}{\partial \omega} = \sum_{i>j}^{n} p_{ij} \left(\frac{1}{\omega} - (t_i - t_j) \right) + \alpha \sum_{i=1}^{n} (T - t_i) e^{-\omega(T-t_i)}. \qquad (2.1.19)$$

Following the approximation $\alpha \sum_{i=1}^{n} (1 - e^{-\omega(T-t_i)}) \approx \alpha n$ suggested in [38] for $\omega << T$, the last two of these equations become

$$\frac{\partial \mathbb{E}(L_{\text{cd}}(\boldsymbol{\theta}))}{\partial \alpha} = \frac{1}{\alpha} \sum_{i>j}^{n} p_{ij} - n, \qquad (2.1.20)$$

$$\frac{\partial \mathbb{E}(L_{\text{cd}}(\boldsymbol{\theta}))}{\partial \omega} = \sum_{i>j}^{n} p_{ij} \left(\frac{1}{\omega} - (t_i - t_j) \right). \qquad (2.1.21)$$

Setting these partial derivatives to 0, the EM-algorithm for the ETAS model can be developed, as seen in much of the literature (e.g. [47, 49, 70, 73]).

2.2 EM-algorithm

The EM-algorithm begins with an initial guess $\boldsymbol{\theta}^{(0)} = (\alpha^{(0)}, \mu^{(0)}, \omega^{(0)})$. The algorithm iterates between an expectation step and a maximisation step [28].

Expectation Step

The expectation step at the $(k+1)$th iteration uses the parameter estimates $\boldsymbol{\theta}^{(k)}$ to estimate the probabilities $p_{ii}^{(k+1)}$ that event i was a background event, and $p_{ij}^{(k+1)}$ that it was caused by event j for $i > j$ as follows

$$p_{ii}^{(k+1)} = \frac{\mu^{(k)}}{\mu^{(k)} + \sum_{j=1}^{i-1} g(t_i - t_j; \boldsymbol{\theta}^{(k)})}, \tag{2.2.1}$$

$$p_{ij}^{(k+1)} = \frac{g(t_i - t_j; \alpha^{(k)}, \omega^{(k)})}{\mu^{(k)} + \sum_{j=1}^{i-1} g(t_i - t_j; \boldsymbol{\theta}^{(k)})}. \tag{2.2.2}$$

Maximisation Step

The maximisation step computes the parameters $\boldsymbol{\theta}^{(k+1)} = (\mu^{(k+1)}, \alpha^{(k+1)}, \omega^{(k+1)})$ which maximise the expected complete data log-likelihood according to the probabilities calculated in the expectation step. These are found as

$$\mu^{(k+1)} = \frac{\sum_{i=1}^{n} p_{ii}^{(k+1)}}{T}, \tag{2.2.3}$$

$$\alpha^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}}{n}, \tag{2.2.4}$$

$$\omega^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}}{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}(t_i - t_j)}. \tag{2.2.5}$$

EM-algorithm

The EM-algorithm alternates between these two steps until some convergence criterion is achieved, for example, until $\max |\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}|$ is below a certain threshold, or the computational budget is exceeded. Local convergence of the algorithm has been proved [19]. Because the problem is not convex, the algorithm suffers from the same drawback as many numerical methods, namely the tendency to converge to local maxima [71].

2.3 Goodness of Fit

Once we have fitted model parameters to a point process, the next logical step is to check the suitability of the model. For a temporal point process this can be done by residual analysis, where we compare a transformed version of the point process with a unit rate Poisson process.

2.3.1 Residual Analysis

The integrated conditional intensity function, also known as the compensator, is given by

$$\Lambda(t) = \int_0^t \lambda^*(s)\,\mathrm{d}s. \tag{2.3.1}$$

The following theorem can be applied as the basis of a test for suitability of the model fitted to the data.

Theorem 1 *If $(t_1, t_2, \dots)$ is a point process with intensity $\lambda^*(t)$, then the transformed sequence $(s_1, s_2, \dots) = (\Lambda(t_1), \Lambda(t_2), \dots)$, known as the residual sequence, is a realisation of a unit rate Poisson process [37].*

Taking advantage of this theorem we can use residual analysis to check whether the model is justified, by comparing the residual sequence $(s_1, s_2, \dots, s_n)$ of the data, with a Poisson process with unit rate.

We can compare in a number of ways. One way to check this visually is through a quantile-quantile plot (Q-Q plot), where we plot the waiting times of the residual sequence $(\Lambda(t_1), \Lambda(t_2) - \Lambda(t_1), \Lambda(t_3) - \Lambda(t_2) \dots)$ against the quantiles from the exponential distribution, with particularly noticeable deviations away from linearity being taken as an indicator against the points coming from the exponential distribution [41].

A more quantitative way of comparing the residuals with an exponential distribution is through a Kolmogorov-Smirnov test. This is carried out by plotting the cumulative distribution of the residual process as a step function, and plotting the exponential distribution on the same diagram. The maximum difference between these two distributions is then compared to the D-value for a Kolmogorov-Smirnov test. If the difference is greater than the D-value for our chosen α level (typically 0.05), we reject the null hypothesis that the residuals do come from the exponential distribution at a 95% confidence interval [34]. In statistics, the null hypothesis is a hypothesis that is rejected at a certain confidence level if the test statistic is in the rejection region [41].

We may also wish to check whether there is correlation between the intervals of $\Lambda(t_i)$. One way of doing this is by plotting the points (U_{i+1}, U_i), where $U_{i+1} = 1 - e^{-(s_{i+1}-s_i)}$. If there is no autocorrelation then we would expect to see an unstructured spread of the points in $(0,1) \times (0,1)$ [52].

Given a point process $(t_1, t_2, \ldots, t_n)$ we may first want to investigate whether a self-exciting point process is justified at all. One way to check whether the point process can be adequately modelled as a homogeneous Poisson point process is by checking the Q-Q plots of the waiting times $(t_1, t_2 - t_1, t_3 - t_2, \ldots)$ against an exponential distribution, and doing a Kolmogorov-Smirnov test of these waiting times. If the data is from a homogeneous Poisson point process then these waiting times will be exponentially distributed [52].

2.3.2 Likelihood-Ratio Test

We may wish to compare a Hawkes process with a Poisson point process in order to infer whether there is self-excitation in the data which needs to be explained with a triggering function. An hypothesis test we can use to do this is a likelihood-ratio test. This compares the difference between the likelihood L_{null} of the null model, in this case a Poisson point process, with the likelihood L_{alt} of the alternative model, in this case the Hawkes model

$$D = 2(\ln(L_{\text{alt}}) - \ln(L_{\text{null}})). \tag{2.3.2}$$

This discrepancy D is compared with a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters of the two competing models. A p-value of less than 0.05 can be taken as strong evidence to reject the null model [41].

2.3.3 Comparison of Competing Models

A more popular way to compare goodness of fit for competing to models is to compute the Akaike Information Criterion (AIC), which is taken to be

$$\text{AIC} = 2k - 2\ln(L), \tag{2.3.3}$$

where k is the number of parameters in the model and L is the likelihood. This statistic measures the fit of a model relative to other models while containing a penalty for more parameters being used, with a lower AIC taken to mean a superior fitting model [40].

Chapter 3

New Triggering Function

3.0.1 Delayed Criminal Reaction Model

We propose an alternative for the triggering function, which we will refer to as the delayed criminal response model (DCR model), namely

$$g(t) = \bar{\alpha}\bar{\omega}^2 t e^{-\bar{\omega}t}. \tag{3.0.1}$$

Unlike the triggering function (2.0.4) which immediately increases the intensity after an event with $g(0) = \alpha\omega$ and shows an exponential decay thereafter, this function has its maximum increase in intensity at a time $1/\bar{\omega}$ after an event, with $g(\frac{1}{\bar{\omega}}) = \bar{\alpha}\bar{\omega}e^{-1}$ the maximum increase in intensity caused by an event, and $\bar{\alpha}$ giving the expected number of events triggered by an event.

We argue that the DCR trigger function (3.0.1) is more realistic than the ETAS version (2.0.4). In a real world setting it may be argued that the risk of further crimes being triggered by a crime will not be increased immediately following the crime, rather the risk increases over a certain period following the crime before starting to decay. For example, it has been noted in the criminology literature that perpetrators of homicide often have a 'cooling-off' period between victims [22]. In terms of burglary several theories have been posited to suggest that repeat victimisation may be more likely to occur after a period of time has passed from the initial incident, including that a burglar will want to wait until the previous victims have replaced their stolen belongings, or that information about a 'successful' crime being committed would take time to pass between criminals [56].

With the triggering function in (3.0.1) we can also quantify the time period following a crime which will give the greatest risk of triggered events, namely this occurs at time $1/\bar{\omega}$ after the initial event. Examples of the triggering functions (2.0.4) and (3.0.1) are shown in Figure 3.1.

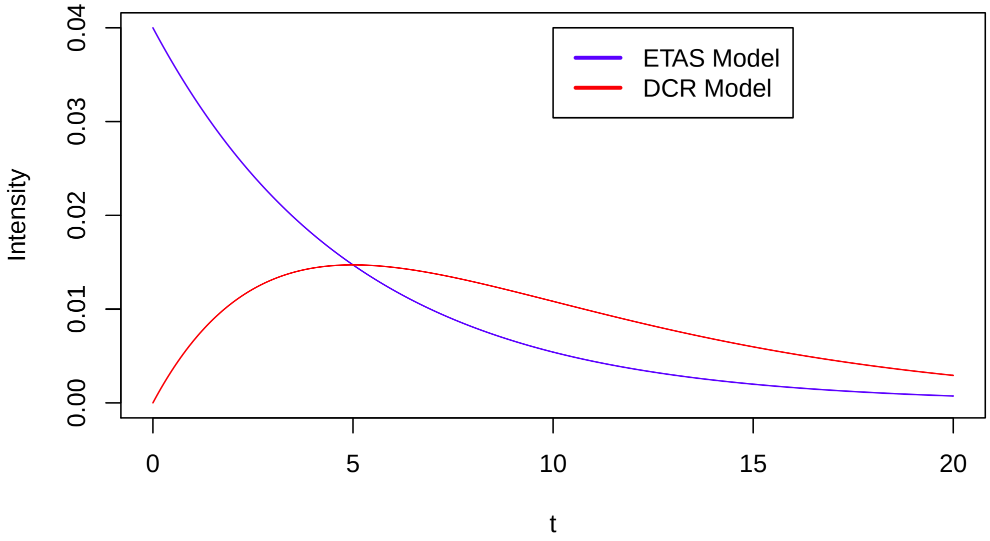

Figure 3.1: Example of the ETAS model (2.0.4) in blue, with $\alpha = 0.2$ and $\omega = 0.2$, and the DCR model (3.0.1) in red, with $\bar{\alpha} = 0.2$ and $\bar{\omega} = 0.2$

3.1 EM-algorithm for DCR model

The expectation step for the DCR model takes the same form as (2.2.1) and (2.2.2) seen for the ETAS model, with the maximisation step taking the form

$$\bar{\mu}^{(k+1)} = \frac{\sum_{i=1}^{n} p_{ii}^{(k+1)}}{T}, \tag{3.1.1}$$

$$\bar{\alpha}^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}}{n}, \tag{3.1.2}$$

$$\bar{\omega}^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} 2p_{ij}^{(k+1)}}{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{i,j}^{(k+1)}(t_i - t_j)}. \tag{3.1.3}$$

3.2 Dealing with Data

The timestamps we observe when dealing with real data can sometimes involve events taking place at the exact same time; a situation which is not possible for point processes. While with certain types of data events may be recorded happening at the exact same time due to limits on how precisely we can timestamp events, with crime data this typically happens either for practical reasons (when an event is recorded a police officer may round to the nearest minute or hour, or a witness may only be able to give a rough estimate for when a crime occurred), or simply due to the fact that in the case of a burglary it might be unclear exactly when the crime took place. Filiminov and Sornette dealt with this issue, where they looked at timestamps of financial transactions which were given to the nearest second, by randomly redistributing each event within each second [26]. Lorenzen urged caution however when using this approach, stating it introduced bias, in particular to the parameters of the exponential triggering function [42].

3.3 Simulation

We will produce some simulated data in order to test how well our EM-algorithms can recover the true parameters of a self-exciting point process. Ogata's modified thinning method is a method for simulating point processes that are specified by their conditional intensities [51]. The algorithm consists of simulating homogeneous Poisson processes with intensities which are too high, and then removing points with probability according to the conditional intensity at that point compared to the intensity of the Poisson process used to simulate [57].

In the algorithm, when we are at time t we wish to find the next point $t_i > t$. We can do this by simulating a homogeneous Poisson process on the interval $[t, T]$, with a constant intensity λ^* which fulfils $\lambda^* > \lambda(s)$ for $s \in [t, T]$. When we find the next point t_i of this Poisson process we have to decide whether to keep it as part of the point process. If $t_i > T$ then the point is not in the interval and the algorithm ends. If t_i is in the interval $[t, T]$ then we keep the point as part of the interval with probability $\lambda(t_i)/\lambda^*$, and then we start again from t_i.

Ogata's modified thinning algorithm

1. Set $t = 0$ and $n = 0$

2. While $t < T$:

a) Compute λ^*

b) Generate independent random variables $s \sim \exp(\lambda^*)$ and $U \sim \text{Unif}([0, 1])$

c) If $t + s > T$, set $t = t + s$

d) Else if $U > \lambda(t + s)/\lambda^*$, set $t = t + s$

e) Else set $n = n + 1$, $t_n = t + s$, $t = t + s$

3. Output is $(t_1, \dots, t_n)$.

To simulate the Hawkes process with triggering function $g(t) = \alpha\omega e^{-\omega t}$ in (2.0.4), we can simply take $\lambda^* = \lambda(t)$ at every starting point t as the intensity function is non-increasing until new points appear.

To simulate the point process with triggering function $g(t) = \bar{\alpha}\bar{\omega}^2 te^{-\bar{\omega}t}$ in (3.0.1), we use the fact that the triggering effect is non-increasing for times $1/\bar{\omega}$ after an event, along with the fact the maximum increase in intensity attributed to an event is $\bar{\alpha}\bar{\omega}e^{-1}$, and take λ^* at point t to be:

$$\lambda^* = \lambda(t) + \sum_{t-\frac{1}{\bar{\omega}}<t_i\leq t} \bar{\alpha}\bar{\omega}e^{-1}$$

3.3.1 Simulation Results

We first generated 1000 realisations of the Hawkes process with the ETAS triggering function with parameter values $\mu = 0.1$, $\alpha = 0.2$, $\omega = 0.2$, and 1000 realisations of the DCR model with $\mu = 0.1$, $\bar{\alpha} = 0.2$, $\bar{\omega} = 0.2$ to attempt to reproduce self-exciting point processes we might observe in real crime data. We simulated these point processes with $T = 730$ and $T = 3650$ to replicate the timescale with which we might analyse crime data over, namely 2 and 10 years.

The parameter estimates of the EM-algorithm for the ETAS and the DCR models for these simulations are shown in Tables 3.1 and 3.2. Table 3.1 shows that for the simulations of the ETAS model, the EM-algorithm estimates the true μ and α values with high accuracy, with the ω parameter tending to be overestimated for the simulations with $T = 730$ and therefore less data, with the true ω level being more accurately estimated for the simulations with $T = 3650$. Similarly Table 3.2 shows that for the simulations of the DCR model, we can accurately estimate the true μ and $\bar{\alpha}$ levels, with $\bar{\omega}$ tending to be overestimated, particularly for the simulations $T = 730$ with less data, albeit they seem to not be as overestimated as the ω parameter was in the ETAS simulations. The results in Tables 3.1 and 3.2 also show that when we use the ETAS EM-algorithm to estimate parameters for the DCR simulations, and the DCR EM-algorithm to estimate parameters for the ETAS simulations, in both cases quite accurate estimations are generated for the background rate μ, and for the number of events which are triggered by prior events.

Table 3.1: 1000 ETAS simulations with parameters $\mu = 0.1$, $\alpha = 0.2$, $\omega = 0.2$

Model Fitted	Horizon	Parameter	Mean Est. Value (s.d)
ETAS	730	μ	0.1037(0.0173)
		α	0.1677(0.1070)
		ω	0.5094(1.1896)
	3650	μ	0.1023(0.0082)
		α	0.1836(0.0537)
		ω	0.2632(0.1541)
DCR	730	μ	0.1034(0.0172)
		$\bar{\alpha}$	0.1712(0.1036)
		$\bar{\omega}$	0.6867(0.8039)
	3650	μ	0.1037(0.0080)
		$\bar{\alpha}$	0.1719(0.0501)
		$\bar{\omega}$	0.5887(0.3694)

Table 3.2: 1000 DCR simulations with parameters $\mu = 0.1$, $\bar{\alpha} = 0.2$, $\bar{\omega} = 0.2$

Model Fitted	Horizon	Parameter	Mean Est. Value (s.d)
ETAS	730	μ	0.1083(0.0181)
		α	0.1312(0.1099)
		ω	0.3971(2.4177)
	3650	μ	0.1045(0.0107)
		α	0.1632(0.0748)
		ω	0.1195(0.0739)
DCR	730	μ	0.1051(0.0181)
		$\bar{\alpha}$	0.1565(0.1137)
		$\bar{\omega}$	0.3479(0.4065)
	3650	μ	0.1026(0.0095)
		$\bar{\alpha}$	0.1790(0.0651)
		$\bar{\omega}$	0.2432(0.1136)

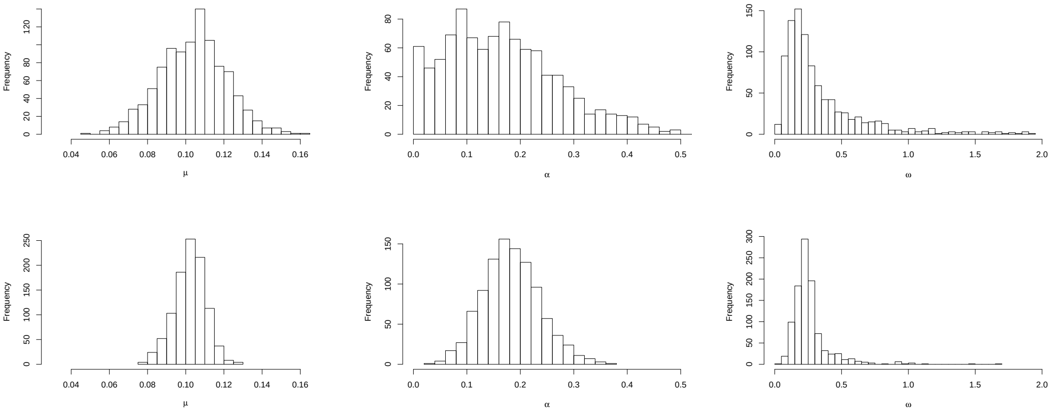

Figure 3.2: Histogram of predictions of the EM algorithm for the ETAS model for 1000 simulated realisations of the Hawkes process with ETAS triggering function and parameters $\mu = 0.1$, $\alpha = 0.2$, $\omega = 0.2$, for $T = 730$ (top) and $T = 3650$ (bottom). The predictions for μ are on the left, α in the middle, and ω on the right. 42 simulations resulted in ω predictions above 2 for $T = 730$, and 1 above 2 for $T = 3650$, which have been omitted from the histogram.

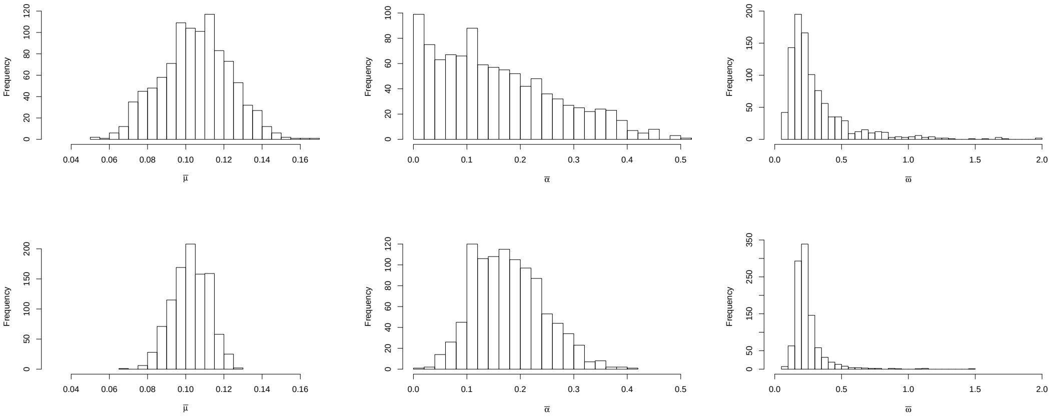

Figure 3.3: Histogram of predictions of the EM algorithm for the DCR model for 1000 simulated realisations of the Hawkes process with the DCR triggering function and parameters $\mu = 0.1$, $\hat{\alpha} = 0.2$, $\hat{\omega} = 0.3$, for $T = 730$ (top) and $T = 3650$ (bottom). The predictions for μ are on the left, $\bar{\alpha}$ in the middle, and $\bar{\omega}$ on the right. 14 simulations resulted in $\bar{\omega}$ predictions above 2 for $T = 730$, which have been omitted from the histogram.

Histograms of the estimated parameters are shown in Figures 3.2 for the simulated

ETAS process with $\mu = 0.1$, $\alpha = 0.2$, $\omega = 0.2$, and in Figure 3.3 for the simulated DCR process with $\mu = 0.1$, $\bar{\alpha} = 0.2$, $\bar{\omega} = 0.3$. It is clear from these histograms that as the length of the simulations increases and therefore more data is generated, the estimations of the parameters become more accurate. For the ETAS simulations, there are a small number of the simulations for which a far larger estimate of ω is observed than the true value of ω from which the simulations were generated from. These large ω values tend to be generated when a small number of events take place extremely close to one another, and the algorithm detects that one of the events trigegred the other with high certainty, resulting in a high ω estimate, and often also a low estimate of α, where the algorithm detects that only the small number of events which happen extremely soon after another have been triggered. As we increase the horizon from $T = 730$ to $T = 3650$ fewer of these very large ω values are detected by the EM-algorithm.

For the DCR simulations, there are far fewer large ω values estimated. From the longer simulations, none of the 1000 simulations result in $\bar{\omega}$ estimations above 2. Although the parameter estimates for $\bar{\omega}$ when estimating parameters for a DCR point process tend to be larger than when estimating ω for an ETAS process on the same data (the equation for $\bar{\omega}$ in (3.1.3) has a factor of 2 compared to the equation for ω in (2.2.5)), we are less likely to observe very large $\bar{\omega}$ values because by construction the DCR function is 0 at $t = 0$, and so an event which happens very close to another is likely to be assumed to be a background event, rather than triggered by the previous event.

3.3.2 Detecting Self-Excitation

We ran simulations to investigate how easily point processes can be identified as having a triggering effect or otherwise. We generated 1000 homogeneous Poisson point processes with $\mu = 0.1$, and with horizons $T = 730$ and $T = 3650$, along with 1000 simulations of self-exciting point processes with ETAS and DCR triggering functions with different parameters. Table 3.3 displays the percentage of the 1000 simulations for which a Kolmogorov-Smirnov test would reject that the data has come from a homogeneous Poisson point process at a 95 % confidence interval. To make the test more realistic, for each simulation we estimated the

Table 3.3: Results of tests on 1000 simulations of homogeneous Poisson point processes with rate $\lambda = 0.1$ - how often a Kolmogorov-Smirnov test will reject that the waiting times are exponentially distributed at a 95% confidence interval, and likelihood-ratio tests showing how often the null hypothesis of the point process being a Poisson point process is rejected in favour of it being from a point process with an ETAS or DCR point process.

Horizon	Poisson rejected by K-S	Like. Test - Poisson rejected for ETAS	Like. Test - Poisson rejected for DCR
730	1.0%	1.8%	2.0%
3650	0.2%	2.1%	1.8%

constant rate λ from the number of points generated, rather than the known ground-truth rate from which the simulations were generated. The rejection rate of the residuals being exponentially distributed for the Kolmogorov-Smirnov test would be closer to 5% if we used the true λ parameter that was used to generate the process. We can see that when a Poisson point process has been generated, the hypothesis that the data is generated via a Poisson point process is rarely rejected by the Kolmogorov-Smirnov test. The percentage for which a likelihood-ratio test would reject a homogeneous Poisson process in favour of the ETAS or DCR models are also displayed, and shows that this test rarely rejects the hypothesis that the Poisson point process generates the data in favour of the ETAS or DCR self-exciting point processes.

Table 3.4 uses data simulated from an ETAS process. The first column of percentages shows how often the Kolmogorov-Smirnov rejects that the data has been generated from a homogeneous Poisson point process. We can see this test generally rejects the data comes from a homogeneous Poisson point process, with the exception being when we have $T = 730$ and $\omega = 0.2$. This shows that the biggest problem this test has with correctly rejecting that data which has been generated from a self-exciting point process could have come from a homogeneous Poisson point process is when there is a smaller amount of data, and when the triggering effect happens over a longer period, i.e. when ω is small. The second column of percentages shows the amount that the likelihood-ratio test correctly rejects a homogeneous Poisson point process in favour of the ETAS model, with likelihoods generated from the estimations of λ for the homogeneous Poisson point

Table 3.4: Results of tests on 1000 simulations of the ETAS triggering function with different parameters - how often a Kolmogorov-Smirnov test will reject that the waiting times are exponentially distributed at a 95% confidence interval, and likelihood-ratio tests showing how often the null hypothesis of the point process being a Poisson point process is rejected in favour of it being from a point process with an ETAS or DCR point process.

α	ω	T	Poisson rejected by K-S	Like. Test - Poisson rejected for ETAS	Like. Test - Poisson rejected for DCR
0.2	0.2	730	8.1%	30.0%	31.0%
0.2	0.2	3650	63.3%	94.2%	91.4%
0.2	5	730	83.1%	99.3%	47.6%
0.2	5	3650	100.0%	100.0%	62.3%
0.4	0.2	730	52.8%	86.3%	85.5%
0.4	0.2	3650	100.0%	100.0%	100.0%
0.4	5	730	100.0%	100.0%	77.3%
0.4	5	3650	100.0%	100.0%	93.7%

Table 3.5: Results of tests on 1000 simulations of the DCR triggering function with different parameters - how often a Kolmogorov-Smirnov test will reject that the waiting times are exponentially distributed at a 95% confidence interval, and likelihood-ratio tests showing how often the null hypothesis of the point process being a Poisson point process is rejected in favour of it being from a point process with an ETAS or DCR point process.

$\bar{\alpha}$	$\bar{\omega}$	T	Poisson rejected by K-S	Like. Test - Poisson rejected for ETAS	Like. Test - Poisson rejected for DCR
0.2	0.2	730	2.0%	10.8%	15.4%
0.2	0.2	3650	9.4%	55.6%	67.9%
0.2	5	730	76.7%	96.1%	63.0%
0.2	5	3650	100.0%	100.0%	84.3%
0.4	0.2	730	7.3%	48.9%	59.6%
0.4	0.2	3650	53.0%	99.8%	99.9%
0.4	5	730	99.8%	100.0%	90.3%
0.4	5	3650	100.0%	100.0%	99.9%

process by the number of points in the simulation, and from the EM-algorithms estimation of the parameters for the ETAS model. This test correctly identifies the data comes from an ETAS process with high accuracy, with the exception being for the tests with $\alpha = 0.2$, $\omega = 0.2$, $T = 730$, i.e. when there are fewer triggered points in the simulation, and when they happen over a longer period. The third column of percentages shows how often the likelihood-ratio test rejects the data has come from a homogeneous Poisson point process in favour of it being generated by a DCR triggering functon (the wrong model). The DCR model is particularly favoured by this test over a homogeneous Poisson point process when there is more triggered data ($\alpha = 0.4$ and $T = 3650$), and when the triggering effect is over a longer time period ($\omega = 0.2$).

Table 3.5 shows the same tests done with data generated from a DCR process with different parameters. We can see in the first column of percentages that the data being generated from a homogeneous Poisson point process is only consistently rejected when $\bar{\omega}$ is larger, i.e. when there is a more pronounced peak in the triggering intensity soon after an initial event. In the third column of percentages we see how often a likelihood-ratio test rejects a homogeneous Poisson point process correctly in favour of the data coming from a DCR process. This happens most consistently when we have a greater amount of triggered data ($T = 3650$ and $\alpha = 0.2$) and when the triggering function has its greatest effect over a shorter period of time ($\bar{\omega} = 5$). In the second column we can see that the likelihood-ratio test consistently rejects a homogeneous Poisson point process in favour of a ETAS triggering function for this data, particularly when $\bar{\omega}$ is high.

Table 3.6 shows average parameters recovered by the EM-algorithm for 1000 simulations of a self-exciting point process with the ETAS triggering function for varying parameter, and Table 3.7 shows results from data with the DCR triggering function with varying parameters. We can see for the ETAS triggering function that as the sample size becomes bigger for larger T the estimates for the parameters by the EM-algorithm become more accurate. We can also see the ω parameter is sometimes greatly overestimated, particularly for small sample size $T = 730$ and for small ω and α. This is likely because there are fewer points which have been triggered in these simulations, resulting in the algorithm only detecting a small number of events have been triggered close to a previous

Table 3.6: Results of the ETAS EM-algorithm on 1000 simulations of the Hawkes process with ETAS triggering function with different parameters, showing the average parameter estimations.

T	μ	Mean est μ (s.d)	α	Mean est α (s.d)	ω	Mean Est ω (s.d)
730	0.1	0.1037(0.0173)	0.2	0.1677(0.1070)	0.2	0.5094(1.1896)
3650	0.1	0.1023(0.0082)	0.2	0.1836(0.0537)	0.2	0.2632(0.1541)
730	0.1	0.0996(0.0128)	0.2	0.2031(0.0587)	5	5.6439(3.6454)
3650	0.1	0.1001(0.0054)	0.2	0.1999(0.0242)	5	5.0471(0.8695)
730	0.1	0.1089(0.0217)	0.4	0.3403(0.1177)	0.2	0.3372(0.4441)
3650	0.1	0.1029(0.0096)	0.4	0.3823(0.0526)	0.2	0.2248(0.0477)
730	0.1	0.1002(0.0113)	0.4	0.3956(0.0603)	5	5.1916(1.1082)
3650	0.1	0.1003(0.0056)	0.4	0.3980(0.0285)	5	5.0393(0.4852)

Table 3.7: Results of the ETAS EM-algorithm on 1000 simulations of the Hawkes process with DCR triggering function with different parameters, showing the average parameter estimations.

T	μ	Mean est μ (s.d)	$\bar{\alpha}$	Mean est $\bar{\alpha}$ (s.d)	$\bar{\omega}$	Mean Est $\bar{\omega}$ (s.d)
730	0.1	0.1051(0.0181)	0.2	0.1565(0.1137)	0.2	0.3479(0.4065)
3650	0.1	0.1026(0.0095)	0.2	0.1790(0.0651)	0.2	0.2432(0.1136)
730	0.1	0.1045(0.0175)	0.2	0.1622(0.0991)	5	3.1582(2.7199)
3650	0.1	0.1024(0.0088)	0.2	0.1805(0.0595)	5	4.1470(1.9463)
730	0.1	0.1128(0.0253)	0.4	0.3092(0.1319)	0.2	0.2737(0.1474)
3650	0.1	0.1055(0.0115)	0.4	0.3651(0.0640)	0.2	0.2221(0.0377)
730	0.1	0.1051(0.0204)	0.4	0.3631(0.1141)	5	4.5046(1.7315)
3650	0.1	0.1001(0.0063)	0.4	0.3976(0.0323)	5	4.9919(0.4484)

Table 3.8: Results on 1000 simulations of the ETAS triggering function with different parameter values showing which model would be preferred based on lowest AIC value. The first two rows show homogeneous Poisson point processes.

T	μ	α	ω	Preferred model by AIC		
				Poisson	ETAS	DCR
730	0.1	0	0	93.1%	3.3%	3.6%
3650	0.1	0	0	92.7%	3.2%	4.1%
730	0.1	0.2	0.2	50.4%	19.2%	30.4%
3650	0.1	0.2	0.2	2.5%	64.3%	33.2%
730	0.1	0.2	5	0.3%	86.2%	13.5%
3650	0.1	0.2	5	0%	97.8%	2.2%
730	0.1	0.4	0.2	6.2%	56.4%	37.4%
3650	0.1	0.4	0.2	0%	89.8%	10.2%
730	0.1	0.4	5	0%	91.8%	8.2%
3650	0.1	0.4	5	0%	99.8%	0.2%

event, and thereby overestimating ω, and also underestimating α. Similarly for the DCR triggering function the EM-algorithm predictions of these parameters become more accurate as the sample sizes become bigger, with the α parameter being more likely to be underestimated when there is less data, and $\bar{\omega}$ tending to be overestimated for $\bar{\omega} = 0.2$, and underestimated when $\bar{\omega} = 5$.

We can see in Tables 3.8 and 3.9 that as the sample sizes get larger, the AIC values tend to be a more accurate barometer of which of the 3 competing models (Homogeneous Poisson point process, self-exciting point process with ETAS triggering function, self-exciting point process with DCR triggering function) have produced the point process. Figure 3.4 shows a QQ-plot of an example simulation of both the waiting times and the residuals, showing a clear deviance away from the exponential distribution for the waiting times of a self-excited point process, indicating that it is indeed not from a homogeneous Poisson point process. Figure 3.5 shows an example of the graph (U_{i+1}, U_i), where $U_{i+1} = 1 - e^{-(s_{i+1} - s_i)}$, with the apparent randomness indicating there is no autocorrelation between successive points.

Table 3.9: Results on 1000 simulations of the DCR triggering function with different parameter values showing which model would be preferred based on lowest AIC value.

T	μ	$\bar{\alpha}$	$\bar{\omega}$	Preferred model by AIC		
				Poisson	ETAS	DCR
730	0.1	0.2	0.2	71.5%	5.9%	22.6%
3650	0.1	0.2	0.2	18.2%	14.7%	67.1%
730	0.1	0.2	5	1.5%	47.7%	50.8%
3650	0.1	0.2	5	0%	21.3%	78.7%
730	0.1	0.4	0.2	26.7%	11.5%	61.8%
3650	0.1	0.4	0.2	0%	9.6%	90.4%
730	0.1	0.4	5	0%	17.5%	82.5%
3650	0.1	0.4	5	0%	0.5%	99.5%

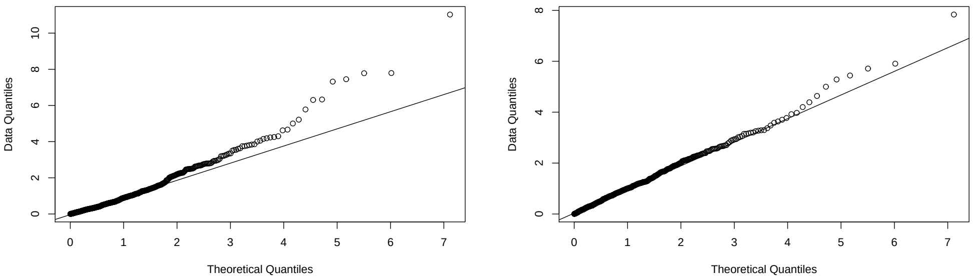

Figure 3.4: On the left is a QQ-plot of the waiting times of an ETAS process generated with parameters $\mu = 0.1$, $\alpha = 0.4$, $\omega = 0.2$, $T = 3650$, and on the right is a QQ-plot of the residuals against the exponential distribution.

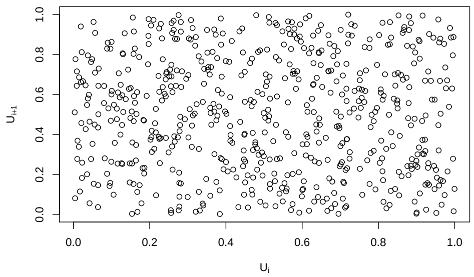

Figure 3.5: An example of the graph (U_{i+1}, U_i), where $U_{i+1} = 1 - e^{-(s_{i+1} - s_i)}$ for an ETAS process with $\mu = 0.1$, $\alpha = 0.4$, $\omega = 0.2$, $T = 3650$.

3.4 Summary

In this Chapter we have shown that for the amount of crime data we plan to make use of, it is feasible to calibrate and compare models from this class. We have shown that the method for detecting parameters for the new DCR function is generally reliable, and that we can consistently identify when point process data has a level of self-excitation.

Chapter 4

Results on Real Data

4.1 Real Crime Data

The City of Chicago has a publicly available dataset consisting of reported crimes in Chicago [4]. The police have split the city in 275 ‘beats’ as shown in Figure 4.1, with a number of beat officers assigned to each beat. We concentrate on burglaries committed in Chicago in the period up until the end of 2015. The aim is to fit self-exciting point processes to these models and successfully predict the occurrence and location of future burglaries. Figure 4.2 shows the number of recorded burglaries which took place in the 10 years between the beginning of 2006 and the end of 2015 which happened in each individual beat. Figure 4.3 displays a choropleth which indicates the number of burglaries in these 10 years which occurred in each beat according to their geographical location. There were 221,725 burglaries which were recorded as taking place between 2006 and 2015, and we can see that some beats appear to be at higher risk of burglary than others.

We fit temporal Hawkes processes with both the ETAS triggering function and the DCR triggering function to the burglary events which happen in each beat in the period before the end of 2015. As the recorded times are often just a best estimate of when a crime was committed, and often recorded to happen exactly on the hour, we add a random noise element $U \sim \text{Unif}([0, 1/24])$ to each time,

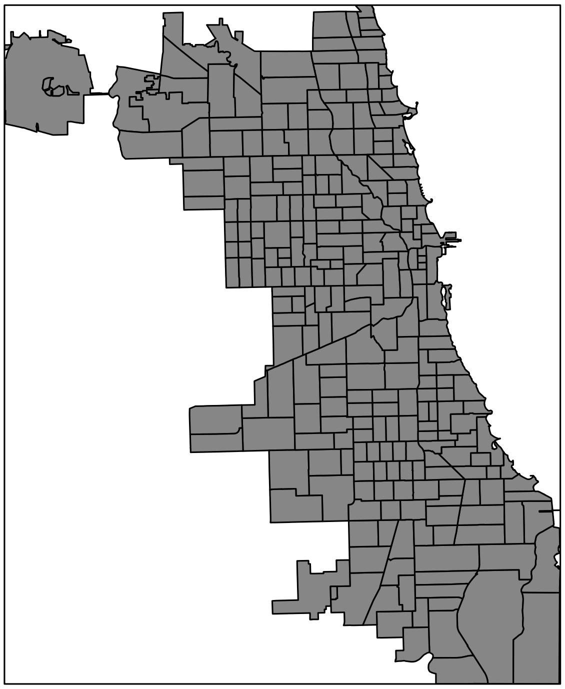

Figure 4.1: Map showing the 275 police beats in the city of Chicago.

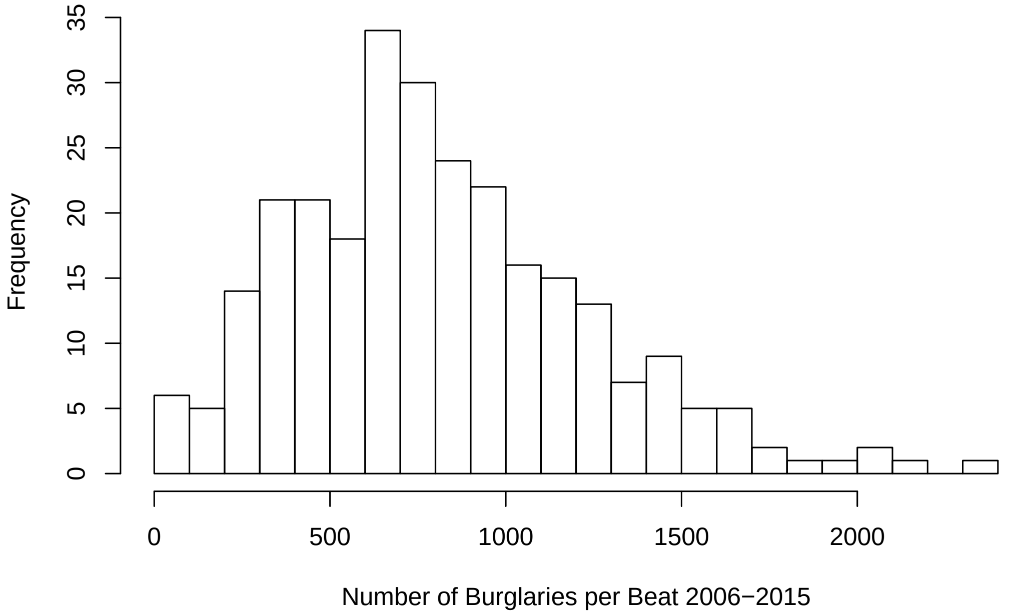

Figure 4.2: Histogram showing the number of burglaries committed in each beat between the beginning of 2006 and the end of 2015.

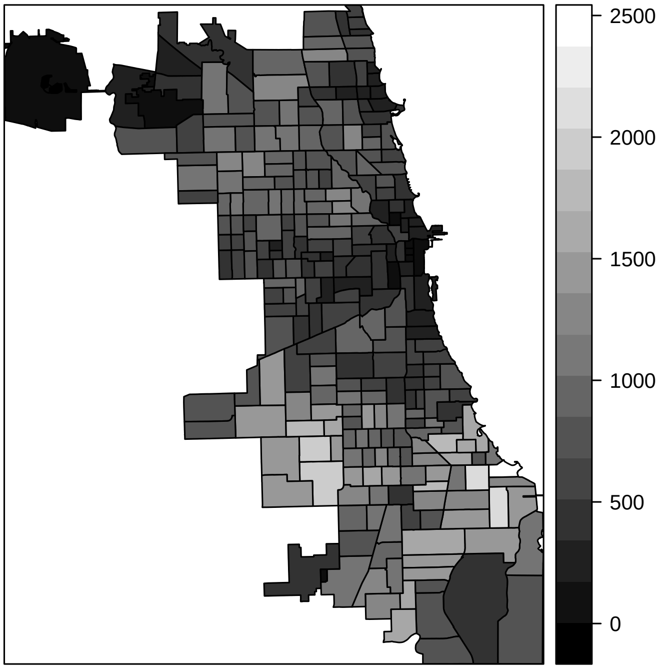

Figure 4.3: A choropleth showing the numbers of burglaries that were committed per beat between 2006 and 2015.

where the time is recorded in days. We fit each model to each beat three times, using the burglary data for the preceding 2 years, 5 years, and 10 years leading up to the end of 2015. Histograms of the μ, α and ω parameters found when fitting the ETAS model are shown in Figure 4.4, and the μ, $\bar{\alpha}$ and $\bar{\omega}$ for the DCR model are shown in Figure 4.5.

As we can see from the histograms in Figure 4.4 and Figure 4.5, when we model burglary events as a temporal point process in each beat, using only data going back two years, both the ETAS and DCR models tend to detect less of a triggering effect in the data than when we model using data over a longer term. The results in Table 4.1 show that as we use a greater amount of data, the Kolmogorov-Smirnov test rejects with greater regularity that the data comes from a Poisson point process, and the likelihood-ratio test rejects that the data is from a Poisson point process in favour of being from one of the two fitted models.

In Table 4.2 we see that while in terms of AIC a simple Poisson point process is preferred for the processes observed in the most beats, as we use more data, the ETAS model is increasingly shown to be the best fit in terms of AIC. In terms of making predictions, the trade-off comes between using enough data where we can reliably detect the underlying triggering effect, and not using data which could potentially be outdated and negatively affect how well we can predict future crimes. Figure 4.6 shows the QQ-plot for the waiting times $(t_1, t_2 - t_1, t_3 - t_2, \dots)$ against an exponential distribution for a sample beat, along with QQ-plots of the waiting times of the residual sequence for the fitted ETAS and DCR processes to the data. The QQ-plot clearly shows that there are more large waiting times than we would expect if the data from this beat came from a homogeneous Poisson point process. When we fit the ETAS and DCR point processes to this data this problem is largely resolved, with similar results being seen in the QQ-plots of the data from other beats. Figure 4.7 shows plots of (U_{i+1}, U_i), where $U_{i+1} = 1 - e^{-(s_{i+1} - s_i)}$ for fitted ETAS and DCR processes on the same beat. We observe a seemingly unstructured spread of the points which we would expect to see if there was no autocorrelation between the waiting time of the residuals.

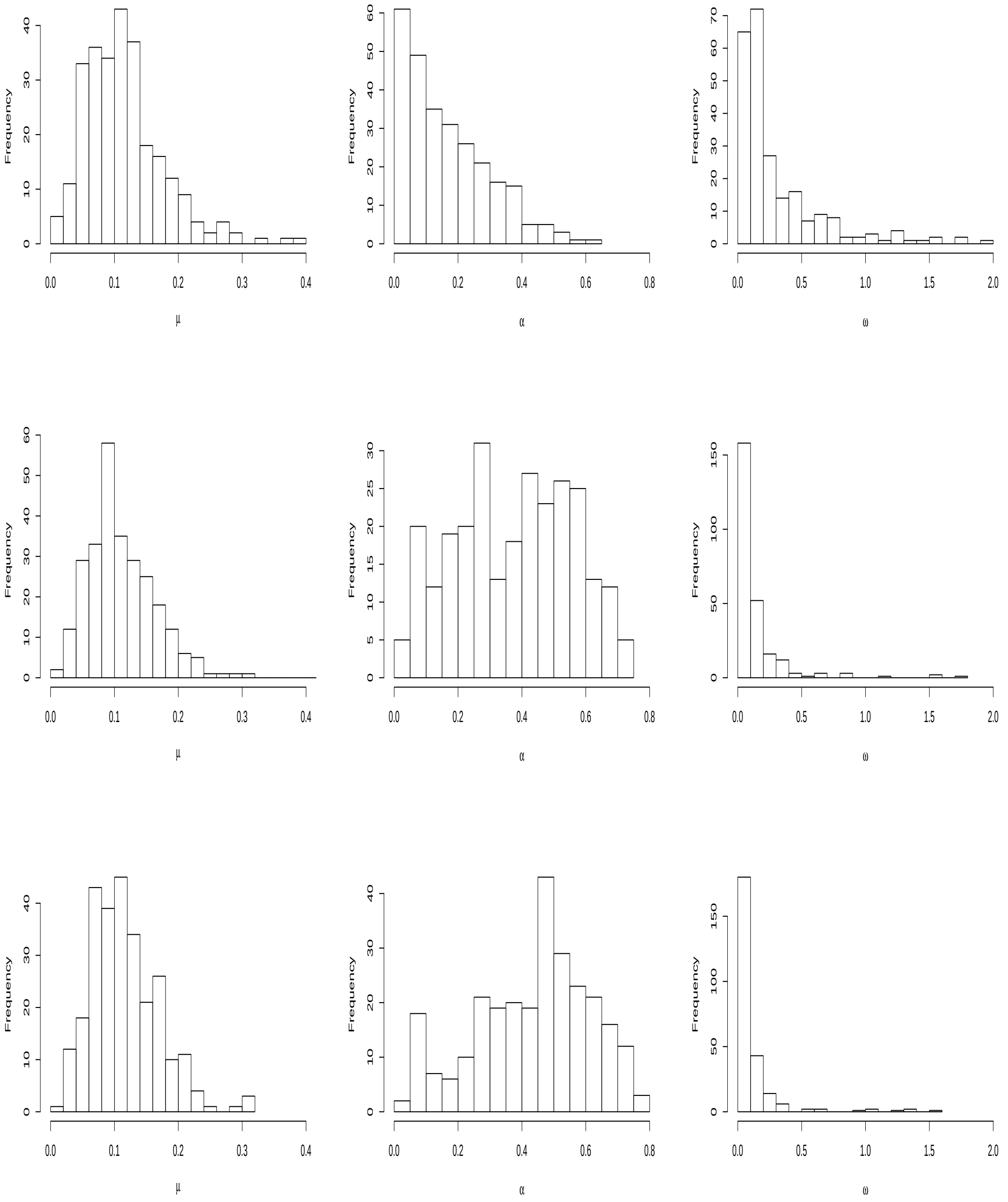

Figure 4.4: Histograms for the μ, α and ω found for each beat when fitting the ETAS model using data for 2 years (top), 5 years (middle) and 10 years (bottom) prior to the end of 2015. 32 of the beats give an ω value over 2 for the data going back 2 years, 17 beats for data going back 5 years, and 15 beats for the data going back 10 years.

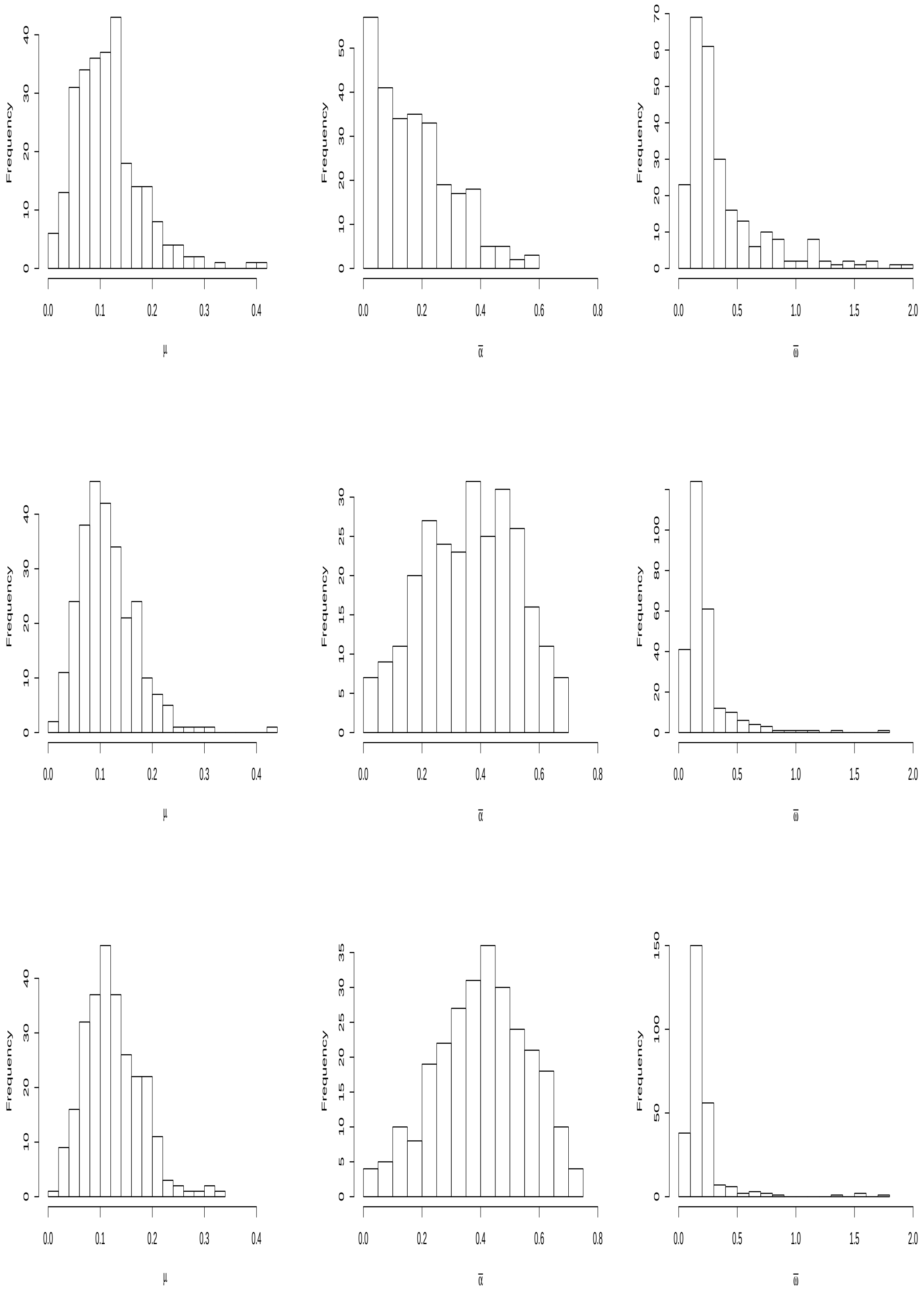

Figure 4.5: Histograms for the μ, $\bar{\alpha}$ and $\bar{\omega}$ found for each beat when fitting the DCR model using data for 2 years (top), 5 years (middle) and 10 years (bottom) prior to the end of 2015. 14 of the beats give an $\bar{\omega}$ value over 2 for the data going back 2 years, 2 beats for data going back 5 years, and for none of the beats for the data going back 10 years.

Table 4.1: Results on burglary data separated into 269 beats over different time periods up until the end of 2015 - how often a Kolmogorov-Smirnov test will reject that the waiting times in the beats are exponentially distributed at a 95% confidence interval, and likelihood-ratio tests showing how often the null hypothesis of the point process being a Poisson point process is rejected in favour of it being from a point process with an ETAS or DCR point process.

Horizon	Poisson rejected by K-S	Like. Test - Poisson rejected for ETAS	Like. Test - Poisson rejected for DCR
730	3.3%	31.2%	24.2%
1826	41.6%	88.5%	83.6%
3652	79.2%	98.5%	95.9%

Table 4.2: Results on the burglary data for 269 beats, indicating which model would be preferred based on lowest AIC value.

T	Preferred model by AIC		
	Poisson	ETAS	DCR
730	49.8%	31.6%	18.6%
1826	4.5%	78.4%	17.1%
3652	0.4%	95.2%	4.5%

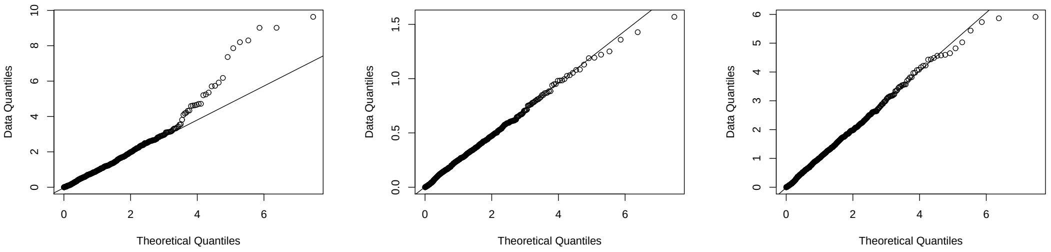

Figure 4.6: On the left is a QQ-plot of the waiting times of the point process of a randomly selected beat, in the middle is a QQ-plot of the residuals for the same beat after the ETAS model is fitted to it, and on the right is a QQ-plot of the residuals after the DCR model is fitted.

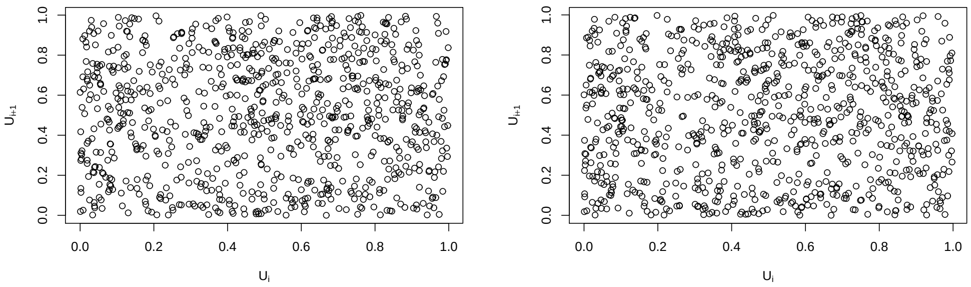

Figure 4.7: On the left is the graph (U_{i+1}, U_i), where $U_{i+1} = 1 - e^{-(s_{i+1}-s_i)}$ for the ETAS process fitted to a randomly selected beat, and on the right is this graph after the DCR model has been fitted to this beat.

To look at the predictive performance of the competing models, starting from the beginning of 2016, after fitting both of the models to each beat, we order the beats in terms of the value of the predicted intensity for each beat using the events which have happened up until that point. We then look at the proportion of overall crimes over the next hour that were correctly 'predicted' by the model at each point of the ranking. For example, using the top 1% of beats (3 beats) in terms of intensity, we record the proportion of the recorded crime in Chicago occurs in these beats over the next hour. We then recalculate the predicted intensity an hour later for each beat and rank the beats again, and look at how well we have predicted the next hours' crime, and so on. After doing this over a period of time we can create a graph as seen in previous papers, e.g. [47],[48], which is analogous to a ROC curve, where we see how many of the total crimes we can predict in relation to how susceptible we have rated each location as being to a future crime. In practical terms, this means if law enforcement only had the resources to police a certain proportion of the city, we will see how well our model would allocate the limited resources to locations where crime is going to take place in the future (making the assumption that having law enforcement resources where crime is going to take place will be productive in either stopping the crime taking place, or apprehending the perpetrators).

We make predictions hourly for the first 3 months of 2016. We look at the proportion of crime which takes place in the top 1% of beats (3 beats) we rank each hour, top 10% (27 beats), and top 20%. Along with those, we also calculate the 'area under the curve'(AUC) which is commonly calculated for ROC curves, where the higher the value, the more successfully the model has predicted future events (this is bounded by 0 and 1). The prediction results are shown in Table 4.3, with an example of the 'ROC curves' shown in Figure 4.8 for the beats using data from the last 5 years for the ETAS and the DCR models, and for a homoegeneous Poisson point process using data for the preceding 2 years. We can see from Table 4.3 that the ETAS and DCR models show the best prediction results when using data from the previous 5 years, with generally the ETAS models predicting slightly better than their DCR counterparts. If you are modelling each beat as a homogeneous Poisson point process, the predictions are more successful while using more recent data (data from the previous month), with these predictions being slightly worse than both the ETAS and the DCR models. We can see also that the area under the curve statistic in the last column of this table shows using less data for the Poisson process gives a worse overall predictive performance, as there are many beats where very few events have taken place over the previous month. Overall the area under the curve statistic does not tell us a great deal about the predictions, as in a practical situation there is far more importance in predicting burglary in the few highest risk areas, as opposed to predicting accurately in beats which are given lower priority.

Table 4.3: Predictive results for burglary using different models and different amounts of data to calibrate the models parameters. The percentage of burglaries which are predicted by the top 1%, 10% and 20% of beats by intensity every hour over the first 3 months of 2016 is shown, along with the total area under the curve.

Model	Data	Top 1%	Top 10%	Top 20%	AUC
ETAS	10 years	3.68%	22.47%	38.54%	0.6447
ETAS	5 years	3.61%	23.00%	38.60%	0.6456
ETAS	2 years	3.84%	22.10%	37.44%	0.6430
DCR	10 years	3.18%	22.37%	38.40%	0.6410
DCR	5 years	3.45%	22.66%	37.54%	0.6425
DCR	2 years	3.45%	22.07%	36.85%	0.6416
Poisson	2 years	3.38%	21.07%	35.98%	0.6370
Poisson	1 year	3.05%	20.87%	36.08%	0.6350
Poisson	1 month	3.15%	22.23%	37.47%	0.6240

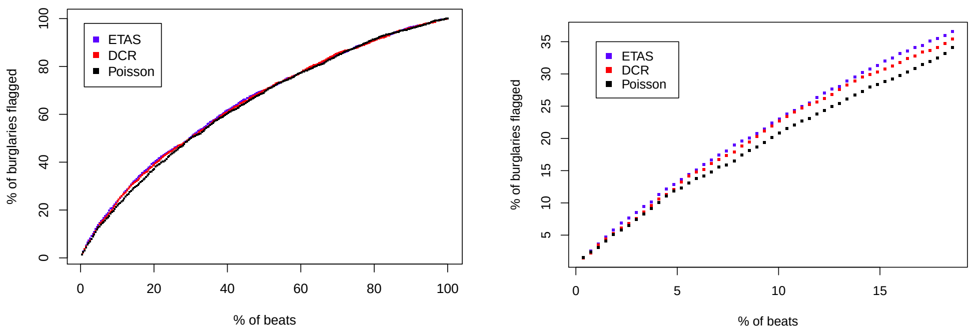

Figure 4.8: ROC like curves showing number of crimes captured if beats targeted every hour according to their intensity with each model, ETAS model in blue, DCR model in red, and a Poisson point process in black.

4.2 Results for Grids

In order to further look at how successfully we can predict the location of future events, we split Chicago into 1km by 1km grids, as shown in Figure 4.9, 710 of which cover the city of Chicago. As we have split Chicago into smaller areas than the beats, on average we have less data for each separate 1km x 1km grid. Using the threshold that we require ten events to have occurred in a grid in order to fit a point process to the data, going back 10 years from the beginning of 2016 we find that 575 of the grids have had 10 or more reported burglaries in that time period.

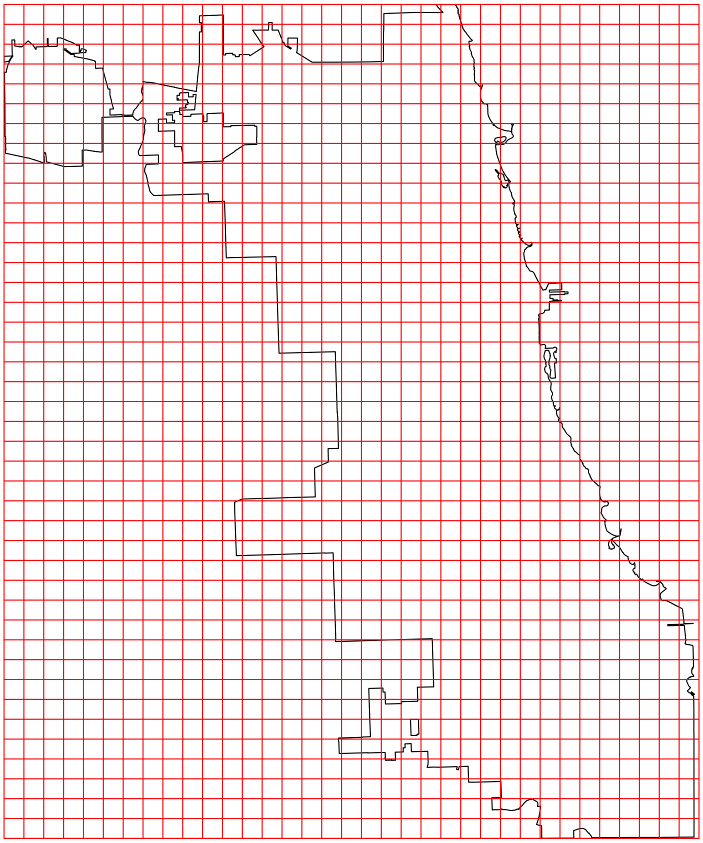

Figure 4.9: Map showing Chicago separated into 1km x 1km grids, 710 of which fall within the boundaries of Chicago.

In an identical way to before, we fit a temporal point process to each grid using historical data, and rank the grids hourly accordingly. For the grids with insufficient data to fit a model to, we rank at random below the 575 grids. The corresponding 'ROC curves' are shown in Figure 4.10, and the predictions in Table 4.4.

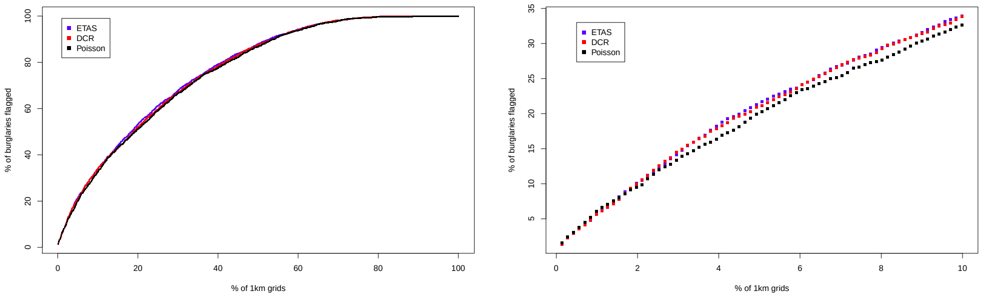

Figure 4.10: ROC like curves showing number of crimes captured if 1km x 1km grids are targeted every hour according to their intensity with each model, with the ETAS model in blue, DCR model in red, and a Poisson point process in black.

Table 4.4: Predictive results for burglary using different models for 1km by 1km grids. The percentage of burglaries which are predicted by the top 1%(7 grids), 10%(71 grids) and 20%(142 grids) of 1km by 1km by intensity every hour over the first 3 months of 2016 is shown, along with the total area under the curve.

Model	Top 1%	Top 10%	Top 20%	AUC
ETAS	5.96%	33.96%	53.18%	0.7686
DCR	5.67%	33.90%	52.02%	0.7653
Poisson(1 year)	6.06%	32.70%	51.19%	0.7615
Poisson(1 month)	5.73%	31.58%	48.67%	0.7207

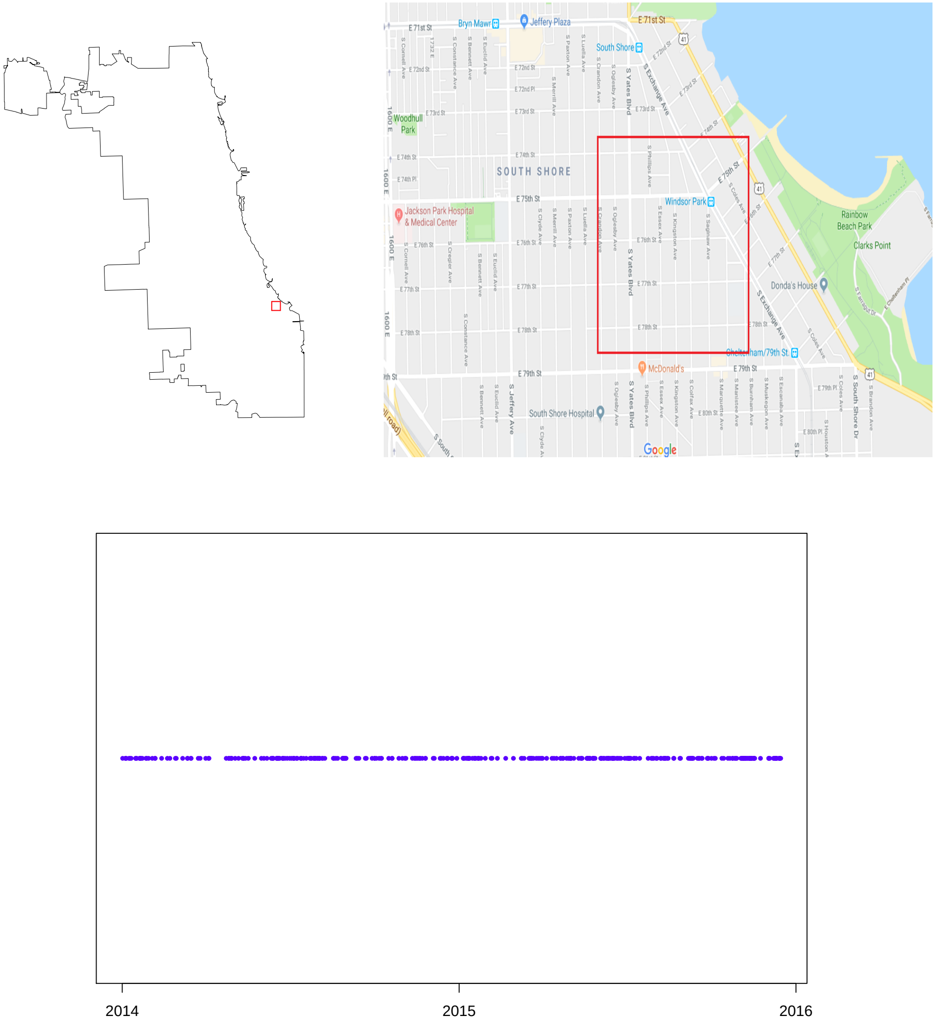

Figure 4.11: The location of the grid with the greatest number of reported burglaries over the 10 years prior to 2016, along with a map displaying location of this grid [6] and a dotplot showing the timing of reported burglaries in 2014 and 2015. Map data ©2020 Google.

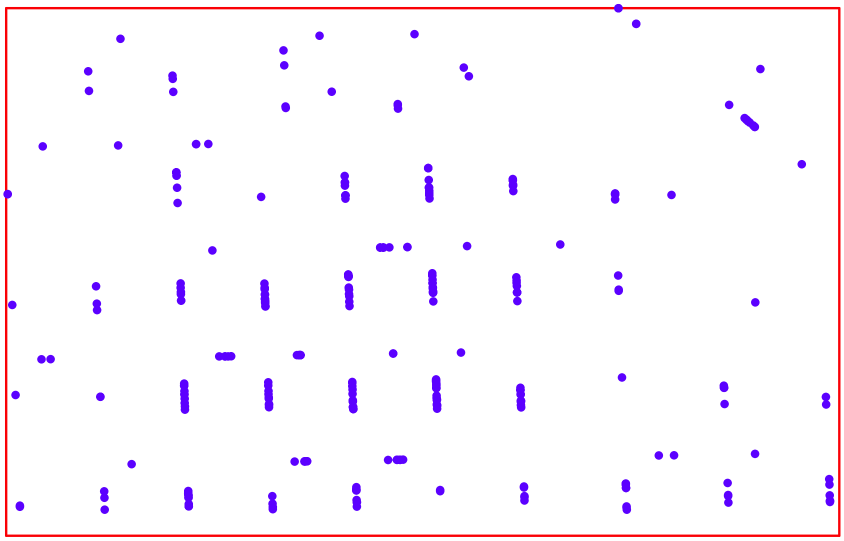

Figure 4.12: Location of events that occurred in 2014 and 2015 in the 1km by 1km grid with the highest burglary levels.

We can see from Table 4.4 and Figure 4.10 that the ETAS model appears to be the most successful at predicting which 1km x 1km grid burglaries will occur in, with the DCR model close behind. Both these self-exciting models are more successful at predicting future crime than a homogeneous Poisson point process.

Figure 4.11 shows the location of the 1km x 1km which has the most recorded burglaries in 2014 and 2015 in Chicago, and a dotplot indicating the timing of each recorded event. In Figure 4.12 we can see that even within the grid with the highest burglary rate in Chicago, the incidents do not appear to be distributed evenly across the area. Refining the grids further to 200m x 200m grids (the length of a typical block in Chicago), we find we can increase the success with which we predict future events of burglary even further. There is a trade off to be made - while refining the grid means we can identify the location of crime at a more local level, it also means we have less data for each grid when attempting to fit a temporal point process. The corresponding ROC curves are shown in Figure 4.13, and predictive results shown in Table 4.5. We can see that again the ETAS model slightly outperforms the DCR model, while both outperform the homogeneous Poisson point process. Both models consistently outperform a homogeneous Poisson point process fitted to each 200m x 200m grid.

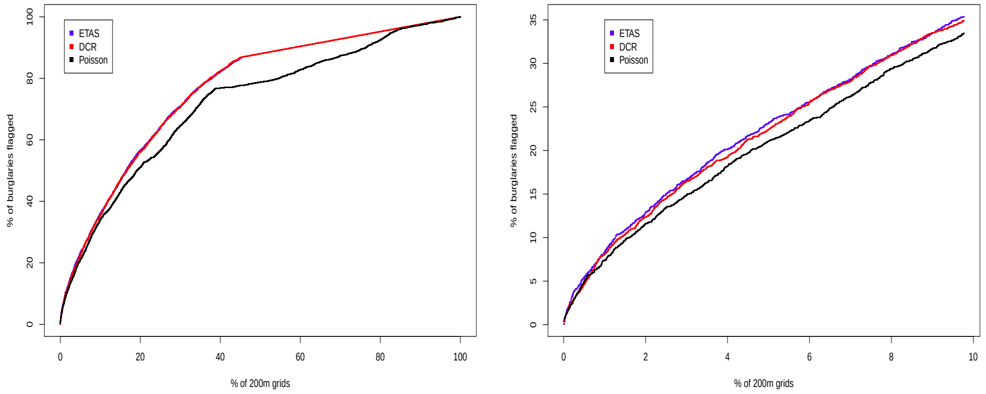

Figure 4.13: ROC like curves showing number of crimes captured if 200m x 200m grids are targeted every hour according to their intensity with each model, with the ETAS model in blue, DCR model in red, and a Poisson point process in black.

Table 4.5: Predictive results for burglary using different models for 200m by 200m grids. The percentage of burglaries which are predicted by the top 1%(156 grids), 5%(779 grids), 10%(1559 grids) and 20%(3117 grids) of 200m by 200m by intensity every hour over the first 3 months of 2016 is shown, along with the total area under the curve.

Model	Top 1%	Top 5%	Top 10%	Top 20%	AUC
ETAS	8.28%	23.06%	35.92%	56.30%	0.7642
DCR	8.08%	22.23%	35.22%	55.93%	0.7635
Poisson(1 year)	7.39%	20.94%	33.83%	50.96%	0.7168
Poisson(1 month)	5.80%	19.32%	22.73%	36.81%	0.7007

4.3 Spatio-Temporal Models

Spatio-temporal point processes [58] are an extension to the temporal point processes seen previously, where spatial coordinates are taken into account. These models have been used previously in crime modelling, most notably in modelling gun-related crime in Chicago [47], and burglary within Los Angeles [48]. The conditional intensity functions for such models are analogous to those for the temporal point process. Given a point process with times t_i and associated locations (x_i, y_i), the conditional intensity function of an unmarked spatial-temporal point process can be defined as

$$\lambda^*(x,y,t) = \lim_{\Delta x, \Delta y, \Delta t \to 0} \frac{\mathbb{E}(N(x+\Delta x, y+\Delta y, t+\Delta t) - N(x,y,t)|\mathcal{H}_t)}{\Delta x \Delta y \Delta t}. \tag{4.3.1}$$

Given a crime data set with times t_i and locations (x_i, y_i), building on the strictly temporal process shown in (2.0.3) a spatial-temporal point process can be defined with the intensity function

$$\lambda(x,y,t) = \mu(x,y) + \sum_{t>t_i} g(t-t_i) f(x-x_i, y-y_i). \tag{4.3.2}$$

If we use the ETAS function to define the exponential decay of the triggering function in time, and we assume f to be Gaussian in space, we have the spatio-temporal triggering function

$$g(t)f(x,y) = \alpha\omega e^{-\omega t} \cdot \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{4.3.3}$$

Following similar logic to that in Section 2.1, the expectation step for such a spatio-temporal triggering function would be

$$p_{ii}^{(k+1)} = \frac{\mu^{(k)}}{\mu^{(k)} + \sum_{j=1}^{i-1} g(t_i - t_j) f(x_i - x_j, y_i - y_j)}, \tag{4.3.4}$$

$$p_{ij}^{(k+1)} = \frac{g(t_i - t_j; \alpha^{(k)}, \omega^{(k)})}{\mu^{(k)} + \sum_{j=1}^{i-1} g(t_i - t_j) f(x_i - x_j, y_i - y_j)}, \tag{4.3.5}$$

and the maximisation step

$$\mu^{(k+1)} = \frac{\sum_{i=1}^{n} p_{ii}^{(k+1)}}{T \cdot A}, \tag{4.3.6}$$

$$\alpha^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}}{n}, \tag{4.3.7}$$

$$\omega^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}}{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}(t_i - t_j)} \tag{4.3.8}$$

$$\sigma^{2(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}((x_i - x_j)^2 + (y_i - y_j)^2)}{\sum_{i=1}^{n} \sum_{j=1}^{i-1} 2p_{ij}(t_i - t_j)}, \tag{4.3.9}$$

where A is the area of the background window.

If we are to model a spatial-temporal process which is Gaussian in space and has the DCR model for time, i.e.

$$g(t)f(x,y) = \bar{\alpha}\bar{\omega}^2 t e^{-\bar{\omega}t} \cdot \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}, \tag{4.3.10}$$

the steps in the EM-algorithm are the same, with the exception that (4.3.9) changes to

$$\bar{\omega}^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} 2p_{ij}^{(k+1)}}{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{i,j}^{(k+1)}(t_i - t_j)}. \tag{4.3.11}$$

4.3.1 Simulating Spatio-Temporal Point Processes

We can simulate a spatial-temporal point process of the type shown in model (4.3.2) by following an algorithm proposed by Zhuang, Ogata and Vere-Jones [72]. The algorithm works by simulating the events generated by the background rate, the so-called first generation. We then simulate the events which are 'triggered' by each event in the first generation, and then simulate the events generated by this generation of events, until the process terminates. The algorithm can be described as follows:

1. Generate the background events according to the background intensity $\mu(x, y)$, and record these events from the first generation as belonging to $G^{(0)}$.

2. Set $j = 0$.

3. For each event (t_i, x_i, y_i) in generation $G^{(j)}$ we simulate its $N^{(i)}$ offspring, $O_i^{(j)} = \{(t_m^{(i)}, x_m^{(i)}, y_m^{(i)}) | m = 1, \ldots, N^{(i)}\}$, where $N^{(i)}$ is a Poisson random variable with mean equal to the expected number of events an event triggers, $t_m^{(i)}$ is generated from the probability distribution $g(t - t_i)$, and $x_m^{(i)}, y_m^{(i)}$ are generated from the probability distribution $f(x - x_i, y - y_i)$.

4. Set $G^{(j+1)} = \bigcup_{i \in G^{(j)}} O_i^{(j)}$.

5. If $G^{(j+1)}$ is not empty, let $j = j{+}1$ and go to step 3. Else return $\bigcup_{k=0}^{j} G^{(k)}$.

Using this algorithm, we simulate 10 spatial-temporal point processes of the form (4.3.3) over an area resembling the city of Chicago, with parameters $\alpha = 0.3$, $\omega = 0.5$, $\sigma = 0.1$, constant background rate $\mu = 0.02$, over time period $T = [0, 365]$. We then use the aforementioned EM-algorithm on these simulations to attempt to recover the model parameters, the results of which are displayed in Table 4.6.

Table 4.6: Average parameters found by the EM-algorithm for 10 simulated spatial-temporal point processes.

Parameter	True Value	Est. Value
μ	0.02	0.0202
α	0.3	0.2913
ω	0.5	0.5116
σ	0.1	0.0990

When assessing the fit of a spatial-temporal point process to data, as well as looking at standard measurements such as AIC, or looking at the predictive accuracy of the model, we can also look at the locations with which the model fits [59]. One way of doing this visually is by inspecting a residual map, where the residual in a certain region responds to the difference between the expected number of events which would be observed in this region over a certain time period, and the actual number of events which take place. For a point process fitted over time

$[0, T]$, the residual R over a grid cell $[x_1, x_2] \times [y_1, y_2]$ is

$$R = N([0, T] \times [x_1, x_2] \times [y_1, y_2]) - \int_0^T \int_{y_1}^{y_2} \int_{x_1}^{x_2} \lambda(x, y, t) \, \mathrm{d}x \, \mathrm{d}y \, \mathrm{d}t, \qquad (4.3.12)$$

where N is a counting process which gives the number of events observed in the cell over the time window, and the integrated intensity giving the expected number of events which would be seen under the fitted model. It has been observed that creating residual plots based on equally sized grid cells can be problematic [15]. In many cases there may be many grid cells which have a very low expected number of events under the model, and a single event happening in one cell results in a large residual compared to neighbouring cells. In such a case a visual inspection is not useful in terms of assessing how well the model fits the data. Attempting to rectify this problem by increasing the size of the grid cells can lead to another problem, whereby areas within a large grid cell which underfit and overfit cancel each other out, and we cannot identify locations where a model may be poorly fitted [15].

4.3.2 Voronoi Residual Analysis

A suggested solution to this problem is to partition the plane into a Voronoi diagram [15], whereby the spatial region is split into Voronoi cells, with each cell containing one event, and being a convex polygon where all locations within are closer to the event within the cell than any other event. For an event (x_i, y_i), the associated Voronoi cell C_i can be defined as

$$C_i = \{(x, y) \mid d((x, y), (x_i, y_i)) \leq d((x, y), (x_j, y_j)) \quad \forall i \neq j\}, \qquad (4.3.13)$$

where d gives the Euclidean distance between two points [53]. The Voronoi diagram produced by one of the previously generated spatial point processes on an area resembling Chicago is shown in Figure 4.14.

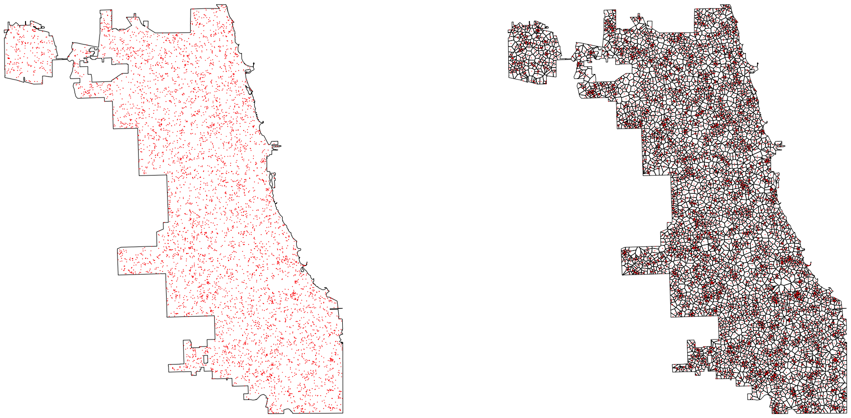

Figure 4.14: Displayed on the left are the locations of events for a simulated point process on a Chicago shaped area, and on the right is the corresponding Voronoi diagram generated by the events.

We can evaluate Voronoi residuals R^i_{vor} for each Voronoi cell i [15], where

$$R^i_{\text{vor}} = 1 - \int_0^T \int \int_{(x,y)\in C_i} \lambda(x,y,t)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t. \tag{4.3.14}$$

By construction, each cell has exactly one event which has taken place within it over the time period. We can numerically integrate the intensity function over the time window within each cell to evaluate the Voronoi residuals. For a homogeneous Poisson point process, the expected area of a typical Voronoi cell is equal to the reciprocal of the intensity of the process, and the area of a Voronoi cell is approximately gamma distributed [15]. It has been shown that these properties still approximately hold in the inhomogeneous case [8]. The second term in (4.3.14) can be approximated by a two-parameter gamma distribution with shape 3.569 and rate 3.569 [68], and so the approximate distribution of the Voronoi residuals R^i_{vor} is given by

$$R^i_{\text{vor}} \sim 1 - \Gamma(3.569, 3.569). \tag{4.3.15}$$

We can then plot the Voronoi residuals on a map, scaling the residuals by

$\Phi^{-1}\{F(R^i_{\text{vor}})\}$, where F is the distribution function of $1-\Gamma(3.569, 3.569)$, and Φ^{-1} is the inverse cumulative distribution function. The resultant Voronoi residual map for a realisation of the previous simulation is shown in Figure 4.15, with a histogram of the residuals being shown in Figure 4.16. The residual map shows a seemingly random spread of the residuals, which we would expect as our simulated point process had a constant background rate, while the histogram shows that the residuals are indeed well estimated by the previously shown distribution.

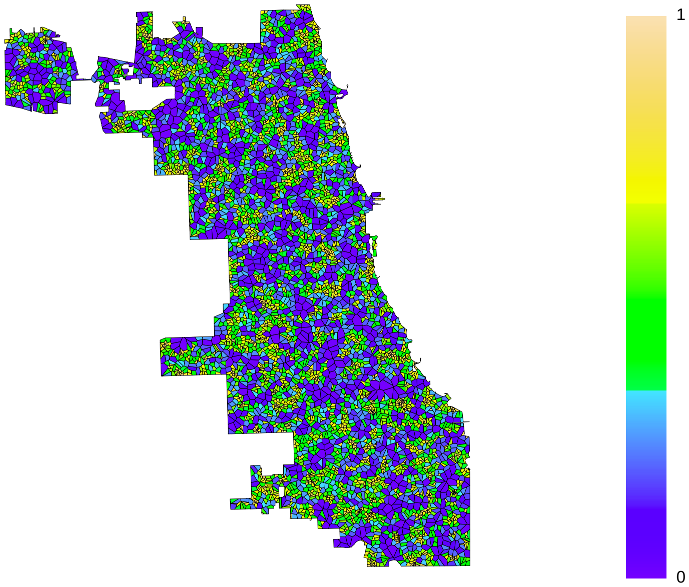

Figure 4.15: A Voronoi residual map on the simulated point process on the city of Chicago.

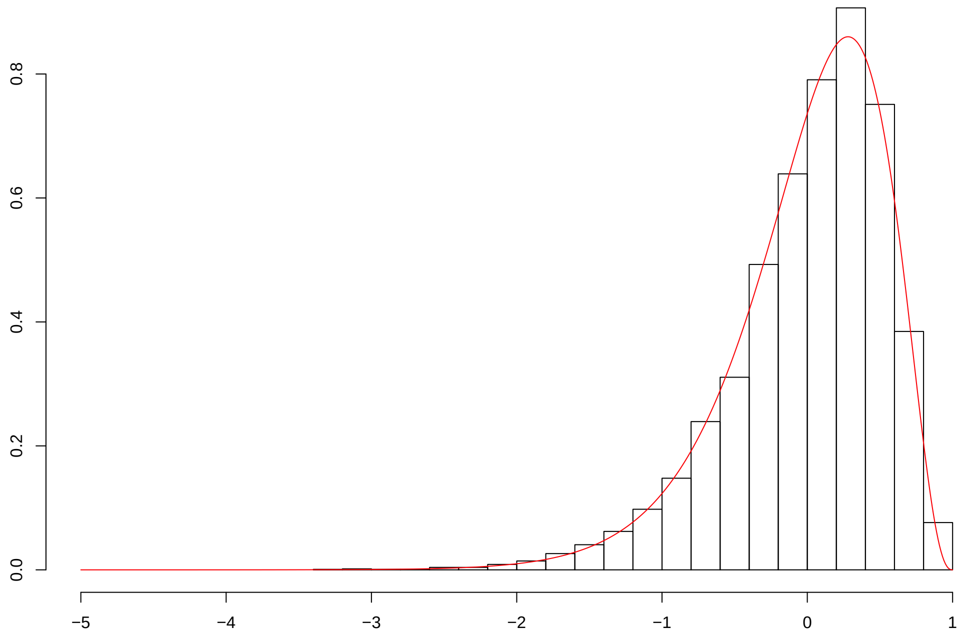

Figure 4.16: Histogram of the Voronoi residuals for the simulated point process, with the $1 - \Gamma(3.569, 3.569)$ distribution plotted in red.

We may also test the residuals using the Probability Integral Transform (PIT). Here we assess the distribution of the values of the residuals under the cumulative distribution function of the proposed model [15]. If the residuals do indeed come from the proposed model, then a histogram of the PIT values should be standard uniform [15]. A plot of the PIT values for a realisation of the simulation is shown in Figure 4.17, along with 95% confidence intervals for the height of the histogram. We can see for the height of this histogram broadly fall within the 95% confidence interval, indicating that the model is a reasonable fit for the data.

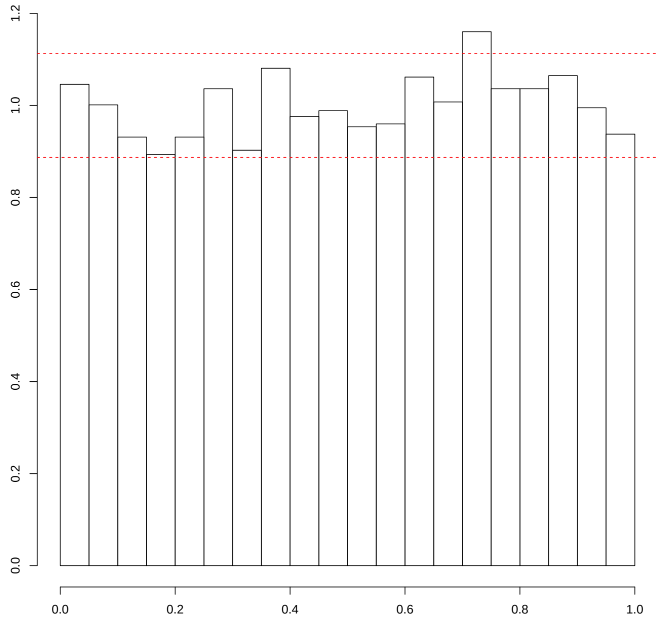

Figure 4.17: Histogram of the PIT values for simulated point process, with a 95% confidence interval for the heights of the histogram shown as a red dotted line.

4.4 Spatio-Temporal Point Process on Chicago Data

We fit spatio-temporal point processes of the form (4.3.3) and (4.3.10) to the burglary data in Chicago. In (4.12) we see the location of events which happened in the 1km x 1km grid with the highest burglary rate in (4.11). We can see how localised the incidents of burglary are, and why modelling the point process as a spatial-temporal process may be advantageous; modelling burglary as separate temporal processes in each grid didn't take into account events which happened in nearby grids, which likely will increase the likelihood of further burglaries in neighbouring grids. Note that the exact location of the burglary is partially redacted in the data so it lies within the same block as the actual incident. We account for this by adding noise in the x or y direction, depending on the cardinal direction of the street the data is recorded. The average block in Chicago is around 200 by 100 metres [2]. For North-South streets we add a random number $U \sim \text{Unif}([-0.1, 0.1])$ to the y-coordinate, and for East-West streets we add a random number $U \sim \text{Unif}([-0.05, 0.05])$ to the x-coordinate. An illustration of how this affects the spatial information is given in Figure 4.18.

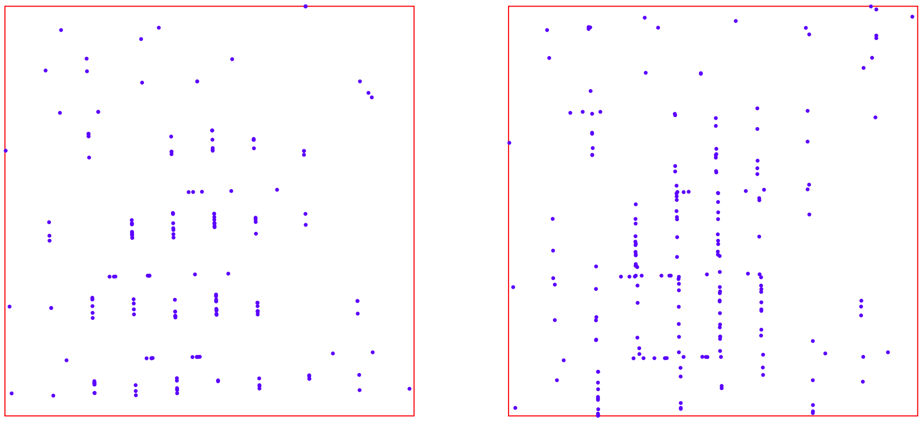

Figure 4.18: Displayed on the left are the locations of recorded burglary in 2016 in the 1km by 1km grid with the highest burglary rate for the original data. On the right are the resultant locations after adding noise to account for the fact that the original data is redacted so as only to lie in the same block as the actual incident.

There are also drawbacks to modelling the entire city as a spatial-temporal point process. As the computational complexity of the EM-algorithm is $O(N^2)$, it means it is not computationally feasible to use data going back as far as we could when we modelled many separate temporal point processes. We therefore only use burglary data from the year of 2015, of which there were 13044 recorded incidents.

Fitting the spatial-temporal process with ETAS triggering function in time and Gaussian in space shown in (4.3.3) to burglary data from the entire year of 2015, we find the parameters $\mu = 0.0160$, $\alpha = 0.7321$, $\omega = 0.0513$, $\sigma^2 = 0.0678$. The DCR triggering function in time and Gaussian in space shown in (4.3.10) generated parameters $\mu = 0.0165$, $\bar{\alpha} = 0.7232$, $\bar{\omega} = 0.1122$, $\sigma^2 = 0.0847$. The prediction results generated by each of the models are shown in Table 4.7.

Table 4.7: Predictive results for the spatial-temporal model with ETAS and DCR triggering function in time. The percentage of burglaries which are predicted by the top 1%(156 grids), 5%(779 grids), 10%(1559 grids) and 20%(3117 grids) of 200m by 200m by intensity every hour over the first 3 months of 2016 is shown, along with the total area under the curve.

Model	Top 1%	Top 5%	Top 10%	Top 20%	AUC
ETAS	7.78%	21.34%	34.23%	51.66%	0.7473
DCR	6.46%	19.95%	32.24%	49.47%	0.7361

We can see that the parameters suggest that there is a far higher triggering rate of burglary when we fit the spatio-temporal point processes to the data, as opposed to when we fit the purely temporal models to grid cells previously. In terms of prediction, the ETAS triggering function which is Gaussian in space predicts slightly better than the DCR triggering function, however neither model gives as good a prediction as the purely temporal models fitted previously.

When we look at the Voronoi residuals generated for the ETAS spatio-temporal model fitted to the data, we can see why these models may not be ideally suited to the data. Figure 4.20 shows the Voronoi residual map generated for the burglary data by the ETAS spatio-temporal model, and Figure 4.21 shows the histogram of the Voronoi residuals.

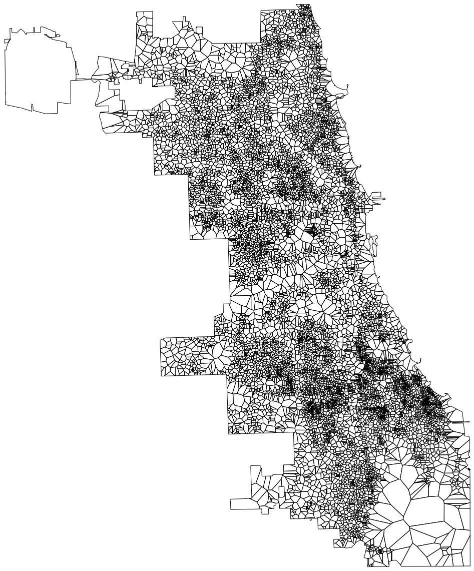

Figure 4.19: Voronoi diagram for the burglary events in 2015.

In Figure 4.20 we can see that there are large areas of Chicago coloured blue for which the model is overpredicting the burglary rate. We can also see in the histogram of the Voronoi residuals in Figure 4.21 that the residuals don't appear to match the theoretical distribution they should if a suitable model has been fitted. This story remains the same for the Voronoi residuals when fitting the DCR model with constant background rate in Figures 4.22 and 4.23.

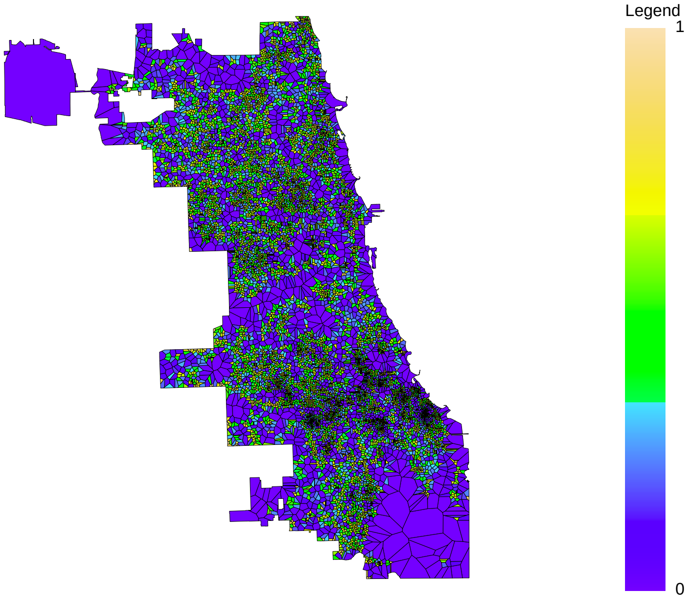

Figure 4.20: Voronoi residual map for the burglary events in 2015 for the fitted ETAS point process.

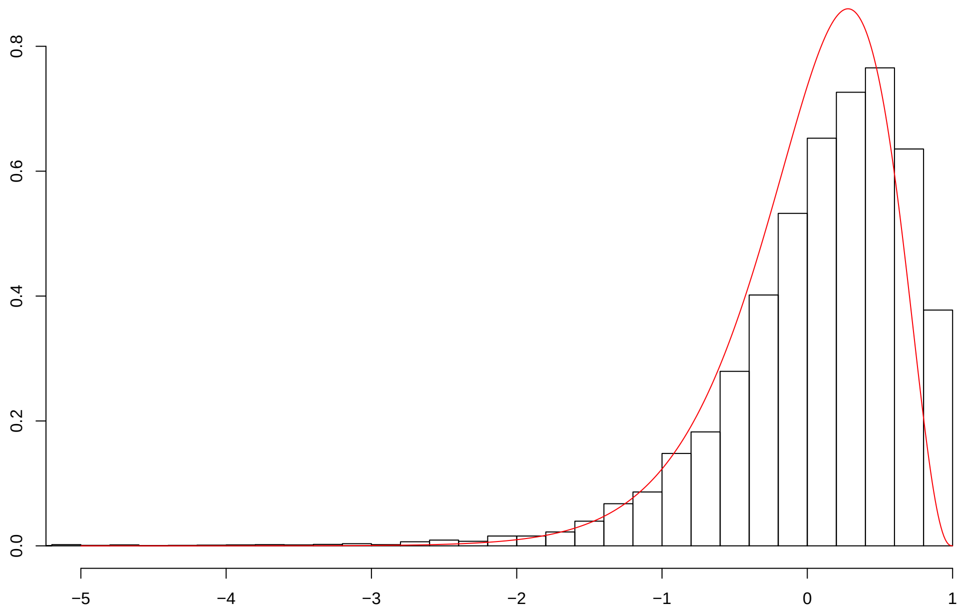

Figure 4.21: Histogram of the Voronoi residuals for the spatio-temporal ETAS model fitted to the burglary events of 2015, with the $1-\Gamma(3.569, 3.569)$ distribution plotted in red.

We can also see when we zoom in to the Voronoi map for the ETAS model in the region with the highest burglary rate in Figure 4.24, that large sections of this also appear to be underpredicted, indicated in yellow. Even within this grid with high levels of burglary, there are areas which appear to be overpredicted. These residual maps give strong evidence that a spatio-temporal point process with constant background rate is not suitable to model burglary in Chicago. This backs up what out intuition might be, that the background rate of burglary is not the same across the entire city. In the residual map in Figure 4.20, we can identify large regions of Chicago for which the model overpredicts burglary; the North-West region, where O'Hare International Airport is situated; a region running south down Chicago and across in a South-Westerly direction in the centre of the map, which coincides with the Chicago River; and in the South of the city, which is largely an industrial area. We obviously would not anticipate seeing high rates of burglary in any of these areas, and so it is not surprising that it is not suitable to model the burglary rate in the same way as we would residential regions.

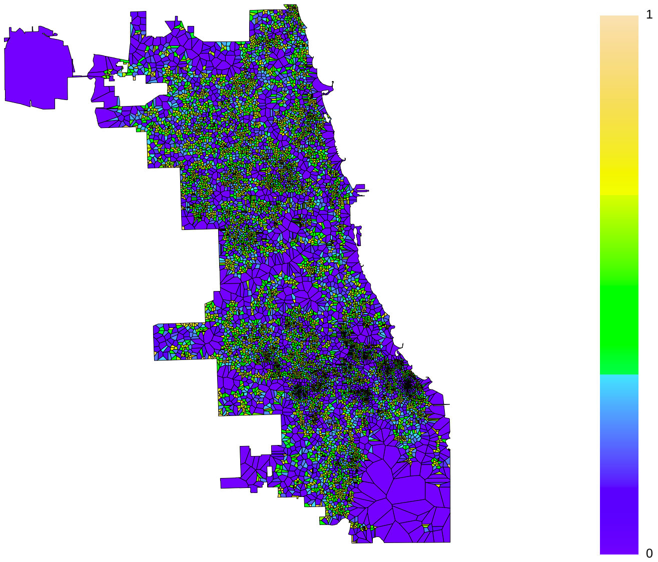

Figure 4.22: Voronoi residuals for the DCR model fitted to burglary events in 2015.

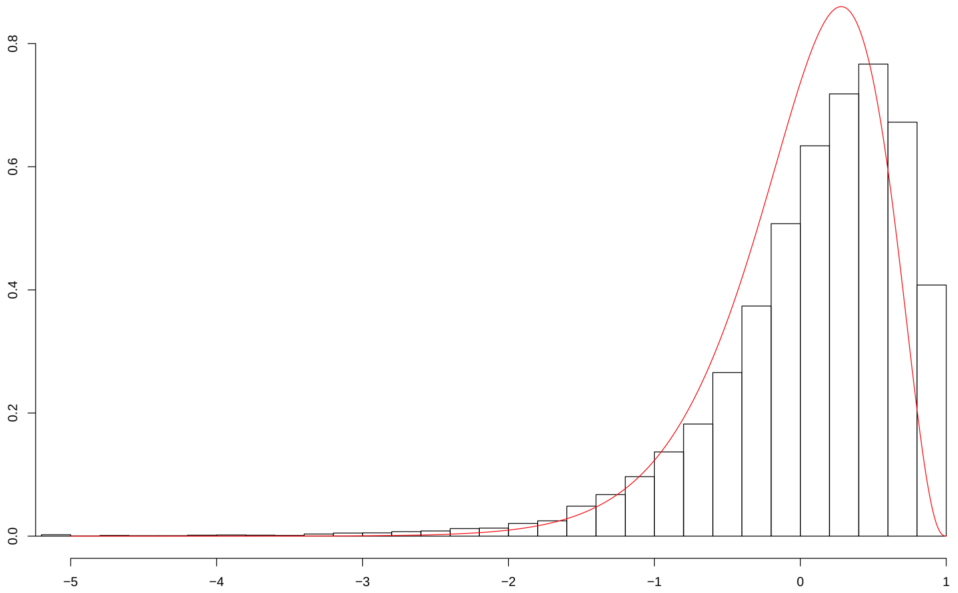

Figure 4.23: Histogram of the Voronoi residuals for the spatio-temporal DCR model fitted to the burglary events of 2015, with the $1-\Gamma(3.569, 3.569)$ distribution plotted in red.

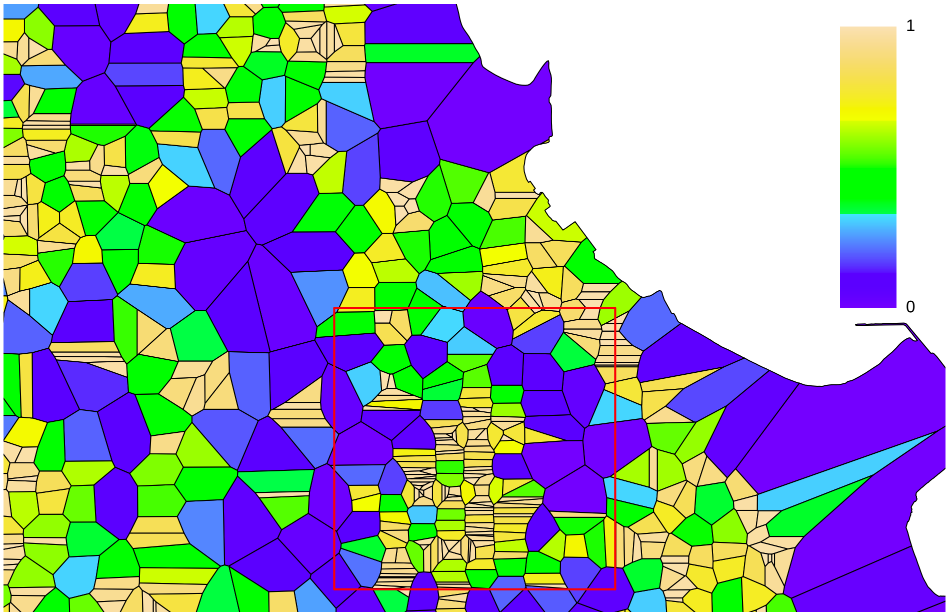

Figure 4.24: Voronoi residual map zoomed in, with the red grid showing the area of Chicago with the highest burglary rate.

The PIT histograms for the ETAS model in Figure 4.25 and for the DCR model in Figure 4.26 confirm this, showing for both models that there are far more Voronoi cells that have underpredicted the level of burglary than we would expect if the models were a good fit for the data. Along with that, we also see there are slightly more cells which have overpredicted the level of burglary than we would expect if the model was well fitted. Modelling burglary in Chicago as a self-exciting point process with a homogeneous background rate is not an appropriate choice; as large areas of Chicago have very little or no burglary, when we estimate the background rate it is 'spreading' the background rate μ over these areas, overpredicting the rate of burglary in those locations, and simultaneously underpredicting the rate of burglary in areas where high numbers of offences have taken place. We propose that this leads to many crimes which have taken place in areas with a high crime rate being modelled as triggered events with high probability instead of possible background events in areas with high crime rates, artificially inflating the triggering parameters we find with the data.

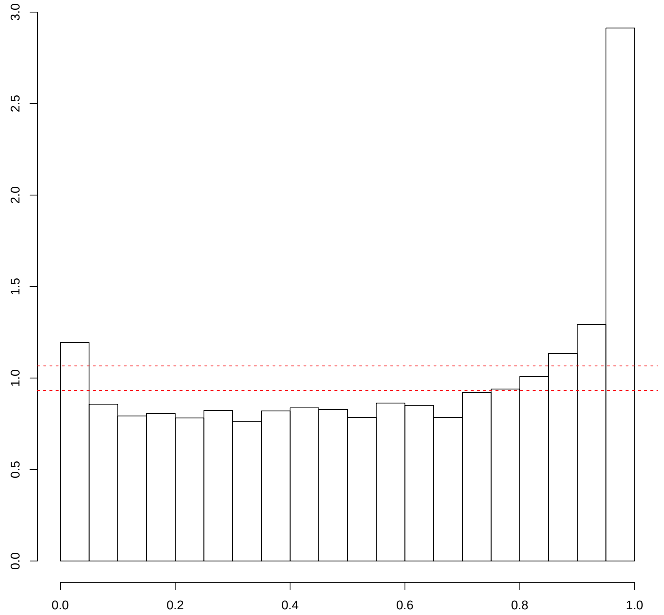

Figure 4.25: Histogram of the PIT values for the ETAS model fitted to burglary in 2015, with a 95% confidence interval for the heights of the histogram shown as a red dotted line.

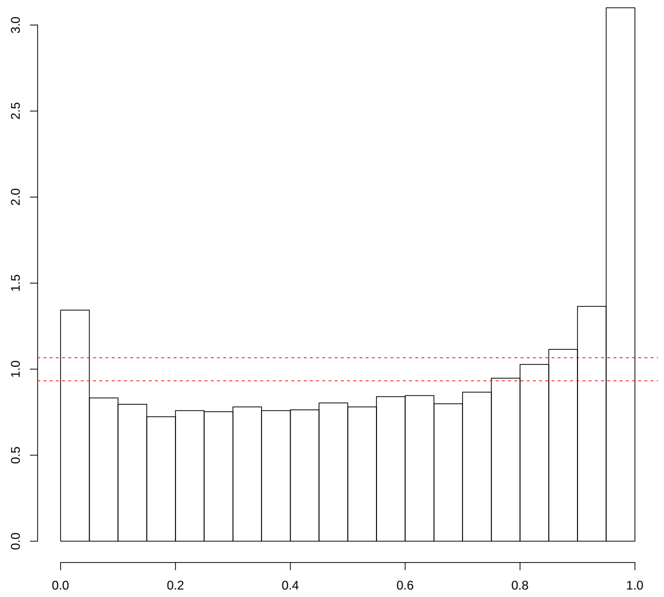

Figure 4.26: Histogram of the PIT values for the DCR model fitted to burglary in 2015, with a 95% confidence interval for the heights of the histogram shown as a red dotted line.

In Figure 4.27 we show the prediction rates with the ETAS and DCR models with homogeneous background rate compared with a Poisson point process in each 200m x 200m grid, where we only take into account the raw amount of burglary in each grid over the last year. We also compare the previous temporal ETAS models made for each grid in Section 4.1 with the spatio-temporal ETAS model with homogeneous background rate.

We can see that the ETAS spatio-temporal model outperforms the DCR spatio-temporal model, which doesn't outperform the Poisson point process. However we see that when compared with the temporal ETAS models that the spatio-temporal model fails to predict burglary as successfully.

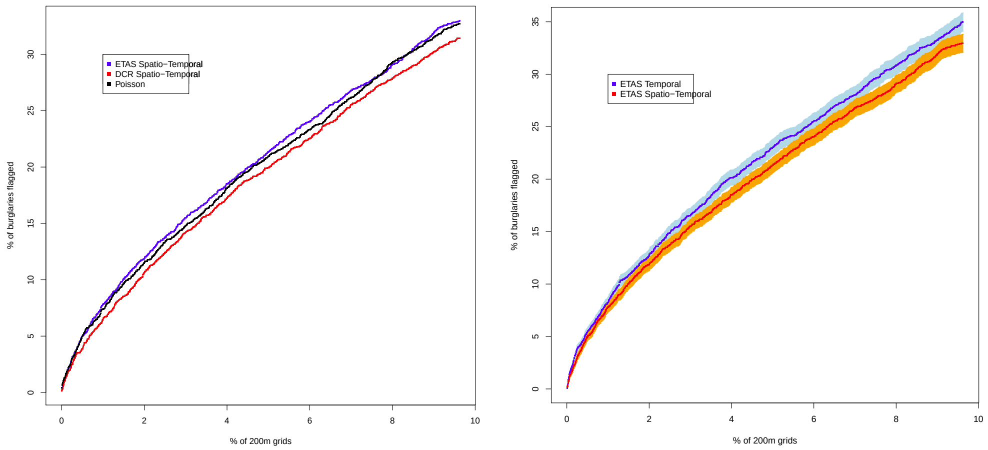

Figure 4.27: On the left is the prediction success of the spatio-temporal models with homogeneous background rate for the ETAS and DCR models, compared to a simple Poisson point process modelled in each grid over one year. On the right is the spatial-temporal ETAS model compared with the solely temporal model, with error bars showing one standard error for each method.

4.5 Summary

In this chapter we have introduced a new potential triggering function for self-exciting point process models. This DCR model outperformed the Poisson point process models that were fitted in discrete 200m x 200m grids, but was not as successful as the ETAS triggering function. However such a triggering function might still be useful in other contexts where there may be a delay before there is the largest self-excitation caused by a previous event. We then fitted spatio-temporal self-exciting point processes with homogeneous background rate to burglary in Chicago, but observed that this may not be a suitable method to model crime occurring within a city. We require a different method to successfully model the inhomogeneity we see in crime rates across the differing geometry of a city.

Chapter 5

Non-Parametric Self Exciting Point Processes

5.0.1 Background Rate of Crime

In the previous chapter we observed that modelling crime as being uniformly randomly distributed in space may not be a suitable approach for burglary within Chicago. That the background rate of crime may not be homogeneous in space is something which is both intuitive (we would expect there to be a higher risk of burglary in a densely populated residential area than in a sparsely populated area, for example), and something which matches the findings of historical criminology research [62].

In 2008 a method was proposed which used a non-parametric approach to estimate the triggering effect in the context of earthquake sequences, which made no prior assumptions on the form of the triggering function [45]. In this chapter we use this methodology in order to first incorporate a non-constant background rate to our model, and then to estimate the triggering effect in a non-parametric fashion.

5.1 Model Independent Stochastic Declustering (MISD)

Marsan and Lengliné proposed a method which could estimate Hawkes processes for earthquakes without making any prior parametric assumptions [45]. The method, named Model Independent Stochastic Declustering (MISD), is similar to the previous iterative EM-algorithm, but differs in using a probability weighted histogram when estimating the triggering function and background rate. Essentially this method simply involves discretising the background observation window into cells and/or discretising the triggering components, and building probability weighted histograms in the maximisation step based on the sum of the probabilities from the expectation step which fall in each discretised region.

5.1.1 EM-algorithm for Non-Constant Background Rate

While the background rate can take quite general forms [70], Veen and Schoenberg divided the spatial observation window into κ spatial grid cells of area A, with constant background intensity μ_m, $m \in \{1, \ldots, \kappa\}$ [70]. We can see that the background intensity for a grid m can be defined as

$$\mu_m = \frac{\text{expected number of background events in grid } m}{(\text{length of time window}) \cdot (\text{area of grid})}. \tag{5.1.1}$$

We can use this to incorporate a non-constant background rate into our modelling of a spatio-temporal point process. To model a self-exciting point process of the form (4.3.2) with triggering function (4.3.3) and non-constant background rates evaluated in grid cells, the expectation step is

$$p_{ii}^{(k+1)} = \frac{\mu^{(k)}_{\{m|i\in m\}}}{\mu^{(k)}_{\{m|i\in m\}} + \sum_{j=1}^{i-1} g(t_i - t_j) f(x_i - x_j, y_i - y_j)}, \tag{5.1.2}$$

$$p_{ij}^{(k+1)} = \frac{g(t_i - t_j; \alpha^{(k)}, \omega^{(k)})}{\mu^{(k)}_{\{m|i\in m\}} + \sum_{j=1}^{i-1} g(t_i - t_j) f(x_i - x_j, y_i - y_j)}, \tag{5.1.3}$$

and the maximisation step is

$$\mu_m^{(k+1)} = \frac{\sum_{i \in m}^{n} p_{ii}^{(k+1)}}{T \cdot A}, \tag{5.1.4}$$

$$\alpha^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}}{n}, \tag{5.1.5}$$

$$\omega^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}}{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}^{(k+1)}(t_i - t_j)} \tag{5.1.6}$$

$$\sigma^{2(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{i-1} p_{ij}((x_i - x_j)^2 + (y_i - y_j)^2)}{\sum_{i=1}^{n} \sum_{j=1}^{i-1} 2p_{ij}(t_i - t_j)}. \tag{5.1.7}$$

5.1.2 Results on Simulated Data

As a proof of principle, we initially tested this algorithm on the spatial-temporal point processes in section 4.3.1, where 10 simulations of the form (4.3.3) were made with parameters $\alpha = 0.3$, $\omega = 0.5$, $\sigma = 0.1$, constant background rate $\mu = 0.02$, over time period $T = [0, 365]$. Note that the algorithm does not assume that the background rate is constant. We attempted to recover the background rate and model parameters while using different sizes of grid: 200m x 200m, 1km x 1km, and 2km x 2km. The average estimates for the parameters of the triggering function for these 10 simulations is shown in Table 5.1 for the different sizes of grid for the non-background rate.

Table 5.1: Average parameters found by the EM-algorithm with non-constant background rate for 10 simulated spatial-temporal point processes with different sizes of grid for the background rate.

Parameter	True Value	200m grid	1km grid	2km grid
α	0.3	0.2551	0.2891	0.2950
ω	0.5	0.6252	0.5157	0.5022
σ	0.1	0.0939	0.0986	0.1000

We can see from Table 5.1 that when we select a smaller grid size to estimate the non-constant background rate, the number of triggered events is underestimated more than when we select a larger grid size. Problems with selecting too small a grid size are displayed in Table 5.2, where we see the average $L2$ error for the estimated background rate over these simulations is far greater for grids of 200m x 200m than 1km x 1km or 2km x 2km. Figure 5.1 illustrates an example of the estimated background rates for one of the simulations with the different selected grid sizes. We can see the 200m x 200m grids tend to overfit the data far more than the larger grids. There is a trade off between not overfitting the background rate, and ensuring the model picks up on events which occur on a very local level. It should also be emphasised that the processes simulated had a constant background rate, so it is not surprising that the model gives a better fit when we use fewer parameters in the spatial background rate.

Table 5.2: Mean $L2$ error of the background rate for the 10 simulations compared with the true constant background rate for the different sizes of grid for the background rate.

Size of Grid	Mean $L2$ error norm
200m	0.971
1km	0.040
2km	0.012

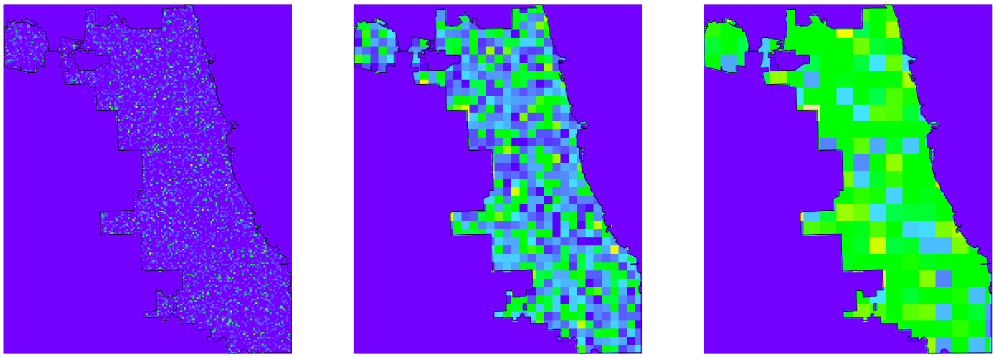

Figure 5.1: Example of the background rate found for one of the simulations with constant background rate when modelled with background grids of 200m x 200m (left), 1km x 1km (middle), and 2km x 2km (right).

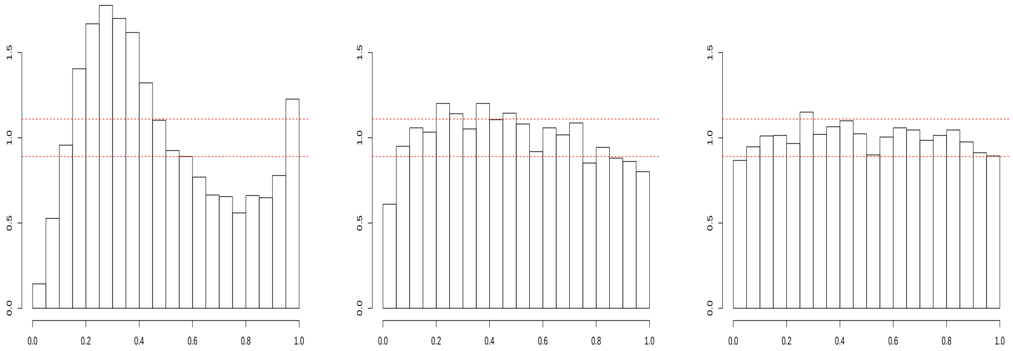

Figure 5.2: Histograms of the PIT values for a simulated point process with non-constant background rate when modelled with background grids of 200m x 200m (left), 1km x 1km (middle), and 2km x 2km (right), with a 95% confidence interval for the heights of the histogram shown as a red dotted line. The known target distribution is uniform.

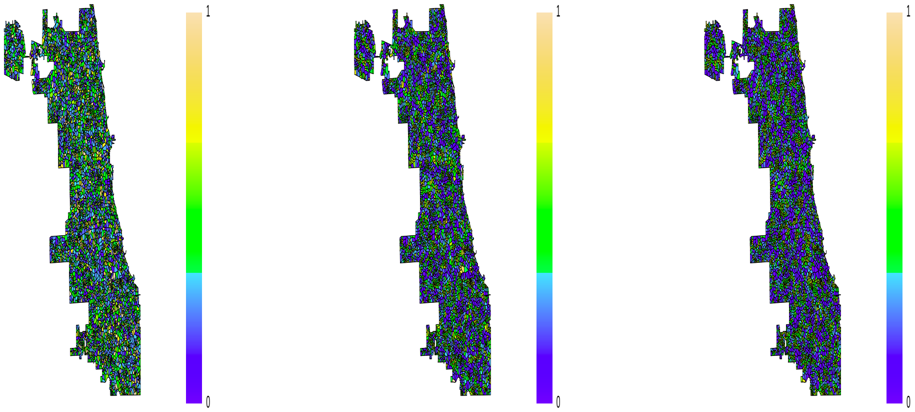

Figure 5.3: Voronoi residual maps for a simulated point process with constant background rate when modelled with background grids of 200m x 200m (left), 1km x 1km (middle), and 2km x 2km (right).

Figure 5.2 shows the histograms of the PIT values for a typical simulation with the different background grid sizes, and Figure 5.3 shows the associated Voronoi residual map. These PIT histograms confirm that when an inappropriate grid size is selected we do not observe the model fitting the process as we would wish. An excess of Voronoi cells underpredict the number of events than we would expect, and too few Voronoi cells overpredict the number of events than we would expect.

5.1.3 Information Gain

One measure to assess how well the point process model we have fitted predicts future crime is the expected information gain [18, 30]. The expected information gain G is defined as

$$G = \frac{1}{N}\log(\frac{L_1}{L_0}), \tag{5.1.8}$$

where L_1 is the likelihood of the model in question, L_0 is the likelihood of some baseline model, which is typically a homogeneous Poisson point process, and N is the number of events which take place in the observation window. We can use this quantity to compare the predictive performance of different models, with a higher information gain being indicative of a superior model in terms of prediction.

5.1.4 Results on Real Data

We fit the model with a triggering function that is ETAS in time and Gaussian in space with non-constant background rate to the burglary events in 2015 as seen in Chapter 4. The parameters found for the data when fitting the non-constant background rate with grids of 200m x 200m, 1km x 1km, and 2km x 2km are displayed in Table 5.3, along with the parameters found when a constant background rate was used in Section 4.4. The background rate estimated for each of these models is displayed in Figure 5.4.

Table 5.3: Parameters estimated for the real data for different background grids, along with the parameters found for a constant background rate in Section 4.4.

Parameter	200m grid	1km grid	2km grid	Constant
α	0.051	0.2184	0.4441	0.7321
ω	0.015	0.0434	0.0437	0.0513
σ^2	2.9×10^{-5}	0.006	0.0231	0.0678

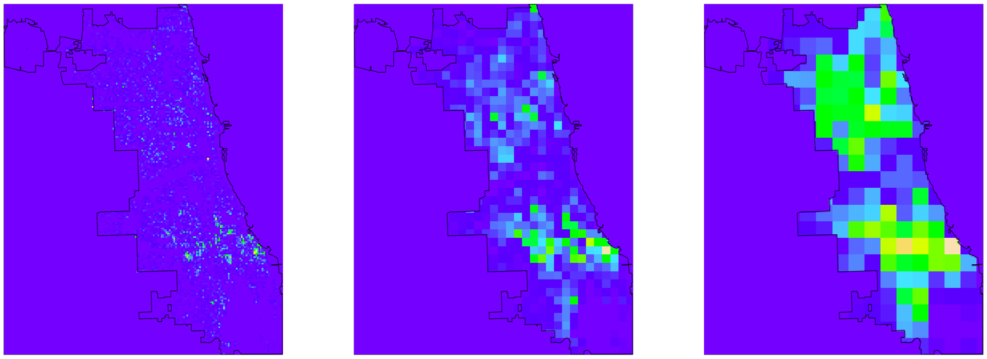

Figure 5.4: Illustration of the estimated background rate found for the Chicago burglary data from 2015 with non-constant background rate when modelled with background grids of 200m x 200m (left), 1km x 1km (middle), and 2km x 2km (right).

We can observe from the parameter values displayed in Table 5.3 that as we refine the background rate into smaller grid cells, less of a triggering effect is detected within the data. Similarly to the problem in Section 4.4 where events which happened in areas with high crime rate were assumed to have been triggered events with high probability when we modelled the background rate as being constant across the whole of Chicago, when we model the background rate at too local a level (with the 200m x 200m grid), events which may have been triggered appear to be assumed to be background events by the algorithm. Along with this, as we estimate the background rate on a more local rate, the estimate of σ^2 also becomes smaller, with triggered events assumed to occur on a far more local level as well.

In Figure 5.5 PIT histograms for the models are displayed. We can see that the pattern of residuals for the 200m x 200m grid is similar to that seen in the simulated data in Figure 5.2 where the background rate was estimated with grid cells which were too small. The other two models with larger grids display more Voronoi cells which have been underpredicted than we would expect if our model was a true reflection of the data.

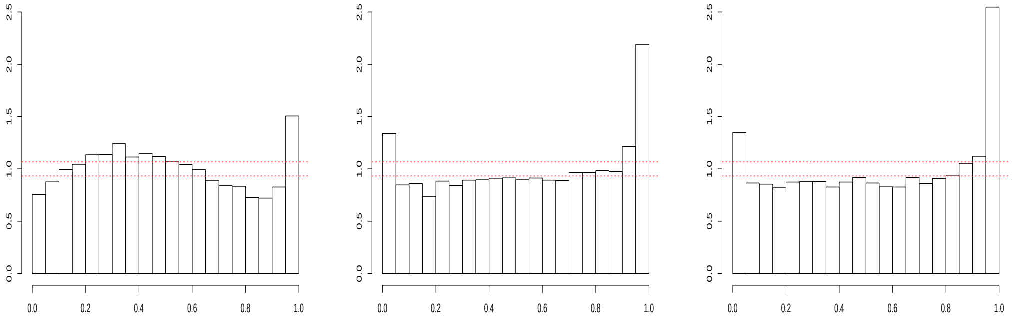

Figure 5.5: Histograms of the PIT values for the Chicago data when modelled with background grids of 200m x 200m (left), 1km x 1km (middle), and 2km x 2km (right), with a 95% confidence interval for the heights of the histogram shown as a red dotted line.

Table 5.4: Predictive results for the spatial-temporal model with ETAS triggering function in time and Gaussian space with different non-constant background measurements, along with the result for a constant background rate. The percentage of burglaries which are predicted by the top 1%(156 grids), 5%(779 grids), 10%(1559 grids) and 20%(3117 grids) of 200m by 200m by intensity every hour over the first 3 months of 2016 is shown, along with the information gain for each model. Note we cannot numerically calculate the information gain when using a 200m x 200m background rate.

Background	Top 1%	Top 5%	Top 10%	Top 20%	Inf. Gain
200m grid	6.62%	19.78%	31.31%	50.13%	-
1km grid	8.97%	22.56%	34.99%	52.32%	0.429
2km grid	8.88%	22.37%	34.63%	52.55%	0.433
Constant	7.78%	21.34%	34.23%	51.66%	0.363

The predictive accuracy of the models are shown in Table 5.4. We can see that in terms of predictive accuracy the best results are achieved when the background rate is modelled at a 1km x 1km level, while the predictive performance dips when the background is modelled with a 200m x 200m grid. Note that we cannot get an accurate measurement of the information gain for this model as many of the events occur at locations where numerically the model predicts $\lambda(t_i, x_i, y_i) = 0$. For the background rates measured with 1km x 1km and 2km x 2km we can see that the predictive performance is greater in terms of burglaries predicted by the hourly ordering shown in the previous chapter, and in terms of information gain when compared to the constant background rate.

5.2 Non-Parametric Triggering Function

In addition to estimating the background rate using non-parametric methods, we can use Marsan and Lengliné's [45] method to estimate the triggering function without making any prior assumptions about the form it takes. We are going to estimate a spatial-temporal point process that has the intensity function

$$\lambda(x, y, t) = \mu(x, y) + \sum_{t > t_i} g(t - t_i) f(x - x_i, y - y_i), \tag{5.2.1}$$

where we will estimate the background intensity μ in the same non-parametric way as previously, and estimate the triggering function while assuming the spatial triggering function is isotropic, $f(x, y) = f(x^2 + y^2)$, as in [27]. In this method we will introduce a probability density function $h(r)$ for the distance r between two events, where we have $f(r) = h(r)/(2\pi r)$, and we assume a.s. there are no points at the origin [17].

We introduce discretisation parameters δt and δr, with A the set of pairs of events such that $u\delta t \leq t_i - t_j \leq (u + 1)\delta t$, and B the set of pairs of events such that $v\delta r \leq \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \leq (v+1)\delta r$. To understand how we can construct g and f in a non-parametric fashion, we can define the height of the triggering function g in the weighted histogram in an analogous to way to how we defined the non-parametric background rate in (5.1.1), with

$$g(u\delta t,(u+1)\delta t)=\frac{\text{expected number of triggered events in }(u\delta t,(u+1)\delta t)}{(\text{total number of events})\cdot(\delta t)},\tag{5.2.2}$$

and similarly the height of the probability density function h defined to be

$$h(v\delta r,(v+1)\delta r)=\frac{\text{expected number of triggered events in }(v\delta r,(v+1)\delta r)}{(\text{expected total number of triggered events})\cdot(\delta r)}.\tag{5.2.3}$$

We can use this to motivate the maximisation step for the EM-algorithm

$$\mu_m^{(k+1)}=\frac{\sum_{i\in m}^{n}p_{ii}^{(k+1)}}{T\cdot A},\tag{5.2.4}$$

$$g_u^{(k+1)}=\frac{1}{N\times\delta t}\sum_{i,j\in A}p_{ij}^{(k+1)},\tag{5.2.5}$$

$$h_v^{(k+1)}=\frac{\sum_{i,j\in B}p_{ij}^{(k+1)}}{\delta r\sum_{1}^{N}\sum_{j=1}^{i-1}p_{ij}^{(k+1)}}.\tag{5.2.6}$$

As before, the expectation step is

$$p_{ii}^{(k+1)}=\frac{\mu_{\{m|i\in m\}}^{(k)}}{\mu_{\{m|i\in m\}}^{(k)}+\sum_{j=1}^{i-1}g(t_i-t_j)f(x_i-x_j,y_i-y_j)},\tag{5.2.7}$$

$$p_{ij}^{(k+1)}=\frac{g(t_i-t_j;\alpha^{(k)},\omega^{(k)})}{\mu_{\{m|i\in m\}}^{(k)}+\sum_{j=1}^{i-1}g(t_i-t_j)f(x_i-x_j,y_i-y_j)}.\tag{5.2.8}$$

5.2.1 Results on Simulated Data

Using the same simulated data as before, we tested the algorithm to find the triggering effect in a non-parametric way. The average results over the 10 simulations can be seen in Figure 5.6.

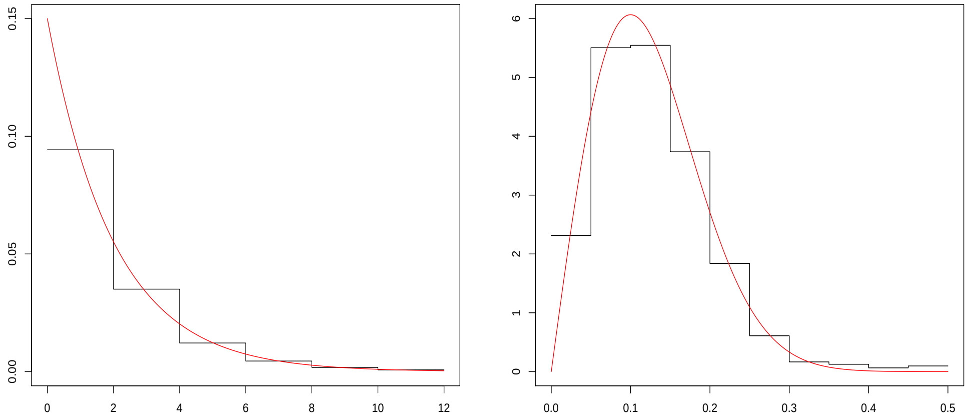

Figure 5.6: Average results for the 10 simulations when the model is fitted using weighted histograms with g on the left, and the probability density function h on the right, with the true model parameters shown in red.

We can see that the model gives a reasonably good fit of the data, albeit the nature of histograms results in not being able to capture the continuity of the ground truth of the triggering effect. Drawbacks of this method are that we have to arbitrarily select how to discretise t and r, along with having to be careful of the estimates we make very close to a previous event, because for events very close to the origin, $f(r) \to \infty$ as $r \to 0$. There are also benefits to using this method over the parametric methods we used previously. Along with the obvious advantage of not making any assumptions of the form which the triggering function takes, this method can also allow us to find information about the triggering effect at a certain distance from previous event after a certain time period. For example, if we wish to investigate whether there is a spike of crime a month after previous crime, we are able to refine the histogram bins for this time period to see whether this is the case.

5.2.2 Non-Parametric Results on Real Data

We fitted the non-parametric model on the burglary data from Chicago in 2015. We chose 1km x 1km grids to estimate the background rate, and then chose to develop the $h(r)$ probability density function using bins every 100m away from the origin up to a distance of 500m. For the $g(t)$ histogram, we selected bins for the first 12 hours and next 12 hours to focus on the immediate aftermath after a burglary where we suspect there might be an especially heightened risk of a repeat event. We also selected bins which focused on the 48 hours which occur every 7 days after the original event, with previous studies suggesting a spike of burglary took place 7 days after the original burglary events for data in Los Angeles [48]. We set a limit on the triggering effect of 60 days, meaning we assumed after 60 days no further crime was triggered. The result for g and h can be seen in Figure 5.7.

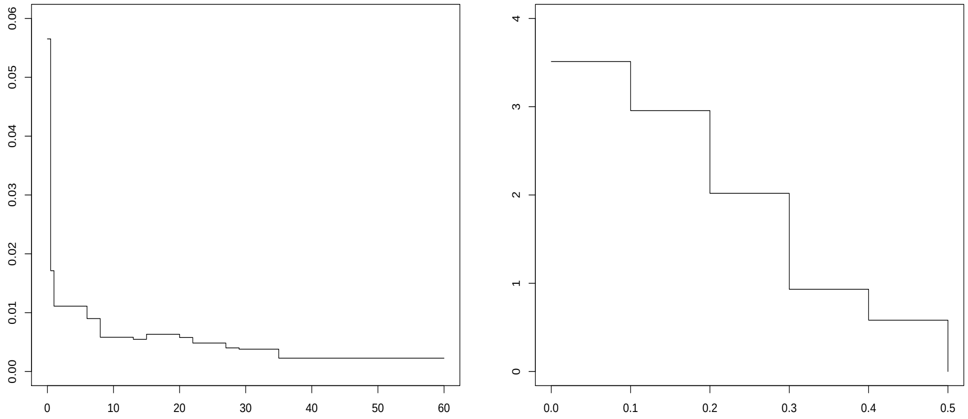

Figure 5.7: Results for the burglary data from 2015 when the model is fitted using weighted histograms with g on the left, and the probability density function h on the right.

It is clear from Figure 5.7 that the greatest risk of a repeat burglary event occurs in the 24 hours after the initial event. Thereafter there is still a risk of a repeat event, albeit much reduced from the original aftermath. The branching ratio of the events an initial event can be expected to trigger under this model is 0.328. We also do not observe any noticeable spike around the 7 days after the initial burglary event. The estimated probability density function h shows that the greatest risk of a repeat event occurs in the 100 metre radius from the initial event, with this risk reducing as you move further away from the initial event. These histograms are compared with the findings for the parametric triggering function found in Section 5.1.4 with the 1km x 1km non-parametric background rate in Figure 5.8. We can see the resulting f found for the Gaussian spatial function found in Section 5.1.4 in Figure 5.9, and for the non-parametric spatial f in Figure 5.10.

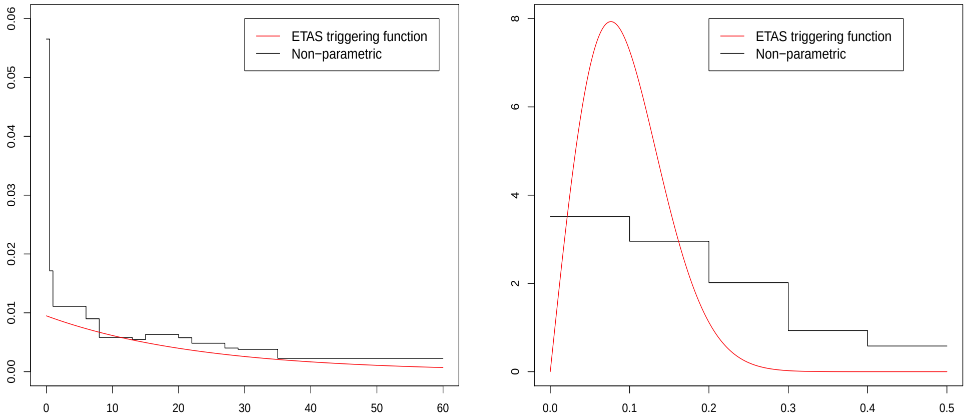

Figure 5.8: Results for the burglary data from 2015 when the model is fitted using weighted histograms with g on the left, and the probability density function h on the right.

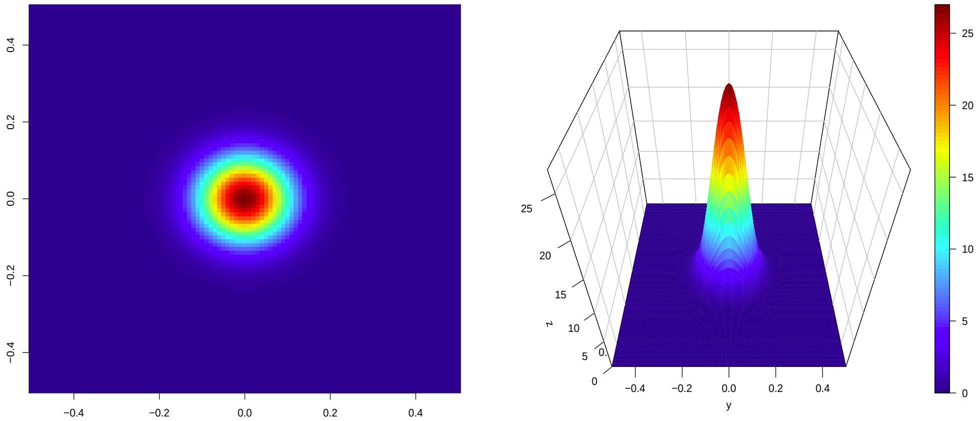

Figure 5.9: Result for f in Chicago when estimated with a Gaussian function in Section 5.1.4.

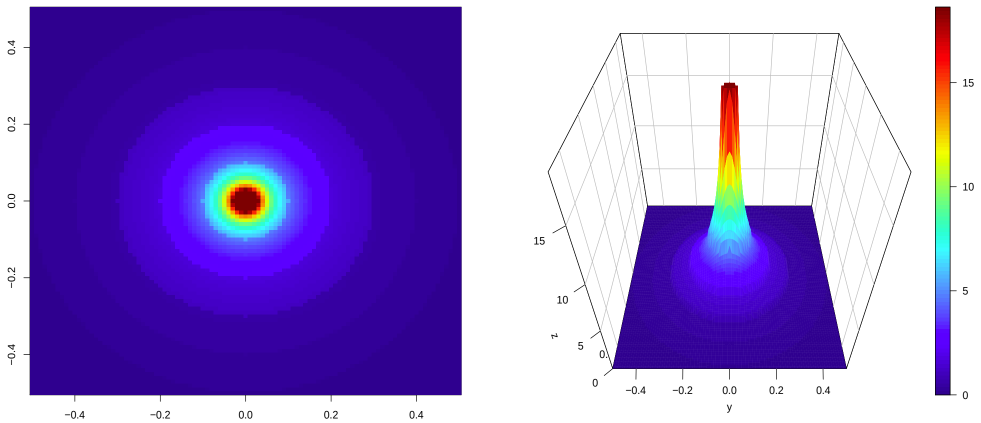

Figure 5.10: Result for f in Chicago when estimated with the non-parametric histograms. For $r < 0.01$ we have taken $f(r) = h(r)/(2\pi \cdot 0.01)$.

It is clear from Figure 5.8 that by finding the triggering function g in a non-parametric way we detect a far higher risk of repeat events in the initial 24 hours after an initial event as opposed to the parameters found with the ETAS function. After this point the non-parametric and ETAS function behave in a similar way. For h we can see that the non-parametric weighted histogram estimates a wider reach for the triggering effect than the Gaussian spatial function detected. Note in Figure 5.10 we took $f(r) = h(r)/(2\pi \cdot 0.01)$ when $r < 0.01$, as we did when making estimates for the predictions to avoid the problem of $f(r) \to \infty$ as $r \to 0$. The prediction results for the non-parametric model are shown in Table 5.5.

Table 5.5: Predictive results for the non-parametric model, with the percentage of burglaries which are predicted by the top 1%(156 grids), 5%(779 grids), 10%(1559 grids) and 20%(3117 grids) of 200m by 200m by intensity every hour over the first 3 months of 2016 shown, along with the information gain.

Top 1%	Top 5%	Top 10%	Top 20%	Inf. Gain
8.61%	22.37%	34.69%	52.98%	0.475

The results in terms of predictive performance in Table 5.5 are broadly in line with those seen when the triggering function was made parametrically in Table 5.4, with a larger information gain for the non-parametric weighted histograms being seen for the test events. This suggests the weighted histograms are fitting the burglary data better than the parametric triggering function, even if we do not observe any great improvements in the predictions over the events tested.

One obvious issue when fitting the weighted histograms is how to select the bins for g and h. We can see what happens when we fit the model as before, only selecting the bins for $h(r)$ to be every 50 metres instead of 100 metres in Figure 5.11.

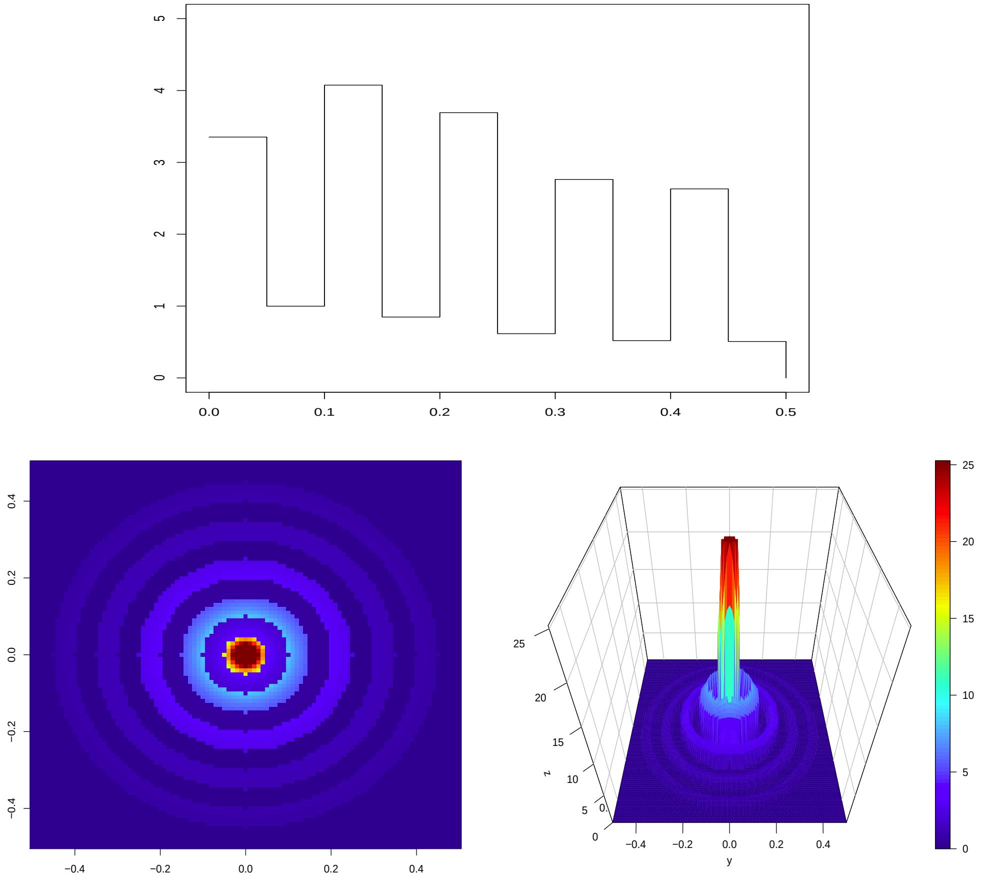

Figure 5.11: Result of the estimated triggering function when we make the discretisation δr to be 50m instead of 100m. The histogram for h is shown on the top, with the bottom two diagrams illustrating the resultant g.

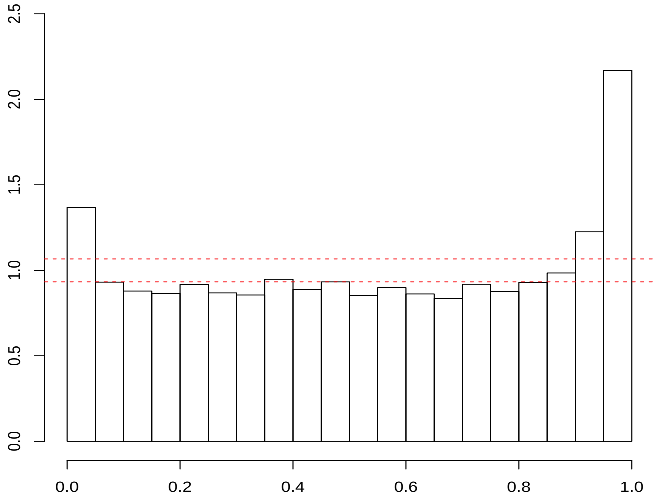

Figure 5.12: Histograms of the PIT values for the Chicago data when estimated with the non-parametric model, with a 95% confidence interval for the heights of the histogram shown as a red dotted line.

We can observe from Figure 5.11 that the spikes seen for h fall approximately where we would expects streets to be in Chicago, i.e. where the events are recorded as having occurred. Figure 5.12 shows the PIT histogram for this model, which indicates that more Voronoi cells have been underpredicted than we would expect if our model was a true reflection of the data. Prediction results for this version of the model is shown in Table 5.6. We can see the information gain for this model has increased, while the predictive performance remains relatively similar. It should be noted that the predictions were made numerically by averaging the estimated intensity in each 200m x 200m grid over 10 randomly selected points within. Therefore while this version of the model may estimate the data more accurately in terms of the exact points where an event will happen, this isn't necessarily reflected in the predictions which are made over 200m x 200m grids. Refining the bins for $h(r)$ to every 50 metres isn't certain to result in better predictions on this scale, even though it may model the data more accurately in terms of the 'information gain' seen for the exact locations where events occur.

Table 5.6: Predictive results for the non-parametric model with h estimated with smaller bins, with the percentage of burglaries which are predicted by the top 1%(156 grids), 5%(779 grids), 10%(1559 grids) and 20%(3117 grids) of 200m by 200m by intensity every hour over the first 3 months of 2016 shown, along with the information gain.

Top 1%	Top 5%	Top 10%	Top 20%	Inf. Gain
8.71%	22.66%	34.59%	52.28%	0.527

5.3 Summary

In this chapter we have estimated the process of burglary events in Chicago in a non-parametric way, first by using Model Independent Stochastic Declustering to estimate the background rate of burglary, and then using this to estimate the triggering effect itself. We have seen by doing this we can develop an effective predictive model for where burglary is likely to occur, along with insight into where and when future burglaries are likely to occur. Our developed model isn't without its drawbacks, with a certain arbitrariness of selecting how to model the background and triggering rate involved. In terms of the background rate, there are some regions which may require us to look even more locally than we did, and others where burglary events are more sparse, and this is not required. Similarly for the triggering function we do not have a fully automated way to select the 'bins' when estimating g and h.

Chapter 6

Models Using Kernel Density Estimation

The models we utilised in the previous chapter using Model Independent Stochastic Declustering allowed us to construct data-driven self-exciting point processes without making assumptions about the form which they took. While there were certain advantages to using these models, there were also drawbacks, including the triggering function not being continuous, and the arbitrariness of how to model the background rate and triggering effect. In this chapter we will seek to address these issues by using kernel density estimation (KDE) to estimate the background and triggering rate.

6.1 Mohler et al. model

The self-exciting point process proposed by Mohler et al. [48] to model residential burglary took the form

$$\lambda(x, y, t|\mathcal{H}_t) = \nu(t)\mu(x, y) + \sum_{t>t_i} g(x - x_i, y - y_i, t - t_i), \tag{6.1.1}$$

where $\mathcal{H}_t$ is the history of all events (x_i, y_i, t_i) up to time t. We will compare (6.1.1) with slightly altered versions of this model in order to show how such modifications could improve the predictive power when applied to real life point processes. In our tests we use this Mohler model with a non-temporal background rate, i.e.,

$$\lambda(x, y, t|\mathcal{H}_t) = \mu(x, y) + \sum_{t>t_i} g(x - x_i, y - y_i, t - t_i). \tag{6.1.2}$$

6.1.1 Kernel Density Estimation

Kernel density estimation provides a framework with which a probability density function can be estimated non-parametrically while providing a continuous estimate and without the dependence on the end points of bins which exist when fitting histograms [67]. If we have n data points x_i, the kernel density estimator $\hat{f}(x)$ of the distribution x is drawn from can be defined as

$$\hat{f}(x) = \frac{1}{n\Delta} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\Delta}\right), \tag{6.1.3}$$

where K is a kernel function and Δ is an associated bandwidth. A common choice for K is the Gaussian kernel

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right). \tag{6.1.4}$$

An example of KDE implemented on simulated data is shown in Figure 6.1. While KDE avoids the question of bin placement and the lack of continuity of histograms, it provides other potential issues, namely in the selection of an appropriate bandwidth. The right hand side of Figure 6.1 displays kernel density estimations on simulated data for three different bandwidths, showing that if too small a bandwidth is selected we can overfit the data, while if too large a bandwidth is selected we can underfit the data. This data was obtained by sampling from two normal distributions with differing means, thus the bimodality of this data.

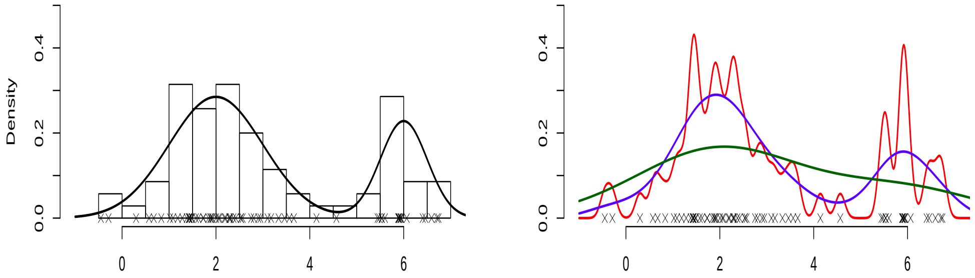

Figure 6.1: On the left is a histogram for simulated data generated from known distribution (black). On the right is kernel density estimations for this data produced when using different bandwidths with a Gaussian kernel. The crosses at the bottom of each plot represent a simulated data point.

6.1.2 Mohler et al. Monte-Carlo Iterative Procedure

Mohler et al.'s model uses variable bandwidth KDE where the bandwidth around each data point is selected according to nearest neighbour distances [48]. The general idea behind this is that data points which occur in close proximity to other data points should use a smaller bandwidth in the kernel density estimator, while data points which exist in 'isolation' from other data points, i.e. outliers, should be subject to a greater bandwidth in their contribution to the estimator.

We recall from Section 2.1 that the EM-algorithm views the point process as an incomplete data problem, where we don't know whether each data point is a background event or a triggered event, but can calculate the associated probabilities for each event [70]. The E-step of the EM-algorithm to find parameters for point processes consists of building an $n \times n$ P-matrix, where n is the number of events which have occurred, the entries p_{ii} represent the probability that event i was a background event, and p_{ij} that it was caused by event j. For the above model (6.1.2) these probabilities are given as

$$p_{ii} = \frac{\mu(x_i, y_i, t_i)}{\lambda(x_i, y_i, t_i)}, \tag{6.1.5}$$

$$p_{ij} = \frac{g(x_i - x_j, y_i - y_j, t_i - t_j)}{\lambda(x_i, y_i, t_i)}. \tag{6.1.6}$$

Given an initial guess of the matrix P, we have $N(N+1)/2$ probabilistic data points $\{(x_i, y_i, t_i, p_{ii})\}_{i=1}^{N}$ and $\{(x_i - x_j, y_i - y_j, t_i - t_j, p_{ij})\}_{i>j}$ [48]. We use this data to estimate μ and g in the so-called M-step, update our P-matrix using these new estimates for μ and g, and continue to iterate between these two steps until convergence.

The fact we have $N(N+1)/2$ data points makes the use of KDE computationally expensive. The procedure by Mohler et al. gets around this problem by taking samples of background events $\{x_i^b, y_i^b\}_{i=1}^{N_b}$ and offspring/parent interpoint distances $\{x_i^o, y_i^o, t_i^o\}_{i=1}^{N_o}$ from the P-matrix, with $N_b + N_o = N$, greatly reducing the number of data points with which to build our kernel density estimators [48].

To describe this process in greater detail, let $P^{(0)}$ be the initial guess of the P-matrix

$$P^{(0)} = \begin{pmatrix} p_{11} & 0 & \dots & 0 \\ p_{21} & p_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ p_{N1} & p_{N2} & \dots & p_{NN} \end{pmatrix}, \tag{6.1.7}$$

where p_{ij} for $i > j$ represents the probability that event i was triggered by event j, and p_{ii} represents the probability that event i was a background event, with $\sum_{j=1}^{N} p_{ij} = 1$. We then sample N_b background events and N_o triggered events from the P-matrix, reducing our number of data points from $N(N+1)/2$ to $N_b + N_o = N$. In effect, based on the probabilities in the P-matrix, we 'randomly' assign each event as either being a background event, or as being an event triggered by a specific previous one (in reality we can not say with certainty whether an event is a background event or has been triggered).

This works as follows: we take a single sample from each row of the P-matrix according to the probabilities, i.e., for the k-th row we take a single sample which with probability p_{kk} will give $\{x_k, y_k\}$ as a background event, with probability p_{k1} will give $\{x_k - x_1, y_k - y_1, t_k - t_1\}$ as an offspring/parent interpoint distance, with probability p_{k2} will give $\{x_k - x_2, y_k - y_2, t_k - t_2\}$ as an offspring/parent interpoint distance, and so on. We are left with a sample of background events $\{x_i^b, y_i^b\}_{i=1}^{N_b}$, and a sample of offspring/parent interpoint distances $\{x_i^o, y_i^o, t_i^o\}_{i=1}^{N_o}$, with $N_b + N_o = N$.

Variable bandwidth KDE is used on these samples in the following way. Firstly, the data $\{x_i^o, y_i^o, t_i^o\}_{i=1}^{N_o}$ is scaled to have unit variance in each coordinate, and based on this data the k'th 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 $\sigma_x, \sigma_y, \sigma_t$ be the sample standard deviation of each coordinate, the triggering function is estimated as

$$\begin{aligned} g(x, y, t) =& \frac{1}{N} \sum_{i=1}^{N_o} \frac{1}{\sigma_x \sigma_y \sigma_t (2\pi)^{(3/2)} D_i^3 K_i} \\ & \times \exp\left(-\frac{(x - x_i^o)^2}{2\sigma_x^2 D_i^2} - \frac{(y - y_i^o)^2}{2\sigma_y^2 D_i^2} - \frac{(t - t_i^o)^2}{2\sigma_t^2 D_i^2} \right), \end{aligned} \tag{6.1.8}$$

which we will hereby refer to as the MM (Mohler Model). K_i is a normalisation constant for each data point i

$$K_i = \int_0^\infty \frac{1}{\sigma_t \sqrt{2\pi} D_i} \exp\left(-\frac{(t - t_i^o)^2}{2\sigma_t^2 D_i^2} \right) \mathrm{d}t = 0.5 \left(1 + \operatorname{erf}\left(\frac{t_i^o}{\sqrt{2} D_i \sigma_t} \right) \right), \tag{6.1.9}$$

which is required due to the fact that we 'lose' the density in the negative time region. This normalisation doesn't fully address the problem associated with there being a jump discontinuity for the time element of the triggering function due to the fact that events can't trigger events to happen in the past - we will look at a way of resolving this in future section 6.2.

It is found that for prediction purposes variable bandwidth KDE on the background μ is less accurate than fixed bandwidth KDE [48]. The background rate is estimated with a two dimensional Gaussian kernel used on the samples of background events at each iteration, with the fixed bandwidth determined using k-fold cross validation [66]. For k-fold cross validation, we randomly split the data into k parts. For each part we build the estimator using the data in the other $(k - 1)$ parts, and then calculate the log-likelihood score of the observations not used to build the estimator. The likelihood score function can be given as

$$\mathrm{CV}(h) = n^{-1} \sum_{i=1}^{n} \log \hat{f}_{-i}(X_i), \tag{6.1.10}$$

where $\hat{f}_{-i}$ represents the density built when omitting the fold containing X_i [66]. We seek the bandwidth h which maximises the value CV, and we locate this via grid search.

We state the full algorithm used by Mohler as follows:

Mohler Algorithm

1. Set $k = 0$, and select the number of nearest neighbours we want to correspond to the bandwidth of the spatial triggering kernel density estimation nn_{trig}.

2. Make an initial guess P_k for the P matrix, 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 $t_i - t_j$, $|x_i - x_j|$ 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 from the P-matrix, with $N_b + N_o = N$.

4. Using k-fold cross validation, find σ_x^b and σ_y^b which maximise the log-likelihood score for unseen data via grid search. The background rate is

$$\mu(x, y) = \frac{1}{T} \sum_{i=1}^{N_b} \frac{1}{\sigma_x^b \sigma_y^b 2\pi} \exp\left(-\frac{(x - x_i^b)^2}{2{\sigma_x^b}^2} - \frac{(y - y_i^b)^2}{2{\sigma_y^b}^2} \right). \tag{6.1.11}$$

5. Using the sampled triggered events (x_i^o, y_i^o, t_i^o), we scale this data to have unit variance in the x, y and t components. Based on this we calculate $D_{i,\text{trig}}$, the nn_{trig}th nearest neighbour three-dimensional Euclidean distance to each data point i. Then, transforming the data back to its original scale, and with σ_x^b, σ_y^b and σ_t^b the standard deviations in the x, y and t directions, we calculate the

triggering function as

$$\begin{aligned} g(x,y,t) =& \frac{1}{N}\sum_{i=1}^{N_o}\frac{1}{\sigma_x\sigma_y\sigma_t(2\pi)^{(3/2)}D_i^3K_i} \\ & \times \exp\left(-\frac{(x-x_i^o)^2}{2\sigma_x^2D_i^2}-\frac{(y-y_i^o)^2}{2\sigma_y^2D_i^2}-\frac{(t-t_i^o)^2}{2\sigma_t^2D_i^2}\right), \end{aligned} \tag{6.1.12}$$

where K_i is the normalisation constant we found previously.

6. With the new μ and g we update our P matrix P_{k+1}, and set $k = k + 1$.
7. If $k <$ itermax, go to step 3.

6.2 Mohler Model with Reflected Time

Rosser and Cheng [60] proposed an alteration to the Mohler model to address the jump discontinuity at $t = 0$, namely reflecting the temporal component of the triggering function about $t = 0$. An example of using a reflected kernel density estimator is shown in Figure 6.2. As no density is 'lost' when doing this, the normalisation constant is not needed, and the triggering function is estimated as

$$\begin{aligned} g(x,y,t) =& \frac{1}{N}\sum_{i=1}^{N_o}\frac{1}{\sigma_x\sigma_y\sigma_t(2\pi)^{(3/2)}D_i^3}\exp\left(-\frac{(x-x_i^o)^2}{2\sigma_x^2D_i^2}-\frac{(y-y_i^o)^2}{2\sigma_y^2D_i^2}\right) \\ & \times\left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2D_i^2}\right)+\exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2D_i^2}\right)\right). \end{aligned} \tag{6.2.1}$$

The model in 6.2.1 will be referred to as the MMrt (Mohler Model with reflected time).

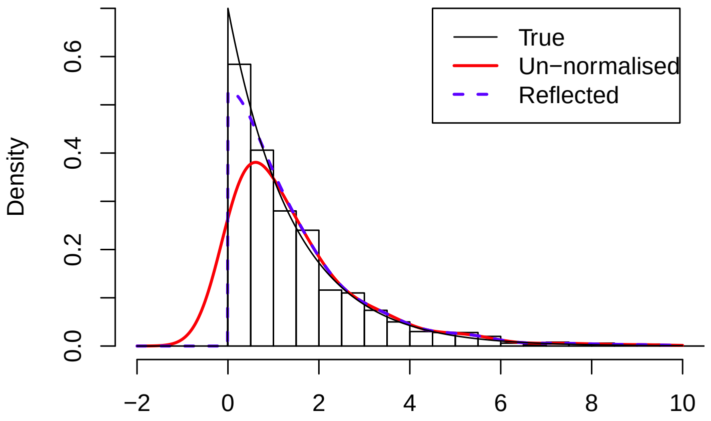

Figure 6.2: An example of KDE on a sample taken from an exponential distribution. The un-normalised result of kernel density estimation is shown in red. The reflected kernel density estimator shown in blue gives a result closer to the true distribution due to there being no values below 0.

6.3 Rosser and Cheng Isotopic Model

Rosser and Cheng also proposed a type of isotropic model as an alternative to the Mohler model. Taking $r = \sqrt{x^2 + y^2}$, the proposed triggering function becomes

$$\begin{aligned} g(r,t) =& \frac{1}{N} \sum_{i=1}^{N_o} \frac{1}{\sigma_t \sqrt{2\pi} K_i} \exp\left(-\frac{(r - r_i^o)^2}{2\sigma_r^2}\right) \\ & \times \left(\exp\left(-\frac{(t - t_i^o)^2}{2\sigma_t^2}\right) + \exp\left(-\frac{(-t - t_i^o)^2}{2\sigma_t^2}\right)\right). \end{aligned} \tag{6.3.1}$$

The model in 6.3.1 will be referred to as the RCM (Rosser and Cheng Model). As we don't have $r < 0$, and we need to normalise the function across the whole plane, we have normalisation constant

$$\begin{aligned} K_i &= \int_{-\pi}^{\pi}\int_0^{\infty} \exp\left(-\frac{(r-r_i^o)^2}{2\sigma_r^2}\right) r\mathrm{d}r\mathrm{d}\theta \\ &= \sigma_r r_i^o \sqrt{2\pi^3}\left(1+\operatorname{erf}\left(\frac{r_i^o}{\sqrt{2}\sigma_r}\right)\right) + 2\sigma_d^2\pi\exp\left(-\frac{r_i^{o2}}{2\sigma_d^2}\right). \end{aligned} \tag{6.3.2}$$

Note this is slightly different to the normalisation constant given by Rosser and Cheng [60]. There is an inherent problem with this implementation of an isotropic function. The kernel density estimator applies it's kernel to the distance r at which an event is said to have taken place at, and this is then applied across the whole plane and normalised. However, if we look at the corresponding probability distribution function of the resultant estimator, we see there is an inherent bias where the expected distance at which the kernel will predict a future event to occur at is greater than the sample it has been given. An example of this is shown in Figure 6.3.

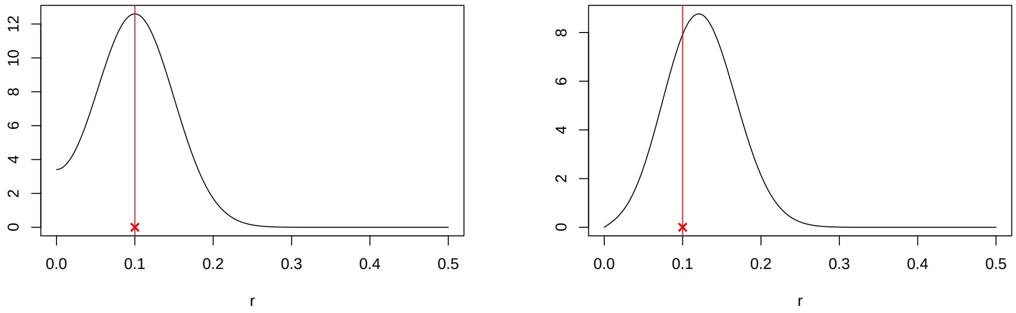

Figure 6.3: On the left is an example of the height of this isotropic function applied to one data point against the distance r from the centre. On the right is the corresponding probability density function for the distance r which we can expect a triggered event to happen at. We can see that there is an inherent bias in the estimator, where the expected value of the distribution is greater than that of the sample event, which may cause problems in terms of predicting future events.

6.4 Proposed New Isotropic Function

To address the potential problems in the isotropic function by Rosser and Cheng, we propose developing such a function in a different way. Fox et al. [27] incorporated an isotropic triggering function by specifying a histogram density estimator of the form $h(r,t) = 2\pi r g(x,y,t)$, since $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y,t)\,dx\,dy = \int_0^{\infty} 2\pi r g(r,t)\,dr = 1$, where the triggering function is assumed to be isotropic, i.e., $g(x,y,t) = g(x^2+y^2,t)$, which we implemented in the previous chapter in Section 5.2. In a similar way we will use the kernel density estimation to approximate the probability density function of the distance r at which events will be expected to happen.

Analogously to the Mohler model, 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

$$K_i = \int_0^{\infty} \frac{1}{\sigma_r D_i \sqrt{2\pi}} \exp\left(-\frac{(r-r_i^o)^2}{2\sigma_r^2 D_i^2}\right) \mathrm{d}r = 0.5\left(1+\operatorname{erf}\left(\frac{r_i^o}{\sqrt{2}\sigma_r D_i}\right)\right), \quad (6.4.1)$$

so that the density of each kernel is 1. We then have the estimator of the probability distribution function $\hat{h}_k$

$$\begin{aligned}\hat{h}_k(r,t) =& \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t\sigma_r D_i^2 2\pi K_i}\left(\exp\left(-\frac{(r-r_i^o)^2}{2\sigma_r^2 D_i^2}\right)\right)\\ &\times\left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right). \end{aligned} \quad (6.4.2)$$

To find the corresponding value of the function $\hat{g}(x,y,t)$, we have $\hat{g}(x,y,t) = \hat{h}_k(r,t)/(2\pi r)$ with $r = \sqrt{x^2+y^2}$, so

$$\begin{aligned}\hat{g}(x,y,t) =& \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t\sigma_r D_i^2 4\pi^2 r K_i}\left(\exp\left(-\frac{(r-r_i^o)^2}{2\sigma_r^2 D_i^2}\right)\right)\\ &\times\left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right). \end{aligned} \quad (6.4.3)$$

The proposed isotropic model in (6.4.3) will be referred to as the PIM. An example of the implementation of this estimator is shown in Figure 6.4. A potential problem associated with this estimator is that, similarly to what we saw in Section 5.2.2, as the kernel density estimator is estimating the pdf for the distance at which an event will occur from the origin, we will typically have $\hat{h}_k(r,t) > 0$ for r very close to 0, leading to $\hat{g}(r,t) \to \infty$ at these points, which may lead to potential problems if we make predictions based on point estimates. We attempt to address these concerns by limiting the value the intensity can take very close to the origin, by making the value of the function for distances within a certain r of the origin, say 0.01, take the value of the function at $r = 0.01$, i.e.,

$$\begin{aligned}\hat{g}(x,y,t) =& \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{0.04\sigma_t\sigma_r D_i^2\pi^2 K_i}\left(\exp\left(-\frac{(0.01-r_i^o)^2}{2\sigma_r^2 D_i^2}\right)\right) \\ &\times\left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)+\exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right), \text{ for } r < 0.01.\end{aligned} \tag{6.4.4}$$

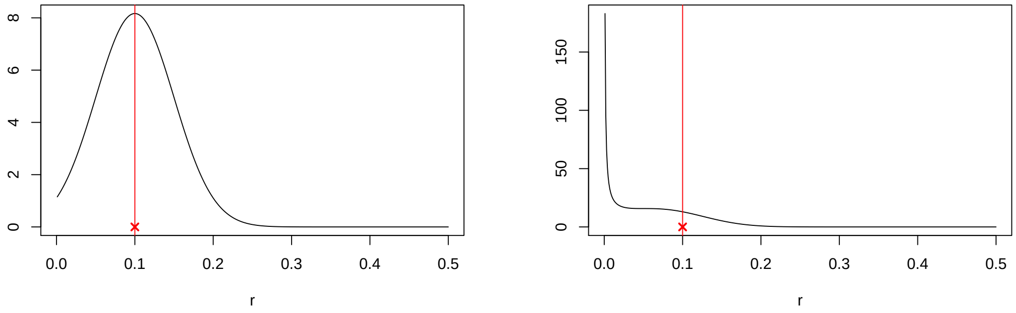

Figure 6.4: On the left is the estimated pdf $\hat{h}_k$ applied to a single event. On the right is the corresponding value of $\hat{g}(r,t)$ for all $r = \sqrt{(x^2+y^2)}$ in the x-y plane, found by $\hat{g}(r,t) = \hat{h}_k(r,t)/(2\pi r)$. A potential problem when it comes to predicting is that as $\hat{h}_k > 0$ for $r \approx 0$, then as r approaches 0 we have $\hat{g}(r,t) \to \infty$.

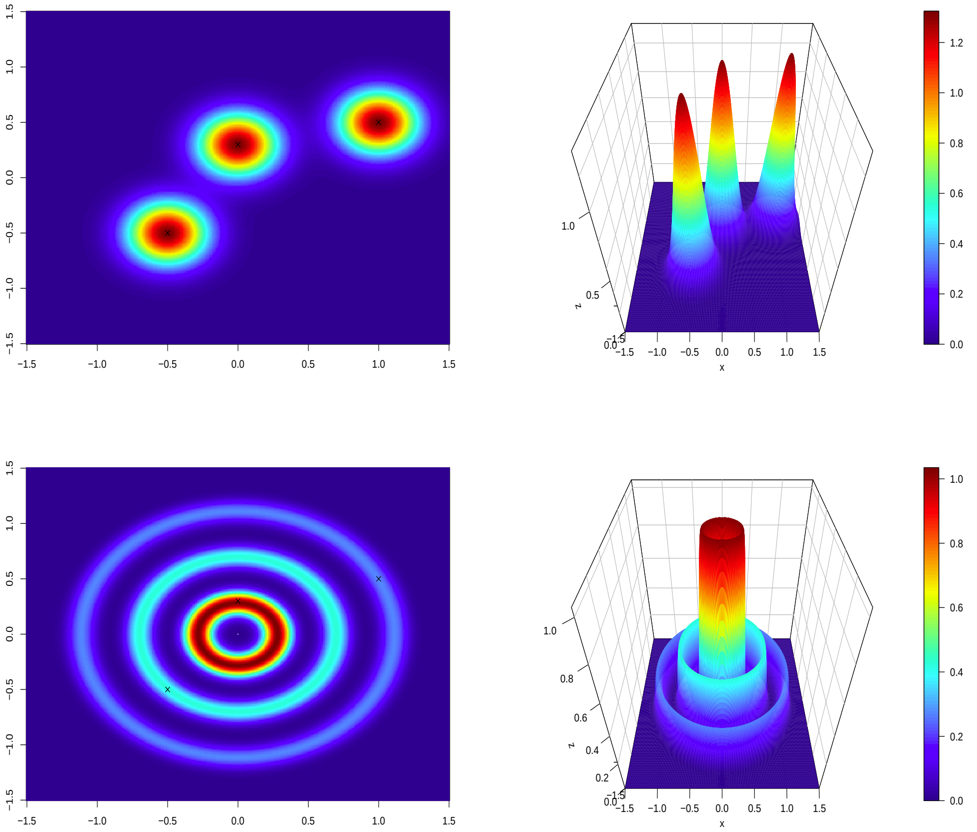

Figure 6.5: The top two figures display an example of the Mohler model in the x-y direction applied to three data points, and the bottom two show the isotropic model applied to the same data, with the locations of the events represented by the crosses.

6.5 Model Based on Manhattan Distance

The road network in the city of Chicago is laid out on a grid plan (see Figure 6.6). This suggests that the distance between spatially located events may be better measured using Manhattan distance rather than Euclidean distance.

With a small alteration to the algorithm we can also build a triggering function which uses the Manhattan distance, an approach which may give us better accuracy when trying to predict future crimes.

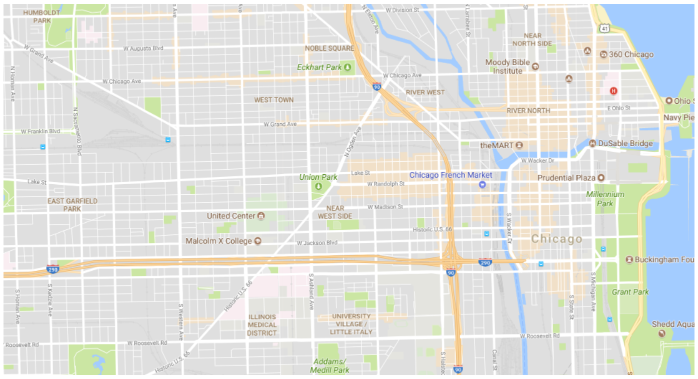

Figure 6.6: Example of the block structure of the road network in Chicago [5]. Map data ©2020 Google.

This is done in an almost identical way to the previous estimator. Using Manhattan distance $d = |x| + |y|$, we have the estimator $\bar{h}_k$ of the pdf

$$\begin{aligned}\bar{h}_k(d,t) =& \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t \sigma_d D_i^2 2\pi K_i} \left(\exp\left(-\frac{(d-d_i^o)^2}{2\sigma_d^2 D_i^2} \right) \right) \\ & \times \left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2} \right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2} \right) \right). \end{aligned} \tag{6.5.1}$$

To find the corresponding value of the function $\bar{g}(x, y, t)$, we have $\bar{g}(x, y, t) = \bar{h}_k(d, t)/4d$, so

$$\begin{aligned}\bar{g}(x,y,t) =& \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t \sigma_d D_i^2 8\pi d K_i} \left(\exp\left(-\frac{(d-d_i^o)^2}{2\sigma_d^2 D_i^2} \right) \right) \\ & \times \left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2} \right) + \exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2} \right) \right). \end{aligned} \tag{6.5.2}$$

This proposed model based on Manhattan distance will be referred to as the PMDM. An example of the implementation of this function is shown in Figure 6.7. Again we will face the same potential issues when looking at point estimates, namely that $\bar{h}_k(d, t) > 0$ for d very close to 0, and $\bar{g}(d, t) \to \infty$ at these points.

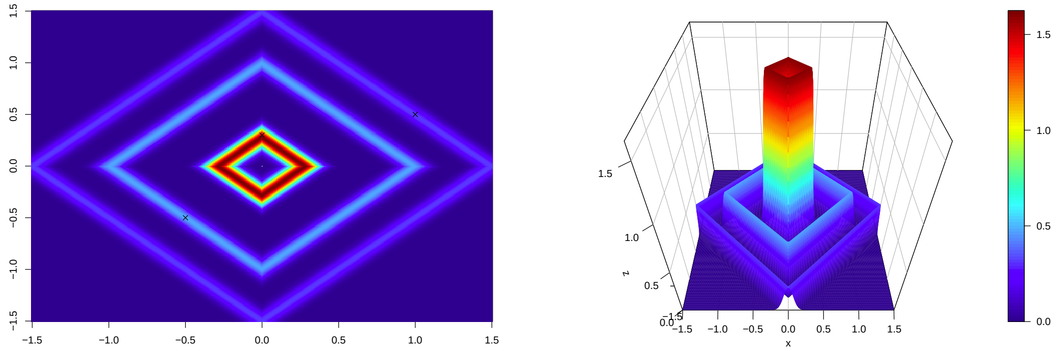

Figure 6.7: An example of the Manhattan function in the x-y direction when applied to three data points.

6.6 Model Based on Maximum Norm

In a similar way we also create a triggering function which just looks at the maximum distance travelled in either the x or y direction, which may be useful at detecting events which are triggered in parallel streets. It can be argued that in a city defined on a grid structure our perception of the distance between two points is dominated by the maximum of the two x-y distances. This model could potentially flag streets which are of increased risk after an initial event. By taking $m = \max(|x|, |y|)$, we have the estimator $\tilde{h}_k$ of the pdf

$$\begin{aligned}\tilde{h}_k(m,t) =& \frac{1}{N}\sum_{i=1}^{N_o} \frac{1}{\sigma_t \sigma_m 2 D_i^2 \pi K_i}\left(\exp\left(-\frac{(m-m_i^o)^2}{2\sigma_m^2 D_i^2}\right)\right)\\ &\times\left(\exp\left(-\frac{(t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)+\exp\left(-\frac{(-t-t_i^o)^2}{2\sigma_t^2 D_i^2}\right)\right). \qquad (6.6.1)\end{aligned}$$

To find the corresponding value of the function $\tilde{g}(x, y, t)$, we have $\tilde{g}(x, y, t) = \tilde{h}_k(m, t)/8m$, so

$$\begin{aligned}\tilde{g}(x, y, t) =& \frac{1}{N} \sum_{i=1}^{N_o} \frac{1}{\sigma_t \sigma_m D_i^2 16\pi m K_i} \left(\exp\left(-\frac{(m - m_i^o)^2}{2\sigma_m^2 D_i^2} \right) \right) \\ & \times \left(\exp\left(-\frac{(t - t_i^o)^2}{2\sigma_t^2 D_i^2} \right) + \exp\left(-\frac{(-t - t_i^o)^2}{2\sigma_t^2 D_i^2} \right) \right), \qquad (6.6.2)\end{aligned}$$

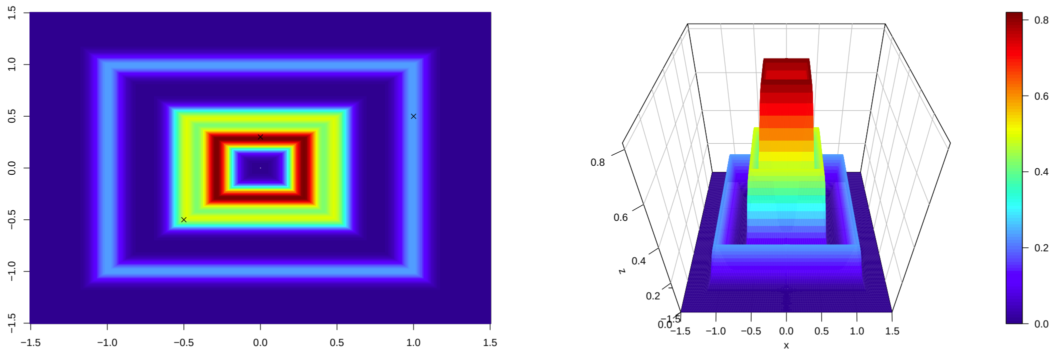

Figure 6.8: An example of the triggering function in the x-y direction when applied to three data points when the distance is based on the maximum norm $m = \max(|x|, |y|)$.

An example of the implementation of this function is shown in Figure 6.8.

6.7 Results on Simulated Data

We initially tested the previously described models on the spatial-temporal point processes in section 4.3.1, where 10 simulations of the form (4.3.3) were made with parameters $\alpha = 0.3$, $\omega = 0.5$, $\sigma = 0.1$, constant background rate $\mu = 0.02$, over time period $T = [0, 365]$. We fitted the Mohler model (MM) and the Mohler model with reflected time (MMrt) while finding D_i corresponding to the 15'th nearest neighbour, and the isotropic model, Manhattan model, and the max norm model with D_i corresponding to the 50'th nearest neighbour as we are working in fewer dimensions. Note that for all the models apart from the MM, time was reflected around $t = 0$.

In Figure 6.9 we show the estimated marginal in the t-direction,

$$g_{\text{marg}}(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y, t)\mathrm{d}x\mathrm{d}y, \tag{6.7.1}$$

for the 10 simulations for the MM model, and the MMrt model, against the true distribution the simulations are generated from. We can clearly see that the estimates for these simulations are superior for the MMrt model where the estimates have been reflected about $t = 0$.

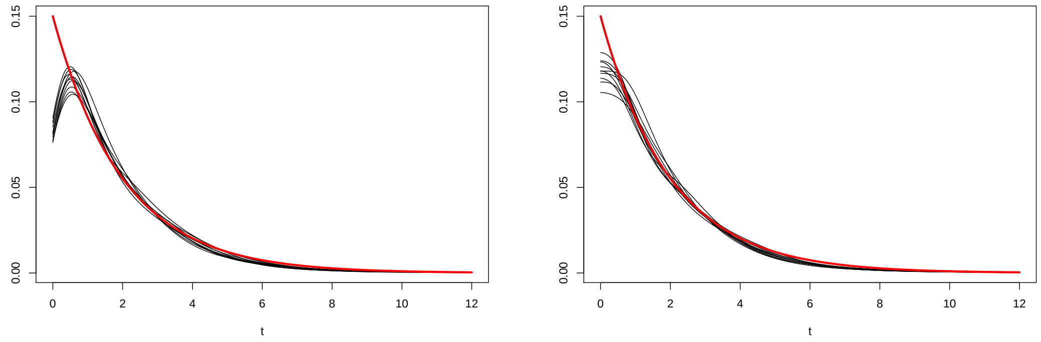

Figure 6.9: On the left is the estimated marginal $g_{\text{marg}}(t)$ for the 10 simulations for the MM model, and on the right is the estimated marginal for the MMrt model, with the true distribution being shown in red.

Figure 6.10 displays the marginal for the probability distribution h for the r component for both the estimates given by the Mohler model, and the isotropic model. While both models give good estimates of the true distribution, we can see the isotropic model slightly better approximates the true distribution the simulated data was generated from. It should be noted that the simulated data is isotropic in nature, and the Mohler model would be preferable if this was not the case.

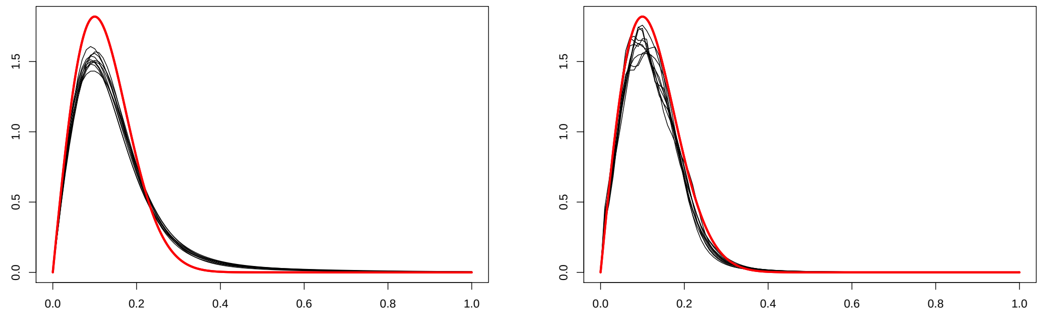

Figure 6.10: Displayed on the left is the estimated probability distribution $h(r)$ generated by the Mohler model with reflected time for the 10 simulations, and on the right is the estimated probability distribution for the isotropic model, with the true distribution shown in red.

Table 6.1: Mean $L2$ error of the triggering parameters for the 10 simulations compared with the true constant triggering rate for the different fitted models.

Model	Mean $L2$ error norm
MM	1.1×10^{-2}
MMrt	6.8×10^{-3}
Isotropic	5.0×10^{-3}
Manhattan	1.0×10^{-2}
Max norm	1.0×10^{-2}

Table 6.1 displays the average $L2$ error for the triggering function estimated by each model compared with the true triggering function the data was generated from. We can see that reflecting time around $t = 0$ gives a large prediction improvement for MMrt as opposed to the MM. We can see that the isotropic outperforms the MMrt for this data. While this is not surprising as the data is indeed isotropic, it does indicate that this model could lead to improved prediction if indeed the true triggering function is isotropic. We can also observe how the Manhattan and max norm model don't estimate the triggering function as well as the MMrt or isotropic model, which is to be expected as they assume this function is not isotropic.

6.8 Results on Real Data

We fitted each of the four aforementioned models (MMrt, Isotropic, Manhattan and Maximum Norm models) to the 13044 burglary events observed in Chicago in 2015. All the models were fitted with a fixed background bandwidth, which is recommended in [48] due to the spatial features of neighbourhoods. We selected $\sigma_x^b = \sigma_y^b = 0.15$, which was found to give good predictive results. All distances displayed are in km unless otherwise stated.

6.8.1 Results with Mohler Model

We fitted the MMrt to the burglary data selecting D_i relating to the 15'th nearest neighbour, as recommended in [48]. We limited triggering effects to only occur within 90 days of the original event and within one kilometre to reduce computational cost. The result found for the burglary events in 2015 is displayed in Figure 6.11.

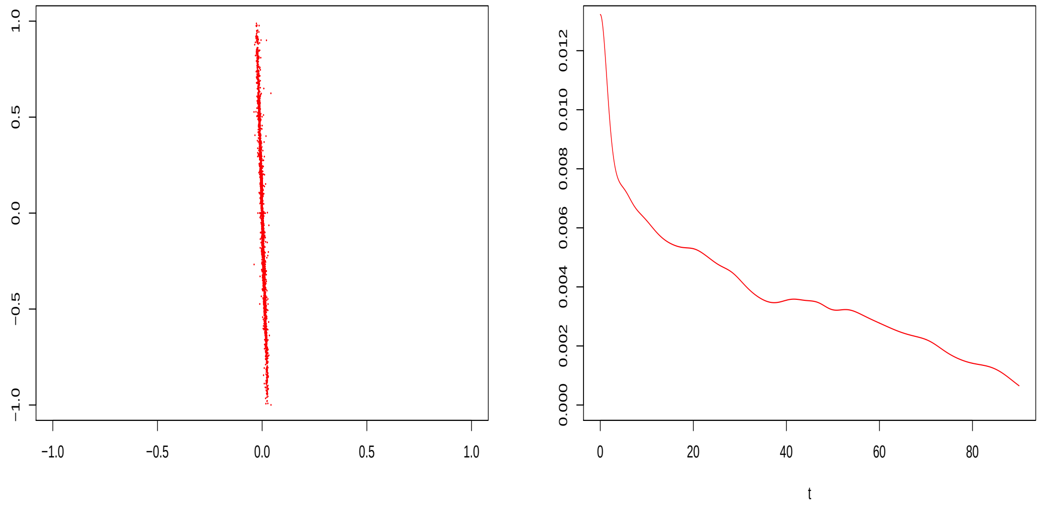

Figure 6.11: On the left are the positions of the assumed triggered events for the MMrt on the 100'th iteration of the EM-algorithm in the x-y direction, in reference to the event assumed to have triggered it at the origin of the graph, with D_i selected relating to the 15'th nearest neighbour. On the right is the estimated marginal in the t-direction.

We can observe in Figure 6.11 that the algorithm assesses that the triggered events almost exclusive occur in the North-South direction of the original crime. The t-marginal shows that the triggering effect is greatest in the days following an events, with this effect dropping off gradually as the time after the initial event grows to 90 days.

A histogram of the PIT values for this fit is shown in Figure 6.12. It displays that the model underpredicts the number of events in far more Voronoi cells than would be expected if the model was a good fit, suggesting that the model is not fitting the burglary data accurately.

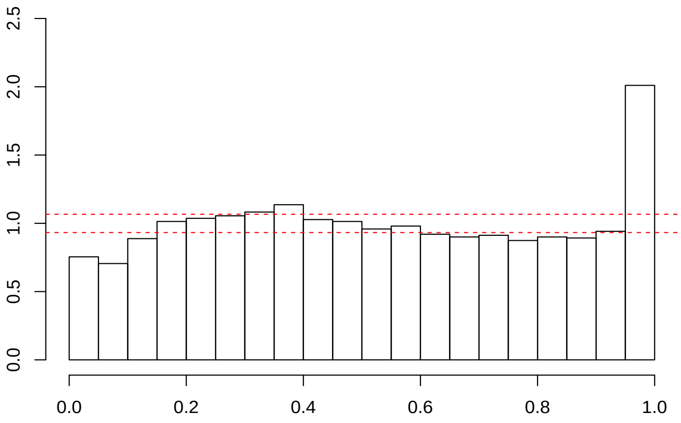

Figure 6.12: Histogram for the PIT values for the MMrt when fitted to the Chicago data with D_i selected relating to the 15'th nearest neighbour, with 95% confidence intervals shown as a red dotted line.

We then fitted the MMrt to the burglary data selecting D_i relating to the 30'th nearest neighbour, to investigate whether this resulted in a better fit to the data. Figure 6.13 shows that the algorithm assesses that triggered events occur predominantly in the North-South direction of the original crime, but not exclusively as when D_i was selected according to the 15'th nearest neighbour. The t-marginal shows that the triggering effect is greatest in the days following an events, with this effect dropping off quite rapidly over the first few weeks after a burglary.

A histogram of the PIT values for this fit is shown in Figure 6.14. It displays that the model underpredicts the number of events in slightly more Voronoi cells than would be expected, although provides a better fit than when D_i was selected according to the 15'th nearest neighbour. We also note that selecting D_i according to the 30'th nearest neighbour as opposed to the 15'th reduces the branching ratio from 0.346 to 0.054.

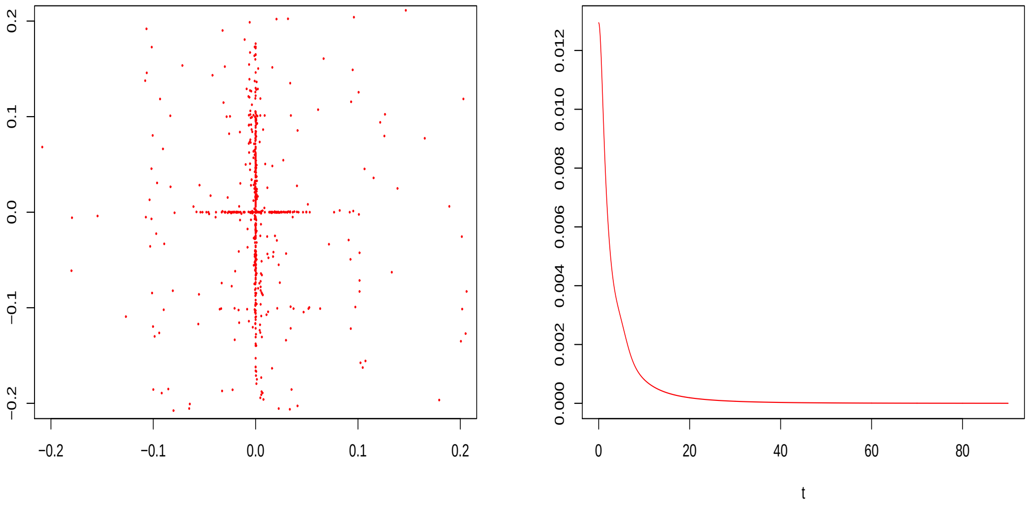

Figure 6.13: On the left are the positions of the assumed triggered events for the MMrt on the 100'th iteration of the EM-algorithm in the x-y direction, in reference to the event assumed to have triggered it at the origin of the graph, with D_i selected relating to the 30'th nearest neighbour. On the right is the estimated marginal in the t-direction.

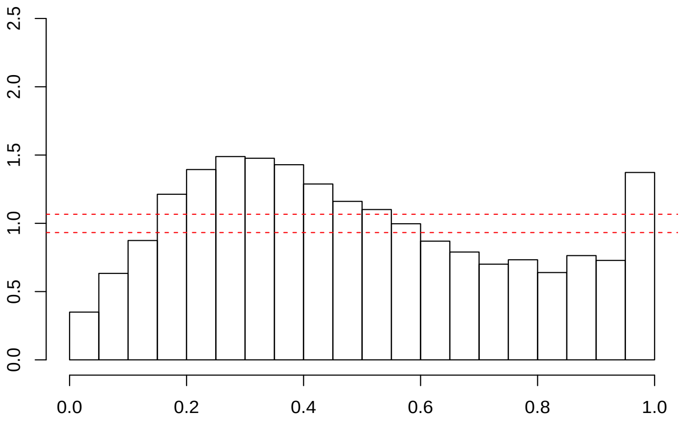

Figure 6.14: Histogram for the PIT values for the MMrt when fitted to the Chicago data with D_i selected relating to the 30'th nearest neighbour, with 95% confidence intervals shown as a red dotted line.

6.8.2 Results with Isotropic Model

We fitted the isotropic model to the burglary data selecting D_i relating to the 50'th nearest neighbour to reflect the reduced dimensionality of g. Again we limited triggering effects to only occur within 90 days of the original event and within one kilometre to reduce computational cost. The result found for the burglary events in 2015 is displayed in Figure 6.15.

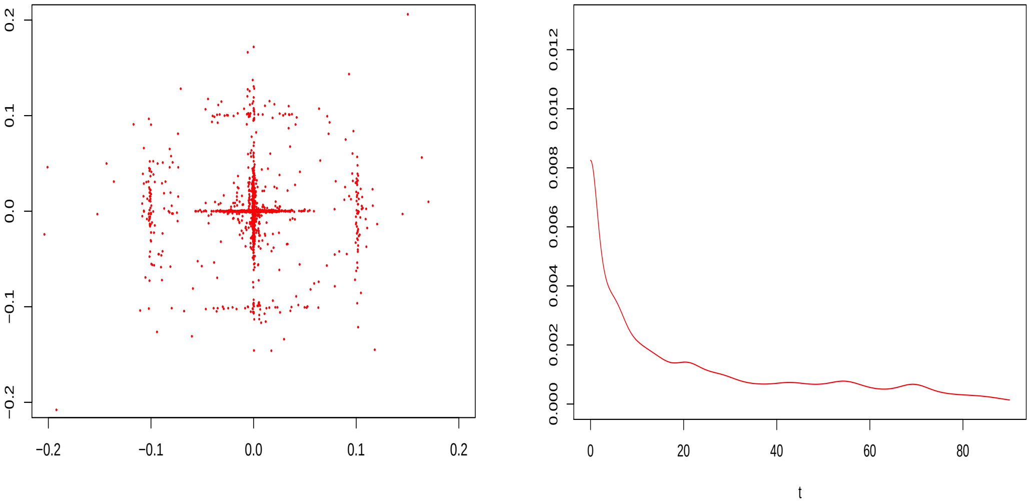

Figure 6.15: On the left are the positions of the assumed triggered events for the isotropic model on the 100'th iteration of the EM-algorithm in the x-y direction, in reference to the event assumed to have triggered it at the origin of the graph. On the right is the estimated marginal in the t-direction.

We can see that the algorithm assesses that a number of events are triggered in close proximity to the original event in both the North-South and East-West directions, with further events triggered in parallel streets to the original event. This is further demonstrated by the probability distribution of $\hat{h}_k$ in the r-direction displayed in Figure 6.16, which shows that along with the high risk of a triggered event occurring very near the original event, there is a clear increase in the risk of a triggered event occurring around $r = 0.1$ from the original event.

A histogram of the PIT values for this model is shown in Figure 6.17. It displays that the model underpredicts the number of events in slightly more Voronoi cells than would be expected, but less so than the MMrt.

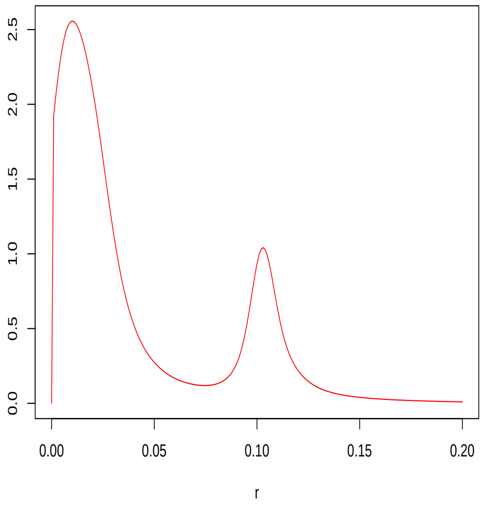

Figure 6.16: This is the estimated probability distribution of $\hat{h}_k$ in the r-direction.

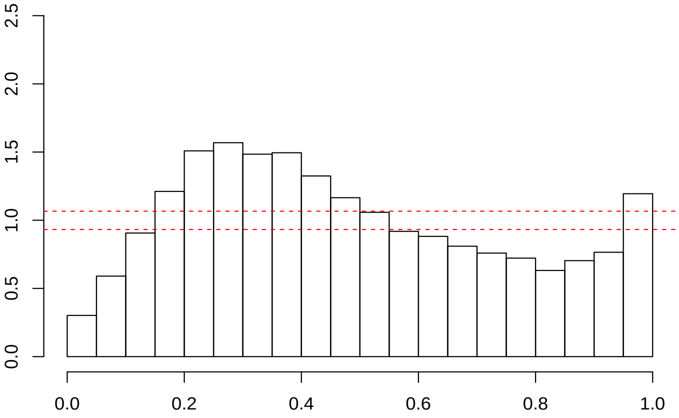

Figure 6.17: Histogram for the PIT values for the isotropic model when fitted to the Chicago data, with 95% confidence intervals shown as a red dotted line.

6.8.3 Results with Manhattan Model

We fitted the Manhattan model to the burglary data selecting D_i relating to the 50'th nearest neighbour. The triggering effects were limited to only occur within 90 days of the original event and within one kilometre of an original event. The result found for the burglary events in 2015 is displayed in Figure 6.18.

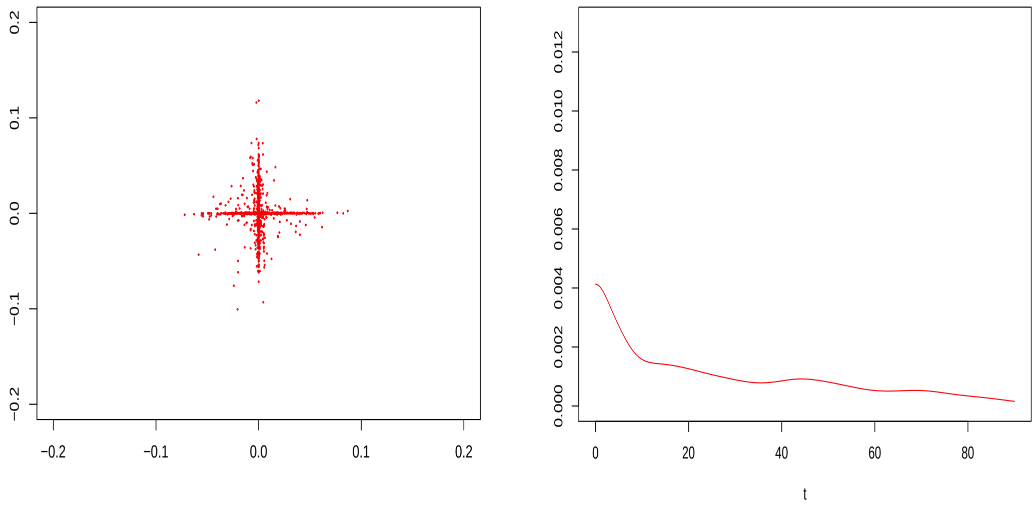

Figure 6.18: On the left are the positions of the assumed triggered events for the Manhattan model on the 100'th iteration of the EM-algorithm in the x-y direction, in reference to the event assumed to have triggered it at the origin of the graph. On the right is the estimated marginal in the t-direction.

As shown by the events assumed to be triggered in the 100'th iteration of the algorithm in Figure 6.18, along with the probability distribution in the d-direction in Figure 6.19, most events which are triggered according to the results occur within 100 metres of the original burglary. The t-marginal on the right of Figure 6.18 shows the greatest risk of a triggered event occurs in the days following a burglary, with this risk dropping fairly gradually up to 90 days following the original event.

Figure 6.20 displays a histogram of the PIT values for the Manhattan model. Similarly to the isotropic model, it displays that the model underpredicts the number of events in slightly more Voronoi cells than would be expected.

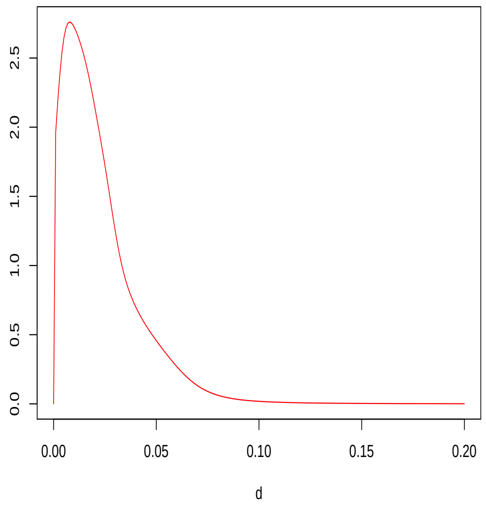

Figure 6.19: This is the estimated probability distribution of $\bar{h}_k$ in the d-direction.

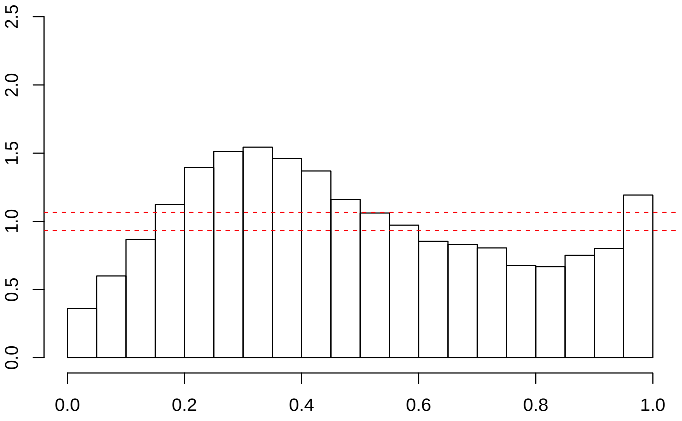

Figure 6.20: Histogram for the PIT values for the Manhattan model when fitted to the Chicago data, with 95% confidence intervals shown as a red dotted line.

6.8.4 Results with Maximum Norm Model

Finally we fitted the model using the maximum norm to the burglary data selecting D_i relating to the 50'th nearest neighbour. The assumed triggered events and marginal in the t-direction are displayed in Figure 6.21.

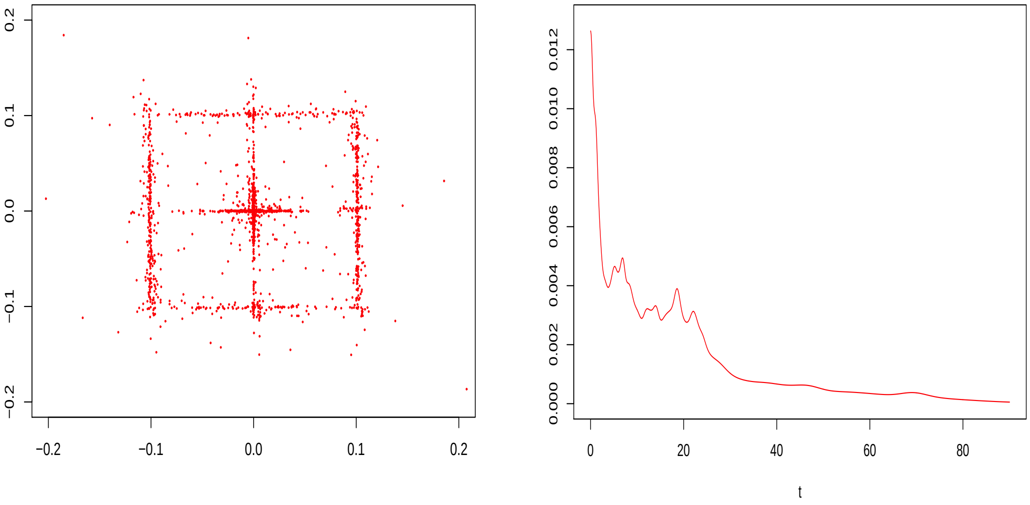

Figure 6.21: On the left are the positions of the assumed triggered events for the maximum norm model on the 100'th iteration of the EM-algorithm in the x-y direction, in reference to the event assumed to have triggered it at the origin of the graph. On the right is the estimated marginal in the t-direction.

We can see that when fitting this model, events are triggered in close proximity to the original event, along with in the streets immediately parallel to the original event. The t-marginal displays the greatest risk of a triggered event occurs in the immediate aftermath of an event, with the risk showing an increase at several periods after the original event, notably including around after 7 days of the original event, in line with previous findings on crime data [48]. The marginal in the m-direction in Figure 6.22 further demonstrates the increased risk of a triggered occurring at around $m = 0.1$ from the initial event.

Figure 6.23 displays a histogram of the PIT values for the maximum norm model. Again it displays that the model underpredicts the number of events in slightly more Voronoi cells than would be expected as with the isotropic and Manhattan models.

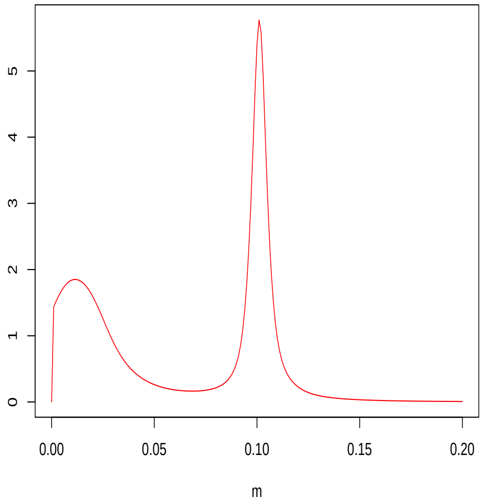

Figure 6.22: This is the estimated probability distribution of $\tilde{h}_k$ in the m-direction.

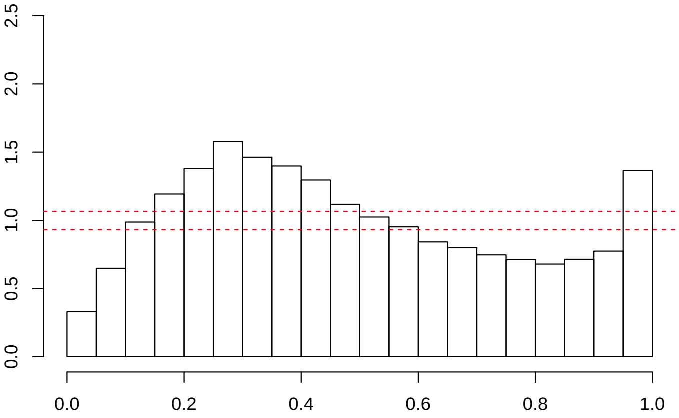

Figure 6.23: Histogram for the PIT values for the maximum norm model when fitted to the Chicago data, with 95% confidence intervals shown as a red dotted line.

6.9 Prediction Results

In this section we will summarise the predictive accuracy of the four models we have fitted, along with the predictive results of a reference model, namely using kernel density estimation on the previous year's events while assuming there is no triggering effect. In Figure 6.24 we display the difference in the assumed triggered events for each model for comparison, and in Figure 6.25 we show the different t-marginals estimated for the four different models. These two figures combine the images in Figures 6.13, 6.15, 6.18, and 6.21.

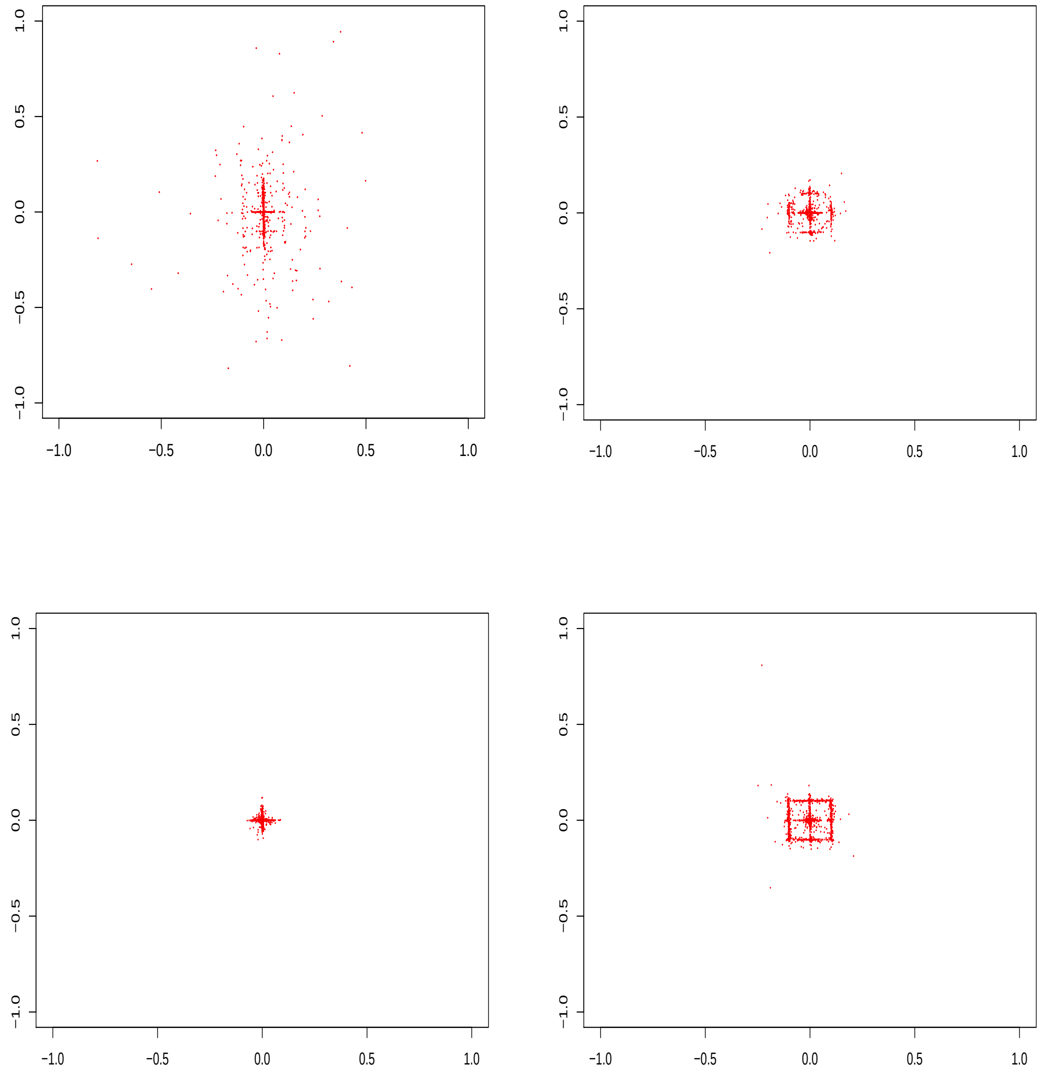

Figure 6.24: These figures display the locations of events which are assumed to be triggered on the 100'th iteration of the EM-algorithm, combining the images in Figures 6.13, 6.15, 6.18, and 6.21. On the top left is the assumed triggered events for the MMrt with D_i selected according to the 30'th nearest neighbour, the top right for the Isotropic model, the bottom right for the model using Manhattan distance, and the bottom right for the model using the maximum norm.

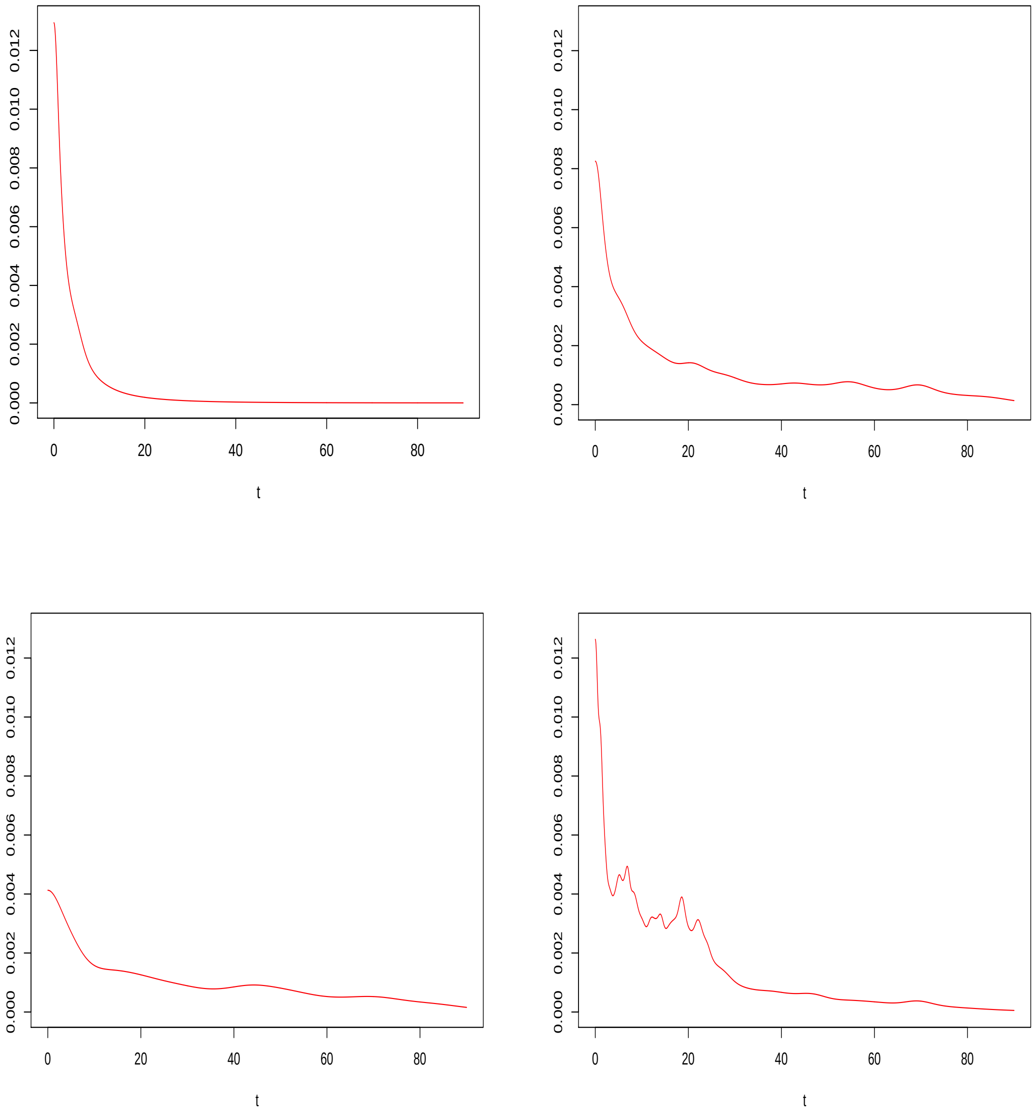

Figure 6.25: Displayed are the estimated marginals in the t-direction for the MMrt with D_i selected according to the 30'th nearest neighbour(top left), the isotropic model(top right), the model using Manhattan distance(bottom left), and the model using the maximum norm(bottom right). These marginals for each model were previously displayed in Figures 6.13, 6.15, 6.18, and 6.21.

Table 6.2: Estimated branching ratio for the four models when fitted with burglary data from 2015.

Model	Branching Ratio
MMrt (15'th nn)	0.346
MMrt (30'th nn)	0.054
Isotropic	0.104
Manhattan	0.089
Max norm	0.131

Table 6.2 displays the branching ratio, or how many events an event can be expected to trigger, of each model. Under the MMrt with D_i selected accorded to the 15'th nearest neighbour, far more of a triggering effect is detected than with the other models. When we take into account the histogram of the PIT values for this model, along with the fact that the algorithm converged to predicting all the triggered events occur exclusively in the North-South directions from the original event, this leads us to speculate that this model is overfitting some particular element of the data, rather than giving us a good fit for prediction purposes.

Table 6.3 displays the prediction results for the four different models, with results for the MMrt when fitted with both selections of D_i, along with the case where we have fitted the model as a spatial Poisson process with no self-excitation, using kernel density estimation for the previous 12 months data, i.e. estimating the intensity as

$$\lambda(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2\pi\sigma^2 T} \exp\left(\frac{-(x - x_i)^2 - (y - y_i)^2}{2\sigma^2} \right), \tag{6.9.1}$$

with $\sigma = 0.15$. The predictions are done hourly based on the ranking of the 200 x 200m grids by intensity as before, where the intensity is estimated by taking the average of 10 randomly selected points within each grid cell. We can see in terms of information gain the MMrt with D_i selected with the 15'th nearest neighbour outperforms the other models, but in terms of prediction is outperformed by the the other 3 models and the MMrt with D_i selected with the 30'th nearest neighbour, along with simple kernel density estimation with no self-excitation. This further suggests the MMrt with D_i selected with the 15'th nearest neighbour is overfitting some element of the data, i.e. the model appears to be learning the specific geometry of the North-South roads in Chicago, without allowing us to predict future events accurately. The other three self-exciting point process models we fitted and the MMrt with D_i selected with the 30'th nearest neighbour give superior prediction results to the simple KDE model, with the model using the maximum norm providing slightly superior results to the other models.

Table 6.3: Predictive results for the four models and the simple KDE model. The percentage of burglaries which are predicted by the top 1%(156 grids), 5%(779 grids), 10%(1559 grids) and 20%(3117 grids) of 200m by 200m by intensity every hour over the first 3 months of 2016 is shown, along with the information gain for each model.

Model	Top 1%	Top 5%	Top 10%	Top 20%	Inf. Gain
MMrt (15'th nn)	5.57%	18.46%	30.28%	50.53%	1.028
MMrt (30'th nn)	7.85%	22.06%	33.83%	53.11%	0.541
Isotropic	7.46%	22.27%	34.86%	54.54%	0.602
Manhattan	7.19%	21.67%	34.26%	53.45%	0.622
Maximum Norm	7.62%	22.56%	35.39%	54.64%	0.627
KDE	7.32%	20.58%	33.27%	53.38%	0.478

In Figure 6.26 we compare the prediction rates with MMrt model with D_i selected according to the 30'th nearest neighbour and the maximum norm model. We can see in terms of prediction the maximum norm model slightly outperforms the MMrt model.

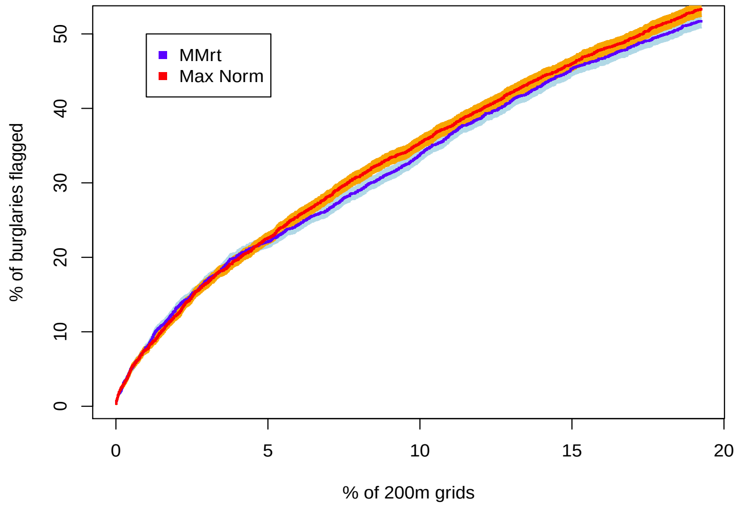

Figure 6.26: The prediction success of the MMrt model with D_i selected according to the 30'th nearest neighbour and the maximum norm model is displayed, with error bars showing one standard error for each method.

6.10 Background Rate Varying in Time

Finally, we fit the four previous models while allowing the background rate to vary in time as in [48], i.e.

$$\lambda(x, y, t|\mathcal{H}_t) = \nu(t)\mu(x, y) + \sum_{t>t_i} g(x - x_i, y - y_i, t - t_i).$$

We estimate $\nu(t)$ with variable bandwidth KDE and $\mu(x, y)$ by selecting a fixed σ_x and σ_y, i.e.,

$$\nu(t) = \frac{1}{N_b} \sum_{i=1}^{N_b} \frac{1}{\sigma_t \sqrt{2\pi} D_i} \exp\left(-\frac{(t - t_i^b)^2}{2\sigma_t^2 D_i^2}\right), \tag{6.10.1}$$

$$\mu(x, y) = \sum_{i=1}^{N_b} \frac{1}{\sigma_x^b \sigma_y^b 2\pi} \exp\left(-\frac{(x - x_i^b)^2}{2{\sigma_x^b}^2} - \frac{(y - y_i^b)^2}{2{\sigma_y^b}^2}\right), \tag{6.10.2}$$

where we select $\sigma_x = \sigma_y = 0.15$, and find the D_i as before, selecting it by using the 100'th nearest neighbour due to the fact ν is only in one dimension. We fitted the MMrt, the isotropic model, Manhattan model, and maximum normal while allowing the background rate to vary in time. All four models gave similar marginals for $\mu(t)$ in time, with an example of this marginal for the isotropic model displayed in Figure 6.27. Here we can see that the models estimate the background in time to vary, with spikes appearing on a roughly weekly basis, perhaps representing the weekly routines of daily life. Indeed the data shows burglary is less likely to happen at the weekend, in line with previous research [50], with the implication that burglaries are more likely to happen when people are at work during the week.

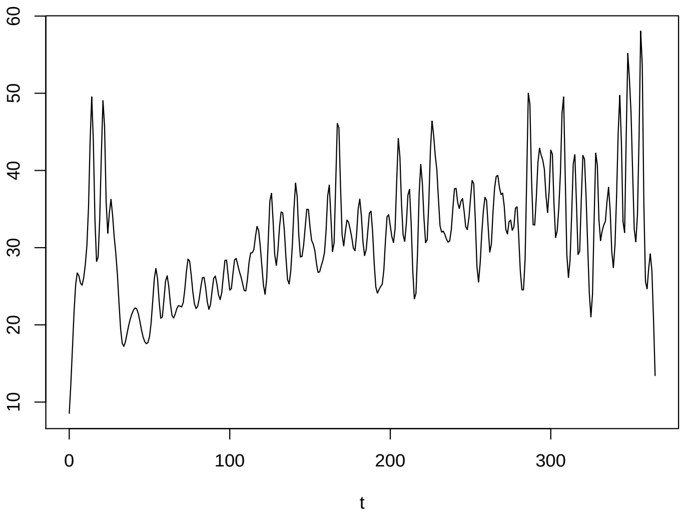

Figure 6.27: This is the marginal for $\nu(t)$ for the burglary data from 2015 fitted with the isotropic model when the background rate it allowed to vary in time (days).

Table 6.4 displays the branching ratio for each model. The isotropic model shows a slightly reduced branching ratio than when estimated without a variable background rate in time, with the MMrt, Manhattan and maximum norm models displaying more modest decreases in terms of the branching ratio. Allowing the background rate to vary in time should allow for slightly improved prediction results.

Table 6.4: Estimated branching ratio for the four models when fitted with burglary data from 2015.

Model	Branching Ratio
MMrt (15'th nn)	0.343
MMrt (30'th nn)	0.048
Isotropic	0.065
Manhattan	0.087
Max norm	0.119

6.11 Summary

In this section we have introduced new methods with which we can estimate a self-exciting point process which give strong results in terms of prediction of future burglary. Along with this, these methods can also reveal fresh insights to the data, such as how far the self-excitation 'travels' in the case of the isotropic model, or in the case of the Manhattan model the distance crime travels according to the gridlike structure of the street network in Chicago. We also introduced a model taking advantage of the maximum norm, which highlighted burglary that was triggered on streets parallel to the original crime.

Chapter 7

Conclusions and Suggestions for Future Work

In this thesis we have concentrated on fitting self-exciting point processes to publicly available crime data from Chicago, with the intention of building on existing models in order to more successfully predict and understand where future crime events may occur. In Chapter 3 we introduced a new parametric form named the DCR model for the triggering function, namely

$$g(t) = \bar{\alpha}\bar{\omega}^2 te^{-\bar{\omega}t}, \tag{7.0.1}$$

as an alternative to the commonly used ETAS model, $g(t) = \alpha\omega e^{-\omega t}$. The DCR model allowed the risk of an event to increase for a period after an initial event before the risk decays, as opposed to the ETAS model where the risk decays immediately after the initial increase when an event occurs. This was a reasonable assumption to make as for some crime, criminals may be more likely to offend after some time has passed from the initial crime as opposed to immediately after, for example a burglar may wish to wait until a victim has replaced their stolen goods, or it may take some time for the information about a potential location to be passed amongst criminals [56].

After demonstrating that the EM-algorithm for the DCR model could recover the true parameters of a process of this nature quite accurately, in Chapter 4 we fitted temporal point processes using both the ETAS and DCR model to burglary events in regions of Chicago. The prediction results displayed in Table 4.5 demonstrated that while the DCR model outperformed a simple Poisson point process modelled in each 200m grid cell, it did not perform quite as well as the ETAS model. While the DCR model should not be a preferable model to the ETAS model in terms of modelling burglary in this setting, it perhaps may still be of use when modelling other types of crimes in situations where the greatest risk of an offence being committed occurs some time after the initial offense, as opposed to immediately after.

Using Voronoi residual analysis we demonstrate that modelling burglary as a spatio-temporal point process with a constant background rate across the whole city of Chicago is inappropriate. In Chapter 5 we show that by using non-parametric methods introduced by Marsan and Lengliné to estimate the background rate and triggering function, we can more accurately model burglary in Chicago.

One challenge with this approach is the selection of bin width for the weighted histograms in the background and triggering rate. To overcome this, in Chapter 6 we use and adapt a non-parametric method proposed by Mohler [48], based on variable bandwidth kernel density estimation, and an estimated triggering function of the form $g(x - x_i, y - y_i, t - t_i)$. We altered the algorithm so the triggering function took an isotropic form, namely $g(r - r_i, t - t_i)$ where $r = \sqrt{x^2 + y^2}$. Using the fact that the streets of Chicago take a grid structure, we also proposed that this triggering function could take the form $g(d-d_i, t-t_i)$ with Manhattan distance $d = |x|+|y|$, and $g(m-m_i, t-t_i)$ where we use the maximum norm $m = \max(|x|, |y|)$. As shown in Figure 6.24, these proposed methods can provide fresh insight into how burglary may be triggered within Chicago, and can also provide improved predictive power, particularly in the case where the maximum norm is utilised.

For the Chicago data, the model using the maximum norm obtained the best predictive results, and this model could be the best to utilise on other locations which have a grid structure. The Manhattan model could also be used on locations with these types of streets, while the isotropic model may be the model of choice when the street structure is more irregular.

For future work, we could look to extend these models to include other types of crime events as leading indicators as seen previously [47], where minor crimes are allowed to trigger more serious offences, along with utilising other types of data to investigate whether this could help us more accurately assess crime risk. By incorporating other spatial information to the model, such as locations where crime cannot possibly take place, we could further improve the predictive power of the model. It may also be of interest to allow the triggering function to vary with the location where the original crime has taken place, to reflect that how a crime event affects future crime risk may be dependent on the location that it has taken place. In all these cases, a balance must be struck between incorporating useful information and overfitting a heavily-parameterized model. Other possible areas of interest include using network science to incorporate the street networks of a city more directly [61]. By defining a self-exciting point process on a network we could perhaps improve the predictive accuracy of our models.

While this research has focused on trying to understand and predict the occurrence of future crime events, more research is required on how police forces can best utilise these predictions in order to prevent or reduce further crime. These studies will need to be carried out while bearing in mind topical concerns about predictive policing, such as the issues of transparency and the possibility of bias.

Bibliography

[1] PredPol. https://www.predpol.com/. [Online; accessed 4-December-2018].

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

[3] Dont even think about it. https://www.economist.com/briefing/2013/07/20/dont-even-think-about-it, 2013. [The Economist, Online; accessed 4-December-2018].

[4] City of Chicago Data Portal. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data, 2016. [Online; accessed 20-November-2016].

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

[6] Google maps. "Location with the highest burglary rate in chicago.". https://www.google.com/maps/@41.7581755,-87.560564,15.25z, 2018. [Online; accessed 1-November-2018].

[7] S. Banerjee. An immune system inspired theory for crime and violence in cities. *Interdisciplinary Description of Complex Systems: INDECS*, 15(2):133–143, 2017.

[8] C. D. Barr and F. P. Schoenberg. On the Voronoi estimator for the intensity

of an inhomogeneous planar Poisson process. *Biometrika*, 97(4):977–984, 2010.

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

[10] H. Berestycki, S. Johnson, J. Ockendon, and M. Primicerio. Criminality. *European Journal of Applied Mathematics*, 21(4-5), 2010.

[11] N. Berg. Predicting crime, lapd-style. `https://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report`, 2014. [The Guardian, Online; accessed 4-December-2018].

[12] A. Bertozzi, S. Johnson, and M. Ward. Mathematical modelling of crime and security: Special issue of EJAM. *European Journal of Applied Mathematics*, 27(3):311–316, 2016.

[13] A. A. Braga. The effects of hot spots policing on crime. *The ANNALS of the American Academy of Political and Social Science*, 578(1):104–125, 2001.

[14] P. J. Brantingham, G. E. Tita, M. B. Short, and S. E. Reid. The ecology of gang territorial boundaries. *Criminology*, 50(3):851–885, 2012.

[15] A. Bray, K. Wong, C. D. Barr, F. P. Schoenberg, et al. Voronoi residual analysis of spatial point process models with applications to California earthquake forecasts. *The Annals of Applied Statistics*, 8(4):2247–2267, 2014.

[16] P. Brémaud and L. Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, pages 1563–1588, 1996.

[17] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*, volume 2. Springer, 2008.

[18] D. J. Daley and D. Vere-Jones. Scoring probability forecasts for point processes: The entropy score and information gain. *Journal of Applied Probability*, 41(A):297–312, 2004.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from

incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

[20] S. Deutsch and C. Malmborg. A dynamic model to forecast incapacitation and deterrence effects. *Applied Mathematical Modelling*, 9(1):53–61, 1985.

[21] R. Di Tella and E. Schargrodsky. Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *American Economic Review*, 94(1):115–133, 2004.

[22] J. E. Douglas, R. K. Ressler, A. W. Burgess, and C. R. Hartman. Criminal profiling from crime scene analysis. *Behavioral Sciences & the Law*, 4(4):401–421, 1986.

[23] M. Egesdal, C. Fathauer, K. Louie, J. Neuman, G. Mohler, and E. Lewis. Statistical and stochastic modeling of gang rivalries in Los Angeles. *SIAM Undergraduate Research Online*, 3:72–94, 2010.

[24] G. Espejo, G. L'Huillier, and R. Weber. A game-theoretical approach for policing decision support. *European Journal of Applied Mathematics*, 27(3):338–356, 2016.

[25] M. Fatehkia, D. O'Brien, and I. Weber. Correlated impulses: Using facebook interests to improve predictions of crime rates in urban areas. *PLOS ONE*, 14(2), 2019.

[26] V. Filimonov and D. Sornette. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108, 2012.

[27] E. W. Fox, F. P. Schoenberg, and J. Seth. Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *The Annals of statistics*, 10(3):1725–1756, 2016.

[28] P. Grindrod. *Mathematical Underpinnings of Analytics*. Oxford University Press, 2015.

[29] E. R. Groff and N. G. La Vigne. Forecasting the future of predictive crime mapping. *Crime Prevention Studies*, 30:29–57, 2002.

[30] D. Harte and D. Vere-Jones. The entropy score and its uses in earthquake forecasting. *Pure and Applied Geophysics*, 162(6-7):1229–1253, 2005.

[31] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[32] R. A. Hegemann, L. M. Smith, A. B. Barbaro, A. L. Bertozzi, S. E. Reid, and G. E. Tita. Geographical influences of an emerging network of gang rivalries. *Physica A: Statistical Mechanics and its Applications*, 390(21-22):3894–3914, 2011.

[33] P. Hunt, J. Saunders, and J. S. Hollywood. *Evaluation of the shreveport predictive policing experiment.* Rand Corporation, 2014.

[34] G. K. Kanji. *100 Statistical Tests.* SAGE Publications, 1999.

[35] T. Karppi. “The computer said so”: On the ethics, effectiveness, and cultural techniques of predictive policing. `https://doi.org/10.1177.2056305118768296`, May 2018. Social Media+ Society.

[36] M. Kelly. Inequality and crime. *Review of economics and Statistics*, 82(4):530–539, 2000.

[37] P. J. Laub, T. Taimre, and P. K. Pollett. Hawkes processes. *arXiv preprint arXiv:1507.02822*, 2015.

[38] E. Lewis and G. Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, 1(1):1–20, 2011.

[39] E. Lewis, G. Mohler, P. J. Brantingham, and A. L. Bertozzi. Self-exciting point process models of civilian deaths in Iraq. *Security Journal*, 25(3):244–264, 2012.

[40] J. Lindsey. *Introductory Statistics: A Modelling Approach.* Oxford University Press, 1995.

[41] J. A. Lindsey. *Mathematical Statistics and Data Analysis.* Duxbury, 3 edition, 2007.

[42] F. Lorenzen. *Analysis of order clustering using high frequency data: A point*

process approach. PhD thesis, Tilburg School of Economics and Management Finance Department, 2012.

[43] N. Malleson and M. A. Andresen. The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42(2):112–121, 2015.

[44] N. Malleson, A. Heppenstall, and L. See. Crime reduction through simulation: An agent-based model of burglary. *Computers, environment and urban systems*, 34(3):236–250, 2010.

[45] D. Marsan and O. Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008.

[46] C. Z. Marshak, M. P. Rombach, A. L. Bertozzi, and M. R. D'Orsogna. Growth and containment of a hierarchical criminal network. *Physical Review E*, 93(2):022308, 2016.

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

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

[49] G. O. Mohler, M. B. Short, S. Malinowski, M. Johnson, G. E. Tita, A. L. Bertozzi, and P. J. Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, 110(512):1399–1411, 2015.

[50] Office for National Statistics. Overview of burglary and other household theft: England and wales. `https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/overviewofburglaryandotherhouseholdtheft/englandandwales`, 2017. [Online; accessed 12-February-2019].

[51] Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

[52] Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

[53] A. Okabe, B. Boots, and K. Sugihara. *Spatial Tessellations Concepts and Applications of Voronoi Diagrams.* Wiley, 1992.

[54] T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

[55] W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood. *Predictive policing: The role of crime forecasting in law enforcement operations.* Rand Corporation, 2013.

[56] N. Polvi, T. Looman, C. Humphries, and K. Pease. The time course of repeat burglary victimization. *The British Journal of Criminology*, 31(4):411–414, 1991.

[57] J. G. Rasmussen. Temporal point processes: the conditional intensity function. `http://people.math.aau.dk/~jgr/teaching/punktproc11/tpp.pdf`, 2011. Course notes for 'rumlige punktprocesser' (spatial point processes), [Online; ed 20-November-2016].

[58] A. Reinhart et al. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018.

[59] A. Reinhart and J. Greenhouse. Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1305–1329, 2018.

[60] G. Rosser and T. Cheng. Improving the robustness and accuracy of crime prediction with the self-exciting point process through isotropic triggering. *Applied Spatial Analysis and Policy*, pages 1–21, 2016.

[61] G. Rosser, T. Davies, K. J. Bowers, S. D. Johnson, and T. Cheng. Predictive crime mapping: Arbitrary grids or street networks? *Journal of quantitative criminology*, 33(3):569–594, 2017.

[62] L. W. Sherman, P. R. Gartin, and M. E. Buerger. Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology*, 27(1):27–56, 1989.

[63] M. B. Short, P. J. Brantingham, A. L. Bertozzi, and G. E. Tita. Dissipation and displacement of hotspots in reaction-diffusion models of crime. *Proceedings of the National Academy of Sciences*, 107(9):3961–3965, 2010.

[64] M. B. Short, P. J. Brantingham, and M. R. Dorsogna. Cooperation and punishment in an adversarial game: How defectors pave the way to a peaceful society. *Physical Review E*, 82(6):066114, 2010.

[65] R. P. Shumate and R. F. Crowther. Quantitative methods for optimizing the allocation of police resources. *J. Crim. L. Criminology & Police Sci.*, 57:197, 1966.

[66] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[67] P. Sprent and N. C. Smeeton. *Applied Nonparametric Statistical Methods*. Chapman & Hall/CRC, 4 edition, 2007.

[68] M. Tanemura. Statistical distributions of poisson voronoi cells in two and three dimensions. *FORMA-TOKYO-*, 18(4):221–247, 2003.

[69] S. Tench, H. Fry, and P. Gill. Spatio-temporal patterns of ied usage by the provisional Irish republican army. *European Journal of Applied Mathematics*, 27(3):377–402, 2016.

[70] A. Veen and F. P. Schoenberg. Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

[71] C. J. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[72] J. Zhuang, Y. Ogata, and D. Vere-Jones. Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth*, 109(B5), 2004.

[73] S. F. C. K. Zipkin, J. and A. Bertozzi. Point-process models of social network interactions: Parameter estimation and missing data recovery. *European Journal of Applied Mathematics*, 27, 2016.