**Carnegie Mellon University**
**Dietrich College of Humanities and Social Sciences**
**Dissertation**

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

**Title**: Point process modeling with spatiotemporal covariates for predicting crime
**Presented by**: Alex Reinhart
**Accepted by**: Department of Statistics & Data Science
**Readers**:

| | |
|---|---|
| JOEL B. GREENHOUSE, ADVISOR | DATE |
| WILLIAM F. EDDY | DATE |
| WILPEN L. GORR | DATE |
| JOHN LEHOCZKY | DATE |
| COSMA ROHILLA SHALIZI | DATE |

Approved by the Committee on Graduate Degrees:

| | |
|---|---|
| RICHARD SCHEINES, DEAN | DATE |

# Point Process Modeling with Spatiotemporal Covariates for Predicting Crime

Alex Reinhart

July 19, 2018

A dissertation submitted in partial fulfillment
of the requirements for the Degree of Doctor of Philosophy

Department of Statistics & Data Science
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213

**Thesis Committee**:
Joel B. Greenhouse, Chair
William F. Eddy
Wilpen L. Gorr
John Lehoczky
Cosma Rohilla Shalizi

# Abstract

Self-exciting point processes are widely used to model events occurring in time and space whose rate depends on the past history of the process, such as earthquake aftershocks, crime, and neural spike trains. By modeling the event rate as the sum of a background (or immigrant) process, often an inhomogeneous Poisson process, and an offspring process consisting of events triggered by previous events, self-exciting point process models naturally account for complex clustering behavior. When the model is physically motivated, as are models of earthquake aftershock sequences, model parameters have direct interpretations in terms of the generative mechanism.

In this thesis, I focus in particular on the application of self-exciting point processes to crime. Crime rates are known to vary greatly in space within a city, as a result of many demographic and economic factors, and crime often exhibits "near-repeats," when one crime is followed by another soon after, either from retaliation or because offenders tend to return to the same areas. Point process models have been used to predict crime, but the available models can be improved: they cannot explicitly account for spatially varying covariates and estimate their effects, and there are no inference tools that could be used to test criminological theories or evaluate interventions.

After extensively reviewing the literature on self-exciting point processes, I introduce a new model which accounts for both spatial covariates and self-excitation, and explore its benefits over simple lagged regressions and other commonly used methods. After discussing computational issues in fitting the model, I use simulations to explore methods for parameter inference, review a set of residual diagnostics and animations, and use these diagnostics to explore the model's behavior under various forms of model misspecification, giving practical advice for the interpretation of model fits. To demonstrate the model's utility, I then analyze large databases of Pittsburgh and Baltimore crime records, linking crime rates to several relevant spatial covariate and leading indicator events, and comparing several model variations.

# Acknowledgments

Many thanks to Daniel Nagin for criminological advice and suggestions, and to Evan Liebowitz for compiling the covariate data used in Chapter 5. My committee—Bill Eddy, John Lehoczky, Wil Gorr, and Cosma Shalizi—were perpetually helpful, getting me out of methodological holes and giving many useful suggestions. And, of course, Joel Greenhouse was an invaluable advisor every step of the way, always offering sage advice, finding opportunities to pursue, and guiding me away from dead ends and time sinks.

The members of my Dissertation Writing Group (sponsored by the CMU Graduate Student Assembly) provided a great deal of support, motivation, and advice throughout, even though we spent most of our time in more of a Dissertation Therapy Group. Thanks to Sangwon Hyun, Jerzy Wieczorek, Lingxue Zhu, Robert Lunde, and Jisu Kim for putting up with the complaints every week.

Thanks to my classmates Nicolás Kim, Peter Elliott, and Kevin Lin for gladly tolerating random statistical questions at all hours of the day and night. Valérie Ventura and Chad Schafer, in advising my ADA project, taught me how to do research and how to give the impression of having done lots of work every week by presenting cool graphs.

Kira Bokalders and Laura Butler helped me snag a National Institute of Justice Graduate Research Fellowship and made sure the funding bureaucracy was satisfied, while Carl Skipper installed packages, fixed servers, and generally kept everything running for my simulation studies and analysis. None of this would have worked without them.

This project was supported by Award No. 2016-R2-CX-0021, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect those of the Department of Justice.

# Contents

# List of Tables

# List of Figures

# *One*

# Introduction

Predictive policing is the science of using historical crime data, and possibly other explanatory variables, to predict the locations and times of future crimes. As police agencies collect ever more data, more opportunities appear to mine these data for insight. Predictive policing models have been used to target interventions aimed at reducing property crime (Hunt, Saunders, & Hollywood, 2014; Mohler et al., 2015) and violent crime (Ratcliffe, Taniguchi, Groff, & Wood, 2011; Taylor, Koper, & Woods, 2011), and to analyze hotspots of robbery (Van Patten, McKeldin-Coner, & Cox, 2009) and shootings (Kennedy, Caplan, & Piza, 2010), among many other applications. Predictive policing methods are now widely deployed, with law enforcement agencies routinely making operational decisions based on them (Perry, McInnis, Price, Smith, & Hollywood, 2013).

Beyond simply predicting crime, other analyses can help answer questions of criminological interest and direct police efforts. For example, there has been a recent focus on the prevalence of near-repeat victimization, which has been analyzed using methods borrowed from epidemiology (Haberman & Ratcliffe, 2012; Ratcliffe & Rengert, 2008). Other efforts, such as Risk Terrain Modeling (Kennedy et al., 2010; Kennedy, Caplan, Piza, & Buccine-Schraeder, 2015), try to find local factors which increase the risk of crime. Spatial analyses may also be used to test ideas such as the "broken windows" theory, which posits that "failure to control minor offenses such as prostitution and disorderly conduct destabilizes neighborhoods by creating a sense of public disorder", leading to more serious crimes, including homicide (Cerdá et al., 2009).

However, previous predictive policing methods have several weaknesses. Tools for evaluating model performance are limited and comparisons between methods are hindered by the use of arbitrary hotspot cutoffs and inappropriate metrics. Further, all existing metrics evaluate fit globally rather than identifying specific problematic areas. Hotspot models, near-repeat models, and models incorporating spatial factors are usually separate and incompatible, so no current method incorporates all these features, resulting in estimates which are confounded by the left-out

features. Also, no existing method has used statistical inference to quantify uncertainty in predictions or model parameters, limiting their usefulness in criminological research.

In this thesis, I develop a single predictive model which incorporates features of hotspot models, near-repeat analysis, and Risk Terrain Modeling. This model has three goals: (1) improved hotspot predictions, by using all the relevant information; (2) rigorous tests of which crimes and features most strongly predict future crimes, using inference tools and model fit diagnostics; and (3) improved understanding of the near-repeat phenomenon, by incorporating it directly into the model. I also develop a Bayesian hierarchical model which extends the analysis to incorporate multiple cities or regions simultaneously, allowing exploration of the differences in crime dynamics between cities.

## 1.1 Predictive Policing Methods

Previous work has used a variety of methods to predict crime and identify high-risk areas. In this section I review prior work in predictive policing, then summarize gaps in these methods in Section 1.2. In Section 1.3, I summarize the work I have done to extend and apply the self-exciting point process models introduced in Section 1.1.3 to our Pittsburgh crime dataset.

### 1.1.1 Hotspot Detection

The most common predictive policing methods focus on hotspots: small geographic areas with high rates of target crimes. These hotspots can be chronic, lasting for years or decades, or temporary, appearing only for a few weeks or months (Gorr & Lee, 2015). They may be detected by spatial kernel density estimates, choropleth maps, standard deviational ellipses, scan statistics, or clustering methods (Chainey, Tompson, & Uhlig, 2008; Levine, 2008); these methods identify hotspots but do not predict crime rates within them or otherwise quantify the risk of crime. Police then choose the top hotspots for intensive patrol or other interventions, such as problem-oriented policing (Taylor et al., 2011).

Hotspot methods often require the use of ad-hoc tuning parameters which must be selected by a trained operator. The commonly used nearest-neighbor hierarchical clustering technique, for example, requires the operator to select the desired number of hotspots in advance (Perry et al., 2013, p. 22); kernel density hotspot mapping software (such as Hotspot Detective) requires users to select a bandwidth and the size of grid cells used for hotspot prediction, providing defaults based on the size of the map instead of the features of the data (Chainey et al., 2008). This means that different operators may produce different hotspot predictions (Hart & Zandbergen,

2014), and tuning parameters are not chosen in a statistically principled way to maximize predictive performance.

### 1.1.2 Risk Terrain Modeling

An alternative to hotspot methods is Risk Terrain Modeling (RTM) (Kennedy et al., 2010; Kennedy et al., 2015), which attempts to identify spatial features that may predict crime: gang territories, bars, dance clubs, residences of recent parolees, foreclosed homes, schools, and so on. This can provide local governments with important information to target law enforcement, social programs, and public works to reduce factors that lead to crime. RTM actually encompasses two related methods, the second a much more sophisticated version of the first. At its most basic, RTM proceeds as follows:

1. By reviewing the relevant literature, consulting with police, and analyzing data as necessary, identify a set of spatial features that are likely relevant to crime in the jurisdiction of interest.
2. Compile maps of each of these risk factors.
3. Create a fixed grid on top of the maps. Inside each grid cell, count the number of distinct risk factors present. This is the risk score.
4. To evaluate the predictive performance of risk scores, fit a logistic regression model and use it to predict the presence of crime in each grid cell during an evaluation period.

Kennedy et al. (2010) did variable selection for this procedure by compiling a $2 \times 2$ contingency table for every risk factor, recording whether cells with that risk factor experienced crime during the evaluation period. After performing $\chi^2$ tests for independence on each table, variables were selected by imposing $p$ value cutoffs. This procedure suffers from a number of methodological flaws: risk factors are implicitly assumed to have equal effects on crime rates, instead of allowing them to have different coefficients; there is no spatial dependence between grid cells; and the variable selection procedure uses marginal significance of variables instead of a statistically motivated procedure, such as using logistic regression with each risk factor as a separate covariate and performing a standard model selection technique.

A revised version of RTM (Kennedy et al., 2015) replaced the summed risk factors with a Poisson regression model, with elastic net regularization. Instead of simple presence/absence variables for each risk factor, there were six variables for each: three binary covariates indicating if the risk factor was present within 426, 852, or 1278 feet, and three binary covariates indicating if the density of the risk factor was two standard deviations above the mean, using the three distances above as

bandwidths. After regressing this large number of variables, the authors were not satisfied with the sparsity of the model, so they then used stepwise regression to further reduce the model while optimizing BIC. It is not clear why they didn't simply adjust the elastic net tuning parameter to induce more sparsity in the penalized regression.

This overcomplicated technique is unsatisfactory for several reasons. The distance cutoffs were chosen arbitrarily based on the size of city blocks in the tested city, in this case Chicago, rather than being determined by any empirical method. There is also no reason not to simply use the distance to the nearest risk factor instead. It is also not clear whether there is any reason to include kernel densities as a separate covariate with three different bandwidths.

Nonetheless, RTM is now implemented in the commercial RTMDx utility, and RTM's predictive performance has been compared to hotspot-based methods, with mixed results (Drawve, 2016; Drawve, Moak, & Berthelot, 2014).

### 1.1.3 Self-Exciting Point Processes

Beyond hotspot methods, some evidence suggests that a target crime (such as violent crime) can be better predicted by taking into account other crimes, such as criminal mischief or liquor law violations (J. Cohen, Gorr, & Olligschlaeger, 2007; Mohler et al., 2015). These crimes of public disorder serve as leading indicators, appearing before more serious crimes; when police are interested in predicting a relatively rare crime, like homicide, leading indicator crimes can provide much-needed data to produce better risk estimates. Leading indicators can also suggest targeted interventions which attack the root causes of crime, rather than simple reactive patrols.

Other research suggests that hotspots are not static in time, requiring a model that can adapt to changing crime rates as patterns of criminal activity shift (Gorr & Lee, 2015). Mohler (2014) developed such a model by building on self-exciting point process models used in earthquake forecasting, known in the seismology literature as epidemic-type aftershock sequence models (Ogata, 1999). Mohler's model allows hotspot estimates to change over time while taking leading indicators into account by separating crime into chronic hotspots, which remain fixed in time, and temporary hotspots, which are caused by increases or changes in crime. (In seismological models, earthquakes are similarly divided into main shocks and aftershocks caused by those main shocks.) Hotspot intensities are modeled with a modification of kernel density smoothing, where leading indicator crimes contribute to the intensity with effects that decay away in time, and the bandwidth parameters are estimated to best fit the data instead of being chosen by the operator. Together, this allows the model to provide better predictions than if it only considered the target crime or assumed hotspots were fixed in time.

Mohler et al. (2015) performed a randomized trial of a simplified version of this model in Los Angeles and Kent, showing that police deployment based on hotspot maps, updated daily with the latest crime data, reduced crime by roughly 7.4% in targeted areas. But the tested model eliminated some features that can improve predictive performance: in particular, it divided space into 150 × 150m grid cells and eliminated the spatial dependence between grid cells.

The details of these models, and the self-exciting point processes on which they are built, will be introduced in Chapter 2, and extensions will be developed in Chapter 3.

## 1.2 Gaps in the Literature

There is no current method that combines spatial covariates, as in RTM, with historical crime data, as in hotspot mapping. Because RTM does not account for near-repeats, its estimates are confounded by self-excitation, leading to bias and potential false positives; this will be illustrated in Section 3.1. It is currently impossible to compare the strength of spatial effects with near-repeat and leading indicator effects, because each type of model is constructed entirely differently, and we cannot disentangle the different modeling choices from the different predictive factors. A unified model could solve this problem and estimate various factors of interest to criminologists, such as the effects of different spatial and temporal covariates, while controlling for the presence of leading indicator crimes.

Previous hotspot prediction methods do not attempt any form of parameter inference. There have been some previous efforts to quantify the near-repeat phenomenon: the increased likelihood of repeat crimes in the vicinity of a recent crime (Ratcliffe & Rengert, 2008; Youstin, Nobles, Ward, & Cook, 2011). Most of these efforts are based on the Near Repeat Calculator (Ratcliffe, 2009), which adapts the Knox test from epidemiology to detect spatiotemporal clustering of events. The Knox test requires the operator to select a threshold distance and a threshold time; crimes which are close together in both distance and time are counted, and a permutation test is used to tell if this number is larger than expected.

This allows a test for the presence of the near-repeat phenomenon, but the threshold distances and times are arbitrary, and there is no satisfactory empirical way to estimate the duration or decay of the near-repeat phenomenon. Distances are often chosen based on the size of city blocks in the analyzed cities, and a variety of times may be tested separately, the operator deciding the duration of the near-repeat effect by which tests appear significant. This means near-repeat estimates are confounded with the statistical power of the studies estimating them. By contrast, the self-exciting point process model directly incorporates near-repeat ef-

fects, and with the addition of inference techniques for these effects, the near-repeat phenomenon can be much better understood.

Additionally, the Knox test suffers from confounding if the spatial risk surface of crime changes over time, for example if some small areas become more attractive to criminals over the course of the observation period (Ornstein & Hammond, 2017). This increased local risk is indistinguishable from the clustering caused by near-repeats unless the spatial features are explicitly accounted for, motivating the development of a model which incorporates both spatial features and self-excitation. Attempts to use the Knox test to determine the distance or length of time over which self-excitation can occur are also confounded with the power of the test, as will be demonstrated in Section 3.2.

Finally, most analysis of spatial covariates or of near-repeats has focused on one city or jurisdiction at a time, perhaps due to the difficulties of obtaining multiple datasets and in combining information across cities. There is not yet a model which can incorporate several cities and determine how similar or different their dynamics are, allowing principled comparisons.

## 1.3 Summary of Contributions

In this thesis, I develop a self-exciting point process model that incorporates spatial covariates and self-excitation, demonstrate how this model may be fit, and apply the model to real crime data. After an extensive review of the properties and uses of self-exciting point process in Chapter 2, I describe the new model in Chapter 3, along with the expectation–maximization algorithm to fit it and tools for simulation and computation based on it. Tools for parameter inference and model diagnostics are introduced in Chapter 4, followed by an extensive illustration using Pittsburgh and Baltimore crime data in Chapter 5 and Chapter 6. Finally, Chapter 7 concludes with a summary and suggested future work.

# *Two*

# Self-Exciting Spatio-Temporal Point Processes

Self-exciting spatio-temporal point processes, an extension of temporal Hawkes processes, model events whose rate depends on the past history of the process.[1] This class of models has proven useful in a wide range of fields: seismological models of earthquakes and aftershocks, criminological models of the dynamics of crime, epidemiological forecasting of the incidence of disease, and many others. In each field, the spatio-temporal distribution of events is of scientific and practical interest, both for prediction of new events and to improve understanding of the process generating the events. We may have a range of statistical questions about the process: does the rate of events vary in space and time? What spatial or temporal covariates may be related to the rate of events? Do events trigger other events, and if so, how are the triggered events distributed in space and time?

Regression is a natural first approach to answer these questions. By dividing space into cells, either on a grid or following natural or political boundaries, and dividing the observed time window into short discrete intervals, we can aggregate events and regress the number of events observed in a given cell and interval against spatial and temporal covariates, prior counts of events in neighboring cells, and so on. This approach has been widely used in applications. However, it suffers several disadvantages: most notably, the Modifiable Areal Unit Problem means that estimated regression coefficients and their variances may vary widely depending on the boundaries or grids chosen for aggregation, and there is no natural "correct" choice (Fotheringham & Wong, 1991).

Instead, we can model the rate of occurrence of events directly, without aggregation, by treating the data as arising from a point process. If the questions of scientific interest are purely spatial, the events can be analyzed using methods for spatial point processes (Diggle, 2014), and their times can be ignored. If time is im-

[1] This chapter has been published as Reinhart (2018). A Review of Self-Exciting Spatio-Temporal Point Processes and Their Applications. *Statistical Science*.

portant, descriptive statistics for the first- and second-order properties of a point process, such as the average intensity and clustering behavior, can also be extended to spatio-temporal point processes (Diggle, 2014, chapter 11).

When descriptive statistics are not enough to understand the full dynamics of the point process, we can use spatio-temporal point process models. These models estimate an intensity function that predicts the rate of events at any spatial location $s$ and time $t$. The simplest case is the homogeneous Poisson point process, where the intensity is constant in space and time. An example of a more flexible inhomogeneous model is the log-Gaussian Cox process, reviewed by Diggle, Moraga, Rowlingson, and Taylor (2013), in which the log intensity is assumed to be drawn from a Gaussian process. With a suitable choice of spatio-temporal correlation function, the underlying Gaussian process can be estimated, although this can be computationally challenging.

Cluster processes, which directly model clustering behavior, split the process in two: cluster centers, generally unobserved, are drawn from a parent process, and each cluster center begets an offspring process centered at the parent (Daley & Vere-Jones, 2003, Section 6.3). The observed process is the superposition of the offspring processes. A common case is the Poisson cluster process, in which cluster centers are drawn from a Poisson process; special cases include the Neyman–Scott process, in which offspring are also drawn from a Poisson process, and the Matérn cluster process, in which offspring are drawn uniformly from disks centered at the cluster centers. Common cluster processes, other spatio-temporal models, and descriptive statistics were reviewed by González, Rodríguez-Cortés, Cronie, and Mateu (2016).

In this chapter, I will focus on *self-exciting* spatio-temporal point process models, where the rate of events at time $t$ may depend on the history of events at times preceding $t$, allowing events to trigger new events. These models are characterized by a *conditional* intensity function, discussed in Section 2.1, which is conditioned on the past history of the process, and has a direct representation as a form of cluster process. Parametrization by the conditional intensity function has allowed a wide range of self-exciting models incorporating features like seasonality, spatial and temporal covariates, and inhomogeneous background event rates to be developed across a range of application areas.

Dependence on the past history of the process is not captured by log-Gaussian Cox processes or spatial regression, but can be of great interest in some applications: the greatest development of self-exciting models has been in seismology, where prediction of aftershocks triggered by large earthquakes is important for forecasting and early warning. However, the literature on theory, estimation, and inference for self-exciting models has largely been isolated within each application, so the purpose of this chapter is to synthesize these developments and place them in context, drawing connections between each application and paving the way for new uses.

Self-exciting models can be estimated using standard maximum likelihood approaches, discussed in Section 2.2.1 below. Once a self-exciting model is estimated, we are able to answer a range of scientifically interesting questions about the dynamics of their generating processes. Section 2.2.2 reviews *stochastic declustering* methods, which attribute events to the prior events which triggered them, or to the underlying background process, using the estimated form of the triggering function. Section 2.2.3 then introduces algorithms to efficiently simulate new data, and Section 2.2.4 discusses methods for estimating model standard errors and confidence intervals. Bayesian approaches are discussed in Section 2.2.5, and general model-selection and diagnostic techniques in Section 2.2.6.

Finally, Section 2.3 introduces three major application areas of self-exciting spatio-temporal point processes: earthquake forecasting, models of the dynamics of crime, and models of infectious disease. These demonstrate the utility of self-exciting models and illustrate each of the techniques described in Section 2.2. Section 2.3.4 introduces a further extension of self-exciting point processes, extending them from spatio-temporal settings to applications involving events occurring on networks.

## 2.1 Basic Theory

### 2.1.1 Hawkes Processes

Consider a temporal simple point process of event times $t_i \in [0, T)$, such that $t_i < t_{i+1}$, and a right-continuous counting measure $N(A)$, defined as the number of events occurring at times $t \in A$. Associated with the process is the history $\mathcal{H}_t$ of all events up to time $t$. We may characterize the process by its *conditional intensity*, defined as

$$\lambda(t \mid \mathcal{H}_t) = \lim_{\Delta t \to 0} \frac{\mathbb{E}\left[N\left([t, t+\Delta t)\right) \mid \mathcal{H}_t\right]}{\Delta t}.$$

The self-exciting point process model was introduced for temporal point processes by Hawkes (1971). Self-exciting processes can be defined in terms of a conditional intensity function in the equivalent forms

$$\begin{aligned}\lambda(t \mid \mathcal{H}_t) &= \nu + \int_0^t g(t-u)\,\mathrm{d}N(u) \\ &= \nu + \sum_{i\,:\,t_i<t} g(t-t_i),\end{aligned}$$

where $\nu$ is a constant background rate of events and $g$ is the triggering function which determines the form of the self-excitation. The process is called "self-exciting" because the current conditional intensity is determined by the past history $\mathcal{H}_t$ of the process. Depending on the form chosen for the triggering function $g$, the process

may depend only on the recent history (if $g$ decays rapidly) or may have longer term effects. Typically, because $\lambda(t \mid \mathcal{H}_t) \geq 0$, we require $g(u) \geq 0$ for $u \geq 0$ and $g(u) = 0$ for $u < 0$.

Hawkes processes have been put to many uses in a range of fields, modeling financial transactions (Bacry, Mastromatteo, & Muzy, 2015; Bauwens & Hautsch, 2009), neuron activity (D. H. Johnson, 1996), terrorist attacks (Porter & White, 2012), and a wide range of other processes. They are particularly useful in processes that exhibit clustering: Hawkes and Oakes (1974) demonstrated that any stationary self-exciting point process with finite intensity may be interpreted as a Poisson cluster process. The events may be partitioned into disjoint processes: a *background process* of cluster centers $N_c(t)$, which is simply a Poisson process with rate $\nu$, and separate *offspring processes* of triggered events inside each cluster, whose intensities are determined by $g$. Each triggered event may then trigger further events. Fig. 2.1 illustrates this separation. The number of offspring of each event is drawn from a Poisson distribution with mean

$$m = \int_0^\infty g(t)\,\mathrm{d}t.$$

Provided $m < 1$, cluster sizes are almost surely finite, as each generation of offspring follows a geometric progression, with expected total cluster size of $1/(1-m)$ including the initial background event. This partitioning also permits other useful results, such as an integral equation for the distribution of the length of time between the first and last events of a cluster (Hawkes & Oakes, 1974, Theorem 5).

### 2.1.2 Spatio-Temporal Form

Spatio-temporal models extend the conditional intensity function to predict the rate of events at locations $s \in X \subseteq \mathbb{R}^d$ and times $t \in [0, T)$. The function is defined in the analogous way to temporal Hawkes processes:

$$\lambda(s, t \mid \mathcal{H}_t) = \lim_{\Delta s, \Delta t \to 0} \frac{\mathbb{E}\left[N\left(B(s, \Delta s) \times [t, t + \Delta t)\right) \mid \mathcal{H}_t\right]}{|B(s, \Delta s)| \Delta t}, \tag{2.1}$$

where $N(A)$ is again the counting measure of events over the set $A \subseteq X \times [0, T)$ and $|B(s, \Delta s)|$ is the Lebesgue measure of the ball $B(s, \Delta s)$ with radius $\Delta s$.

A *self-exciting* spatio-temporal point process is one whose conditional intensity is of the form

$$\lambda(s, t \mid \mathcal{H}_t) = \mu(s) + \sum_{i\,:\,t_i < t} g(s - s_i, t - t_i), \tag{2.2}$$

where $\{s_1, s_2, \ldots, s_n\}$ denotes the observed sequence of locations of events and $\{t_1, t_2, \ldots, s_n\}$ the observed times of these events. Generally the triggering function $g$ is nonnegative, and is often a kernel function or power law decay function; often, for simplicity,

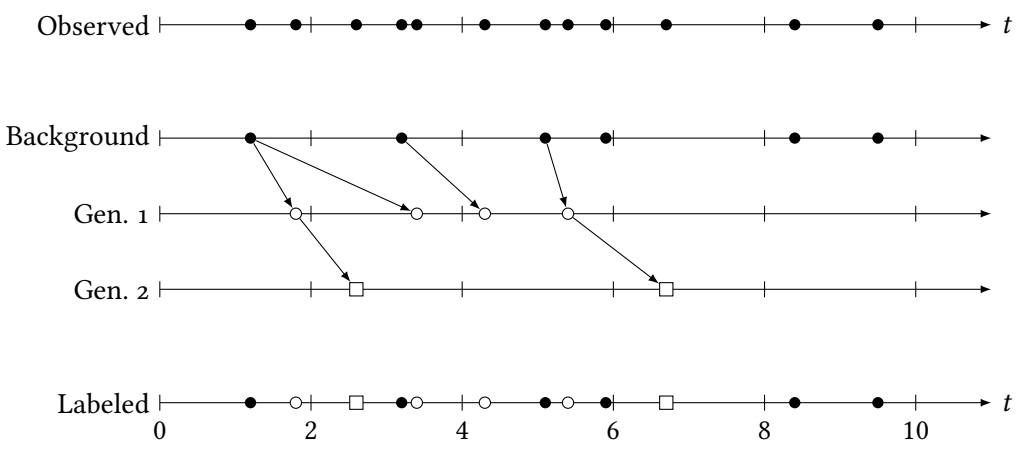


*Figure 2.1:* At top, a hypothetical observed self-exciting point process of events from $t = 0$ to $t = 10$. Below, the separation of that process into a background process and two generations of offspring processes. The arrows indicate the cluster relationships of which events were triggered by which preceding events; solid circles are background events, and open circles and squares are triggered events. At bottom, the combined process with generation indicated by shapes and shading. This cluster structure is not directly observed, though it may be inferred with the methods of Section 2.2.2.

it is taken to be separable in space and time, so that $g(s - s_i, t - t_i) = f(s - s_i)h(t - t_i)$, similar to covariance functions in other spatio-temporal models (Cressie & Wikle, 2011, Section 6.1). Sometimes a general nonparametric form is used, as in the model described in Section 2.2.2.

For ease of notation, the explicit conditioning on the past history $\mathscr{H}_t$ will be omitted for the rest of this dissertation, and should be read as implied for all self-exciting conditional intensities.

As with Hawkes processes, spatio-temporal self-exciting processes can be treated as Poisson cluster processes, with the mean number of offspring

$$m = \int_X \int_0^T g(s, t) \, \mathrm{d}t \, \mathrm{d}s. \tag{2.3}$$

The triggering function $g$, centered at the triggering event, is the intensity function for the offspring process. Properly normalized, it induces a probability distribution for the location and times of the offspring events. The cluster process representation will prove crucial to the efficient estimation and simulation of self-exciting

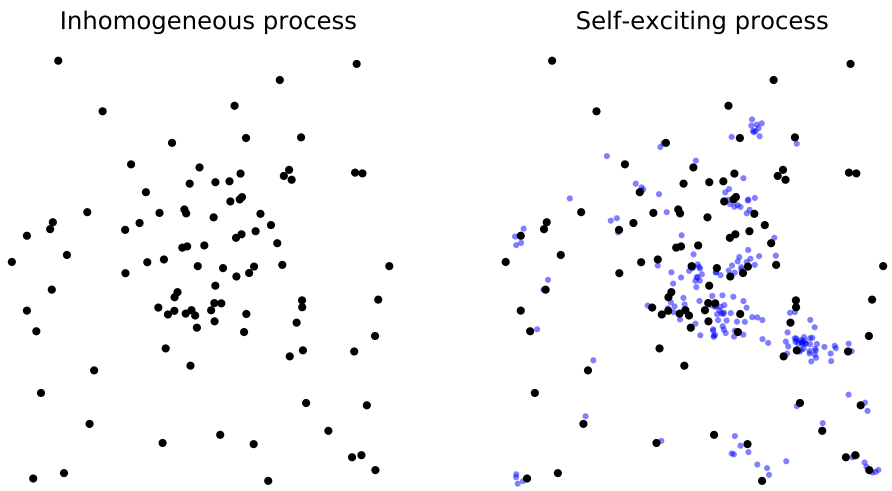


*Figure 2.2:* At left, a realization of an inhomogeneous Poisson process, in which the intensity is higher inside a central square and lower outside. At right, a self-exciting process with average total cluster size of 4, using the inhomogeneous Poisson process as the background process. Excited events are shown in blue. The cluster structure of the process is clearly visible, with clumps emerging from the self-excitation.

processes, and the estimation of the cluster structure of the process will be the focus of Section 2.2.2.

To illustrate the cluster process behavior of spatio-temporal self-exciting processes, Fig. 2.2 compares a simulated realization of a spatio-temporal inhomogeneous Poisson process against a self-exciting process using the same Poisson process realization as its background process. The self-exciting process, simulated using a Gaussian triggering function with a short bandwidth, shows clusters (of expected total cluster size 4) emerging from the Poisson process. The simulation was performed using Algorithm 2.5, to be discussed in Section 2.2.3, which directly uses the cluster process representation to make simulation more efficient.

### 2.1.3 Marks

Point processes may be *marked* if features of events beyond their time or location are also observed (Daley & Vere-Jones, 2003, Section 6.4). For example, if earthquakes are treated as a spatiotemporal point process of epicenter locations and times, the magnitude of each earthquake is an additional observed variable that is an important part of the process: the number and distribution of aftershocks may depend upon it. A marked point process is a point process of events $\{(s_i, t_i, \kappa_i)\}$, where $s_i \in X \subseteq \mathbb{R}^d$, $t_i \in [0, T)$, and $\kappa_i \in \mathscr{K}$, where $\mathscr{K}$ is the *mark space* (e.g. the space of earthquake

magnitudes). A special case is the *multivariate point process*, in which the mark space is a finite set $\{1, \dots, m\}$ for a finite integer $m$. Often the mark in a multivariate point process indicates the type of each event, such as the type of crime reported.

Marks can have several useful properties. A process has *independent marks* if, given the locations and times $\{(s_i, t_i)\}$ of events, the marks are mutually independent of each other, and the distribution of $\kappa_i$ depends only on $(s_i, t_i)$. Separately, a process has *unpredictable marks* if $\kappa_i$ is independent of all locations and marks $\{(s_j, t_j, \kappa_j)\}$ of previous events ($t_j < t_i$).

A marked point process has a *ground process*, the point process of event locations and times without their corresponding marks. Using the ground process conditional intensity, $\lambda_g(s, t)$, we can write the marked point process's conditional intensity function as

$$\lambda(s, t, \kappa) = \lambda_g(s, t) f(\kappa \mid s, t), \tag{2.4}$$

where $f(\kappa \mid s, t)$ is the conditional density of the mark at time $t$ and location $s$ given the history of the process up to $t$. In general, the ground process may depend on the past history of marks as well as the past history of event locations and times. For simplicity of notation, the following sections will largely consider point processes without marks, except where noted, but most methods apply to marked and unmarked processes alike.

### 2.1.4 Log-Likelihood

The likelihood function for a particular parametric conditional intensity model is not immediately obvious: given the potentially complex dependence caused by self-excitation, even the distribution of the total number of events observed in a time interval is difficult to obtain, and the spatial distributions of this varying number of events must also be accounted for. Instead, for a realization of $n$ points from a point process, we start with its Janossy density (Daley & Vere-Jones, 2003, Section 5.3). For a temporal point process, where a realization is the set of event times $\{t_1, t_2, \dots, t_n\}$ in a set $T$, the Janossy density is defined by the Janossy measure $J_n$,

$$J_n(A_1 \times \cdots \times A_n) = n! p_n \Pi_n^{\text{sym}}(A_1 \times \cdots \times A_n),$$

where the total number of events is $n$, $p_n$ is the probability of a realization of the process containing exactly $n$ events, $(A_1, \dots, A_n)$ is a partition of $T$ where $A_i$ represents possible times for event $i$, and $\Pi_n^{\text{sym}}(\cdot)$ is a symmetric probability measure determining the joint distribution of the times of events in the process, given there are $n$ total events. The Janossy measure is not a probability measure: it represents the sum of the probabilities of all $n!$ permutations of $n$ points. It is nonetheless useful, as its density $j_n(t_1, \dots, t_n)\, \mathrm{d}t_1 \cdots \mathrm{d}t_n$ has an intuitive interpretation as the probability that

there are exactly $n$ events in the process, one in each of the $n$ infinitesimal intervals $(t_i, t_i + \mathrm{d}t_i)$.

This interpretation connects the Janossy density to the likelihood function, which can be written as (Daley & Vere-Jones, 2003, Definition 7.1.II)

$$L_T(t_1, \dots, t_n) = j_n(t_1, \dots, t_n \mid T) \tag{2.5}$$

for a process on a bounded Borel set of times $T$; for simplicity in the rest of this section, we'll consider times in the interval $[0, T)$. Here $j_n(t_1, \dots, t_n \mid T)$ denotes the *local* Janossy density, interpreted as the probability that there are exactly $n$ events in the process before time $T$, one in each of the infinitesimal intervals.

The likelihood can be rewritten in terms of the conditional intensity function, which is usually easier to define than the Janossy density, by connection with survival and hazard functions. Consider the conditional survivor functions $S_k(t \mid t_1, \dots, t_{k-1}) = \Pr(t_k > t \mid t_1, \dots, t_{k-1})$. Using these functions and the conditional probability densities $p_k(t \mid t_1, \dots, t_{k-1})$ of event times, we can write the Janossy density recursively as

$$\begin{aligned} j_n(t_1, \dots, t_n \mid T) = p_1(t_1) p_2(t_2 \mid t_1) \cdots p_n(t_n \mid t_1, \dots, t_{n-1}) \times \\ S_{n+1}(T \mid t_1, \dots, t_n). \end{aligned} \tag{2.6}$$

Additionally, we may define the hazard functions

$$\begin{aligned} h_k(t \mid t_1, \dots, t_{k-1}) &= \frac{p_k(t \mid t_1, \dots, t_{k-1})}{S_k(t \mid t_1, \dots, t_{k-1})} \\ &= -\frac{\mathrm{d} \log S_k(t \mid t_1, \dots, t_{k-1})}{\mathrm{d}t}. \end{aligned} \tag{2.7}$$

The hazard function has a natural interpretation as the conditional instantaneous event rate—which means the conditional intensity $\lambda(t)$ can be written directly in terms of the hazard functions:

$$\lambda(t) = \begin{cases} h_1(t) & 0 < t < t_1 \\ h_k(t \mid t_1, \dots, t_{k-1}) & t_{k-1} < t \leq t_k, k \geq 2. \end{cases}$$

This allows us to write the likelihood from eq. (2.5) in terms of the conditional intensity function instead of the Janossy density. Observe that from eq. (2.7) we may write

$$S_k(t \mid t_1, \dots, t_{k-1}) = \exp\left( - \int_{t_{k-1}}^{t} h_k(u \mid t_1, \dots, t_{k-1}) \, \mathrm{d}u \right)$$

Substituting eq. (2.7) into eq. (2.6), replacing the hazard function with the conditional intensity, and combining terms leads to the likelihood, for a complete parameter vector $\Theta$, of (Daley & Vere-Jones, 2003, Proposition 7.2.III)

$$L(\Theta) = \left[\prod_{i=1}^{n} \lambda(t_i)\right] \exp\left(-\int_0^T \lambda(t)\,\mathrm{d}t\right).$$

By treating spatial locations as marks, we may extend this argument to spatio-temporal processes and obtain the log-likelihood (Daley & Vere-Jones, 2003, Proposition 7.3.III):

$$\ell(\Theta) = \sum_{i=1}^{n} \log\left(\lambda(s_i, t_i)\right) - \int_0^T \int_X \lambda(s, t)\,\mathrm{d}s\,\mathrm{d}t, \tag{2.8}$$

where $X$ is the spatial domain of the observations. For spatio-temporal marked point processes with intensity defined as in eq. (2.4), the log-likelihood is written in terms of the ground process, and has an extra mark term (Daley & Vere-Jones, 2003, Proposition 7.3.III):

$$\begin{aligned}\ell(\Theta) = &\sum_{i=1}^{n} \log\left(\lambda_g(s_i, t_i)\right) + \sum_{i=1}^{n} \log\left(f(m_i \mid s_i, t_i)\right) \\ &- \int_0^T \int_X \lambda_g(s, t)\,\mathrm{d}s\,\mathrm{d}t.\end{aligned}$$

In unmarked processes, the first term in eq. (2.8) is easy to calculate, assuming the conditional intensity is straightforward, but the second term can require computationally expensive numerical integration methods.

There are several approaches to evaluate this integral. The spatial domain $X$ can be arbitrary—e.g. a polygon defining the boundaries of a city—so Meyer, Elias, and Höhle (2012) (see Section 2.3.3) used two-dimensional numeric integration via cubature, as part of a numerical maximization routine. This requires an expensive numeric integration at every step of the numerical maximization, making the procedure unwieldy.

Schoenberg (2013) observed that, for some conditional intensities, it may be much easier to analytically integrate over $\mathbb{R}^2$ instead of an arbitrary $X$. Hence the approximation

$$\int_0^T \int_X \lambda(s, t)\,\mathrm{d}s\,\mathrm{d}t \le \int_0^\infty \int_{\mathbb{R}^2} \lambda(s, t)\,\mathrm{d}s\,\mathrm{d}t$$

may reduce the integral to a form that may be evaluated directly. The approximation is exact when the effect of self-excitation is contained entirely within $X$

and before $t = T$, and overestimates otherwise; because overestimation decreases the calculated log-likelihood, Schoenberg argued that likelihood maximization will avoid parameter values where overestimation is large. Lippiello, Giacco, Arcangelis, Marzocchi, and Godano (2014) argued that the temporal approximation biases parameter estimates more than the spatial one, and advocated only approximating $X$ by $\mathbb{R}^2$. This approximation was used by Mohler (2014), discussed in Section 2.3.2. Lippiello et al. (2014) also proposed a more accurate spatial approximation method based on a transformation of the triggering function to polar coordinates.

## 2.2 Estimation and Inference

Suppose now we have observed a realization of a self-exciting point process, with event locations $\{s_1, s_2, \dots, s_n\}$ and times $\{t_1, t_2, \dots, t_n\}$ over a spatial region $X$ and temporal window $[0, T)$. We have a model for the conditional intensity function and would like to be able to estimate its parameters, perform inference, and simulate new data if needed. This section discusses common approaches to these problems in the literature, focusing largely on maximum likelihood estimation, though with a brief discussion of Bayesian approaches in Section 2.2.5.

Fitting conditional intensity functions is not the only way to approach spatio-temporal point processes; there is also extensive literature that primarily uses descriptive statistics, such as first and second order moments of the process. I will not delve into this literature here, as it is less useful for understanding self-exciting processes; nonetheless, Vere-Jones (2009) gives a brief review, and more thorough treatments are available from González et al. (2016) and Diggle (2014).

### 2.2.1 Maximum Likelihood

Self-exciting point process models are most commonly fit using maximum likelihood. This is usually impossible to perform analytically: the form of the log-likelihood in eq. (2.8) involves a sum of logarithms of conditional intensities, which themselves involve sums over previous points, making analytical maximization intractable. Numerical evaluation of the intensity takes $O(n^2)$ time, and the log-likelihood can be nearly flat in large regions of the parameter space, causing problems for numerical maximization algorithms and making convergence extremely slow; in some examples explored by Veen and Schoenberg (2008), numerical maximization may fail to converge altogether. Nonetheless, for small datasets where the log-likelihood is computationally tractable to evaluate, numerical maximization is often used.

Alternately, Veen and Schoenberg (2008) showed the likelihood can be maximized with the expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 2008) by introducing a latent quantity, $u_i$, for each

event $i$, which indicates whether the event came from the background ($u_i = 0$) or was triggered by a previous event $j$ ($u_i = j$). This follows naturally from the cluster process representation discussed in Sections 2.1.1 and 2.1.2: if $u_i = 0$, event $i$ is a cluster center, and otherwise it is the offspring (directly or indirectly) of a cluster center.

Veen and Schoenberg (2008) derived the complete-data log-likelihood for a specific earthquake clustering model. More generally, consider a model of the form given in eq. (2.2). If the branching structure $u_i$ is assumed to be known, the complete-data log-likelihood for a parameter vector $\Theta$ can be written as

$$\begin{aligned}\ell_c(\Theta) = &\sum_{i=1}^{n} \mathbb{1}(u_i = 0) \log\left(\mu(s_i)\right) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}(u_i = j) \log\left(g(s_i - s_j, t_i - t_j)\right) \\ &- \int_0^T \int_X \lambda(s,t)\,\mathrm{d}s\,\mathrm{d}t,\end{aligned}$$

where $\mathbb{1}(\cdot)$ is the indicator function, which is one when its argument is true and zero otherwise. The branching structure dramatically simplifies the log-likelihood, as each event's intensity comes only from its trigger (the background or a previous event); this is analogous to the common EM approach to mixture models, where the latent variables indicate the underlying distribution from which each point came.

To complete the E step, we take the expectation of $\ell_c(\Theta)$. This requires estimating the triggering probabilities $\Pr(u_i = j) = \mathbb{E}[\mathbb{1}(u_i = j)]$ for all $i$, $j$, based on the current parameter values $\hat{\Theta}$ for this iteration. We can calculate these probabilities as

$$\Pr(u_i = j) = \begin{cases} \frac{g(s_i - s_j, t_i - t_j)}{\lambda(s_i, t_i)} & t_j < t_i \\ 0 & t_j \geq t_i \end{cases} \tag{2.9}$$

$$\Pr(u_i = 0) = 1 - \sum_{j=1}^{i-1} P(u_i = j) = \frac{\mu(s_i)}{\lambda(s_i, t_i)}. \tag{2.10}$$

This leads to the expected complete-data log-likelihood

$$\begin{aligned}\mathbb{E}[\ell_c(\Theta)] = &\sum_{i=1}^{n} \Pr(u_i = 0) \log\left(\mu(s_i)\right) \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n} \Pr(u_i = j) \log\left(g(s_i - s_j, t_i - t_j)\right) \\ &- \int_0^T \int_X \lambda(s,t)\,\mathrm{d}s\,\mathrm{d}t,\end{aligned}$$

which is much easier to analytically or numerically maximize with respect to each parameter in the M step. Once new parameter estimates are found, the procedure returns to the E step, estimating new triggering probabilities, and repeats until the log-likelihood converges, or until the estimated parameter values change by less than some pre-specified tolerance.

The EM algorithm has several advantages over other numerical maximization methods. Introducing the branching structure avoids the typical numerical issues encountered by other maximization algorithms, making the maximization at each iteration much easier, and the triggering probabilities also have a dual use in stochastic declustering algorithms, discussed in the next section.

One important warning must be kept in mind, however. If we have observed only data in the region $X$ and time interval $[0, T)$, but the underlying process extends outside this region and time, our parameter estimates will be biased by boundary effects (Zhuang, Ogata, & Vere-Jones, 2004). Unobserved events just outside $X$ or before $t = 0$ can produce observed offspring which may be incorrectly attributed to the background process, and observed events near the boundary can produce offspring outside it, biasing downward estimates of the mean number of offspring $m$ (see eq. (2.3)). Boundary effects can also bias the estimated intensity $\lambda(s,t)$ in ways analogous to the bias experienced in kernel density estimation (Cowling & Hall, 1996), but these effects are not well characterized for common self-exciting models.

### 2.2.2 Stochastic Declustering

For some types of self-exciting point processes, the background event rate $\mu(s)$ is fit nonparametrically from the observed data, for example by kernel density estimation or using splines (Ogata & Katsura, 1988). This could be fit by maximum likelihood—Mohler (2014) fit the background as a weighted kernel density via maximum likelihood, for example—but in some cases, we would like to estimate $\mu(s)$ using events from the background process only, and not using events which were triggered by those events. We may also want to analyze the background process

intensity separately from the triggered events, since the background process may have an important physical interpretation. This requires a procedure which can separate background events from triggered events, as illustrated in Fig. 2.1: stochastic declustering.

*Model-Based Stochastic Declustering.*

This version of stochastic declustering, introduced by Zhuang, Ogata, and Vere-Jones (2002), assumes that the triggering function $g$ has a parametric form, but that the background $\mu(s)$ should be estimated nonparametrically from only background events. Estimating the background requires determining whether each event was triggered by the background, but to do so requires $g$, so the procedure is iterative, starting with initial parameter values and alternately updating the background estimate and $g$ until convergence.

Consider the total spatial intensity function, defined as (Zhuang et al., 2002)

$$m_1(s) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \lambda(s,t)\,\mathrm{d}t, \tag{2.11}$$

where $T$ is the length of the observation period. The function $m_1(s)$ does not require declustering to estimate, since it sums over all events, including triggered events; by replacing the limit in eq. (2.11) with a finite-data approximation and substituting in eq. (2.2), we obtain

$$\begin{aligned} m_1(s) &\approx \frac{1}{T} \int_0^T \mu(s) + \sum_{i\,:\,t_i<t} g(s-s_i, t-t_i)\,\mathrm{d}t \\ &= \mu(s) + \frac{1}{T} \int_0^T \sum_{i\,:\,t_i<t} g(s-s_i, t-t_i)\,\mathrm{d}t. \end{aligned}$$

We hence obtain the relation

$$\begin{aligned} \mu(s) &\approx m_1(s) - \frac{1}{T} \sum_{i\,:\,t_i<t} \int_0^T g(s-s_i, t-t_i)\,\mathrm{d}t \\ &= m_1(s) - \gamma(s). \end{aligned}$$

We can now use a suitable nonparametric technique, such as kernel density estimation, to form $\hat{m}_1(s)$:

$$\hat{m}_1(s) = \frac{1}{T} \sum_{i=1}^{n} k(s-s_i),$$

where $k$ is a kernel function. It may also be desirable to estimate $\gamma(s)$ the same way. To do so, we use the same latent quantity $u_i$ defined and estimated in Section 2.2.1. We can estimate the cluster process by, for example, a weighted kernel density estimate, using

$$\hat{\gamma}(s) = \frac{1}{T} \sum_{i=1}^{n} \Pr(u_i \neq 0) k(s - s_i).$$

This leads to the estimator

$$\begin{aligned} \hat{\mu}(s) &= \hat{m}_1(s) - \hat{\gamma}(s) \\ &= \frac{1}{T} \sum_{i=1}^{n} (1 - \Pr(u_i \neq 0)) k(s - s_i). \end{aligned} \tag{2.12}$$

We now need to iteratively estimate parameters of the triggering function $g$. Provided these can be found by maximum likelihood, Zhuang et al. (2002) suggested the following algorithm:

**Algorithm 2.1.** Let $\hat{\mu}(s) = 1$ initially.

1. Using maximum likelihood (see Section 2.2.1), fit the parameters of the conditional intensity function
$$\lambda(s, t) = \hat{\mu}(s) + \sum_{i \,:\, t_i < t} g(s - s_i, t - t_i).$$
2. Calculate $\Pr(u_i \neq 0)$ for all $i$ using the parameters found in step 1 and eq. (2.10).
3. Using the new branching probabilities, form a new $\hat{\mu}^*(s)$ using eq. (2.12).
4. If $\max_s |\hat{\mu}(s) - \hat{\mu}^*(s)| > \epsilon$, for a pre-chosen tolerance $\epsilon > 0$, return to step 1. Otherwise, terminate the algorithm.

We can now perform stochastic declustering by thinning the process. With the final estimated $\hat{\mu}(s)$, we recalculate $\Pr(u_i \neq 0)$ and keep each event with probability $1 - \Pr(u_i \neq 0)$; the rest of the events are considered triggered events and deleted. We are left with those identified as background events.

In the original implementation of this algorithm, Zhuang et al. (2002) used an adaptive kernel function $k$ in eq. (2.12) whose bandwidth was chosen separately for each event, rather than being uniform for the whole dataset. After choosing an integer $n_p$ between 10 and 100, for each event they found the smallest disk centered at that event which includes at least $n_p$ other events (forced to be larger than some

small value $\epsilon$, chosen on the order of the observation error in locations). The radius of this disk was used as the bandwidth for the kernel at each event. This method was chosen because, in clustered datasets, any single bandwidth oversmooths in some areas and is too noisy in others. A method to estimate kernel parameters from the data will be introduced in Section 2.2.2.

Zhuang et al. (2002) also adapted the declustering algorithm to produce a "family tree": a tree connecting background events to the events they trigger, and so on from each event to those it triggered. The algorithm considers each pair of events and determines whether one should be considered the ancestor of the other:

**Algorithm 2.2.** Begin with the final estimated $\hat{\mu}(s)$ from Algorithm 2.1.

1. For each pair of events $i$, $j$ (with $t_i > t_j$), calculate $\Pr(u_i = j)$ and $\Pr(u_i = 0)$.
2. Set $i = 1$.
3. Generate a uniform random variate $R_i \sim \text{Uniform}(0, 1)$.
4. If $R_i < \Pr(u_i = 0)$, consider event $i$ to be a background event.
5. Otherwise, select the smallest $J$ such that $R_i < \Pr(u_i = 0) + \sum_{j=1}^{J} \Pr(u_i = j)$. Consider the $i$th event to be a descendant of the $J$th event.
6. When $i = N$, the total number of events, terminate; otherwise, set $i = i + 1$ and return to step 3.

Though the thinning algorithm and family tree construction are stochastic and hence do not produce unique declusterings, Zhuang et al. (2002) argue this is an advantage, as uncertainty in declustering can be revealed by running the declustering process repeatedly and examining whether features are consistent across declustered processes. These methods have been used to answer important scientific questions in seismology, as discussed in Section 2.3.1.

*Forward Likelihood-based Predictive approach.*

In a semiparametric model, where the background $\mu(s)$ is estimated nonparametrically from background events, the nonparametric estimator (such as a kernel smoother) may have tuning parameters which need to be adapted to the data. The model-based stochastic declustering procedure discussed above uses an adaptive kernel in $\mu(s)$, but we may wish to use a standard kernel density estimator with bandwidth estimated from the data. However, if we follow Algorithm 2.1, adjusting the bandwidth with maximum likelihood at each iteration, the bandwidth would go to zero, placing a point mass at each event.

To avoid this problem, Chiodi and Adelfio (2011) introduced the Forward Likelihood-based Predictive approach (FLP). Rather than directly maximizing the likelihood, consider increments in the log-likelihood, using the first $k$ observations to predict the $(k+1)$th:

$$\delta_{k,k+1}(\Theta \mid \mathscr{H}_{t_k}) = \log \lambda(s_{k+1}, t_{k+1} \mid \Theta, \mathscr{H}_{t_k}) - \int_{t_k}^{t_{k+1}} \int_X \lambda(s, t \mid \Theta, \mathscr{H}_{t_k}) \, \mathrm{d}s \, \mathrm{d}t,$$

where the past history $\mathscr{H}_{t_k}$ explicitly indicates that the intensity experienced by point $k+1$ depends only on the first $k$ observations (i.e. the estimate of $\mu(s)$ only includes the first $k$ points). A parameter estimate $\hat{\Theta}$ is formed by numerically maximizing the sum

$$\mathrm{FLP}(\hat{\Theta}) = \sum_{k=k_1}^{n-1} \delta_{k,k+1}(\hat{\Theta} \mid \mathscr{H}_{t_k}),$$

where $k_1 = \lfloor n/2 \rfloor$. Adelfio and Chiodi (2015a) and Adelfio and Chiodi (2015b) developed the FLP method into a semiparametric method following an alternated estimation procedure similar to Algorithm 2.1. The procedure splits the model parameters into the nonparametric smoothing parameters $\Sigma$ and the triggering function parameters $\Theta$, and iteratively fits them in the following steps:

**Algorithm 2.3.** Begin with a default estimate for $\Sigma$, for example by Silverman's rule for kernel bandwidths (Silverman, 1986). Use this to estimate $\mu(s_i)$ for each event $i$.

1. Using the estimated values of $\mu(s_i)$ and holding $\Sigma$ fixed, estimate the triggering function parameters $\Theta$ via maximum likelihood.
2. Calculate $\Pr(u_i = 0)$ for each event $i$ using the current parameter estimates.
3. Estimate the smoothing parameters by maximizing $\mathrm{FLP}(\hat{\Sigma})$, holding $\Theta$ fixed.
4. Calculate new estimates of $\mu(s_i)$ for each event $i$, using a weighted estimator with the weights calculated in step 2.
5. Check for convergence in the estimates of $\Sigma$ and $\Theta$ and either terminate or return to step 1.

Adelfio and Chiodi (2015b) applied this method to a large catalog of earthquakes in Italy, using the earthquake models to be discussed in Section 2.3.1, finding improved performance over a version of the model where smoothing parameters were fixed solely with Silverman's rule.

*Model-Independent Stochastic Declustering.*

Marsan and Lengliné (2008) proposed a model-independent declustering algorithm (MISD) for earthquakes which removed the need for a parametric triggering function $g(s, t)$, instead estimating the shape of $g(s, t)$ from the data. They assumed a conventional conditional intensity with constant background rate $\lambda_0$,

$$\lambda(s, t) = \lambda_0 + \sum_{i\,:\,t_i<t} g(s - s_i, t - t_i),$$

but $g(s, t)$ was simply assumed to be piecewise constant in space and time, with the constant for each spatial and temporal interval estimated from the data. Marsan and Lengliné (2010) showed their method can be considered an EM algorithm, following the same steps as in Section 2.2.1: estimate the probabilities $\Pr(u_i = j)$ in the E step and then maximize over parameters of $g(s, t)$ and $\lambda_0$ in the M step, eventually leading to convergence and final estimates of the branching probabilities.

Fox, Schoenberg, and Gordon (2016) extended this method to the case where the background $\lambda_0$ is not constant in space by assuming a piecewise constant background function $\mu(s)$ or by using a kernel density estimate of the background, then quantified uncertainty in the background and in $g(s, t)$ by using a version of the parametric bootstrap method to be discussed in Section 2.2.4. This can be considered a general nonparametric approach to spatio-temporal point process modeling as well as a declustering method, since with confidence intervals for the nonparametric triggering function, useful inference can be drawn for the estimated triggering function's shape.

### 2.2.3 Simulation

It is often useful to simulate data from a chosen model. For temporal point processes, a range of simulation methods are described by Daley and Vere-Jones (2003, section 7.5). Several spatio-temporal methods are based on a thinning procedure which first generates a large quantity of events, then thins them according to their conditional intensity, starting at the first event and working onward so history dependence can be taken into account. The basic method was introduced for nonhomogeneous Poisson processes by P. A. W. Lewis and Shedler (1979).

Ogata (1998) proposed a two-stage algorithm for general self-exciting processes which requires thinning fewer events and is hence more efficient. Events are generated sequentially, and the time of each event is determined before its location. To generate times, we require a version of the conditional intensity which is only a

function of time, having integrated out space:

$$\lambda_X(t) = \nu_0 + \sum_{j\,:\,t_j<t} \nu_j(t)$$
$$\nu_0 = \int_X \mu(s)\,\mathrm{d}s$$
$$\nu_j(t) = \int_X g(s,t)\,\mathrm{d}s.$$

This allows us to simulate times of events before simulating their locations. The algorithm below, though apparently convoluted, amounts to drawing the waiting time until the next event from an exponential distribution, drawing its location according to the distribution induced by $g$, and repeating, rejecting (thinning) some proposed times proportional to their intensities $\lambda_X$:

**Algorithm 2.4.** Start with $a = b = c = 0$ and $i = 1$.

1. Set $s_a = 0$ and generate $U_b \sim \mathrm{Uniform}(0,1)$. Let $\Lambda_c = \nu_0$ and $u_a = -\log(U_b)/\Lambda_c$.
2. If $u_a > T$, stop. Otherwise, let $t_i = u_a$, let $J = 0$, and skip to step 7.
3. Let $b = b + 1$ and $a = a + 1$. Generate $U_b \sim \mathrm{Uniform}(0,1)$ and let $u_a = -\log(U_b)/\Lambda_c$.
4. Let $s_a = s_{a-1} + u_a$. If $s_a > T$, stop; otherwise let $b = b + 1$ and generate $U_b \sim \mathrm{Uniform}(0,1)$.
5. If $U_b > \lambda_X(s_a)/\Lambda_c$, set $c = c + 1$ and let $\Lambda_c = \lambda_X(s_a)$, then go to step 3.
6. Let $t_i = s_a$, set $b = b + 1$, generate $U_b \sim \mathrm{Uniform}(0,1)$, and find the smallest $J$ such that $\sum_{j=0}^{J} \nu_j(t_i) > U_b \lambda_X(t_i)$.
7. If $J = 0$ then generate $s \in X$ from the non-homogeneous Poisson intensity $\mu(s)$ and go to step 10.
8. Otherwise, set $b = b + 1$, then set $s_i$ by drawing from the normalized spatial distribution of $g$ centered at $s_J$.
9. If $s_i$ is not in $X$, return to step 3.
10. Otherwise, set $i = i + 1$ and return to step 3.

This can be computationally expensive. The intensity $\lambda_X$ must be evaluated at each candidate point, involving a large sum, and the thinning in step 5 means multiple candidate times will often have to be generated. Another method, developed for earthquake models, directly uses the cluster structure of the self-exciting process, eliminating the need for thinning or repeated evaluation of $\lambda(s, t)$ (Zhuang et al., 2004):

**Algorithm 2.5.** Begin with a fully specified conditional intensity $\lambda(s, t)$.

1. Generate events from the background process using the intensity $\mu(s)$, by using a simulation method for nonhomogeneous stationary Poisson processes (P. A. W. Lewis & Shedler, 1979). Call this catalog of events $G^{(0)}$.

2. Let $l = 0$.

3. For each event $i$ in $G^{(l)}$, simulate its $N^{(i)}$ offspring, where $N^{(i)} \sim \text{Poisson}(m)$ (with $m$ defined as in eq. (2.3)), and the offspring's location and time are generated from the triggering function $g$, normalized as a probability density. Call these offspring $O_i^{(l)}$.

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

5. If $G^{(l)}$ is not empty, set $l = l + 1$ and return to step 3. Otherwise, return $\bigcup_{j=0}^{l} G^{(j)}$ as the final set of simulated events.

This algorithm has been widely used in the seismological literature for studies of simulated earthquake catalogs. However, both methods suffer from the same edge effects as discussed in Section 2.2.1: if the background is simulated over a time interval $[0, T)$, the offspring of events occurring just before $t = 0$ are not accounted for. Similarly, if events occurred just outside the spatial region $X$, they can have offspring inside $X$, which will not be simulated. This can be avoided by simulating over a larger space-time window and then only selecting simulated events inside $X$ and $[0, T)$. Møller and Rasmussen (2005) developed a perfect simulation algorithm for temporal Hawkes processes which avoids edge effects, but its extension to spatio-temporal processes remains to be developed.

### 2.2.4 Asymptotic Normality and Inference

Ogata (1978) demonstrated asymptotic normality of maximum likelihood parameter estimates for temporal point processes, and showed the covariance converges to the inverse of the expected Fisher information matrix, suggesting an estimator based on

the Hessian of the log-likelihood at the maximum likelihood estimate. This estimator has been frequently used for spatio-temporal models in seismology; however, Wang, Schoenberg, and Jackson (2010), comparing it with sampling distributions found by repeated simulation, found that standard errors based on the Hessian can be heavily biased for small to moderate observation period lengths, suggesting the finite-sample behavior is poor.

Rathbun (1996) later demonstrated that for spatio-temporal point processes, maximum likelihood estimates of model parameters are consistent and asymptotically normal as the observation time $T \longrightarrow \infty$, under regularity conditions on the form of the conditional intensity function $\lambda(s, t)$. An estimator for the asymptotic covariance of the estimated parameters is

$$\hat{\Sigma} = \left( \sum_{i=1}^{n} \frac{\Delta(s_i, t_i)}{\lambda(s_i, t_i)} \right)^{-1}, \tag{2.13}$$

where $\Delta(s_i, t_i)$ is a matrix-valued function whose entries are

$$\Delta_{ij}(s, t) = \frac{\dot{\lambda}_i(s, t)\dot{\lambda}_j(s, t)}{\lambda(s, t)}$$

and $\dot{\lambda}_i(s, t)$ denotes the partial derivative of $\lambda(s, t)$ with respect to the $i$th parameter. From $\hat{\Sigma}$ we can derive Wald tests of parameters of interest, and by inverting the tests we can obtain confidence intervals for any parameter.

Rather than relying on asymptotic normality, another approach is the parametric bootstrap, which has been used for temporal point process models in neuroscience (Sarma et al., 2011). The parametric bootstrap, though computationally intensive, is conceptually simple:

**Algorithm 2.6.** Using the parameter values $\hat{\Theta}$ from a previously fitted model, and starting with $i = 1$:

1. Using a simulation algorithm from Section 2.2.3, simulate a new dataset in the same spatio-temporal region.

2. Fit the same model to this new data, obtaining new parameter values $\hat{\Theta}^{(i)}$.

3. Repeat steps 1 and 2 with $i = i + 1$, up to some pre-specified number of simulations $B$ (e.g 1000).

   (Alternately, the algorithm can be adaptive, by checking the confidence intervals after every $b$ steps and stopping when they seem to have converged.)

4. Calculate bootstrap 95% confidence intervals for each parameter by using the 2.5% and 97.5% quantiles of the estimated $\hat{\Theta}^{(i)}$.

This is straightforward to implement, relies on minimal assumptions, and is asymptotically consistent in some circumstances. However, just as asymptotically normal standard errors may be biased for finite sample sizes, the bootstrap has no performance guarantees for small samples. Wang et al. (2010) tested neither the parametric bootstrap nor the estimator of Rathbun (1996) in their simulations, so no direct comparison is possible here, and those intending to use the bootstrap should test its performance in simulation.

It is sometimes desirable to estimate only a subset of the parameters in a model, either because full estimation is intractable or because some covariates are unknown. Dropping terms from the conditional intensity results in a *partial* likelihood, and parameter estimates obtained by maximizing the partial likelihood may differ from those obtained from the complete likelihood. Schoenberg (2016) explored the circumstances under which the parameter estimates are not substantially different, finding that partial likelihood estimates are identical under assumptions about the separability of the omitted parameters, and are still consistent in more general additive models under assumptions that the omitted parameters have relatively small effects on the intensity. In either case, the maximum partial likelihood estimates still have the asymptotic normality properties discussed above.

### 2.2.5 Bayesian Approaches

Rasmussen (2013) introduced two methods for Bayesian estimation for self-exciting temporal point processes: direct Markov Chain Monte Carlo (MCMC) on the likelihood, using Metropolis updates within a Gibbs sampler, and a method based on the cluster process structure of the process. Loeffler and Flaxman (2017) recently adapted MCMC to fit a version of the self-exciting crime model discussed in Section 2.3.2, using the Stan modeling language (Stan Development Team, 2016) and Hamiltonian Monte Carlo to obtain samples from the posteriors of the parameters. Ross (2016), however, working with the seismological models discussed in Section 2.3.1, argued that direct Monte Carlo methods are impractical: a sampling method involving repeated rejection requires evaluating the likelihood many times, an $O(n^2)$ operation, and the strong correlation of some parameters can make convergence difficult.

Instead, building on the cluster process method suggested by Rasmussen (2013), Ross (2016) proposed taking advantage of the same latent variable formulation introduced for maximum likelihood in Section 2.2.1. If the latent $u_i$s are known for all $i$, events in the process can be partitioned into $N + 1$ sets $S_0, \ldots, S_N$, where

$$S_j = \{t_i \mid u_i = j\}, \qquad 0 \leq j < N.$$

Events in each set $S_j$ can be treated as coming from a single inhomogeneous Poisson process, with intensity proportional to the triggering function $g$ (or to $\mu(s)$, for $S_0$). This allows the log-likelihood to be partitioned, reducing dependence between parameters and dramatically improving sampling performance. The algorithm now involves sampling $u_i$ (using the probabilities defined in eqs. (2.9)–(2.10)), then using these to sample the other parameters, in a procedure very similar to the expectation maximization algorithm for these models.

### 2.2.6 Model Selection and Diagnostics

In applications, model selection is usually performed using the Akaike information criterion (AIC) or related criteria like the Bayesian information criterion (BIC) and the Hannan–Quinn criterion. J. Chen, Hawkes, Scalas, and Trinh (2017) compared the performance of these methods in selecting the correct model in a range of settings and sample sizes, finding AIC more effective in small samples and less in larger samples. A variety of tests and residual plots are available for evaluating the fit of spatio-temporal point process models. Bray and Schoenberg (2013) provide a comprehensive review focusing on earthquake models; I will give a brief summary here.

First, we observe that any process characterized by its conditional intensity $\lambda(s, t)$ may be thinned to obtain a homogeneous Poisson process (Schoenberg, 2003), allowing examination of the fit of the spatial component of the model. We define $b = \inf_{s,t} \lambda(s, t)$, and for each event $i$ in the observed process, calculate the quantity

$$p_i = \frac{b}{\lambda(s_i, t_i)}.$$

Retain event $i$ with probability $p_i$. If this is done with an estimated intensity $\hat{\lambda}(s, t)$ from the chosen model, the thinned process (now ignoring time) will be Poisson with rate $b$, and can be examined for homogeneity, for example with the $K$-function (Ripley, 1977), which calculates the proportion of events per unit area which are within a given distance. This will detect if the thinned process still has clustering not accounted for by the model.

If $b$ is small, the thinned process will contain very few events, making the test uninformative. Clements, Schoenberg, and Veen (2012) propose to solve this problem with "super-thinning", which superimposes a simulated Poisson process. We choose a rate $k$ for the super-thinned process, such that $b \le k \le \sup_{s,t} \lambda(s, t)$, and thin with probabilities

$$p_i = \min\left\{\frac{k}{\lambda(s_i, t_i)}, 1\right\}.$$

We add to the thinned process a simulated inhomogeneous Poisson process with rate $\max\{k - \lambda(s, t), 0\}$. The sum process is, if the estimated model is correct, homogeneous with rate $k$.

Graphical diagnostics are also available. For purely spatial point processes, Baddeley, Turner, Moller, and Hazelton (2005) developed a range of residual diagnostic tools to display differences between the fitted model and the data, demonstrating further properties of these residuals in Baddeley, Møller, and Pakes (2007) and Baddeley, Rubak, and Møller (2011). Zhuang (2006) showed these tools could be extended directly to spatio-temporal point processes, producing residual maps which display the difference between the predicted number of events and the actual number, over grid cells or some other division of space. Bray, Wong, Barr, and Schoenberg (2014) argued that a grid is a poor choice: if grid cells are small, the expected number of events per cell is low and the distribution of residuals is skewed, but if grid cells are large, over- and under-prediction within a single cell can cancel out. Instead, they proposed using the Voronoi tesselation of space: for each event location $s_i$, the corresponding Voronoi cell consists of all points that are closer to $s_i$ than to any other event. This generates a set of convex polygons. By integrating the conditional intensity over a reasonable unit of time and over each Voronoi cell, we obtain a map of expected numbers of events, which we can subtract from the true number in each cell (which is 1 by definition). This produces a map which can be visually examined for defects in prediction.

As an example, Fig. 2.3 is a Voronoi residual map of the self-exciting point process previously shown in Fig. 2.2, produced following the procedure suggested by Bray et al. (2014). A model was fit to the simulated point process data that does not account for the inhomogeneous background process, instead assuming a constant background rate, and a spatial pattern in the residuals is apparent, with positive residuals (more events than predicted) in areas where the background rate is higher and negative residuals outside those areas.

## 2.3 Applications

This section will review four major applications of self-exciting point processes: earthquake models, crime forecasting, epidemic infection forecasting, and events on networks. This is by no means an exhaustive list—self-exciting point process models have been applied to problems as disparate as wildfire occurrence (Peng, Schoenberg, & Woods, 2005) and civilian deaths in Iraq (E. Lewis, Mohler, Brantingham, & Bertozzi, 2011). The selected applications illustrate the features that make self-exciting point processes valuable: parameters of the triggering function $g$ have important physical interpretations and can be used to test scientific hypotheses about the event triggering process, while the background $\mu$ flexibly incorporates spatial and temporal covariates whose effects can be estimated. Purely descriptive methods, or methods such as log-Gaussian Cox processes, do not permit the same inference about the event triggering process.

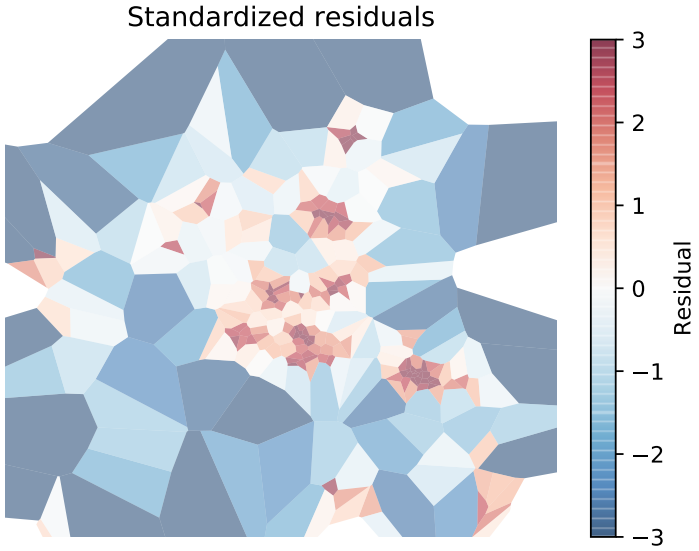


*Figure 2.3:* A Voronoi residual map of the self-exciting point process shown in Fig. 2.2. The model was fit assuming a constant background intensity and does not account for the inhomogeneous rate, leading to positive residuals in the center area and negative residuals outside. Residual values are standardized according to an approximate distribution given by Bray, Wong, Barr, and Schoenberg (2014).

### 2.3.1 Earthquake Aftershock Sequence Models

After a large earthquake, a sequence of smaller aftershocks is typically observed in the days and weeks afterwards, usually near the epicenter of the main shock (Freed, 2005). These tremors are triggered by the seismic disturbance of the main shock, and the distribution of their magnitudes and arrival times has proven to be relatively consistent, allowing the development of models for their prediction and analysis.

Sequences of earthquakes and aftershocks show rich behavior, such as spatial and temporal clustering, complex spatial dependence, and gradual shifts in overall seismicity. Self-exciting point processes are a natural choice to model this behavior, as they can directly capture spatio-temporal aftershock triggering behavior and can incorporate temporal trends and spatial inhomogeneity. The Epidemic-Type Aftershock Sequence (ETAS) model, developed and expanded over several decades, provides a flexible foundation for modeling this behavior, and has been widely applied to earthquake sequences in Japan, California, and elsewhere. A comprehensive review is provided by Ogata, 1999.

The initial ETAS model was purely temporal, modeling the rate of earthquakes at time $t$ as a superposition of a constant rate of background seismicity and of aftershocks triggered by these background events:

$$\lambda(t) = \mu + \sum_{i\,:\,t_i<t} \frac{K_i}{(t - t_i + c)^p}$$

Here $\mu$ is the background seismic activity rate and $K_i$ is related to the recorded magnitude $M_i$ of earthquake $i$ by the relationship

$$K_i = K_0 e^{\alpha(M_i - M_0)},$$

where $M_0$ is the minimum magnitude threshold for earthquakes to be recorded in the dataset, and $K_0$, $\alpha$, and $p$ are constants. Earthquake magnitudes are treated as unpredictable marks. The functional form of the triggering function, known as the modified Omori formula, was determined empirically by studies of aftershock sequences.

The temporal ETAS model was soon extended to a spatio-temporal model of the form in eq. (2.2). A variety of triggering functions $g$ were used, ranging from bivariate normal kernels to more complicated exponential decay functions and power laws; some triggering functions allow the range of spatial influence to depend on the earthquake magnitude. The inhomogeneous background $\mu(s)$, which represents spatial differences in fault structure and tectonic plate physics, can be obtained by a simple kernel density estimate (Musmeci & Vere-Jones, 1992) or by the stochastic declustering methods discussed in Section 2.2.2.

Zhuang et al. (2004) demonstrated that stochastic declustering can be used to test model assumptions. They applied the ETAS model and stochastic declustering to a catalog of 19,139 earthquakes compiled by the Japanese Meteorological Agency, then used the declustered data to test assumptions typically used in modeling earthquakes; for example, the distribution of earthquake magnitudes is assumed to be the same for main shocks and aftershocks, and both mainshocks and aftershocks trigger further aftershocks with the same spatial and temporal distribution. By identifying main shocks and aftershocks and connecting them with their offspring, it was possible to test each assumption, finding that some do not hold and leading to a revised model (Ogata & Zhuang, 2006).

Further, by using AIC, different triggering functions have been compared to improve understanding of the underlying triggering mechanisms. For example, spatial power law triggering functions were found more effective than normal kernels, suggesting aftershocks can be triggered at long ranges, and the rate of aftershock triggering depends on the magnitude of the mainshock. This has led to improved earthquake forecasting algorithms based on the ETAS model (Zhuang, 2011). Harte

(2012) explored the effects of model misspecification and boundary effects on model fits, finding that a good fit for the background component is also essential, as a poor background fit tends to bias the model to consider background events as triggered events instead, overestimating the rate of triggering and the expected number of offspring events, $m$.

Some research suggests that the parameters of the ETAS model are not spatially homogeneous, and that a more realistic model would allow the parameters to vary in space. Ogata, Katsura, and Tanemura (2003) introduced a method which allows parameters to vary in space, linearly interpolated between values defined at the corners of a Delaunay triangulation of the space defined by the earthquake locations. To ensure spatial smoothness in these values, a smoothness penalty term was added to the log-likelihood. Nandan, Ouillon, Wiemer, and Sornette (2017) took a similar approach, partitioning the region $X$ drawing $q$ points uniformly at random within $X$, obtaining the Voronoi tesselation, and allowing each Voronoi cell to have a separate set of parameters. No spatial smoothness was imposed, and the number of points $q$ was selected via BIC.

Similar concerns apply to temporal nonstationarity. Kumazawa and Ogata (2014) considered two approaches to model changes in parameters over time: a change-point model, in which parameters are fitted separately to events before and after a suspected change point, and a continuously varying model in which several parameters, including the triggering rate, were assumed to be first-order spline functions in time. Temporal smoothness was enforced with a penalty term in the log-likelihood, and AIC was used to compare the fits in series of earthquakes recorded in Japan, finding evidence of nonstationarity in an earthquake swarm.

### 2.3.2 Crime Forecasting

After the development of ETAS models, Mohler, Short, Brantingham, Schoenberg, and Tita (2011) drew an analogy between aftershock models and crime. Criminologists have demonstrated that near-repeat victimization is common for certain types of crime—for example, burglars often return to steal from the same area repeatedly (Bernasco, Johnson, & Ruiter, 2015; Short, D'Orsogna, Brantingham, & Tita, 2009; Townsley, Homel, & Chaseling, 2003), and some shootings may cause retaliatory shootings soon after (Loeffler & Flaxman, 2017; Ratcliffe & Rengert, 2008), typically within just a few hundred meters. These can be treated as "aftershocks" of the original crime.

Similarly, several criminological theories suggest the background rate of crime can be expected to widely vary by place. Routine activities theory (L. E. Cohen & Felson, 1979) states that criminal acts require three factors to occur together: likely offenders, suitable targets, and the absence of capable guardians. These factors vary widely in space depending on socioeconomic factors, business and residential de-

velopment, and the activities of police or other guardians (e.g. vigilant neighbors). Rational choice theory (Clarke & Cornish, 1985) considers criminals making rational decisions to commit offenses based on the risks and rewards they perceive—and the availability of low-risk high-reward crime varies in space. Weisburd (2015), using crime data across several cities, argued for a *law of crime concentration*, stating that a large percentage of crime occurs within just a few percent of street segments (lengths of road between two intersections) in a given city. Bolstering this, Gorr and Lee (2015) demonstrated that a policing program based on both chronic hot spots and temporary flare-ups can be more effective than a program based on only one or the other.

These theories suggest a model of crime that assumes the conditional intensity of crime occurrence can be divided into a chronic background portion, which may vary in space depending on a variety of factors, and a self-exciting portion which accounts for near-repeats and retaliations (Mohler et al., 2011):

$$\lambda(s,t) = \nu(t)\mu(s) + \sum_{i\,:\,t_i<t} g(s-s_i, t-t_i),$$

where $g$ is a triggering function and $\nu(t)$ reflects temporal changes from weather, seasonality, and so on. Initially, $\nu$, $\mu$, and $g$ were determined nonparametrically following Algorithm 2.1, though weighted kernel density estimation was too expensive to perform on the full dataset of 5,376 residential burglaries, so they modified the algorithm to subsample the dataset on each iteration. An alternate approach, requiring no subsampling, would be to use a fast approximate kernel density algorithm to reduce the computational cost (Gray & Moore, 2003).

Mohler (2014) introduced a parametric approach intended to simplify model fitting and also incorporate "leading indicators"—other crimes or events which may be predictive of the crime of interest. In a model forecasting serious violent crime, for example, minor offenses like disorderly conduct and public drunkenness have proven useful in predictions, since they may reflect behavior that will escalate into more serious crime (J. Cohen et al., 2007). The intensity is simplified to make the background constant in time ($\nu(t) = 1$), and to incorporate leading indicators, the background is based on a weighted Gaussian kernel density estimate, in which $\nu(t) = 1$ and

$$\mu(s) = \sum_{i=1}^{n} \frac{\alpha_{M_i}}{2\pi\eta^2 T} \exp\left(-\frac{\|s-s_i\|^2}{2\eta^2}\right),$$

where $T$ is the length of the time window encompassed by the dataset, $s_i$ and $t_i$ the location and time of crime $i$, $M_i$ is a mark giving the *type* of crime $i$ (where $M_i = 1$ by convention for the crime being predicted), and $\alpha$ is a vector of weights determining

the contribution of each event type to the background crime rate. The sum is over all crimes, avoiding the additional computational cost of stochastic declustering. The marks are treated as unpredictable, and only the ground process is estimated, not the conditional distribution of marks.

Similarly to $\mu(s)$, the triggering function $g$ is Gaussian in space with an exponential decay in time:

$$g(s, t, M) = \frac{\theta_M}{2\pi\omega\sigma^2} \exp(-t/\omega) \exp\left(-\frac{\|s\|^2}{2\sigma^2}\right).$$

$\theta$ performs a similar function to $\alpha$, weighting the contribution of each type of crime to the conditional intensity. The bandwidth parameters $\sigma^2$ and $\eta^2$ determine the spatial influence of a given crime type, while $\omega$ determines how quickly its effect decays in time. In principle, different spatial and temporal decays could be allowed for each type of crime, but this would dramatically increase the number of parameters.

Mohler (2014) fit the parameters of this model on a dataset of 78,852 violent crimes occurring in Chicago, Illinois between 2007 and 2012. The crime of interest was homicide, using robberies, assaults, weapons violations, batteries, and sexual assaults as leading indicators. The resulting model was used to identify "hotspots": small spatial regions with unusually high rates of crime. Previous research has suggested that directing police patrols to hotspots can produce measurable crime reductions, with results varying by the type of policing intervention employed (Braga, Papachristos, & Hureau, 2014). To test the self-exciting model's effectiveness in this role, Mohler (2014) compared its daily predictions to true historical records of crime, finding that it outperforms methods that consider only fixed hotspots (equivalent to setting $\theta_i = 0$ for all $i$) and those that only consider near-repeats ($\alpha_i = 0$ for all $i$).

### 2.3.3 Epidemic Forecasting

Forecasting of epidemics of disease, such as influenza, typically rely on time series data of infections or infection indicators (such as physician reports of influenza-like illness, without laboratory confirmation), and hence often rely on time series modeling or compartment models, such as the susceptible–infectious–recovered model (Nsoesie, Brownstein, Ramakrishnan, & Marathe, 2013). This data does not typically include the location and time of individual infections, instead containing only aggregate rates over a large area.

When individual-level data is available, however, point processes can model the clustered nature of infections. Spatial point processes have been widely used for this purpose (Diggle, 2014, chapter 9), and when extended to spatio-temporal analysis, self-exciting point processes are a natural choice, with excitation representing the

transmission of disease. Again following the ETAS literature, Meyer et al. (2012) introduced a self-exciting spatio-temporal point process model adapted for predicting the incidence of invasive meningococcal disease (IMD), a form of meningitis caused by the bacterium *Neisseria meningitidis,* which can be transmitted between infected humans and sometimes forms epidemics. Unaffected carriers can retain the bacterium in their nasopharynx, suggesting that observed cases of IMD can be divided into "background" infections, transmitted from an unobserved carrier to a susceptible individual, and triggered infections transmitted from this individual to others.

In a dataset of 636 infections observed in Germany from 2002–2008, each infection's time, location (by postal code), and finetype (strain) was recorded. The model includes unique features: rather than empirically estimating the background function, it is composed of a function of population density and of a vector of covariates (in this case, the number of influenza cases in each district of Germany, hypothesized to be linked to IMD). The resulting conditional intensity function is

$$\begin{aligned}\lambda(s,t) = \rho(s,t)\exp\left(\beta' z(s,t)\right) \\ + \sum_{j\in I^*(s,t)} e^{\eta_j} g(t-t_j) f(\|s-s_j\|),\end{aligned}$$

where $I^*(s,t)$ is the set of all previous infections within a known fixed distance $\delta$ and time $\epsilon$. Here $\rho(s,t)$ represents the population density, $z(s,t)$ the vector of spatio-temporal covariates, and $\eta_j = \gamma_0 + \gamma' m_j$, where $m_j$ is a vector of unpredictable marks on each event, such as the specific strain of infection. The spatial triggering function $f$ is a Gaussian kernel, and the temporal triggering function $g$ is assumed to be a constant function, as there were comparatively few direct transmissions of IMD in the dataset from which to estimate a more flexible function.

The results were promising, showing that the self-exciting model can be used to estimate the epidemic behavior of IMD. The unpredictable marks $m_j$ included patient age and the finetype (strain) of bacterium responsible. Comparisons between finetypes revealed which has the greatest epidemic potential, and the age coefficient allowed comparisons of the spread behavior between age groups.

Meyer and Held (2014) then proposed to replace $f$ with a power law function, previously found to better model the long tails in the movement of people (Brockmann, Hufnagel, & Geisel, 2006). Using the asymptotic covariance estimator given in eq. (2.13), they also produced confidence intervals for their model parameters, though without verifying the necessary regularity assumptions on the conditional intensity function (Meyer, 2010, section 4.2.3). A similar modeling approach was used to test if psychiatric hospital admissions have an epidemic component, via a permutation test for the parameters of the epidemic component of the model (Meyer, Warnke, Rössler, & Held, 2016).

Schoenberg, Hoffman, and Harrigan (2017) introduced a recursive self-exciting epidemic model in which the expected number of offspring $m$ of an event is not constant but varies as a function of the conditional intensity, intended to account for the natural behavior of epidemics: when little of the population has been exposed to the disease, the rate of infection can be high, but as the disease becomes more prevalent, more people have already been exposed and active prevention measures slow its spread. The model takes the form

$$\lambda(s,t) = \mu + \int_X \int_0^t H\left(\lambda(s',t')\right) g(s-s',t-t')\,\mathrm{d}N(s',t'),$$

where $g$ is a chosen triggering function and $H$ is the *productivity function*, determining the rate of infection stimulated by each event as a function of its conditional intensity. Schoenberg et al. (2017) took $H(x) = \kappa x^{-\alpha}$, with $\kappa > 0$, to model decreasing productivity, and fit to a dataset of measles cases in Los Angeles, California with maximum likelihood to demonstrate the effectiveness of the model.

### 2.3.4 Events on Social Networks

The models discussed so far have considered events in two-dimensional space (e.g. latitude and longitude coordinates of a crime or infection). Recently, however, self-exciting point processes have been extended to other types of events, including events taking place on social networks.

Fox, Short, Schoenberg, Coronges, and Bertozzi (2016) considered a network of officers at the West Point Military Academy. Each officer is a node on the network, and directed edges between officers represent the volume of email sent between them. Fox, Short, et al. (2016) developed several models, the most general of which models the rate at which officer $i$ sends email as

$$\lambda_i(t) = \nu_i \mu(t) + \sum_j \sum_{r_k^{ij} < t} \theta_{ij}\omega_i e^{-\omega_i (t - r_k^{ij})}.$$

Here $r_k^{ij}$ represents the time of the $k$th message sent from officer $j$ to officer $i$, $\omega_i$ is a temporal decay effect for officer $i$, and $\theta_{ij}$ models a pairwise reply rate for officer $i$'s replies to officer $j$. The background rate $\mu(t)$ is allowed to vary in time to model time-of-day and weekly effects, with a offset $\nu_i$ for each officer. The model is fit by expectation maximization and standard errors found by parametric bootstrap.

Zipkin, Schoenberg, Coronges, and Bertozzi (2015) considered the same dataset, but instead of modeling a self-exciting process for each officer, they assigned one to each edge between officers, which enabled them to develop methods for a missing-data problem: can the sender or recipient be inferred if one or both are missing from

a given message? The self-exciting model had promising results, and they suggested a possible application in inferring participants in gang violence.

Taking an alternate approach, Green, Horel, and Papachristos (2017) modeled the contagion of gun violence through social networks in Chicago. The network nodes were all individuals who had been arrested by Chicago police during the study period, connected by edges for each pair of individuals who had been arrested together, assumed to indicate strong pre-existing social ties. Rather than predicting the rate on edges, as Fox, Short, et al. (2016) did, this study modeled the probability of each individual being a victim of a shooting as a function of seasonal variations (the background) and social contagion of violence, as the probability of being involved in a shooting is assumed to increase if someone nearby in the social network was recently involved as well.

This is formalized in the conditional intensity for individual $k$,

$$\lambda_k(t) = \mu(t) + \sum_{t_i<t} \phi_{k_i,k}(t - t_i),$$

where $\mu(t)$ represents seasonal variation and the self-excitation function $\phi_{k_i,k}$ is composed of two pieces, a temporal decay $f_\beta(t)$ and a network distance $g_\alpha(u, v)$:

$$\begin{aligned} f_\beta(t) &= \beta e^{-\beta t} \\ g_\alpha(u, v) &= \begin{cases} \alpha \operatorname{dist}(u, v)^{-2} & \text{when } \operatorname{dist}(u, v) \le 3 \\ 0 & \text{otherwise} \end{cases} \\ \phi_{u,v}(t) &= f_\beta(t) g_\alpha(u, v), \end{aligned}$$

where $\operatorname{dist}(u, v)$ is the minimum distance (number of edges) between nodes $u$ and $v$. The model was fit numerically via maximum likelihood, and a form of declustering performed by attributing each occurrence of violence to the larger of the background $\mu(t)$ or the sum of contagion from previous events, rather than using a stochastic declustering method as discussed in Section 2.2.2.

## 2.4 Summary

This chapter introduced self-exciting point process models and their varied applications. Next, Chapter 3 introduces a new model adapted for the spatio-temporal modeling of crime, extending the models discussed in Section 2.3.2.

# *Three*

# The Extended Model

Our model builds on the model introduced by Mohler (2014), discussed in Section 2.3.2.[1] Mohler's model used a nonparametric estimate of the background intensity by using observed crime data. Instead, we would like to incorporate relevant spatial covariates, like population density, poverty rates, socioeconomic and demographic variables, and attractors of crime like pawn shops and liquor stores, so we may estimate their effects on local crime rates while also accounting for self-excitation. In this chapter we introduce our revised crime model incorporating these features.

## 3.1 Why Not Just Use Regression?

Before I get into the details, I should answer a fairly simple question: why is a full self-exciting point process model with covariates necessary? Methods like Risk Terrain Modeling (Kennedy et al., 2010; Kennedy et al., 2015) use regression methods to predict the number of crimes in an area using spatial covariates, and a straightforward Poisson regression seems adequate for the task.

However, in the presence of self-excitation, spatial regression cannot accurately estimate covariate effects. There is an intuitive explanation for this: if some crimes trigger other crimes, for example by causing retaliation, the point process will exhibit additional clustering which is not accounted for by the covariates, and more events than can be explained by the covariates. In attempting to fit a model which does not include self-excitation, we will naturally obtain biased results.

A simulation demonstrates this effect. I simulated events occurring on a spatial grid, using the self-exciting point process model defined in the next section, with two spatial covariates which varied separately over the grid (shown in Figure 3.1). The amount of self-excitation varied from crimes never triggering other crimes to crimes nearly *always* triggering another crime. The simulated crimes were then

[1] Portions of this chapter have been published as Reinhart and Greenhouse (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C*. doi:10.1111/rssc.12277

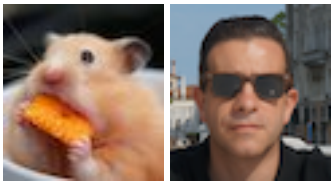

*Figure 3.1:* Two covariates were used in the simulation: a hamster eating a Cheez-It, and Alessandro Rinaldo (right). Values were obtained by extracting the grayscale brightness value of each pixel. Each is 66 × 60 pixels, each pixel representing a grid cell.

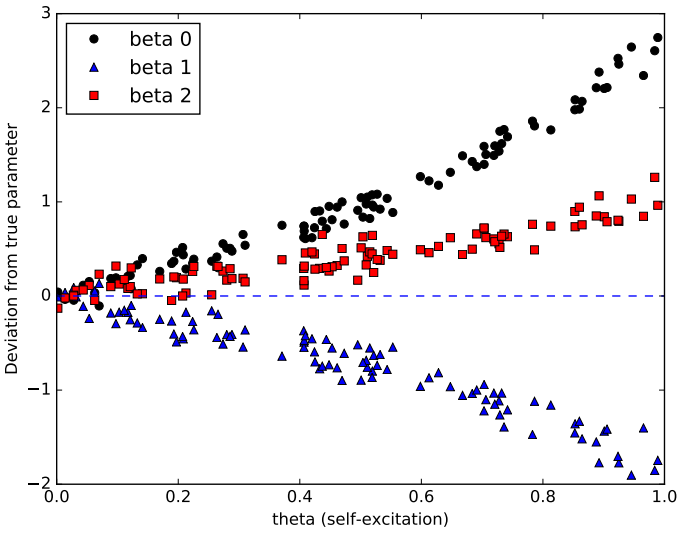


*Figure 3.2:* As the amount of self-excitation increases from 0 crimes triggered to 1 crime triggered for every observed crime, spatial Poisson regression coefficients gradually become more and more biased.

gridded and a Poisson generalized linear regression model fit to the counts of crimes in each grid cell. I compared the regression coefficients obtained in this model fit to the true coefficients I set in the simulation.

The results are shown in Figure 3.2. As the amount of self-excitation increases, the regression coefficients become more and more systematically biased. The intercept, $\beta_0$, increases to account for the additional crimes; the covariate coefficient $\beta_1$ decreases from its true value of 4.8, and $\beta_2$ increases from its true value of −2.3.

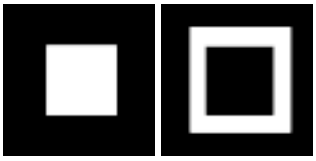

*Figure 3.3:* Two synthetic covariates. The covariates have value 1 in the white areas and zero elsewhere. The covariate on the left has a true coefficient of zero in the simulations, while the covariate on the right has a positive true effect. The spatial decay distance is $\sigma = 5$ pixels, so the effect of the right covariate spreads to the area of the left covariate.

Notably, this means both covariate coefficients shrink towards zero in the presence of self-excitation, and the magnitude of this effect is large compared with their absolute size.

In certain circumstances, self-excitation can cause increases in coefficients instead of decreases. For example, Figure 3.3 shows two synthetic spatial covariates. One is nonzero in a center square, the other in a ring around that square. Only the first covariate has a true nonzero coefficient, but because the clustering produces crimes outside the square, its effect "leaks" to the outer ring, causing the second covariate to appear to have a positive coefficient, as shown in the simulation results in Figure 3.4.

A common way to avoid these problems is to regress with lags. Figure 3.5 shows the results of simulations in which events were simulated over two months, using the covariates shown in Figure 3.1, and counts taken for every five-day period in the interval. Along with the spatial covariates, the counts from the previous three periods were also included as covariates, potentially allowing the Poisson regression to account for self-excitation. The temporal decay constant was $\omega = 10\,\mathrm{d}$, so the three lags should have accounted for most of the self-excitation. Nonetheless, Figure 3.5 still shows the bias effect for high values of $\theta$, only slightly less bias than in the previous simulation without lags.

This is likely because lagged counts within each grid cell are not sufficient to account for self-excitation, since events triggered by a crime in one cell may occur in other nearby cells. Further simulations demonstrate that if the self-excitation is forced to always trigger events within the same grid cell, the bias does indeed decrease, but is not eliminated, suggesting that a simple linear dependence on past lags is insufficient to account for the self-excitation. Further, if triggered events are known to occur within a certain distance $d$, we would want to include as covariates the lagged counts of events in all grid cells within distance $d$, so their effect could also be accounted for. But because regression provides no direct way to estimate $d$,

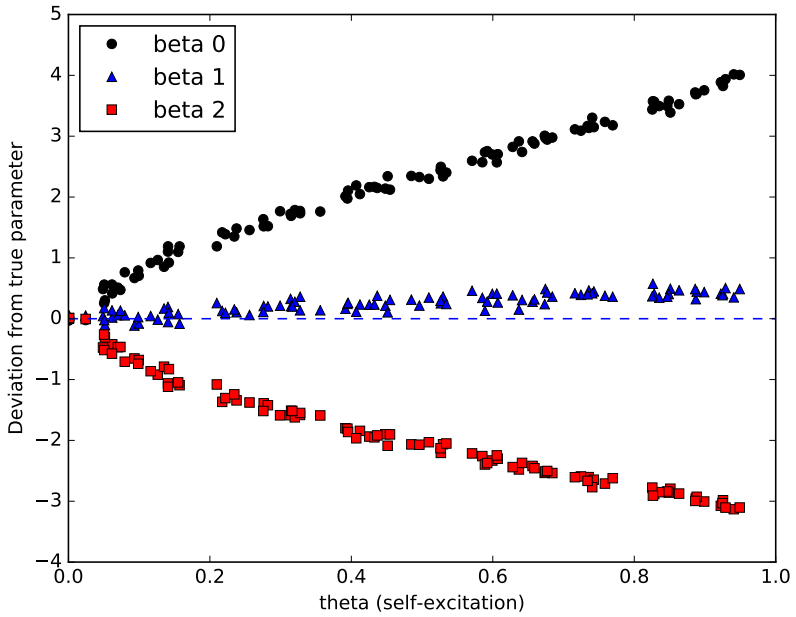


*Figure 3.4:* As the amount of self-excitation increases, the coefficient $\beta_1$ (the left covariate in Figure 3.3) increases from zero, despite its true value being zero. $\beta_2$ shrinks toward zero for the same reason as in Figure 3.2.

we do not know which lagged counts must be included and must err on the side of including too many.

This bias effect is generic and does not depend on the choice of regression model or the form of the self-excitation. (A similar effect would occur if crime *suppressed* future crime, instead of stimulating it.) We can see this more clearly in the causal diagram presented in Figure 3.6, which presents a simplified situation in which we observe crime in a grid cell $i$ at two times, $t$ and $t-1$. A crime at $t$ may be caused by a crime at $t-1$, or may be caused by two separate covariates. Crucially, because there is a causal path from the covariates through $t-1$ to $t$, estimates of the direct effect of the covariates on crime at $t$ are confounded unless the crime at $t-1$ is observed and controlled for.

By using a point process model which explicitly accounts for both covariates and self-excitation, and does not require arbitrary aggregation of data into grid cells and hence avoids the Modifiable Areal Unit Problem mentioned in Chapter 2, we can avoid the biases shown above and more explicitly model each component of the underlying process, including directly estimating the self-excitation distance and time decay.

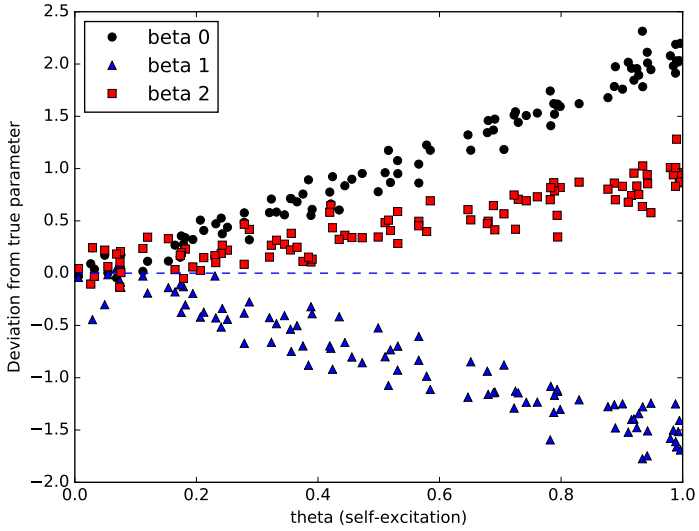


*Figure 3.5:* By including counts in three previous five-day windows as covariates, the Poisson regression model can attempt to account for self-excitation. However, the bias as $\theta$ increases is still present, only slightly reduced from Figure 3.2.

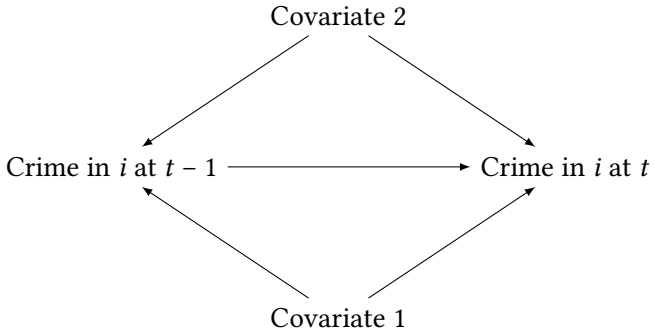


*Figure 3.6:* A simplified causal diagram of crime observed in a grid cell $i$ at two times, $t$ and $t-1$, when there are two covariates which may affect the rate of crime.

## 3.2 Why Not Knox?

We can also ask a related question: when analyzing near-repeat behavior in crime data, why not simply use the Knox test for spatio-temporal clustering? The Knox test (Knox, 1964) is conceptually simple: the statistician selects a threshold distance $\Delta s$ and threshold time $\Delta t$, and counts all pairs of events which are within this distance and time of each other. This test statistic is compared against a null distribution, obtained by Monte Carlo by permuting the times of all events, to produce a $p$ value, testing whether the spatio-temporal clustering is stronger than expected by chance. There can be false positive problems if the background event rate varies over time in small areas (Ornstein & Hammond, 2017), which induces apparent clusters which are not actually self-exciting, but the test seems otherwise sound.

In numerous criminological applications, the range of self-excitation—the distance over which the influence of a recent crime spans—has been determined by selecting various thresholds $\Delta s$ and determining at which distance the Knox $p$ value is no longer statistically significant. Sometimes this involves a variation of the Knox test in which pairs are binned into discrete categories of distance and time, such as pairs within 100 m, 200 m, and so on (S. D. Johnson et al., 2007; Townsley et al., 2003).

Regardless, these variations suffer from a common flaw. Determining the range of self-excitation by determining which Knox test is significant is, in fact, merely a determination of whether the sample size is sufficiently large enough to give power to reject the null. If the range of self-excitation is $\Delta s = 1$ in some arbitrary unit, excess clustering should be observed in a Knox test done with a threshold of $\Delta s = 10$ provided the sample size is large enough. This is illustrated in Figure 3.7, which shows the results of simulations done with the same self-excitation distance and increasing sample sizes, using different Knox cutoffs. Short distance cutoffs have the highest power, but as the sample size increases, longer cutoffs gain power, making the self-excitation distance appear to be longer.

A similar problem occurs for efforts to find the self-excitation time $\Delta t$ by successive Knox tests. Because the null hypothesis is false for all values of $\Delta t$, significance at each chosen cutoff reflects only the statistical power of the test, not the actual self-excitation range. Self-excitation must be directly modeled and fit to data to understand its dynamics and parameters.

## 3.3 Adding Covariates

We start with the basic conditional intensity form given in (2.2). Our background function $\mu(s)$ will be a log-linear predictor based on covariates. We assume that the observation domain $X$ is divided into cells $c$ of arbitrary shape, inside of which a

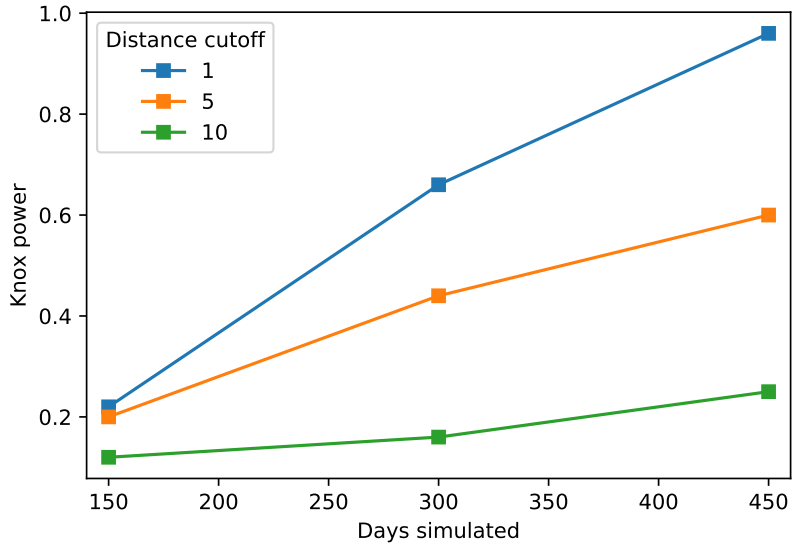


*Figure 3.7:* The power of the Knox test to detect true clustering. The simulated clustering is on the length scale $\Delta s = 1$, but we see that as the sample size (length of time period over which crimes are simulated) increases, the power of the Knox test to detect clustering at longer length scales increases, leading to false conclusions about the range of self-excitation.

covariate vector $X_c$ (including an intercept term) is known. Our model is then

$$\lambda(s,t) = \exp\left(\beta X_{C(s)}\right) + \sum_{i\,:\,t_i<t} g(s-s_i, t-t_i, M_i), \tag{3.1}$$

where $C(s)$ is the index of the covariate cell containing $s$ and

$$g(s,t,M) = \frac{\theta_M}{2\pi\omega\sigma^2}\exp(-t/\omega)\exp\left(-\frac{\|s\|^2}{2\sigma^2}\right),$$

as was used by Mohler (2014). We let $g(s,t,M) = 0$ for $s < \delta$, for an arbitrary short distance $\delta$, to prevent crimes which occur at exactly the same location from enticing the model to converge to $\sigma = 0$.

In principle, this model could be built with covariates which vary continuously in space, defined by a function $X(s)$. This would increase the generality of the model. However, in practice, this generality is not necessary: most socioeconomic, demographic, or land use variables are observed only in cells such as city blocks, census blocks, or neighborhoods. As we will see in Section 3.4 and Section 3.5, piecewise

constant covariates make estimation and simulation computationally tractable, and so the small loss in generality is worth the substantial gain in practicality.

We may also reasonably ask about the form of the triggering function $g$, which specifies an exponential decay in time and a Gaussian kernel in space. Meyer and Held (2014), for example, analyzing the spread of infectious disease, proposed a power law kernel to account for long-range flows of people. Unfortunately, most alternate spatial kernels make the expectation maximization strategy described in the next section more difficult, by making analytical maximization on each iteration impossible. These kernels could still be used, but with the additional computational cost of numerical maximization.

## 3.4 Expectation Maximization

The log-likelihood given in (2.8) could be maximized by any numerical optimization method, but given the natural interpretation of this model as a mixture model, where crimes arise from a mixture of the static background and self-excited foreground components of the model, an expectation maximization (EM) approach is simple to implement and reasonably effective. We use the basic EM approach described in Section 2.2.1, with a few modifications.

The complete-data log-likelihood, following Section 2.2.1, is

$$\begin{aligned}\ell_c(\Theta) = &\sum_{i=1}^{n} \mathbb{1}(u_i = 0)\beta X_{C(s)} \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{1}(u_i = j)\log\Big(g(s_i - s_j, t_i - t_j)\Big) \\ &- \int_0^T \int_X \lambda(s,t)\,\mathrm{d}s\,\mathrm{d}t,\end{aligned}$$

where $\Theta$ is the complete vector of parameters.

We first attack the integral term, which I'll call $C$. We have

$$\begin{aligned}C &= \int_0^T \int_X \lambda(s,t)\,\mathrm{d}s\,\mathrm{d}t \\ &= \int_0^T \int_X \exp\Big(\beta X_{C(s)}\Big) + \sum_{i\,:\,t_i<t} \frac{\theta_{M_i}}{2\pi\omega\sigma^2}\exp\big(-(t-t_i)/\omega\big)\exp\left(-\frac{\|s\|^2}{2\sigma^2}\right)\mathrm{d}s\,\mathrm{d}t.\end{aligned}$$

We split the integral in two. The first portion is

$$\begin{aligned}\int_0^T \int_X \exp\left(\beta X_{C(s)}\right) \mathrm{d}s \, \mathrm{d}t &= T \int_X \exp\left(\beta X_{C(s)}\right) \mathrm{d}s \\ &= T \sum_{\text{cells } i} A_i \exp(\beta X_i),\end{aligned}$$

where the sum is over covariate cells and $A_i$ is the area of covariate cell $i$. If covariates were not piecewise constant in space, we would have to evaluate the integral numerically instead, a significant computational cost.

The second portion is more difficult, and requires that we let $X$ be all of $\mathbb{R}^2$ instead of just the bounding box of the crimes or the jurisdiction. (This means we neglect boundary effects, which may have consequences at the edge of our jurisdiction. See Section 3.7.) We need not let $T \longrightarrow \infty$, which, as discussed in Section 2.1.4, harms accuracy more than the approximation of $X$ as $\mathbb{R}^2$. With this approximation, we can integrate out:

$$\begin{aligned}&\int_0^T \int_X \sum_{i\,:\,t_i<t} \frac{\theta_{M_i}}{2\pi\omega\sigma^2} \exp\left(-(t-t_i)/\omega\right) \exp\left(-\frac{\|s\|^2}{2\sigma^2}\right) \mathrm{d}s \, \mathrm{d}t \\ &= \frac{1}{2\pi\omega\sigma^2} \sum_{\text{crimes } i} \theta_{M_i} \int_{t_i}^T \exp\left(-(t-t_i)/\omega\right) \int_X \exp\left(-\frac{\|s\|^2}{2\sigma^2}\right) \mathrm{d}s \, \mathrm{d}t \\ &\le \frac{1}{\omega} \sum_{\text{crimes } i} \theta_{M_i} \int_{t_i}^T \exp\left(-(t-t_i)/\omega\right) \mathrm{d}t \\ &= \sum_{\text{crimes } i} \theta_{M_i} \left(1 - e^{-(T-t_i)/\omega}\right).\end{aligned}$$

Putting it together, we get

$$C \le T \sum_{\text{cells } i} A_i \exp(\beta X_i) + \sum_{\text{crimes } i} \theta_{M_i} \left(1 - e^{-(T-t_i)/\omega}\right),$$

where the first sum is over grid cells and the second over *all* crimes (not just response crimes).

This gives us the approximate *expected* complete-data log-likelihood of

$$\begin{aligned}\mathbb{E}[\ell_c(\Theta)] \le &\sum_{\text{responses } i} \left( P(u_i = 0)\beta X_{C(x_i,y_i)} + \sum_{t_j<t_i} P(u_i = j) \log\left(\frac{\theta_{M_j}}{2\pi\omega\sigma^2} e^{-(t_i-t_j)/\omega} \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right)\right)\right) \\ &- T \sum_{\text{cells } i} A_i \exp(\beta X_i) - \sum_{\text{crimes } i} \theta_{M_i} \left(1 - e^{-(T-t_i)/\omega}\right)\end{aligned}$$

This is what we need to maximize.

### 3.4.1 M Step

In the following sections we derive the update rules necessary to perform the M step of EM, after the E step (calculation of $P(u_i = j)$ for all $i, j$, following eqs. (2.9) and (2.10)) has already been performed.

*Covariates*

To maximize, consider only the terms that depend on $\beta$:

$$\mathbb{E}[\ell_c(\Theta)] \propto \sum_{\text{responses } i} P(u_i = 0)\beta X_{C(x_i,y_i)} - T \sum_{\text{cells } i} A_i \exp(\beta X_i)$$

If we try taking the partial derivative (which is a vector, with respect to each component of $\beta$), we obtain:

$$\frac{\partial \mathbb{E}[\ell_c(\Theta)]}{\partial \beta} = \sum_{\text{responses } i} P(u_i = 0)X_{C(x_i,y_i)} - T \sum_{\text{cells } i} A_i X_i \exp(\beta X_i)$$

$$\sum_{\text{responses } i} P(u_i = 0)X_{C(x_i,y_i)} = T \sum_{\text{cells } i} A_i X_i \exp(\beta X_i).$$

This cannot be directly solved for $\beta$, as it is a high-order polynomial in $\exp(\beta)$. However, the original likelihood expression is convex in $\beta$ and easy to differentiate, so we use a standard numerical maximization routine to find a maximum on each iteration.

*Spatial Decay*

For $\sigma^2$ we extract the only component of the expected log-likelihood which depends on it:

$$\mathbb{E}[\ell_c(\Theta)] \propto \sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)\left(\log\left(\frac{\theta_{M_j}}{2\pi\omega\sigma^2}\right) - \frac{d_{ij}^2}{2\sigma^2}\right).$$

We take the partial derivative and obtain

$$\frac{\partial \mathbb{E}[\ell_c(\Theta)]}{\partial \sigma^2} = \sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)\left(\frac{d_{ij}^2}{2\sigma^4} - \frac{1}{\sigma^2}\right).$$

Setting this to zero and solving for $\sigma^2$ yields

$$0 = \sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)\left(d_{ij}^2 - 2\sigma^2\right)$$

$$\sigma^2 = \frac{\sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)d_{ij}^2}{2\sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)}.$$

*Temporal Decay*

For $\omega$ we extract three terms from the log-likelihood:

$$\mathbb{E}[\ell_c(\Theta)] \propto \sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)\left(\log\left(\frac{\theta_{M_j}}{2\pi\omega\sigma^2}\right) - (t_i - t_j)/\omega\right) - \sum_{\text{crimes } i} \theta_{M_i}\left(1 - e^{-(T-t_i)/\omega}\right).$$

Taking the partial derivative, we obtain

$$\frac{\partial\,\mathbb{E}[\ell_c(\Theta)]}{\partial\omega} = \sum_i \sum_{t_j<t_i} P(u_i = j)\left(\frac{t_i - t_j}{\omega^2} - \frac{1}{\omega}\right) - \sum_i \theta_{M_i} e^{-(T-t_i)/\omega}(t_i - T)/\omega^2,$$

which we set to zero and solve to obtain

$$0 = \sum_i \sum_{t_j<t_i} P(u_i = j)\left((t_i - t_j) - \omega\right) - \sum_i \theta_{M_i} e^{-(T-t_i)/\omega}(t_i - T)$$

$$\omega = \frac{\sum_i \sum_{t_j<t_i} P(u_i = j)(t_i - t_j) - \sum_i \theta_{M_i} e^{-(T-t_i)/\omega}(t_i - T)}{\sum_i \sum_{t_j<t_i} P(u_i = j)}.$$

Note that we have not actually solved for $\omega$—we simply used the previous version of $\omega$ on the right-hand side instead of explicitly maximizing, since the maximum cannot be found analytically here. This was the strategy used by Mohler (2014), who found that this fixed point iteration approach was adequate for maximizing $\omega$ with less computational difficulty.

*Foreground Coefficients*

Extracting the terms involving $\theta$, we obtain

$$\mathbb{E}[\ell_c(\Theta)] \propto \sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)\log(\theta_{M_j}) - \sum_{\text{crimes } i} \theta_{M_i}\left(1 - e^{-(T-t_i)/\omega}\right).$$

Taking the derivative with respect to a chosen $\theta_L$, we get the update rule

$$\frac{\partial\,\mathbb{E}[\ell_c(\Theta)]}{\partial\theta_L} = \sum_{\text{responses } i} \sum_{t_j<t_i} \frac{P(u_i = j)}{\theta_L}\,\mathbb{1}(M_j = L) - \sum_{\text{crimes } i} \mathbb{1}(M_i = L)\left(1 - e^{-(T-t_i)/\omega}\right)$$

$$0 = \frac{1}{\theta_L} \sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)\,\mathbb{1}(M_j = L) - K_L + \sum_{\text{crimes } i} \mathbb{1}(M_i = L)e^{-(T-t_i)/\omega}$$

$$\theta_L = \frac{\sum_{\text{responses } i} \sum_{t_j<t_i} P(u_i = j)\,\mathbb{1}(M_j = L)}{K_L - \sum_{\text{crimes } i} \mathbb{1}(M_i = L)e^{-(T-t_i)/\omega}}.$$

### 3.4.2 Termination Criterion

The E and M steps are applied repeatedly until the fit reaches convergence. As convergence criterion, we calculate the full log-likelihood on every iteration, and halt when the relative change in log-likelihood is less than $\epsilon$, which is typically chosen to be $10^{-10}$. Calculating the log-likelihood is not a significant computational cost if implemented correctly: it depends on the intensity $\lambda(s_i, t_i)$ at each event $i$, but those intensities are already required for the calculations of $P(u_i = j)$ in the E step, and hence may be re-used to minimize the cost.

## 3.5 Simulation System

As discussed in Section 2.2.3, a good strategy for simulation is very useful in testing statistical properties of our model and evaluating its behavior when assumptions are violated. I implemented a simulation system based on Algorithm 2.5. The covariates model makes this particularly fast and efficient, because the background coefficients are piecewise constant in space, being defined by polygons from census block shapes or other administrative divisions. We hence do not need a general method to simulate from an inhomogeneous Poisson process: background events in each polygon are simulated from a homogeneous Poisson process, simply by sampling uniformly within the polygon (via rejection sampling). A future improvement could sample within polygons by triangulating them and directly sampling within the triangles, avoiding the need for rejection, but this was not deemed necessary for our purposes, since simulation is already fast for datasets of thousands of events.

I also worked to lower the computational cost. To save on memory allocation overhead, the size of each new generation of offspring events is calculated in advance. Each event has a Poisson($m$)-distributed number of offspring, so if there are $n$ events in the previous generation, the total number of offspring is Poisson($nm$). An array of this size is allocated, and the number of these offspring which come from each in the previous generation is drawn from a multinomial with appropriate weights.

The simulation system can simulate from the model specified by eq. (3.1), but can also simulate various violations of assumptions: the spatial distribution of offspring can be $t$ with arbitrary degrees of freedom, instead of Gaussian, and their temporal distribution can be drawn from a Gamma distribution with arbitrary parameters. The framework is flexible and allows additional distributions to be chosen easily, so we can test performance under unusual conditions.

## 3.6 Fast Dual-Tree Intensities

The EM algorithm described in Section 3.4 is computationally intensive: on each iteration, $\lambda(s, t)$ must be calculated at each of $K_0$ crimes, and since each requires a sum over $O(K)$ crimes, this step takes $O(KK_0)$ time. Even with intensities calculated in parallel on an 8-core machine, this step takes a large portion of the time of each EM iteration.

This is analogous to kernel density estimation, where obtaining an exact kernel density estimate at each of $N$ points requires $O(N^2)$ operations. Gray and Moore (2003) observed that it is possible to dramatically reduce computation time in the kernel density case by computing *approximate* densities. For any given point $s$, most points contribute very little to the density, because they are too far away in space. By carefully organizing the data with a space-splitting $k$-d tree, we can exploit this knowledge to approximate the contributions of large groups of points as a single point, in time conjectured to be $O(N)$.

Here I adapt this method to approximate calculation of $\lambda(s, t)$, with the additional complication that each crime $i$ has different weights $\theta_{M_i}$, and that the foreground influence decays in time as well as space.

### 3.6.1 $k$-d Trees

Bentley (1975) introduced $k$-d trees as a data structure to store $k$-dimensional data while supporting $O(\log N)$ queries for points in specific regions of the space. ($k$-d trees also have $O(\log N)$ performance for insertion or deletion of points, but this is less relevant for us.) A $k$-d tree is a binary tree that repeatedly splits the space along each coordinate axis, until leaf nodes contain some small number of points in a small region of the space.

There are a variety of ways to implement $k$-d trees, based on different splitting heuristics and storage methods. We need two implementations: a base `QueryTree` class representing a tree of two-dimensional points with timestamps, and a subclass `DataTree` which records the *types* of the crimes contained in each node: while calculating the intensity, we must know the type of every crime to look up the appropriate $\theta_{M_i}$ parameter to determine its contribution to the intensity.

To build either type of tree, we start with a large set of crimes. Then:

1. Choose a coordinate axis (either $x$, $y$, or $t$) arbitrarily and find the median value of that coordinate among all crimes.

2. Split the crimes into two groups, separated by the median. One group will be designated the left child of this node, the other the right child.

3. Recursively repeat this procedure on the two child nodes, stopping when the node contains fewer than a small number $M$ of nodes. We typically choose $M = 200$.

There are several heuristics to choose the optimal axis along which to split, but we simply split along each axis in turn; splitting along the axis with the largest range of values is also common, but showed no benefit for our data.

Crucially, each node stores its bounding box: the minimum and maximum values of $x$, $y$, and $t$ for all the crimes it contains. This structure makes queries easy: to find nodes in a specific region, recurse through the tree, only looking at a child node if its bounding box intersects the query region. As we will see in the next section, this also enables fast intensity calculations, as we can quickly estimate upper and lower bounds of the intensity contributed by points inside a node by using its bounding box.

### 3.6.2 Dual-Tree Intensity Algorithm

The dual-tree algorithm proposed by Gray and Moore (2003) builds upon $k$-d trees to accelerate approximate kernel density estimation. Our adaptation is presented in Algorithm 3.1 and uses a formulation of the algorithm presented by Lang (2004, Section 3.3.5). It uses two $k$-d trees, one `QueryTree` to store the points at which the intensity is to be evaluated, and one `DataTree` to store the crime data generating the intensity.

The basic principle is to gradually refine intensity estimates by progressing deeper into the $k$-d trees. We start with the root nodes of both trees, and calculate the upper and lower bounds of intensity on all points in the query tree node contributed by all points in the data tree node. These bounds are obtained by treating all points in each node as being at a single point, then finding the minimum and maximum possible distances between these two points by using the node bounding boxes. Hence these bounds can be found without summing over individual points.

If the upper and lower bounds are too far apart, we must refine our estimates by considering the child nodes of the roots. We push (query node, data node) pairs onto a priority queue, with their priority calculated from the number of points contained in each node and the gap between bounds, so that the least precise estimates with the most points will be refined first. We then loop through the queue, refining our estimates with child nodes and pushing their children onto the queue as necessary, until the bounds are within acceptable tolerances everywhere and we can stop.

A more rigorous treatment is given in Algorithm 3.1. The algorithm uses several utility functions. The PriorityQueue, Enqueue, and Dequeue functions create, push, and pop from a priority queue, respectively; in this case we are pushing pairs of nodes, and the last argument to Enqueue is the priority. Dequeue returns the

| Variable | Average difference | SE |
|---|---|---|
| $\sigma^2$ | $3.0278 \times 10^{-10}$ | $2.9235 \times 10^{-9}$ |
| $\omega$ | $1.6795 \times 10^{-4}$ | $1.6529 \times 10^{-3}$ |
| $\theta$ | $3.2329 \times 10^{-10}$ | $2.3021 \times 10^{-9}$ |
| $\beta_0$ | $-4.2464 \times 10^{-9}$ | $5.5147 \times 10^{-8}$ |
| $\beta_1$ | $4.8978 \times 10^{-9}$ | $7.3175 \times 10^{-8}$ |

*Table 3.1:* Differences between parameter values from exact and inexact fits to the same data.

node pair with the highest priority. We use a convenience function Zeros to initialize an array of zeros of a given size. Bounds calculates the lower and upper bounds on the contribution to $\hat{\lambda}(s, t)$ at $q$ from the points in $d$, and AddBounds adds $l$ and $u$ to lower and upper for all points contained in $q$. Finally, Intensity performs the exact calculation for the intensity at a point $q$ contributed by the crime at point $p$.

Note that Algorithm 3.1 is *not* an anytime algorithm, and the upper and lower bound variables are not valid bounds if the algorithm is interrupted at any stage, unlike the proposal of Gray and Moore (2003). The anytime algorithm requires initializing the upper bound at the largest possible value; in our case, despite using 64-bit floating point numbers, the difference in scale between the upper and lower bounds caused floating-point errors that rapidly accumulated, eventually rendering the bounds nonsensical. (Small numbers subtracted from the upper bound were often lost to floating-point error.)

### 3.6.3 Validation

Because the dual-tree approach produces *approximate* intensities, we must demonstrate that the approximation does not harm the model fit. A series of 100 simulated datasets, using random parameter values, a single covariate, and the simulation algorithm described in Section 3.5, were generated and used to fit with both the exact and approximate intensity algorithms. The average difference between the exact and approximate fits is shown in Table 3.1, and is quite small: on the order of $10^{-9}$ or smaller for all parameters but $\omega$, which sees an average difference on the order of $10^{-4}$ seconds. This shows that the approximation does not appreciably harm model fits.

### 3.6.4 Performance

The dual-tree intensity approach is intended to speed up the repeated calculation of intensities during expectation maximization, which otherwise would be the limiting factor for the speed of model fits. However, the naive $O(N^2)$ exact calculation is

**Algorithm 3.1** Dual-tree intensity approximation algorithm

```
1:  function DualTree(Q, D, ϵ)        ▷ Q and D are k-d trees, ϵ the maximum error
2:      P ← PriorityQueue
3:      N ← NumPointsIn(Q)
4:      lower ← Zeros(N)
5:      upper ← Zeros(N)
6:      Enqueue(P, Q, D, 0)
7:      while not Empty(P) do
8:          q, d ← Dequeue(P)
9:          l, u ← Bounds(q, d)
10:         if (u − l) ≤ 2ϵ min(lower[q])/N then
11:             AddBounds(q, lower, upper, l, u)
12:         else if Leaf(q) and Leaf(d) then
13:             ExactIntensity(q, d, lower, upper)
14:         else
15:             for qchild ∈ Children(q) do
16:                 for dchild ∈ Children(d) do
17:                     p ← Priority(q, l, u)
18:                     Enqueue(P, qchild, dchild, p)
19:     return (lower + upper) / 2
20: function ExactIntensity(q, d, lower, upper)
21:     for qpoint ∈ q do
22:         for dpoint ∈ d do
23:             c ← Intensity(qpoint, dpoint)
24:             lower[qpoint] += c
25:             upper[qpoint] += c
26: function Priority(q, l, u)
27:     n ← NumPointsIn(q)
28:     return n × (u − l)
29: function Children(node)
30:     if Leaf(node) then
31:         return {node}
32:     else
33:         return {node.left, node.right}
```

both trivially parallelizable (the intensity at each point can be computed in parallel) and easy to speed up—a large fraction of its time is spent calculating the pairwise Euclidean distances between events, which do not change from iteration to iteration and can hence be precomputed and stored in a matrix. The dual-tree approach avoids considering every single data point, but it is not easily parallelizable and requires minimum and maximum Euclidean distance calculations as each pair of nodes is examined.

In practice, then, a parallel implementation of the exact $O(N^2)$ calculation proves to be faster than the dual-tree algorithm, while also being easier to understand. There may be circumstances when the dual-tree algorithm would perform better—in a very large city, for example, when it can discard most points, or if it could be parallelized by allowing multiple threads to access the priority queue simultaneously. We did not pursue these possibilities in this thesis.

## 3.7 Boundary Effects

Section 2.2.1 briefly discusses the problem of boundary effects: if crimes are only observed in the region $X$ and time interval $[0, T)$, but also occur outside $X$ and at $t < 0$ or $t \geq T$, parameter estimates can be biased by boundary effects.

The nature of the boundary effects can be seen clearly from the parameter updates in the M step of the EM algorithm (Section 3.4.1). For example, the foreground update for $\theta_L$ is

$$\theta_L = \frac{\sum_{\text{responses } i} \sum_{t_j < t_i} P(u_i = j)\, \mathbb{1}(M_j = L)}{K_L - \sum_{\text{crimes } i} \mathbb{1}(M_i = L) e^{-(T - t_i)/\omega}},$$

which can be interpreted as a weighted average: for all crimes of type $L$, sum up their contributions to response crimes (measured by $P(u_i = j)$), and take the average. An average of 0.5, for example, says a crime of type $L$ can be expected to contribute to about 0.5 future response crimes. The denominator also contains a temporal boundary correction term which is negligible when $T$ is very large.

Suppose, however, that many crimes of type $L$ occur near the boundary of the observation region $X$, and trigger response crimes that occur outside of $X$. These response crimes will not be included in the sum in the numerator, and hence $\theta_L$ will be biased downward. Updates for $\sigma^2$ and $\omega$ can also be interpreted as weighted averages, and are subject to similar biases.

Harte (2012) explored the effects of these biases on the ETAS models discussed in Section 2.3.1. One common workaround to reduce the bias is to introduce a region $X_0 \subset X$, chosen so that events inside $X_0$ have triggered offspring that mostly occur within $X$. All events in $X$ contribute to the intensity $\lambda(s, t)$, but the weighted

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $5.436 \times 10^5$ | 6.292 | d |
| $\sigma^2$ | 9.736 | 3.12 | ft |

| Predictor | Foreground |
|---|---|
| Self-excitation | 0.3353 |

*Table 3.2:* Average parameter values from 50 simulations where true parameters are $\theta = 0.5$, $\omega = 7$ d, and $\sigma = 4$ ft. The grid is $66 \times 60$ ft and no boundary correction was applied, resulting in the biases above. Note that $\theta$ is biased too low, since events triggered outside the grid were not observed, and both $\omega$ and $\sigma$ are also too small.

averages in the M step only average over events inside $X_0$: that is, to update $\theta_L$, we average over events of type $L$ within $X_0$, counting their contributions to any response crimes within $X$. Since most of their offspring will be within $X$ by construction, the average will not leave much out.

The same subsetting is also done in time, so only events in the interval $[0, T_0)$ are considered, where $T_0 < T$. This eliminates bias caused by events at $t$ close to $T$ triggering offspring that occur after $T$ and are hence not observed.

Of course, averaging over events only in a subset of space and time reduces the effective sample size of the fit, introducing additional variance to parameter estimates. It does, however, dramatically reduce bias. To demonstrate this, Table 3.2 shows parameter values obtained from repeated simulations from a model with known parameter values, with covariates defined by the images shown in Figure 3.1. It is apparent from the table that parameters are heavily biased in the fit. However, Table 3.3 shows fits obtained when an 8-pixel boundary was established around the images, so $X_0$ was the inner $50 \times 46$ box; the simulated events occurred over the course of two years, of which the last thirty days were also left out. These fits suffer from much less bias.

An additional danger in our model is in estimates of $\beta$. If $\beta$ changes dramatically across the boundary of $X$, effects of covariates just outside of $X$ can "leak" inside $X$ and bias $\hat{\beta}$, in the same way as the covariate configuration in Figure 3.3 causes a false positive in spatial regression. A covariate inducing high event rates just outside of $X$ will induce moderately higher event rates inside of $X$ because of self-excitation, which will falsely be attributed to covariates and self-excitation from inside $X$. The subsetting procedure above does not address this problem, although a similar form of subsetting that skips background cells near the boundary may help, and deserves further research.

Additionally, boundaries have similar effects on estimates of $\lambda(s, t)$ as they do

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $6.128 \times 10^5$ | 7.092 | d |
| $\sigma^2$ | 14.75 | 3.841 | ft |

| Predictor | Foreground |
|---|---|
| Self-excitation | 0.4770 |

*Table 3.3:* Average parameter values from simulations from the same model as in Table 3.2, but where boundary correction was applied with an 8 ft buffer around all edges. The biases are substantially reduced.

in kernel density estimation (e.g., Cowling & Hall, 1996). In kernel density estimation, biases occur when the density being estimated has bounded support; near the boundaries, the kernel estimator underestimates the density because it "sees" the area outside the support with no events. This can be avoided with a variety of strategies that modify the kernel near the boundary or add "pseudodata" outside it. The self-excitation component of $\lambda(s, t)$ can also be seen as a weighted kernel density estimate, and though in this case the problem is censoring (data is unavailable outside the domain $X$), the effect is much the same, and $\lambda(s, t)$ is systematically biased downward near the boundaries.

This problem can be avoided using the same strategy as above—establishing a buffer area around the domain $X$ and not calculating $\lambda(s, t)$ inside this buffer. It could also potentially be addressed by weighted or modified triggering functions $g(s, t)$, which account for the fraction of the kernel's support contained outside $X$, along the lines of modified kernels used in density estimation, though any such strategy would make expectation maximization of the log-likelihood much more difficult. We leave this problem to further research.

## 3.8 Summary

This chapter introduced a new self-exciting point process model incorporating spatial covariates, which could be used for modeling the spread of crime, along with estimation and simulation tools for its use. Before we use it, however, we must have tools for diagnosing model fit and performing parameter inference, which will be introduced in Chapter 4.

# *Four*

# Inference and Model Diagnostics

Once we have fit a predictive policing model, we are interested in quantifying the uncertainty in its parameters and in its predictions.[1] Section 4.1 considers confidence intervals for model parameters, which previous crime models have not included. These intervals improve the interpretability of the model and aid in its use to answer criminological questions about factors which influence crime.

After examining the model parameters, we need tools to evaluate its predictive performance. Along with the methods discussed in Section 1.1.1, Section 1.1.2, and Section 1.1.3, previous researchers have developed metrics to evaluate their performance, which I review in Section 4.2. However, these tools have a number of flaws, discussed in Section 4.2.3, leading us to propose a new method based on ROC curves in Section 4.2.4. Section 4.3 also discusses the possibility of using proper scoring rules, such as the log score, for comparing predictive performance between models . In addition, in Section 4.4 I propose diagnostic tools specifically adapted to our self-exciting point process model, enabling more detailed understanding of the accuracy of its predicted crime rates.

Finally, these diagnostic and evaluation tools are used to explore the model's behavior under simulations of various forms of misspecification in Section 4.5, including various alternate forms of the triggering function $g$ and omitted but relevant spatial covariates.

## 4.1 Confidence Intervals and Coverage

Section 2.2.4 discusses several approaches to inference on the parameters of a self-exciting point process model: the Hessian of the log-likelihood at the MLE (Ogata, 1978), an estimator based on the conditional intensity function (Rathbun, 1996), and the parametric bootstrap.

[1] Portions of this chapter have been published as Reinhart and Greenhouse (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C*. doi:10.1111/rssc.12277

There are not clear theoretical results to indicate which method should perform best in producing confidence intervals and statistical tests of parameters; all three methods are asymptotic, the first two relying on asymptotic normality results for point process models, and it's not obvious which estimator of the asymptotic covariance matrix should perform best.

To settle the issue, I implemented both asymptotically normal methods. (For large datasets, the parametric bootstrap simply proves impractically slow, because of the simulation and refitting cycle.) To calculate the Hessian of the log-likelihood, I used Theano (Bergstra et al., 2010), a Python package for describing computations which automatically generates fast C code and automatically computes all necessary derivatives, meaning I did not need to explicitly derive each manually. (Theano also supports computation using the GPU, which may significantly speed up this calculation, but I have not yet tested this feature.) With the full estimated covariance matrix, I calculated standard errors for each estimator, and produced confidence intervals from these.

The estimator proposed by Rathbun (1996) (given in eq. (2.13)) was easily implemented by analytically deriving gradients of the conditional intensity function, making this estimator particularly fast. The parametric bootstrap uses the simulation method described in Section 2.2.3, Algorithm 2.5.

I ran a series of simulations to determine if either asymptotic confidence interval method attains its nominal coverage level. Each used two covariates (those shown in Figure 3.1) and simulated over a period of two years, averaging one background event per day. All model parameters were randomly selected for each simulation, and a boundary correction buffer (as described in Section 3.7) was used to limit the distortion caused by boundary effects.

Table 4.1 shows the results for the observed information estimate and for Rathbun's estimator. Coverage is worst for the self-excitation parameter $\theta$, which is affected by any remaining boundary effect not compensated for by the buffer region; Rathbun's covariance estimator achieves nearly nominal coverage for $\beta$, which is less affected. Overall, Rathbun's estimator achieves 88% coverage and is closest to its nominal 95% coverage.

It's also worth checking the assumption of asymptotic normality used in the observed information estimator and Rathbun's estimator. I ran 350 simulations averaging 6478 events each, using the two background covariates shown in Figure 3.1 and a fixed arbitrary set of parameters, and fit to each simulated dataset. The resulting sampling distribution of parameter values was collected into probability plots to compare against the normal distribution; these plots are shown in Figure 4.1, and show the shape of most sampling distributions is close to the expected normal distribution, apart from a slight right skew in several distributions. This suggests that the asymptotically normal confidence intervals should be accurate.

| Variable | Hessian (%) | Rathbun (%) |
|---|---|---|
| $\sigma^2$ | 86 | 88 |
| $\omega$ | 87 | 91 |
| $\theta$ | 82 | 63 |
| $\beta_0$ | 77 | 83 |
| $\beta_1$ | 89 | 92 |
| $\beta_2$ | 86 | 89 |
| **Average** | 85 | 88 |

*Table 4.1:* A comparison of the coverage rates of nominal 95% confidence intervals generated using the observed information matrix estimated from the Hessian or by using Rathbun's estimator, in a series of 500 simulations of two years of data. Simulations averaged around 3,000 events each, with a maximum of over 15,000.

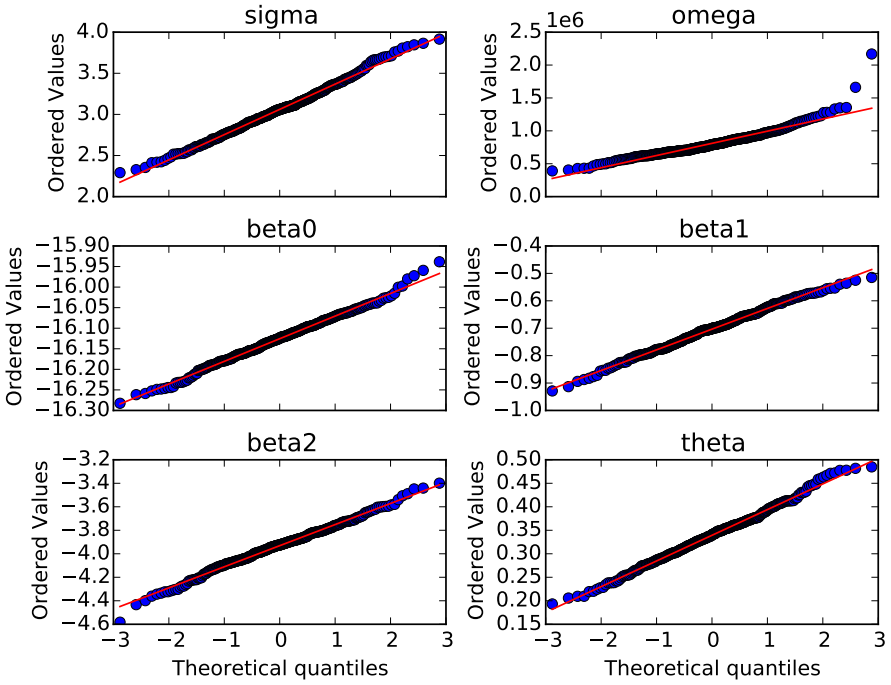


*Figure 4.1:* Probability plots of parameter estimates on simulated datasets against the reference normal distribution. The red lines are least-squares lines of best fit.

## 4.2 Hotspot-Based Hit Rate Metrics

The performance of predictive policing models is typically evaluated by producing a single hotspot map and measuring the fraction of crimes in a subsequent time period, such as several months or a year, contained in areas predicted as hotspots. For a kernel density map, hotspots are chosen by evaluating the density on a grid and marking all grid cells over a certain cutoff as hotspots; the cutoff is arbitrarily chosen, and practitioners have used the top 10% of cells with values two standard deviations above the mean (Drawve, 2016), the top 20% of cells (Levine, 2008; Van Patten et al., 2009), cells 1.96 standard deviations above the mean (Hart & Zandbergen, 2014), and other schemes. For covariate-based models like RTM, grid cells with the highest risk scores are similarly selected.

Once these hotspots are selected, a variety of metrics may be calculated from them, as discussed in the following sections.

### 4.2.1 Search Efficiency Rate

Bowers, Johnson, and Pease (2004) proposed the Search Efficiency Rate (SER), a very simple metric to compare procedures:

$$\text{SER} = \frac{\text{number of test crimes successfully predicted}}{\text{area of hot spots (km}^2\text{)}}.$$

Procedures that predict a high number of crimes in a small area are hence ranked better than those that require a larger area. This, they argue, has a clear advantage over the hit rate (total fraction of crimes predicted), since police cannot practically patrol very large hotspots. It is more useful to know the crime density within the hotspots, since a high density implies a patrolling officer can have a greater effect in a smaller area.

### 4.2.2 Prediction Accuracy Index

Chainey et al. (2008) pointed out that the SER does not adequately handle comparisons between different study areas:

> "For example, a study area that is 10km$^2$ in area may have determined certain areas where crimes are predicted to occur from which a Search Efficiency Rate of 20 crimes per km$^2$ has been calculated. A study area that is 50 km$^2$ in size may have experienced the same volume of crime as the smaller study area and also have the same Search Efficiency Rate of 20 crimes per km$^2$ for the areas where crimes are predicted to occur. Yet in the larger study area there is more space where no crime has been predicted to occur, meaning that the predicted areas that have

> been identified cover a smaller relative area than the predicted areas determined in the smaller study area, and provide a more useful basis from which to target resources, relative to the entire study area's size."

To replace SER, they introduced the Prediction Accuracy Index (PAI), which balances the hit rate of the hotspot map with the fraction of the total map area designated as hotspots, rewarding models that predict a large fraction of crimes without marking a large fraction of the map as hotspots:

$$\text{PAI} = \frac{\text{hit rate}}{\text{area of hotspots/total map area}}.$$

Hence a model which predicts 100% of future crime by designating the entire map a hotspot would have PAI = 1, while a map that predicts 90% of crime by selecting only 10% of the map would have PAI = 9. Reported PAIs vary widely; for vehicle theft, for example, they range from 2.32 in London with kernel density estimates (Chainey et al., 2008) to a spectacular 459.15 in Houston using nearest-neighbor hierarchical clustering (Levine, 2008).

However, the PAI does not solve the problems it claims to. Though Chainey et al. (2008) do not make the connection explicit, we can rewrite the PAI in terms of the SER:

$$\text{PAI} = \frac{\text{SER}}{\text{average crime density}}.$$

Scaling by crime density (crimes per unit area) makes intuitive sense: if the number of crimes doubles across the entire city, we should be able to predict twice as many crimes. However, this does not make PAI comparable between different study areas, since it is strongly sensitive to the *distribution* of crime, not just its average density. In a city where crime is clustered in several small hotspots, it is easy for any hotspot method to obtain a large PAI, because the denominator can be very small. In a city where crime is more evenly distributed, a high PAI is difficult to obtain, even if the city has the same *average* crime density. This implies that the different reported PAIs reflect differences in crime distribution as much as differences in the hotspot methods, and to compare different methods we must test them in a single city and single study area.

### 4.2.3 Other Flaws

Each of these metrics suffers from the weakness that hotspot thresholds are arbitrary, and hence model performance may vary widely depending on the chosen threshold. This makes it difficult to compare results obtained from different hotspot methods, which may use very different default hotspot thresholds. I know of no previous work which has considered the effects of varying thresholds, though my own analysis suggests the threshold can affect PAI by a factor of two or more.

Prediction evaluations are also often ad-hoc. Different investigators use data from different cities and time periods to evaluate their methods, based on whatever data they have agreements to access, and there have been few systematic attempts to understand the effect of tuning parameters (number of hotspots, smoothing bandwidth, grid size, etc.) on predictive performance (see Hart & Zandbergen, 2014, for one example). No systematic study has been made of the properties or utility of metrics like the PAI, and there are no diagnostic tools to assess why models do not fit well to particular datasets or which specific regions they do not fit to.

These prediction evaluations also do not match how hotspot maps are used in practice. Rather than being produced once and used for months, maps are typically updated daily or weekly with new data (e.g. Mohler et al., 2015), and so any evaluation method needs to work with regularly updated predictions. Also, to compare predictive models, it is essential to understand performance variation between cities and time periods; prior practice has been to use an evaluation in a single city for a single time period and treat it as authoritative.

### 4.2.4 ROC-Based Metrics

Rather than arbitrarily defining hotspots using the conditional intensity and then computing the PAI, as discussed in the previous sections, I have built new performance evaluation methods which do not require arbitrary cutoffs. Adapting techniques from statistics, and following the suggestion of Gorr (2009), I applied Receiver Operating Characteristic (ROC) curves to our model (Fawcett, 2006; Lasko, Bhagwat, Zou, & Ohno-Machado, 2005). Additionally, instead of evaluating on a single long time period, I use one-ahead predictions that more accurately reflect how hotspot models would be used: the model parameters are fit to training data, a prediction map is made, and new data is added periodically to update the map. Police may update their maps weekly or monthly as crime data arrives. Model parameters are held fixed; it is assumed that these change fairly slowly, and so they can be refit infrequently.

To calculate ROC curves, first evaluate the conditional intensity function on a fine grid. Next, record the presence or absence of the target crime in each grid cell in the subsequent time period. Treating the conditional intensity as a test statistic in a classifier, vary the hotspot cutoff from taking only the single highest intensity cell to taking every cell, recording the number of true and false positive cell classifications on the way. Figure 4.2 shows two example ROC curves, computed by making weekly predictions of burglaries over six months in 2012 and updating the intensity map with new predictions at the end of each week. One model uses only population density as a covariate, while the other includes three additional covariates. (Both models will be discussed further in Chapter 5.)

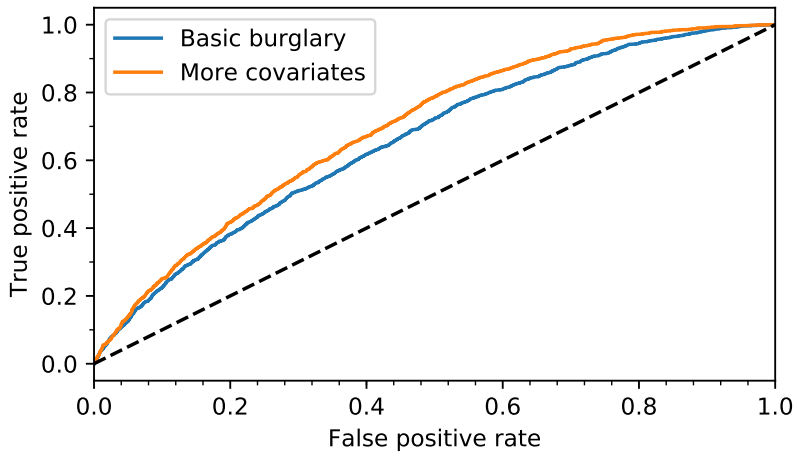


*Figure 4.2:* ROC curves for weekly predictions of all burglaries in the last six months of 2012, with or without additional demographic covariates. The covariates increase predictive performance in the middle of the range. Without covariates, the AUC is 0.66; with covariates, it increases to 0.70.

As a more interpretable alternative metric for users, I also created hit rate curves, which plot the hit rate (fraction of crimes included in selected hot spots) against the fraction of the map selected as hot spots. Figure 4.3 shows one example, using the same models, indicating that the additional covariates improve performance only when selecting a fairly large portion of the map as hotspots to be patrolled.

## 4.3 Predictive Scores

The hit rate metrics discussed in the previous section are necessary for typical hotspot methods because the hotspot methods make purely dichotomous predictions: each grid cell or map area is either part of a hotspot or it isn't. We then make evaluations based on if these hotspots capture a large fraction of crimes while only selecting a small land area. But in a self-exciting point process model of crime, we have much more information than simply a dichotomous prediction: we have a predicted crime rate $\lambda(s, t)$ at every point and time. How can we test whether this rate is well-calibrated?

Vere-Jones (1998) considered this problem for ETAS models (Section 2.3.1) and related earthquake forecasting models, which also produce conditional intensities, by drawing a connection to proper scoring rules (Gneiting & Raftery, 2007). Scoring rules evaluate probabilistic forecasts of events: a score $S(P, x)$ returns the score of a predictive distribution $P$ when outcome $x$ occurs. A simple example is the

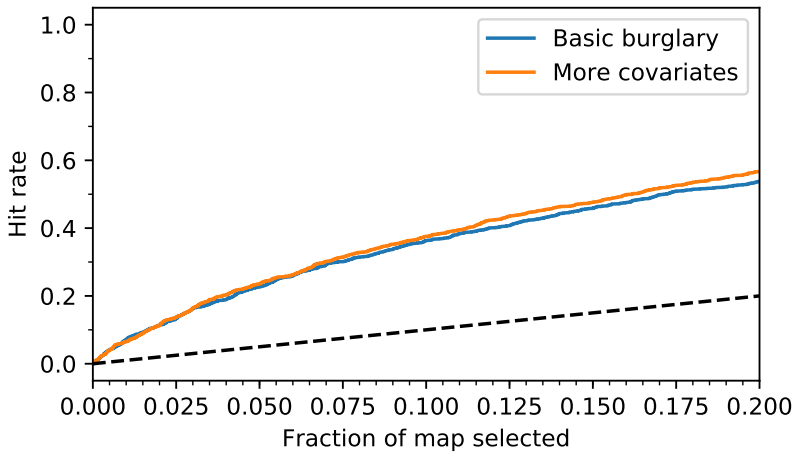


*Figure 4.3:* Hit rate curve for weekly predictions of burglaries, with and without the additional demographic covariates.

logarithmic score,

$$S_{\text{logarithmic}}(P, x) = \log p_x,$$

where $p_x$ is the forecast probability of event $x$ occurring. A prediction method which maximizes the expected score is desirable. There are many different scoring rules; a scoring rule is *proper* if the expected value of the score, under the predictive distribution $P$, is maximized by predicting $P$, meaning a forecaster has no incentive to choose any other predictive distribution than their true belief.

Vere-Jones (1998) considered defining events $x$ such as "an earthquake occurs within this forecast interval." To obtain $p_x$ from a conditional intensity model, he proposed simulating repeatedly from the model, then calculating the fraction of simulated datasets which contain an event within the chosen forecast interval. Since outcomes are binomial, one can use a binomial score

$$S_{\text{binomial}}(p, x) = x \log p + (1 - x) \log(1 - p),$$

where $x \in \{0, 1\}$ and $p$ is the forecast probability that $x$ will be 1. But despite the scoring, this still dichotomizes outcomes. Harte and Vere-Jones (2005) takes the logic further by connecting it to the entropy of the predictive distribution $P$, and defining the relative entropy

$$I^* = \mathbb{E}_P \log \frac{p_x}{\pi_x},$$

where $\pi$ is some baseline predictive distribution, such as a homogeneous Poisson process model, against which all models are compared. Because the predictive distribution $P$ is conditional on the past history of the point process, this quantity is

random, depending on the particular realization of the process; the expected information gain $G = \mathbb{E}[I^*]$ averages over all possible realizations, and numerically quantifies the intrinsic predictability of the process.

A connection soon becomes apparent. A score which sums up the logarithm of the predictive probabilities of each event is just the log-likelihood of the model; the relative entropy $I^*$ is just a log-likelihood ratio. Hence, the expected information gain $G$ is estimated by the log-likelihood ratio on an observed dataset, which converges to $G$ as the number of events grows to infinity (provided the process is stationary):

$$\hat{G} = \frac{1}{T} \log\left(\frac{L_1}{L_0}\right), \tag{4.1}$$

where $L_0$ is the baseline model likelihood and $L_1$ the likelihood of the model of interest. The likelihood ratio between two models estimates the difference in score between them, in the form of the relative entropy. (The theoretical aspects here were reviewed in more depth by Daley and Vere-Jones (2004).)

This justifies the use of the Akaike Information Criterion (AIC) to compare models as (approximately) a comparison of scores, with penalty for the number of parameters in the model. Evaluation can be performed on a separate test time period to prevent overfitting.

## 4.4 Residuals

Beyond goodness-of-fit tests and overall hit rate metrics, it is useful to be able to determine *where* the model fits: what types of systematic deviations are present, where covariates may be lacking, what types of crimes are over- or under-predicted, and so on. Eq. (2.1) suggests we can produce these detailed analyses: because the point process model predicts a conditional intensity at each location, we can calculate the expected number of crimes within each region in a certain period of time, and compare this against the true occurrences over the same time, producing a residual map. These residuals are defined to be (Daley & Vere-Jones, 2008, chapter 15)

$$R(h) = \int_{\mathbb{R}\times\mathbb{R}^2} h(s,t)\left[N(\mathrm{d}t \times \mathrm{d}s) - \lambda(s,t)\,\mathrm{d}t\,\mathrm{d}s\right],$$

where $N(\cdot)$ is the counting measure of events in the given region, and $h(s,t)$ is a bounded window function (typically an indicator function for a chosen region).

To calculate $R(h)$, a typical approach is to choose a time window—say, a particular week or month—and integrate the conditional intensity over this window. Then the spatial region $X$ is divided appropriately (see Section 4.4.1 below) and the intensity is integrated over each subdivision, then compared against the number of events in that subdivision during that time window.

Our chosen conditional intensity function (eq. (3.1)) is fairly easily integrated with respect to time, leaving a function that must be integrated over the chosen spatial regions:

$$\begin{aligned}\lambda(s) &= \int_{t_1}^{t_2} \lambda(s,t)\,\mathrm{d}t \\ &= (t_2 - t_1)\exp\left(\beta X_{C(s)}\right) + \int_{t_1}^{t_2} \sum_{i\,:\,t_i<t} g(s-s_i, t-t_i, M_i)\,\mathrm{d}t \\ &= (t_2 - t_1)\exp\left(\beta X_{C(s)}\right) + \\ &\quad \sum_{i\,:\,t_i<t_2} \frac{\theta_{M_i}}{2\pi\sigma^2}\exp\left(-\frac{\|s_i - s\|^2}{2\sigma^2}\right)\left(\exp\left(\frac{t_i - \max(t_1, t_i)}{\omega}\right) - \exp\left(\frac{t_i - t_2}{\omega}\right)\right)\end{aligned}$$

### 4.4.1 Residual Maps

Choosing spatial subdivisions for residuals requires care. The obvious choice is a discrete grid, but the right size is elusive: small grid cells produce skewed residuals with high variance (as most cells have no crimes), and positive and negative residual values can cancel each other out in large cells. Bray et al. (2014) suggest instead using the Voronoi tessellation of the plane, which produces a set of convex polygons, known as Voronoi cells, each of which contains exactly one crime and all locations that are closer to that crime than to any other.

Given this tessellation, the raw Voronoi residuals $\hat{r}_i$ for each cell $C_i$ are

$$\hat{r}_i = 1 - \int_{C_i} \hat{\lambda}(s)\,\mathrm{d}s.$$

The choice of Voronoi cells ensures that cell sizes adapt to the distribution of the data, and Bray et al. (2014) cite extensive simulations by Tanemura (2003) indicating that the Voronoi residuals of a homogeneous Poisson process have an approximate distribution given by

$$\hat{r}_i \sim 1 - X; \qquad X \sim \text{Gamma}(3.569, 3.569),$$

so that $\mathbb{E}[\hat{r}_i] = 0$. (Here the gamma distribution is parametrized by its shape and rate.) But because our model is not a homogeneous Poisson process, we performed similar simulations for random parameter values, then used maximum likelihood to fit the approximate distribution $X \sim \text{Gamma}(3.389, 3.400)$ to the 1,332,546 simulated residuals.

After each $\hat{r}_i$ is found, using Monte Carlo integration over $C_i$, the Voronoi cells can be mapped with colors corresponding to their residual values. To ease interpretation, colors are determined by $-\Phi^{-1}(F(1 - \hat{r}_i))$ where $F$ is the cumulative distribution function of the approximate distribution of $X$ and $\Phi^{-1}$ is the inverse normal cdf.

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $4.511 \times 10^6$ | 52.21 | d |
| $\sigma^2$ | $2.664 \times 10^5$ | 516.1 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −31.63 | $1.8 \times 10^{-14}$ |
| Population | 31.66 | $5.6 \times 10^{13}$ |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 2892 | 0.7640 |

*Table 4.2:* Parameters of a fit to one year of Pittsburgh burglary data, using population density (per square meter) as a covariate for each Census block.

Positive residuals indicate more observed crime than was predicted, and negative residuals less.

These residual maps provide much more detailed information than previous global measures of hotspot fit, and can indicate areas with unusual patterns of criminal activity. For example, consider a model that predicts homicides using leading indicators such as assault and robbery; this model may perform well in an area which experiences gang-related violence, but would systematically over-predict homicides in a commercial area full of bars and nightclubs, where most assaults are drunken arguments rather than signs of gang conflict.

To demonstrate the use of residual maps, consider a fit predicting burglaries using the Pittsburgh crime data to be introduced in Chapter 5. A model was fit using one year of data, using population density as the sole covariate. Fit parameters are shown in Table 4.2 and show that the spatial bandwidth of self-excitation is about 540 feet, over 51 days. This is reflected in maps of burglaries, which show hotspots appearing and disappearing from month to month.

Figure 4.4 shows the effect of this hotspot behavior on residual plots. A hotspot of burglary in the Oakland neighborhood disappears over several weeks; the model, expecting the high rate of crime to excite further burglaries, over-estimates the burglary rate as the hotspot comes to an end.

The example map does illustrate one weakness of Voronoi residual maps. We would expect areas with large positive residuals (red, in the map) to have a higher crime density than areas with large negative residuals (blue), since positive residuals indicate more crimes occurred than were expected. Hence areas with positive residuals tend to have smaller Voronoi cells than areas with negative residuals, and the map is visually dominated by large cells with negative residuals. Closer inspec-

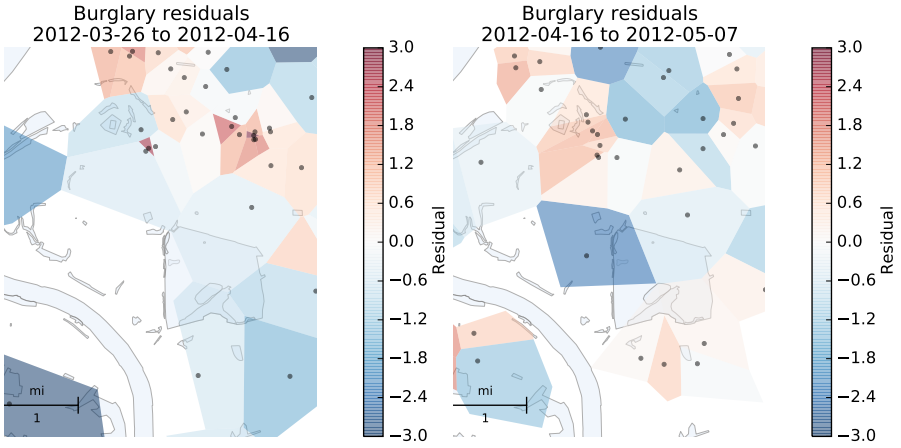


*Figure 4.4:* Burglary residuals in the Oakland neighborhood of Pittsburgh in two separate three-week periods in 2012. A cluster of burglaries near the upper right of the map is apparent in early April (*left*), containing over a dozen burglaries. By late April and early May (*right*), the cluster has shifted west, and negative residuals appear, showing that the model expected the cluster to continue.

tion reveals clusters of very small cells containing large positive residuals; these are the locations of new crime hotspots. Users should be aware of this problem when interpreting residual maps.

Previous applications of residual maps have focused on their use in visualizing individual models, but we can easily extend them to compare the fits of two different models fit to the same dataset. We might, for example, use a residual map to discover a flaw in one model, which we hypothesize could be fixed with the use of an additional covariate, then want to determine whether the additional covariate indeed fixed the problem. We calculate the residuals $\hat{r}_i^a$ and $\hat{r}_i^b$ in the same way as before for both models $a$ and $b$, using a common set of Voronoi cells, then calculate

$$\hat{\delta}_i = |\hat{r}_i^a| - |\hat{r}_i^b|.$$

This indicates the degree of improvement obtained by model $b$ over model $a$: when $\hat{\delta}_i$ is positive, model $b$'s prediction in Voronoi cell $i$ is closer to the truth than model $a$'s.

We can also compare residuals to covariate values at the location of the event defining the Voronoi cell, in a fashion analogous to a plot of residuals versus a covariate in ordinary multiple regression. This could be useful in identifying transformations that may be needed for covariates, or for exploratory analysis to find covariates that should be included in a model. Figure 4.5 shows an example plot for

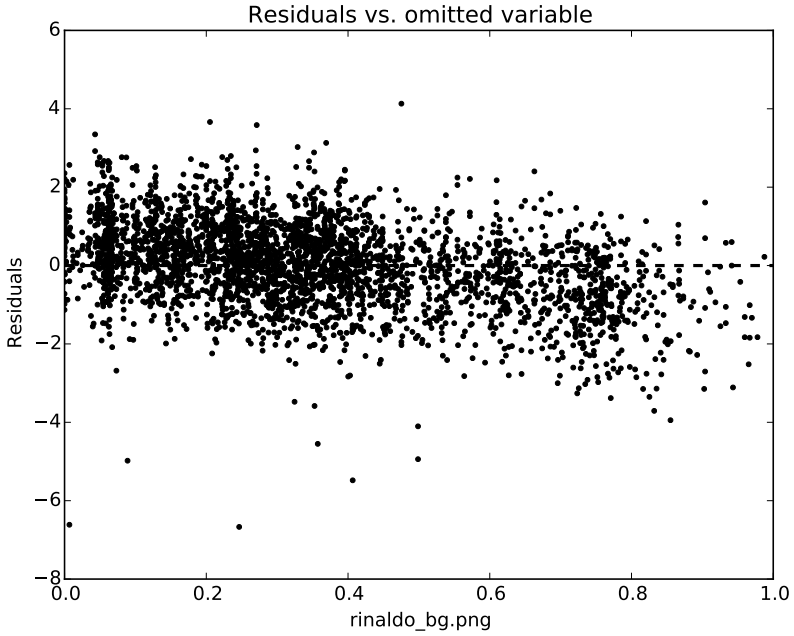


*Figure 4.5:* Residuals of a simulated fit, plotted against the values of a second covariate which is relevant to the event rate but which was not included in the fit. The visible trend indicates that the covariate should be added to the model.

a simulated dataset in which one of the relevant covariates was omitted from the fit; plotted against this covariate, the residuals show a clear trend, demonstrating that it should be added to the model.

A purely temporal residual analysis can be useful to illustrate the calibration of the model over time. Consider plotting the index $i$ of each event versus the quantity

$$\tau_i = \int_0^{t_i} \int_X \lambda(s, t) \, \mathrm{d}s \, \mathrm{d}t,$$

the expected number of events in the interval $[0, t_i)$. This is an extension of the standard transformation property of point processes: if the model is correct, the resulting process $\{\tau_i\}$ will be a stationary Poisson process with intensity 1 (Papangelou, 1972). Hence the plotted points will fall on the diagonal, and by plotting the deviation from the diagonal, poor calibration becomes obvious. Similar diagnostics have previously been used for seismological models (e.g. Ogata, 1988). An example of this diagnostic will be shown in Section 4.5.2, demonstrating its use in detecting some forms of model misspecification.

### 4.4.2 Accelerated Residual Calculation

Monte Carlo integration of $\lambda(s)$ over each Voronoi cell $C_i$ is computationally expensive, and so residual maps can take considerable time to calculate. Fortunately, the integrated intensity can be calculated approximately using the same dual-tree approach described in Section 3.6. The INTENSITY function in Algorithm 3.1 is replaced with one calculating the integrated intensity $\lambda(s)$, the $k$-d tree is generated over two-dimensional space without including the time coordinate, and BOUNDS is replaced with a version calculating bounds in integrated intensities.

In practice, rejection sampling points in each Voronoi cell for the Monte Carlo integration is more computationally expensive than the actual calculation of $\lambda(s)$, so if residual calculations prove to be a burden (e.g. for interactive visualization tools), rejection sampling could be replaced with a faster method, such as triangulating the polygons and uniformly sampling within the triangles. For applications such as the residual videos discussed in the next section, sampled points could be retained from one frame to the next, instead of being redrawn each time.

### 4.4.3 Residual Videos

Residual comparisons between time periods, as shown in Figure 4.4, can be useful to understand the temporal dynamics of hotspots and self-excitation. To automate this process, I introduced residual videos, which animate residual maps over time. In each frame of the video, residuals are calculated for a specific time window, then mapped, and the window advances with each frame.

Videos pose an additional challenge because the Voronoi cells used change with every frame of the video, as the crimes contained within the time window change. Left unchecked, this would produce a nearly incoherent video, as the shapes and colors of cells change rapidly from frame to frame. Instead, the animation works on a series of time windows, one per frame. For each window, we calculate the Voronoi tessellation of crimes occurring in that window and the corresponding residuals $\hat{r}_j$. These residuals, and the times of the events defining each cell, are used to build a smoothed residual field similar to that suggested by Baddeley et al. (2005). The residual value at each animation frame and each point in space is determined by a kernel smoother, using an exponential kernel in time and a Gaussian kernel in space, with the same structure as the triggering function $g(s, t)$. This eliminates jarring changes in the map and makes it easier to interpret.

## 4.5 Robustness to Model Misspecification

Using extensive simulations and the tools discussed in this chapter, we can begin to explore the behavior of the point process model under various types of model misspecification. Several modeling choices, such as the Gaussian spatial triggering

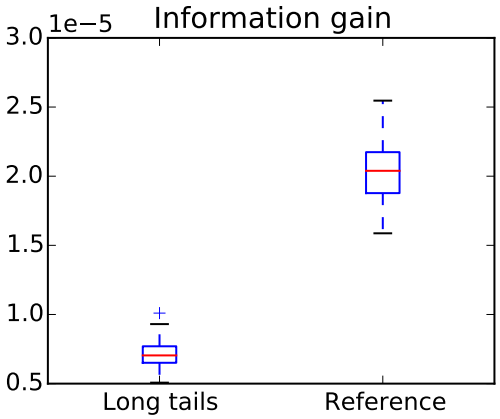


*Figure 4.6:* Boxplot of log-likelihood ratios (eq. (4.1)) obtained from fits to simulated data with Cauchy-distributed offspring (left) or Gaussian offspring (right). The poor fit from model misspecification is noticeable.

and exponential time decay in the triggering function $g$, will not exactly hold in real data, so we should verify that the model still behaves well in such circumstances.

### 4.5.1 Triggering Function

Consider two simulations: one in which event offspring locations are drawn from the Gaussian triggering function $g$ used in fitting our model, and one in which their locations are drawn from a Cauchy distribution, giving them a heavy tail which is not accounted for by the model. Running 100 simulations under each condition and calculating the log-likelihood ratios of fits to each, I obtained the ratios shown in Figure 4.6, which demonstrate this method's ability to detect poor model fit. In this situation, the disturbance in model fit is limited to the self-excitation parameters $\theta$ and $\omega$ ($\sigma^2$ is not meaningful to compare here), along with the intercept $\beta_0$; the estimates of $\beta$ for the simulated covariates are unaffected, suggesting that misspecification of the triggering function need not harm inference about the spatial covariates.

Similarly, we can test whether the exponential time decay assumption in the triggering function can be violated. The assumption is equivalent to drawing waiting times to each offspring from an $\text{Exp}(\omega)$ distribution, so as an alternative specification, I simulated data from a $\text{Gamma}(3, \omega)$ distribution, giving the alternate shape shown in Figure 4.7. The spatial offspring distribution remained Gaussian. Despite the dramatically different temporal triggering, after 100 simulations averaging 3000 events each, there was no significant bias in estimates of $\theta$ or $\beta$ (apart from the

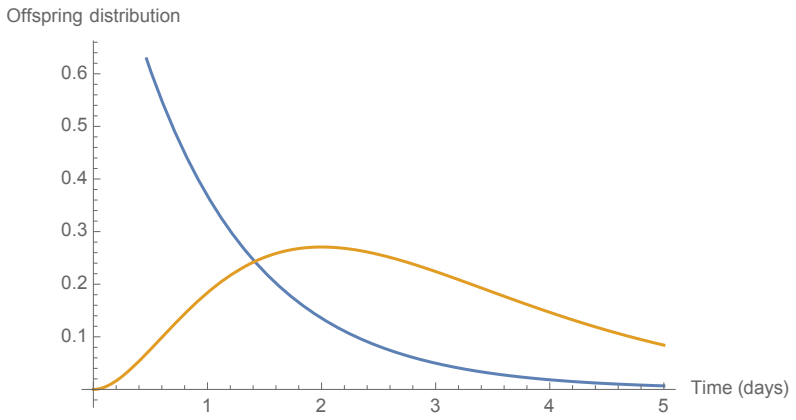


*Figure 4.7:* In blue, the exponential time decay function assumed by $g$. In orange, a Gamma-distributed decay function from which simulated data is drawn.

intercept $\beta_0$), though $\sigma^2$ was systematically underestimated; most surprisingly, $\omega$ appeared to be correctly estimated on average as well.

### 4.5.2 Omitted Variables and Confounding

Another type of misspecification concerns the covariates used in fitting the model. Section 3.1 discussed the inherent confounding which can occur when estimating the effect of spatial covariates on crime without accounting for self-excitation. Figure 3.6 demonstrated that this confounding is generic, occurring whenever there are covariates which affect crime over time. By building a self-exciting point process model which accounts for self-excitation and covariates, we can account for both and avoid the confounding.

We must, however, be aware of other types of confounding that can creep in. The most common is an unobserved covariate: there are many causal factors which can influence crime rates, and it is unlikely we can directly measure all of them. Figure 4.8 demonstrates the danger. A covariate may be causally related to another covariate as well as to crime rates, and if it is not observed and accounted for, the other covariate's estimated effect will be confounded. This is directly analogous to the situation in ordinary regression, when unobserved predictors may confound regression coefficient estimates.

On the other hand, if the two covariates are *not* correlated in any way, omitting one does not bias estimates of the other's effect; in traditional regression its mean effect is simply added to the intercept and the individual effects simply add to the error variance. However, in the more complicated self-exciting point process

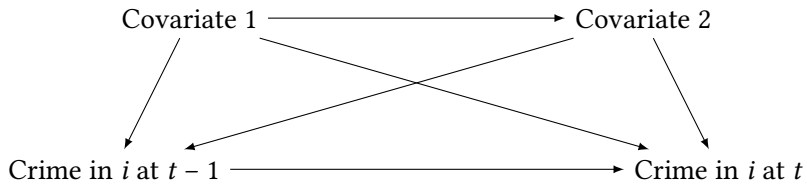


*Figure 4.8:* A simplified causal diagram depicting potential confounding: covariate 1 has a causal relationship with both covariate 2 and crime rates, and so if it is unobserved, estimates of covariate 2's effect will be confounded.

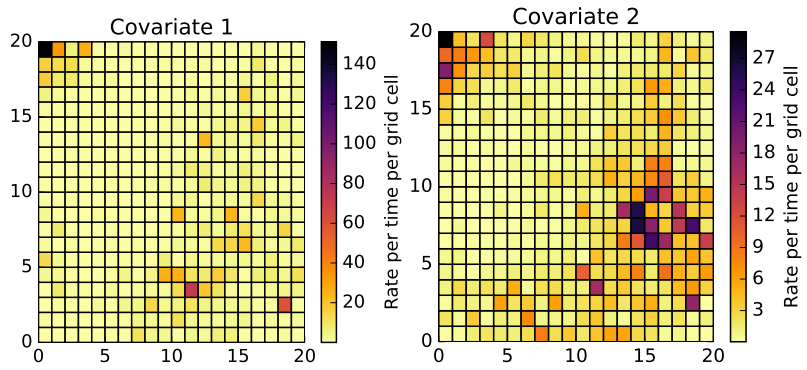


*Figure 4.9:* The rate induced (that is, $\exp(\beta X)$, where $\beta = 1$ for simplicity and $X$ is the covariate) by two Gaussian process covariates on a $20 \times 20$ grid. The second covariate is dependent upon the first. Notice the spatial structure of the Gaussian process.

model, omitted covariates may have other detrimental effects. A series of simulations demonstrate this.

To simulate two uncorrelated covariates, I generated covariates on a grid, drawing the covariate values from a Gaussian process with squared exponential covariance function to ensure there was some spatial structure. Each covariate was an independent draw from the Gaussian process. To simulate confounded covariates, I drew the first covariate from the Gaussian process and defined the second to be the sum of the first and a new Gaussian process draw, so that the second was correlated with the first. Both covariates had effects on the crime rate. Sample correlated covariates are shown in Figure 4.9.

With uncorrelated covariates, I ran 100 simulations (each with new Gaussian process draws), fitting models with both covariates included and with the second covariate omitted. Simulations were performed with random true parameter values,

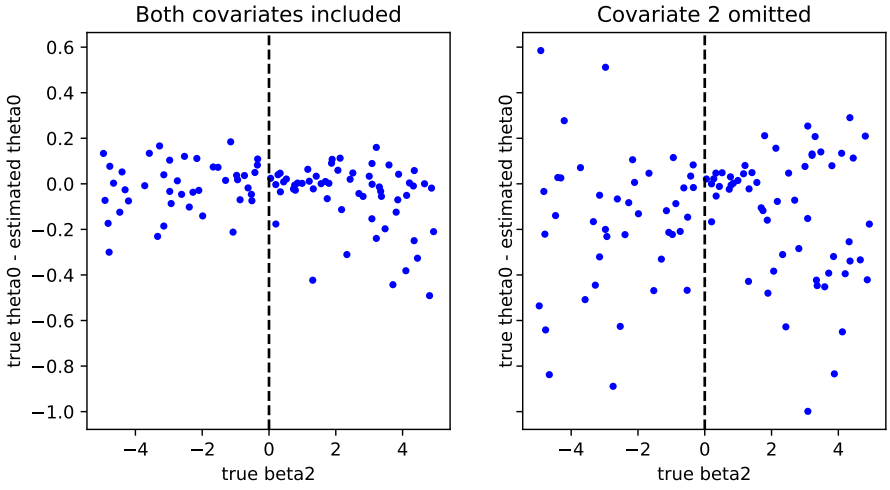


*Figure 4.10:* The difference between the true value of $\theta$ and the estimated value, as a function of the coefficient $\beta_2$. On the left, fits made when $\beta_2$ is accounted for; on the right, when it is not. Notice the odd behavior around $\beta_2 = 0$: when the omitted covariate does not matter, $\theta$ is estimated to be close to its true value, but when it has a larger effect, $\hat{\theta}$ has much higher variance.

and these values were recorded, along with the fits. It is apparent from the results that estimates of $\hat{\theta}$ are affected by the missing covariate: Figure 4.10 shows the fits, as a function of the true value of $\beta_2$ used in the simulation, and a distinct pattern can be seen when the second covariate is omitted from the fit, with $\hat{\theta}$ having larger variance for larger values of $|\beta_2|$. On average, the estimated $\hat{\theta}$ with a missing covariate is larger than the true $\theta$ by 0.18 (95% CI $[0.10, 0.27]$).

Overestimation of $\theta$ has other consequences. For example, Figure 4.11 shows a temporal residual plot (see Section 4.4.1) for a fit to a simulated dataset with an omitted covariate. An obvious calibration problem is present: by the time the 500th event occurred, the conditional intensity function predicted 150 fewer events than occurred. Near $t = 0$, $\lambda(s, t)$ cannot predict the observed events because there is little past history of events; near $t = T$, a long past history and overestimated $\theta$ causes $\lambda(s, t)$ to overestimate the intensity and "catch up" in the cumulative predicted number of events.

Additionally, the time decay parameter $\omega$ is also overestimated by 70% on average. Together, these biases suggest that the clustering induced by the unobserved covariate is being accounted for by increasing self-excitation and by allowing the effects of self-excitation to last longer in the model.

Next, I simulated causally confounded covariates, following the causal model in

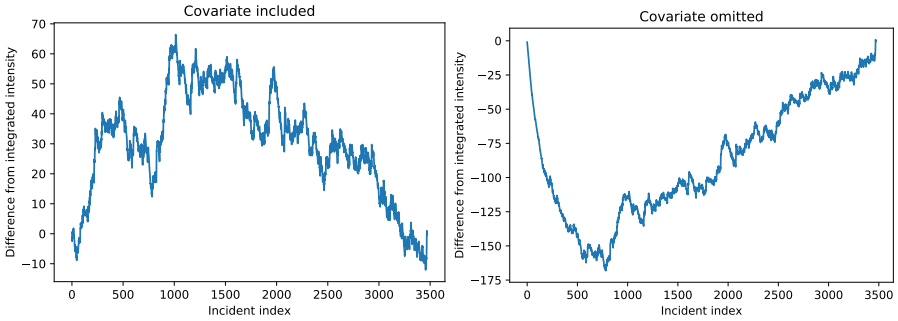


*Figure 4.11:* At left, a temporal residual plot for a fit to a simulated dataset with one covariate, showing normal variation in the residuals. At right, a temporal residual plot for a fit to the same data which omits the covariate, demonstrating the effect of the overestimated $\theta$. Note the difference between the maximum deviations from 0 in both plots.

Figure 4.8. Covariate 1 was drawn from a Gaussian process, as before, and Covariate 2 was defined to be the average of Covariate 1 and a separate independent Gaussian process. This gave an average correlation of $r = 0.66$ between the covariates. Data was simulated from these covariates (with random coefficients) and then models fit with and without Covariate 2 included. Figure 4.12 demonstrates the bias in estimates of $\beta_1$ which ensues when the effect of $\beta_2$ is not accounted for, similar to the biases that can occur in ordinary linear regression when covariates are confounded. The confounding also affects $\hat{\theta}$ and $\hat{\omega}$ in a similar way as in the previous simulation, with bias as $|\beta_2|$ increases.

Together, these simulations demonstrate two important caveats of self-exciting point process models:

1. Omitted spatial covariates, whether or not they are confounded with observed covariates, can bias estimates of the self-excitation parameter $\theta$, making it seem as though events are more likely to trigger offspring events.

2. Omitted spatial covariates can also bias estimates of the temporal decay parameter $\omega$, making it seem as though self-excitation or near-repeat effects occur over a longer timescale than they really do.

3. If there is a confounding relationship between covariates, such as that shown in Figure 4.8, unobserved covariates can bias estimates of observed covariate effects ($\hat{\beta}$) as well as of self-excitation.

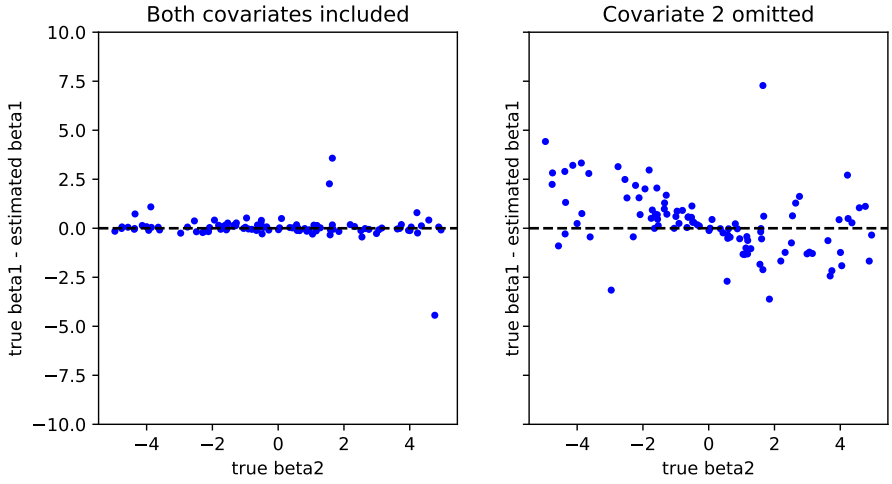


*Figure 4.12:* Bias observed in estimated values of $\beta_1$ when $\beta_2$ is also estimated (left) or is omitted from the fit (right).

The first two points are particularly concerning, since in practical applications it is unlikely that all covariates could ever be accounted for—there will always be unmeasured spatial differences in base rates, or imperfectly measured covariates. This suggests that previous applications of self-exciting point process models may have overestimated the amount and time scale of self-excitation in the process, unless their background estimator was able to capture all spatial variation in base rates.

In some cases, it may be possible to detect when there is an important unobserved spatial covariate. Residual maps, introduced in Section 4.4.1, can make systematic deviations from the predicted event rate visible, and careful examination of the maps may suggest variables that need to be included. Chapter 5 gives several examples of this in Pittsburgh crime data.

General approaches to account for unobserved covariates are more difficult. One strategy, sometimes used in spatial regressions, is to include a spatial random effect term intended to account for the unobserved covariates. However, at least in spatial regression, this method does not achieve its goal: a spatial random effect can bias coefficients of the observed covariates in arbitrary ways, particularly if the unobserved covariate is spatially correlated with any of the observed covariates (Hodges & Reich, 2010). Given the causal diagram in Figure 3.6, it does not seem possible for any one adjustment to account for an unobserved covariate and give unbiased estimates of the effects of the other covariates. Users of spatial regression and the self-exciting point process model introduced here need to be aware of their limitations in the presence of unobserved confounders, and interpret results

carefully.

## 4.6 Summary

This chapter introduced a range of diagnostic and inference tools for self-exciting point processes, building on inference and residual methods previously used in other areas of application. Using these tools, Section 4.5 presented a comprehensive simulation study of the effects model misspecification. These results, and the diagnostic tools used to obtain them, can now be put to use analyzing real-world crime data.

# *Five*

# Application to Pittsburgh Crime Data

## 5.1 Pittsburgh Data

This chapter analyzes a database of 205,485 police incident records filed by the Pittsburgh Bureau of Police (PBP) between June 1, 2011, and June 1, 2016, specifying the time and type of each incident and the city block on which it occurred.[1] (Privacy regulations prevent PBP from releasing the exact addresses or coordinates of crimes, so PBP provides only the coordinates of the block containing the address.) The records include crimes from very minor incidents (such as 38 violations of Pittsburgh's ordinance against spitting) to violent crimes, such as homicides and assaults. Only crimes reported to PBP are included, so the dataset does not include records from the police departments of Pittsburgh's several major universities, such as the University of Pittsburgh, Carnegie Mellon University, Chatham University, or Carlow University.

Because the database contains only incident reports, offense types are preliminary. Charges listed in the reports may be downgraded or dropped, suspects acquitted, or new charges filed. The reports represent only the charges reported by the initial investigating officers, so they may not correspond with final FBI Uniform Crime Report data or other sources. While this limits the accuracy of our data, it is also the only practical approach—final charges may not be known for months, so predictions based on them would be hopelessly out of date.

Rather than dealing with the numerous sections and subsections of the Pennsylvania Criminal Code represented in the incident data, we used the FBI Uniform Crime Report hierarchy, which splits incident types into a common hierarchy comparable across states and jurisdictions. Among so-called "part I" crimes, homicide, assault, and rape are at the top of the hierarchy, followed by other crimes like theft,

[1] Portions of this chapter have been published as Reinhart and Greenhouse (2018). Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C*. doi:10.1111/rssc.12277

| Hierarchy | Crime | Count |
|---|---|---|
| 1 | Homicide | 300 |
| 2 | Forcible rape | 893 |
| 3 | Robbery | 5884 |
| 4 | Aggravated assault | 5900 |
| 5 | Burglary | 11 943 |
| 6 | Larceny/theft | 37 487 |
| 7 | Motor vehicle theft | 3892 |
| 8 | Arson | 0 |

*Table 5.1:* The part I crime hierarchy prescribed by the FBI Uniform Crime Report system, along with counts of each type of offense in the Pittsburgh dataset. (Arson appears to have been miscoded in this dataset, making it falsely appear to contain no arson incidents.)

burglary, and so on. If an incident involves two distinct types of crime (e.g. a burglary involving an assault on a homeowner), we use the type higher in the hierarchy, following the FBI's "Hierarchy Rule" (FBI, 2004). The hierarchy of offenses is shown in Table 5.1. In our analysis we focused on crimes in these categories, though other "part II" crimes, such as simple assault and vandalism, are also available in the dataset, along with every other offense type recorded by the Pittsburgh Bureau of Police.

To supplement the crime data, PBP also provided 1,027,056 records from its Computer Aided Dispatch (CAD) system, which records both 911 calls and other officer-initiated incidents. (For example, an officer may call in to dispatch to record a "police park & walk" when parking to patrol on foot.) From this data we extracted broad groupings of calls related to assaults, gunshots, drug incidents, and other types considered relevant to violent crime by previous leading indicator studies.

All analyses are performed in the Pennsylvania South State Plane Coordinate System (SRID 2272), with coordinates in feet.

## 5.2 Spatial Covariates

With the assistance of Evan Liebowitz, I obtained shapefiles of geographic covariates for the city of Pittsburgh containing, for each U.S. census block,

- The fraction of residents who are male from age 18–24
- The fraction of residents who are black

- The fraction of homes which are occupied by their owners, rather than rented
- The total population
- Population density (per square meter)
- The fraction of residents who are black or Hispanic.

Some city blocks have no population (e.g. in commercial areas with no residents), so an additional dummy variable was used to record whether each block had a population. In all models that follow, population-based covariates only enter the models when the block has a nonzero population.

Additional variables were obtained from the American Community Survey, Pittsburgh land parcel records, and other public Pittsburgh GIS data, and were recorded at the census block group level:

- The fraction of residents without a high school diploma
- The fraction of residents living under the poverty line
- Number of bus stops in the block group
- Numbers of bars, banks, and retail stores, from land parcel data.

The census block variables were also aggregated to the block group level, so they are available when analyzing with block group covariates. The analysis software does not yet have the ability to mix covariates recorded on different shapefiles, though this is technically possible with some extra harmonization work.

The business data is least reliable, since it is based on land parcel records and reflects only the owner of the land, not any lessees or secondary uses. Office buildings with shops on the first floor may not be recorded as containing retail stores, for example. We did not use this data for our primary analysis, focusing instead on the demographic and socioeconomic variables.

## 5.3 Dealing with Aggregated Data

The Pittsburgh dataset does not contain exact locations of every incident. Instead, each incident's location is the center of the city block containing the incident. This aggregation causes some problems. As discussed in Section 3.3, crimes closer than a short distance $\delta$ are not permitted to contribute to the intensity, implying that if two crimes occur at exactly the same location, one cannot have "caused" the other through self-excitation in the foreground process. Hence crimes which occur within the same block cannot have caused each other.

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $5.199 \times 10^6$ | 60.18 | d |
| $\sigma^2$ | $5.359 \times 10^5$ | 732.1 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −31.95 | $1.3 \times 10^{-14}$ |
| Population | 31.77 | $6.2 \times 10^{13}$ |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 2682 | 0.8913 |

*Table 5.2:* Predicting part I violent crimes using only self-excitation and background effects, with no jitter, from June 1, 2011 to June 1, 2012.

This limits the utility of the self-exciting component of the model, since it cannot account for self-excitation within the same city block, where self-excitation could be expected to have the strongest effect. One possible solution is *jitter*, which adds independent random noise to the location of each crime. However, the most obvious form of jitter, simply adding independent bivariate random normal numbers with mean zero to each crime's coordinates, proves to introduce artifacts to the model fit.

In particular, jittering aggregated data leads to clusters: if we use a standard deviation of ten feet, we will create small clusters which are well-fit with a model with $\hat{\sigma}^2 \approx 10\,\text{ft}$. To illustrate this, I used part I violent crime data (hierarchy levels 1–4); Table 5.2 shows the fit without jitter, and Table 5.3 shows the fit with ten-foot normal jitter, both using population density as a background covariate. There are several crucial differences. The original fit has a log-likelihood of −83377.1, compared to −78459.1 for the jittered fit, an improvement of 4818. The self-excitation effect goes from a coefficient of 0.8913 to 0.9492. But the self-excitation bandwidth is $\hat{\sigma} = 12.4\,\text{ft}$, a sign the self-excitation is fitting to the clusters I artificially introduced.

Instead, I adopted a jittering approach that does not introduce artificial clustering, and is more "honest" about the accuracy of our aggregated data. Using a shapefile containing the boundaries of every city block in Pittsburgh, I independently and uniformly drew the location of each crime from the block containing it. (To find the block, I used R-trees (Guttman, 1984), an efficient data structure for searching spatial polygon data.) Because the draw is uniform, this does not cause clustering; because it is within the block to which the crime was aggregated, it uses exactly the precision available to us and no more.

A block-jittered fit to the same data is shown in Table 5.4. $\hat{\sigma}^2$ is no longer artifi-

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $2.286 \times 10^7$ | 264.6 | d |
| $\sigma^2$ | 153.7 | 12.4 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −31.18 | $2.9 \times 10^{-14}$ |
| Population | 31.83 | $6.7 \times 10^{13}$ |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 2682 | 0.9492 |

*Table 5.3:* Predicting part I violent crimes using only self-excitation and background effects, with ten feet of jitter, from June 1, 2011 to June 1, 2012.

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $7.788 \times 10^6$ | 90.14 | d |
| $\sigma^2$ | $1.471 \times 10^5$ | 383.5 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −31.93 | $1.4 \times 10^{-14}$ |
| Population | 31.60 | $5.3 \times 10^{13}$ |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 2682 | 0.9771 |

*Table 5.4:* Predicting part I violent crimes using only self-excitation and background effects, jittered within city blocks, from June 1, 2011 to June 1, 2012.

cially fitting to clusters introduced by jittering, and self-excitation is stronger than the un-jittered model in Table 5.2, showing the effect of allowing crimes within the same block to excite each other.

Repeated fits with jitter (i.e. starting from the original unaltered data, jittering with new random numbers, and refitting) shows that the jitter has only a small effect on parameter values—less than 1% in most cases.

## 5.4 Predicting Burglary

I began analysis of the Pittsburgh crime data with burglary. Selecting only the first year of data, I fit two models, one using only population density as a covariate

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $4.511 \times 10^6$ | 52.21 | d |
| $\sigma^2$ | $2.664 \times 10^5$ | 516.1 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −31.63 | $1.8 \times 10^{-14}$ |
| Population | 31.66 | $5.6 \times 10^{13}$ |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 2892 | 0.7640 |

*Table 5.5:* A model predicting burglary using self-excitation and population density (persons per square meter).

and the other using additional covariates. The model fits are shown in Table 5.5 and Table 5.6. The additional covariates improve the model AIC from 179 750 to 179 319, an improvement of about 431 units. Notice the relative consistency of the self-excitation parameters $\hat{\omega}$ and $\hat{\sigma}^2$ between fits, and that, as expected from the discussion in Section 4.5.2, $\hat{\theta}$ decreases when additional covariates are added.

Interpretation of the model with all covariates (Table 5.6) is straightforward. High population densities predict higher risks of burglary, as there are more residences to burgle; higher proportions of young men also indicate a higher risk, in agreement with previous criminological research suggesting this is the demographic most likely to commit crime. Home ownership, rather than renting, correlates with decreases in burglary risk, while a higher fraction of black residents is correlated with higher burglary rates; these last two factors are likely confounded with measures of poverty and unemployment, which also likely have strong relationships with crime, but are not included in this model.

Predictive evaluations of these models were used as examples in Section 4.2.4, in Figure 4.2 and Figure 4.3, showing the small but detectable improvement in predictive performance coming from the use of additional covariates. Residual maps were used as examples in Figure 4.4, illustrating the nature of burglary hotspots. For a larger view of Pittsburgh, Figure 5.1 shows an overall residual map of Pittsburgh over two months. Several trends appear, suggesting inadequacies in the available covariates and the presence of boundary effects: commercial areas such as downtown (at the confluence of the two rivers) have fewer burglaries than predicted, and the presence of the University of Pittsburgh and Carnegie Mellon University also results in negative residuals, as each has its own police department whose records are not included in our dataset. Note that, as discussed in Section 4.4.1, negative

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $4.061 \times 10^6$ | 47 | d |
| $\sigma^2$ | $2.194 \times 10^5$ | 468.4 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −33.15 | $4 \times 10^{-15}$ |
| Population | 25.50 | $1.2 \times 10^{11}$ |
| is_positive(TotalPopul) | 2.49 | 12 |
| is_positive(TotalPopul):PercentMal | −0.69 | 0.5 |
| is_positive(TotalPopul):PercentBla | 0.75 | 2.1 |
| is_positive(TotalPopul):PercentOwn | −1.14 | 0.32 |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 2892 | 0.5893 |

*Table 5.6:* A model predicting burglary using self excitation and multiple covariates: population density, fraction of residents who are 18–24 year old males (PercentMal), fraction of residents who are black (PercentBla), and fraction of homes occupied by their owners (PercentOwn).

(blue) residuals visually dominate, because areas with lower-than-expected crime hence have larger Voronoi cells; also note the presence of several clusters of small cells with large positive residuals, at the locations of temporary burglary hotspots.

Further exploration of the ROC and hit rate curves (Figure 4.2 and Figure 4.3) is illuminating. Despite the 221 unit improvement in AIC from the addition of demographic covariates, the predictive performance gain is fairly small. To explore the reasons for this, I performed two simulations. In the first, I used the model parameters in Table 5.6, along with the Pittsburgh demographic covariates, to generate a synthetic burglary dataset, fit to that dataset with and without the covariates, and compared the ROC curves on simulated data; the curves were essentially identical. On the other hand, if I fit a model in which I set $\theta = 0$, so there was no self-excitation and only the covariates mattered (essentially a spatial regression), I obtained Figure 5.2.

This suggests that, at least in this context, self-excitation accounts for much more of the predictive power of the model than the covariates; $\hat{\theta} = 0.5893$, which leads to an expected total cluster size of 2.43, so there are almost twice as many offspring crimes as there are crimes arising from the background. To confirm this intuition, I simulated another burglary dataset using the parameters in Table 5.6, except I artificially set $\theta = 0.1$, giving an expected total cluster size of 1.11 and ensuring

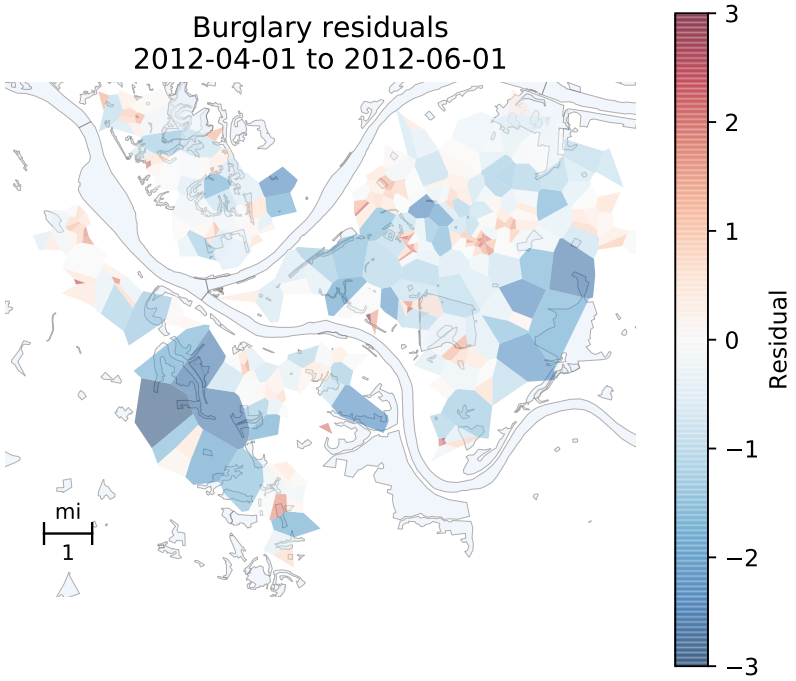


*Figure 5.1:* Residual map from the fit shown in Table 5.5, over two months of burglaries.

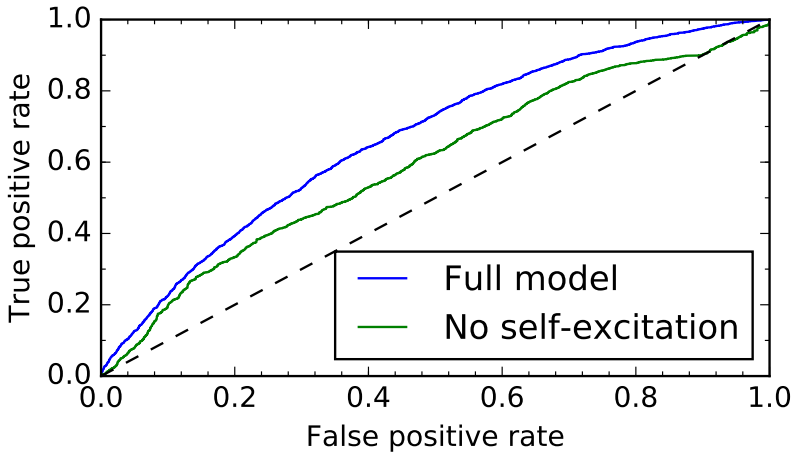


*Figure 5.2:* ROC curves for a full model with all covariates and one with no self-excitation, on simulated burglary data.

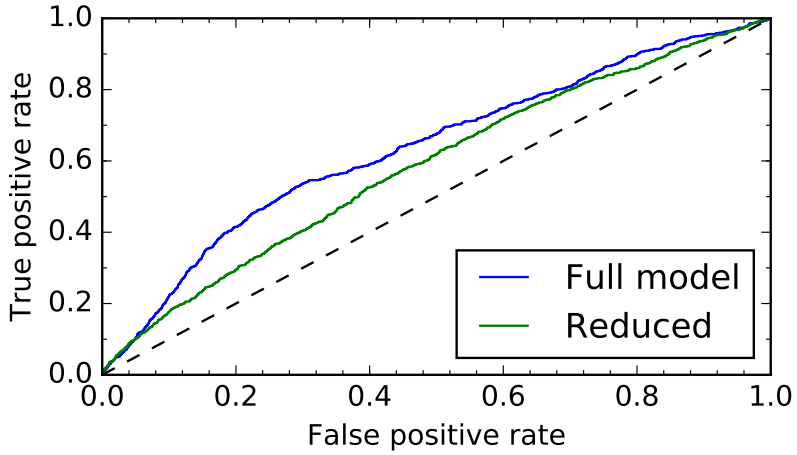


*Figure 5.3:* ROC curves for a full model with all covariates and one with no self excitation, on simulated burglary data with $\theta = 0.1$ set artificially.

most crimes arose from the background directly. In this simulation, omitting covariates caused a large reduction in predictive performance, as shown in Figure 5.3.

Interpreting these results requires care. At first glance it appears the covariates have little effect on crime rates; however, the estimated background crime rate $\mu(s)$ varies by four orders of magnitude between city blocks, so it is not the case that covariates do not matter at all. The large value of $\hat{\theta}$ instead suggests that either burglary is naturally very highly self-excited, giving it a stronger predictive effect than covariates, or that there are additional relevant covariates that have not been accounted for—Section 4.5.2 showed that omitting relevant factors can significantly increase $\hat{\theta}$, as locally high event rates are accounted for with self-excitation instead of background features. We must be careful interpreting $\hat{\theta}$ to mean, for example, that burglars commit 3.05 burglaries on average before being caught or moving elsewhere, because that number is biased upward by unknown factors. We can, at best, consider it an upper bound on the average burglary cluster size.

Finally, we can look at the addition of leading indicators. Using the same covariates as in Table 5.6, we add in larceny/theft and motor vehicle theft as possible leading indicators, and obtain the fit shown in Table 5.7. The covariate coefficients change slightly, and motor vehicle theft seems to predict burglaries better than larceny/theft. The AIC of the fit is 179 201, an improvement of a further 118 units—a smaller improvement than the addition of the original covariates, but nonetheless substantial. As could be expected, the self-excitation decreases again to roughly 0.5, as motor vehicle theft and larceny account for some previously unaccounted-for

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $3.551 \times 10^6$ | 41.1 | d |
| $\sigma^2$ | $1.619 \times 10^5$ | 402.3 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −33.90 | $1.9 \times 10^{-15}$ |
| Population | 25.19 | $8.7 \times 10^{10}$ |
| is_positive(TotalPopul) | 3.00 | 20 |
| is_positive(TotalPopul):PercentMal | −0.85 | 0.43 |
| is_positive(TotalPopul):PercentBla | 0.94 | 2.5 |
| is_positive(TotalPopul):PercentOwn | −1.00 | 0.37 |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 2892 | 0.4480 |
| Larceny/theft | 7382 | 0.0632 |
| Motor vehicle theft | 824 | 0.1167 |

*Table 5.7:* A fit to 1 year of burglary data using other types of property crime as leading indicators.

clustering.

## 5.5 Predicting Violent Crime

Violent crime is frequently the target of police attention and interventions, and so models to direct police resources to lower violent crime rates are of high interests. From the Pittsburgh dataset I selected hierarchy levels 1–4: homicide, forcible rape, robbery, and aggravated assault, totaling 12 975 offenses over five years.[2]

As an initial exploration, I fit a model using all five years of data and covariates on the census block group level, including indicators of poverty and education. The result is shown in Table 5.8. Population density continues to have a strong positive relationship with violent crime, as do poverty and low education levels; surprisingly, the fraction of residents who are males aged 18–24 has a *negative* correlation with crime, the opposite of what we would usually expect. This may occur because the parts of Pittsburgh with the highest concentrations of young men are near the campuses of the University of Pittsburgh and Carnegie Mellon University, where

[2] Table 5.1 lists 12 977 such offenses, but two had invalid or missing coordinates and could not be used.

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $2.444 \times 10^7$ | 282.9 | d |
| $\sigma^2$ | $5.186 \times 10^4$ | 227.7 | ft |
| **Covariate** | **Coefficient** | **exp(Coef)** | |
| Intercept | −31.92 | $1.4 \times 10^{-14}$ | |
| Population | 213.21 | $4 \times 10^{92}$ | |
| BlockGr_13 | 1.64 | 5.2 | |
| BlockGr_12 | 0.20 | 1.2 | |
| PercentMal | −5.54 | 0.0039 | |
| PercentOwn | −1.17 | 0.31 | |
| **Predictor** | **N** | **Foreground** | |
| Self-excitation | 12 960 | 0.8658 | |

*Table 5.8:* A fit to five years of part I violent crime data. BlockGr_13 is the fraction of population under the poverty line; BlockGr_12 is the fraction without a high school diploma. As before, PercentMal is the fraction of residents who are males age 18–24 and PercentOwn the fraction of homes occupied by their owners.

education levels are high and crimes are reported to campus police departments and do not appear in our dataset.

The self-excitation in this model is roughly consistent with previous research (Haberman & Ratcliffe, 2012; Ratcliffe & Rengert, 2008), which has found near-repeat effects over a scale of about one city block, or around 400 feet. The decay time is nearly ten months, however, which is much longer than most previous studies have found—the usual period used in Knox tests is about two weeks. Further investigation is needed to understand why there is such a dramatic difference in time periods.

If we fit to the same dataset without the socioeconomic and demographic covariates, using only population density as a covariate, we obtain broadly similar self-excitation parameters, as shown in Table 5.9. The self-excitation rate $\hat{\theta}$ is higher, as we would expect from Section 4.5.2, but the spatial and temporal decays from self-excitation are nearly the same, and the direction of the effect of population density is similar as well. The decline in AIC from removing the socioeconomic and demographic variables is 305, suggesting they are indeed quite important to the model fit.

Together, these two examples illustrate the use of the extended self-exciting model of crime for crime analysis, and demonstrate its potential for quantifying the

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $2.333 \times 10^7$ | 270.1 | d |
| $\sigma^2$ | $6.045 \times 10^4$ | 245.9 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −32.62 | $6.8 \times 10^{-15}$ |
| Population | 186.45 | $9.5 \times 10^{80}$ |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 12 960 | 0.9214 |

*Table 5.9:* A fit to the same data as Table 5.8, but without most of the spatial covariates.

effects of leading indicators and spatial covariates. The analysis here is not complete, and further criminological analysis with many more interesting covariates is possible, though out of the scope of this thesis. Instead, in the next chapter we will demonstrate the model's applicability to cities other than Pittsburgh, by analyzing crime in Baltimore.

# *Six*

# Application to Baltimore Crime Data

## 6.1 Baltimore Data

This chapter, intended to provide a point of comparison against Chapter 5's analysis of Pittsburgh crime data, analyzes a database of crime records released by the Baltimore Police Department covering Part 1 crime (offense types listed in Table 5.1). The dataset is publicly available through the city's Open Baltimore service and frequently updated with the latest crime data; I extracted two subsets of this data, one of Part 1 Violent crime (hierarchy levels 1–4) and one of burglary (with larceny/theft and motor vehicle theft as leading indicators), both between July 1, 2015 and July 1, 2017.

Similar caveats apply to this dataset: it represents preliminary reports, before the data is validated and submitted to the final FBI Uniform Crime Report, and information may change as incidents are investigated. Only Baltimore Police Department data is included, not any other police agencies which may have jurisdiction in the area. Also, the data released by the Baltimore Police Department is geocoded to the nearest city block coordinates, and so, following Section 5.3, I used Census block boundaries to uniformly jitter the events within city blocks. All analyses were conducted in the Maryland State Plane coordinate system (SRID 3582), in feet.

Spatial covariates were obtained from the Vital Signs 15 dataset produced by the Baltimore Neighborhood Indicators Alliance, which aggregates city and census data about each of Baltimore's 55 neighborhoods. A large range of variables are available; from these, I used the following (descriptions quoted from BNIA data pages):

- Household density (number of households per square mile), based on the total number of households variable
- Percent of residents aged 18–24
- Percent of family households living below the poverty line

- Percent of the population unemployed
- Percent of the population age 25+ with less than a high school diploma/GED
- High school dropout/withdrawal rate
- Percent of 9th–12th graders who are chronically absent

Because Baltimore is split into only 55 neighborhoods, these covariates are rather coarse-grained, and finer variation in population density, poverty, and other demographics cannot be captured by this covariate data. This is an important difference versus the Pittsburgh data, which is recorded at the block level but has less rich covariates.

## 6.2 Predicting Burglary

I began analysis of the Baltimore data with burglary, following the same steps as with the Pittsburgh data. Selecting the first year of data (July 1, 2015 to July 1, 2016), I fit a model containing the covariates listed above. The resulting fit is shown in Table 6.1. The coefficients on the covariates appear to be small, but the scaling matters here; a 10 percentage point change in the proportion of residents age 18–24, for example, correlates with a 1.76× increase in background crime rate, for example.

Comparing against Table 5.7, self-excitation is higher, perhaps because this fit did not use additional leading indicators, but spatial and temporal decays are similar. Unfortunately it's difficult to evaluate the consistency in covariate coefficients between cities because a consistent set of covariates is not available across both; further work would be required to extract Census data in identical ways for both cities to enable a meaningful comparison.

Residuals of this fit for one two-week period are shown in Figure 6.1. Several serious burglary hotspots are visible, as well as a range of residual values that appear to be related to the hotspots. To evaluate the predictive performance of the model, it was tested on weekly predictions for the following year of data (July 1, 2016 to July 1, 2017) in the same way as in Figure 4.2 and Figure 4.3 for Pittsburgh data, giving the curves shown in Figure 6.2 and Figure 6.3. The curves are very similar to those for Pittsburgh data, with the ROC curve's AUC of 0.68 nearly matching the AUC of 0.70 for Pittsburgh, suggesting the predictive performance is similar between cities (though again, the difference in covariates makes direct comparison impossible).

## 6.3 Predicting Violent Crime

Again following the same steps as with the Pittsburgh data, I analyzed Part 1 violent crime data (hierarchy levels 1–4) from July 1, 2015 to July 1, 2016, a total of 17 973

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $4.242 \times 10^6$ | 49.1 | d |
| $\sigma^2$ | $1.969 \times 10^5$ | 443.7 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −33.01 | $4.6 \times 10^{-15}$ |
| I(hhs15 / (area / 5280 / 5280)) | 0.00 | 1 |
| age18_15 | 0.06 | 1.1 |
| hhpov15 | −0.01 | 0.99 |
| unempr15 | 0.01 | 1 |
| lesshs15 | 0.03 | 1 |
| drop15 | −0.02 | 0.98 |
| abshs15 | −0.01 | 0.99 |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 7565 | 0.7424 |

*Table 6.1:* Fit to one year of Baltimore burglary data with covariates and leading indicators. The covariates, in turn, are households per square mile, percent age 18–24, percent of households under the poverty line, percent unemployed, percent with less than a high school education, high school dropout rate, and percent of 9th–12th graders who are chronically absent.

incidents. For the sake of comparison, I produced two separate fits, one with the full set of covariates used in Table 6.1 and one using only household density and self-excitation.

The fits are shown in Table 6.2 and Table 6.3. The self-excitation rate is much higher than for burglary, and as we'd expect from Section 4.5.2, it is higher without covariates. AIC improves from 1 056 176 to 1 055 457, or about 718 units, with the addition of the covariates, suggesting they are quite important despite their apparently small magnitude. (Note again that, as the covariates are percentages, they vary over a large range.) The time scale of self-excitation ($\hat{\omega}$) is longer than many previous results on near-repeats would suggest, though not as extreme as the Pittsburgh results in Section 5.5; this suggests there is more to the near-repeat phenomenon than has been discovered so far.

It is particularly interesting that self-excitation parameter $\theta$ varies in magnitude so greatly between burglary and violent crime, since previous evaluations based on Knox tests have not had any easily comparable measure of effect size, only the significance of the test. The consistency of this difference between Pittsburgh and

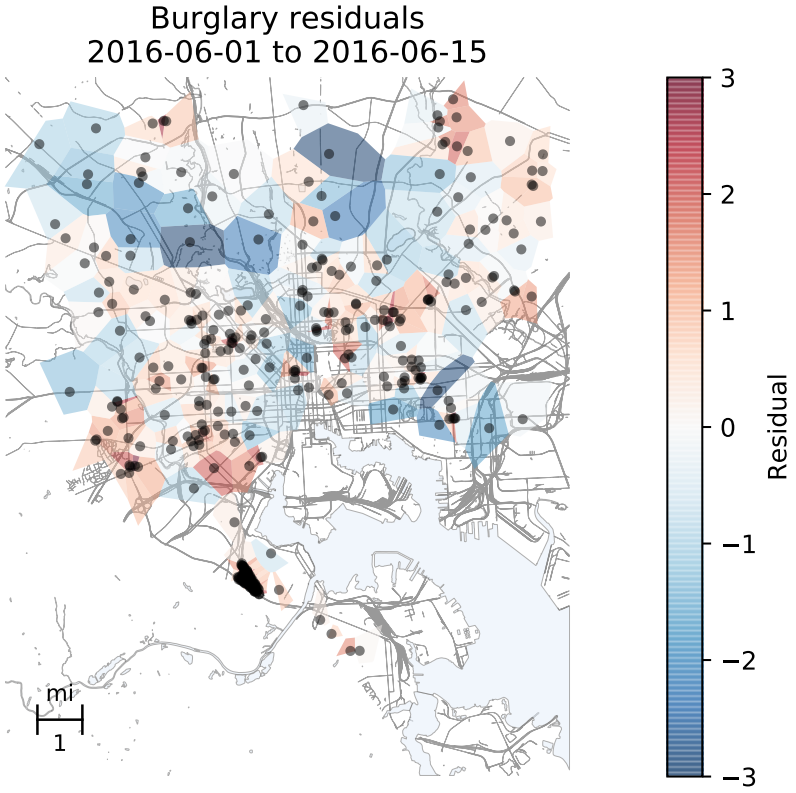


*Figure 6.1:* Residual map for the fit shown in Table 6.1. Several clusters are visible, one to the southwest of downtown and several directly around it.

| Parameter | Value | | Interpretation |
|---|---|---|---|
| $\omega$ | $6.61 \times 10^6$ | 76.51 | d |
| $\sigma^2$ | $7.756 \times 10^4$ | 278.5 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −31.40 | $2.3 \times 10^{-14}$ |
| I(hhs15 / (area / 5280 / 5280)) | 0.00 | 1 |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 17 973 | 0.9452 |

*Table 6.2:* Fit to Baltimore violent crime data without covariates, apart from households per square mile.

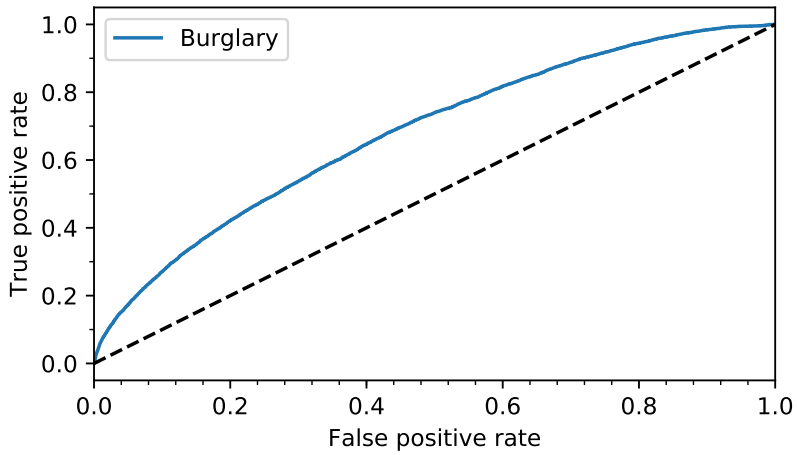


*Figure 6.2:* ROC curve for weekly predictions of burglaries in Baltimore. Compare against Figure 4.2. The AUC is 0.68.

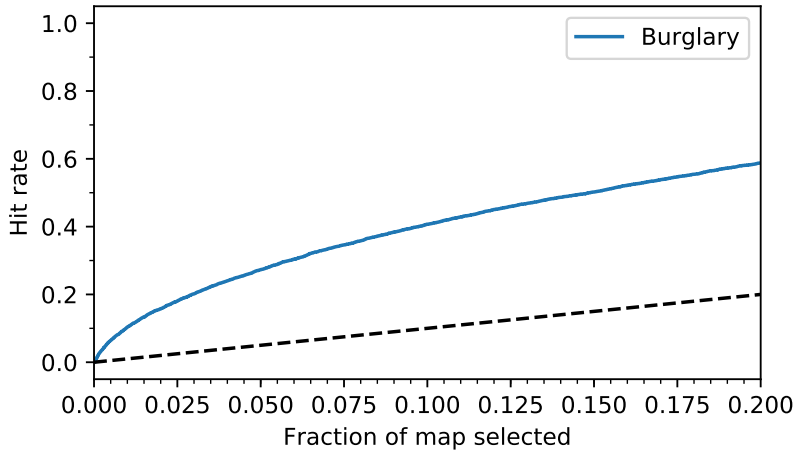


*Figure 6.3:* Hit rate for weekly predictions of burglaries in Baltimore. Compare against Figure 4.3.

| Parameter | Value | Interpretation | |
|---|---|---|---|
| $\omega$ | $7.888 \times 10^6$ | 91.3 | d |
| $\sigma^2$ | $5.859 \times 10^4$ | 242 | ft |

| Covariate | Coefficient | exp(Coef) |
|---|---|---|
| Intercept | −33.66 | $2.4 \times 10^{-15}$ |
| I(hhs15 / (area / 5280 / 5280)) | 0.00 | 1 |
| age18_15 | 0.02 | 1 |
| hhpov15 | −0.01 | 0.99 |
| unempr15 | 0.03 | 1 |
| lesshs15 | 0.00 | 1 |
| drop15 | −0.10 | 0.9 |
| abshs15 | 0.06 | 1.1 |

| Predictor | N | Foreground |
|---|---|---|
| Self-excitation | 17 973 | 0.8799 |

*Table 6.3:* Fit to Baltimore violent crime data with all covariates.

Baltimore suggests an underlying phenomenon is at work; the difference could be explained by different crime dynamics for different types of crime, as well as differences in the efficacy of the covariates—perhaps the covariates available to us are much better at predicting burglary than at predicting violent crime, and the addition of other as-yet-unknown covariates would result in the estimated self-excitation for violent crime dropping to a similar level. Further criminological research is required.

## 6.4 Summary

Analysis of Pittsburgh and Baltimore crime data has demonstrated the practical use of the model introduced in Chapter 3 and illustrated the diagnostics discussed in Chapter 4. Though the models are intended more for demonstration than as rigorous tests of criminological theories, the preliminary results are already interesting, suggesting differences between the dynamics of different types of crime and giving better estimates of near-repeat behavior than were previously possible. Naturally, the results suggest a great deal of future criminological work that could be possible with more comprehensive covariates, more detailed maps, and more extensive records from additional cities, so we now turn to consider potential future work.

# *Seven*

# Conclusions and Future Work

This thesis develops a self-exciting spatio-temporal point process model of crime, building on previous work by simultaneously accounting for spatial features, leading indicators, and past crime history, and by combining useful diagnostics and inference tools to make analysis practical. A series of simulations and practical applications to Pittsburgh and Baltimore demonstrate the model's usefulness.

Many further extensions to the model are possible, and this chapter reviews several that would have immediate practical uses.

## 7.1 Distance Metrics

One obvious modification to the self-exciting point process model of crime is the choice of distance metric. The triggering function $g$ uses the Euclidean distance between crimes to determine influence on each other, but in a city with rivers, highways, bridges, and other complicated geography, Euclidean distance is likely not the best measure of how influential one crime may be on another. It may be more reasonable to design a triggering function that increases the risk of crimes in areas close in *travel distance* to the crime, rather than in Euclidean space. This choice also makes it simple to avoid predicting crime in locations which are physically close but cannot experience crime, such as the middle of a lake.

One version of this is advocated by Rosser, Davies, Bowers, Johnson, and Cheng (2017), who argue in favor of using distances along the street network. Crimes typically occur along streets (or residences adjacent to streets), and street distance is a good proxy measure for how easy it is to get from one point to another. Using a spatio-temporal kernel density method, they show that hotspot predictions performed on the street network outperform predictions in Euclidean space.

If provided a dataset of events geocoded to street segments, along with a current city road map, it would be reasonably simple to adapt the self-exciting point process model of crime to operate on the network. The triggering function would be modified to use network distance instead of Euclidean distance, and the spatial integral in the log-likelihood would integrate over the road network instead of over

a domain $X \subset \mathbb{R}^2$. Mapping and visualizing the intensity $\lambda(s,t)$ would be more difficult, since it would only be defined on the road network and not arbitrary points $s \in X$, but predictions on the network could be performed as usual. Residuals would be defined in terms of sets of road segments instead of Voronoi cells.

It may be valuable in future work to explore this option, and to explore other applications where events occur on a network. Other possible extensions include tests for difference between different types of network edges, such as the effects of public transit lines making certain routes much easier to travel than others, using weighted graph traversal methods.

## 7.2 Spatio-Temporal Covariates

Another extension is the possibility of including covariates that change as a function of time. Rather than the observation domain $X$ being divided into spatial cells, each of which has a fixed constant covariate vector, we could imagine the covariate vectors being allowed to change over time. Meyer et al. (2012), for example, discussed in Section 2.3.3, used each district's recent count of influenza cases as a covariate for predicting invasive meningococcal disease. For crime prediction, other spatiotemporal covariates could include the weather, which is known to influence crime (Brunsdon, Corcoran, Higgs, & Ware, 2009; Field, 1992; Mares, 2013), weekday and weekend effects, long-term changes in population density or police activity patterns, and so on.

For covariates that change at discrete intervals, such as daily or weekly, the model can be fairly easily extended, and the same EM fitting algorithm applied, with minor modifications to the update steps in Section 3.4 to sum over space-time covariate cells, rather than spatial cells. However, covariates that are allowed to vary continuously in time or space pose a problem: the integral of $\lambda(s,t)$ over all space and time in the log-likelihood (see Section 2.1.4) can no longer easily be done analytically, but must be done numerically for the specific spatio-temporal form of the covariates. This would dramatically slow down model fitting, except perhaps in special cases that can still be done analytically.

## 7.3 Leading Indicator Suppression

The self-exciting point process model can incorporate additional types of events, such as misdemeanor offenses or 911 calls, that are not the target of prediction but which may indicate locally higher risks of the target crime types. There could also be types of leading indicators that locally *suppress* the crime rate—for example, a police foot patrol or the arrest of a repeat offender. Other applications of the model outside of crime may have analogous suppression effects that need to be modeled.

As the model is written, suppression is difficult to account for, because of the need to ensure that $\lambda(s, t) \geq 0$. An event that contributes a negative rate to the sum may violate this constraint. One possible approach is given by S. Chen, Shojaie, Shea-Brown, and Witten (2017), which replaces the cluster process representation described in Section 2.1.2 with a "thinning process representation" allowing each event "to increase or decrease the occurrence of future events". This representation is not yet widely used, and adopting it would mean we cannot use the current expectation-maximization method to fit the model; future work should explore this representation and estimation strategies to practically apply it to our model.

## 7.4 Modeling Police Responses

Another concern is the change in crime dynamics that results from police *using* a predictive policing model to direct their effects. We could imagine this happening in several different ways. For example, perhaps police implement a burglary intervention program intended to prevent burglary near-repeats, hence stopping hotspots before they become hotspots; if the program is guided by a predictive policing model and the intervention is effective, then a successful model will fail to predict burglary, since the burglaries it predicts are successfully prevented.

On the other hand, perhaps a model intended to predict drug-related street crime could be used to direct police patrols and lead to *higher* incident rates, as police search more suspicious persons, make more drug arrests, and observe drug transactions that otherwise would have gone unreported. A successful predictive model hence excites more reported crime.

Accounting for either effect would be challenging, and would require extensions to the model to account for police activity. One approach would be to include police activity as a leading indicator that can either excite or inhibit crime, requiring the extensions discussed in the previous section; another might be to allow police activity to be a spatio-temporal covariate in the background process.

## 7.5 Bayesian Modeling

Recently there has been interest in developing Bayesian versions of self-exciting spatio-temporal point processes, as an alternate model fitting approach to maximum likelihood estimation. As discussed in Section 2.2.5, there have been several recent advances in Bayesian inference for these models that make it dramatically more computationally tractable, hence making Bayesian inference practical for much larger datasets than were previously possible to use.

The crucial advance was made by Rasmussen (2013) and Ross (2016), who applied the same conditioning process used during expectation maximization (Section 2.2.1). Conditioning on the branching structure allows the log-likelihood to be

separated into pieces, following the cluster structure described in Section 2.1.1; this makes it much easier to sample from the posterior of the model parameters, as we shall see below. I have made some preliminary steps towards adapting this method to the extended model introduced in Chapter 3, and this promises to allow hierarchical Bayesian estimation that can fit to several cities or neighborhoods at the same time, accounting for heterogeneity across or between cities.

### 7.5.1 The Partitioned Likelihood

Following the setup used in Chapter 3, we consider events to be triples $(s_i, t_i, M_i)$, where $s_i$ is a location in $\mathbb{R}^2$, $t_i$ is a time in $[0, T)$, and $M_i$ is the type of event. By convention, $M_i = 0$ for the response event type and a positive integer for leading indicators. Let $K_l$ be the number of events with $M_i = l$. Suppose, for the response variables with $M_i = 0$, we have the latent variables $u_i$, such that $u_i = 0$ if event $i$ arose from the background and $u_i = j$ if event $i$ was triggered by event $j$, with $t_j < t_i$. ($M_j$ need not be 0.) Partition the *response* events into sets $S_0, \ldots, S_n$ such that

$$S_j = \{i \mid u_i = j\}, \qquad M_i = 0 \text{ and } 0 \le j < n,$$

so $S_0$ contains the indices of the background events and $S_j$ is the set of indices of events triggered by event $j$. Note that some of these sets may be empty, as some events do not trigger any offspring events.

We know that the likelihood of a realization of a spatio-temporal point process with intensity $\lambda(s, t \mid \mathscr{H}_t)$ and parameter vector $\Theta$ is (eq. (2.8))

$$L(\Theta) = \left[\prod_{i=1}^{n} \lambda(s_i, t_i \mid \mathscr{H}_t)\right] \exp\left(-\int_0^T \int_X \lambda(s, t \mid \mathscr{H}_t)\, \mathrm{d}s\, \mathrm{d}t\right).$$

Also, we know that once events are partitioned into the sets $S_j$, each set is *independent* of the others—that's the key of the cluster process representation. So we may calculate the likelihoods of each set separately.

The background component, $S_0$, has intensity $\mu(s) = \exp\big(\beta X(s)\big)$, where $X(s)$ is a covariate function that is piecewise constant in space. The background component can then be broken up into separate covariate cells (each cell is a region in which the covariate function is constant), each with its own Poisson process, resulting in the likelihood

$$L_0(\Theta) = \left[\prod_{j=1}^{J} \exp\left(|S_0 \cap C_j| \beta X_j\right)\right] \exp\left(-T \sum_{j=1}^{J} |C_j| \exp\left(\beta X_j\right)\right) \tag{7.1}$$

where there are $J$ total covariate cells, $|C_j|$ is the area of cell $j$, $|S_0 \cap C_j|$ is the count of background events in cell $j$, and $X_j$ is the covariate vector in cell $j$.

Next, for each $j > 0$, we have independent nonhomogeneous Poisson process clusters, where the intensities are given by the triggering function

$$g(s, t, M) = \frac{\theta_M}{2\pi\omega\sigma^2} \exp(-t/\omega) \exp\left(-\frac{\|s\|^2}{2\sigma^2}\right),$$

centered at $(s_j, t_j)$. The product term in the likelihood is

$$\begin{aligned}\prod_{i \in S_j}^{n} g(s_i - s_j, t_i - t_j, M_j) &= \prod_{i \in S_j} \frac{\theta_{M_j}}{2\pi\omega\sigma^2} \exp\left(-(t_i - t_j)/\omega\right) \exp\left(-\frac{\|s_i - s_j\|^2}{2\sigma^2}\right) \\ &= \left(\frac{\theta_{M_j}}{2\pi\omega\sigma^2}\right)^{|S_j|} \exp\left(-\sum_{i \in S_j} (t_i - t_j)/\omega\right) \exp\left(-\sum_{i \in S_j} \frac{\|s_i - s_j\|^2}{2\sigma^2}\right).\end{aligned}$$

The integral is

$$\begin{aligned}\int_{t_j}^{T} \int_X g(s - s_j, t - t_j, M_j)\, \mathrm{d}s\, \mathrm{d}t &= \int_{t_j}^{T} \int_X \frac{\theta_M}{2\pi\omega\sigma^2} \exp\left(-(t - t_j)/\omega\right) \exp\left(-\frac{\|s - s_j\|^2}{2\sigma^2}\right) \mathrm{d}s\, \mathrm{d}t \\ &= \theta_M \left(1 - e^{-(T - t_j)/\omega}\right).\end{aligned}$$

So, in cluster $j$, we have $|S_j|$ events and likelihood

$$\begin{aligned}L_j(\Theta) &= \exp\left(-\theta_{M_j}\left(1 - e^{-(T - t_j)/\omega}\right)\right) \prod_{i \in S_j} \frac{\theta_{M_j}}{2\pi\omega\sigma^2} \exp\left(-(t_i - t_j)/\omega\right) \exp\left(-\frac{\|s_i - s_j\|^2}{2\sigma^2}\right) \\ &= \exp\left(-\theta_{M_j}\left(1 - e^{-(T - t_j)/\omega}\right)\right) \left(\frac{\theta_{M_j}}{2\pi\omega\sigma^2}\right)^{|S_j|} \exp\left(-\sum_{i \in S_j} (t_i - t_j)/\omega\right) \exp\left(-\sum_{i \in S_j} \frac{\|s_i - s_j\|^2}{2\sigma^2}\right).\end{aligned}$$

Note that we have approximated the integral on $[0, T) \times X$ with one on $[0, T) \times \mathbb{R}^2$. If we further approximate to $[0, \infty) \times \mathbb{R}^2$, we get

$$L_j(\Theta) \approx \exp\left(-\theta_{M_j}\right) \left(\frac{\theta_{M_j}}{2\pi\omega\sigma^2}\right)^{|S_j|} \exp\left(-\sum_{i \in S_j} (t_i - t_j)/\omega\right) \exp\left(-\sum_{i \in S_j} \frac{\|s_i - s_j\|^2}{2\sigma^2}\right).$$

Also note that when $|S_j| = 0$, this reduces to

$$L_j(\Theta) = \exp\left(-\theta_{M_j}\left(1 - e^{-(T - t_j)/\omega}\right)\right) \approx \exp\left(-\theta_{M_j}\right).$$

Since no parameter appears in both the background and triggered likelihoods, the components can be fit separately, when we condition on knowledge of the branching structure $S_j$.

### 7.5.2 Conditioning on the Branching Structure

The above likelihoods were computed on the assumption that the branching structure ($u_i$ for all crimes $i$ with $M_i = 0$) is known. In expectation maximization, we calculate the expected values and require the quantities $P(u_i = j)$; in the sampling procedure suggested by Ross (2016), we draw directly from the conditional distribution of $u_i$, then condition on these to sample from the posterior.

Most of the computational work goes into this conditioning: drawing $u_i$ for each crime with $M_i = 0$ requires calculating the intensity at each such crime (based on the current parameter values) and then drawing at random from the contributors to that intensity, following the stochastic declustering procedure in Algorithm 2.2. For efficiency, this part of the sampling algorithm was written in Cython, and the rest in Python. Stochastic declustering is $O(n^2)$ in practice, though $u_i$ for each $i$ can be drawn in parallel.

### 7.5.3 A Hierarchical Model

Once we have a purely Bayesian self-exciting point process model of crime, it makes sense to ask a further question: how do the dynamics of crime vary between cities, or even within a single city? Do the parameters of the model vary widely or do cities have broadly similar crime dynamics? A hierarchical model could answer this question by allowing the model parameters to vary from city to city, being drawn from prior distributions whose parameters are allowed to vary within a hyperprior distribution.

A fully Bayesian hierarchical point process model has not been developed before, though some existing models do allow their parameter values to vary in space. For example, Ogata and Katsura (1988) allowed some parameters to vary smoothly in space, with their variation controlled by a roughness penalty. Rather than following this approach, we propose a hierarchical Bayesian model that lets parameters vary between different spatial units of analysis.

The city-level conditional intensity function remains the same as in our original model, but we apply city-level priors drawn from fixed hyperpriors to each parameter. The priors are introduced below.

#### *Prior Specification – Normal Case*

We chose to assign a separate parameter for each $\beta_i$, in effect allowing estimates of a single covariate's effect to be pooled between cities but not pooling the covariates together in any way. Consider a single coefficient $\beta_i$ in a particular city $i$ (the

coefficient index is dropped for simplicity). We let

$$\begin{aligned}\beta_i \mid \mu_\beta, \sigma_\beta &\sim \text{Normal}(\mu_\beta, \sigma_\beta^2)\\ \sigma_\beta &\sim \text{Uniform}(0, A)\\ \mu_\beta &\sim \text{Normal}(M, V).\end{aligned}$$

Hence the covariate $\beta_i$ for each city is estimated by the model, along with the mean value for all cities $\mu_\beta$ and the inter-city variance $\sigma_\beta^2$; only $A$, $M$, and $V$ are fixed by hyperprior specification.

In this case, we draw from the posterior of $\mu_\beta$ and $\sigma_\beta^2$ by the following procedure. Conditioning on the data and the current value of $\sigma_\beta^2$, $\mu_\beta$ has a conjugate posterior, and can be drawn from

$$\mu_\beta \mid \{\beta_i\}, \sigma_\beta^2 \sim \text{Normal}\left(\frac{1}{D}\left(\frac{M}{V} + \frac{\sum \beta_i}{\sigma_\beta^2}\right), \frac{1}{D}\right)$$

where

$$D = \frac{1}{V} + \frac{n}{\sigma_\beta^2}.$$

This step is followed by a Metropolis sample from the posterior of $\sigma_\sigma^2$, conditioning on the data and the new draw of $\mu_\sigma$. The Metropolis likelihood ratio is

$$\begin{aligned}\frac{L(\sigma_\beta^{2*})}{L(\sigma_\beta^2)} &= \prod_{i=1}^{n} \frac{\sqrt{2\pi\sigma_\beta^2}}{\sqrt{2\pi\sigma_\beta^{2*}}} \exp\left(-\frac{(\beta_i - \mu_\beta)^2}{2\sigma_\beta^{2*}} + \frac{(\beta_i - \mu_\beta)^2}{2\sigma_\beta^2}\right)\\ &= \left(\frac{\sigma_\beta}{\sigma_\beta^*}\right)^{n/2} \exp\left(\left(-\frac{1}{2\sigma_\beta^{2*}} + \frac{1}{2\sigma_\beta^2}\right)\sum_{i=1}^{n}(\beta_i - \mu_\beta)^2\right)\end{aligned}$$

*Prior Specification – Log-Normal Case*

The remaining parameters—$\theta$, $\sigma^2$ and $\omega$—must be nonnegative, so a normal prior does not make sense for them. Instead, we use a log-Normal; for example, for $\omega_i$ in a particular city $i$, we let

$$\begin{aligned}\omega_i \mid \sigma_\omega, \mu_\omega &\sim \text{LogNormal}(\mu_\omega, \sigma_\omega^2)\\ \sigma_\omega &\sim \text{Uniform}(0, A)\\ \mu_\omega &\sim \text{Normal}(M, V).\end{aligned}$$

Note that $\mu_\omega$ and $\sigma_\omega^2$ are the mean and variance of the *logarithm* of $\omega_i$, and so $A$, $M$, and $V$ must all be specified on the scale of its logarithm.

In this case, we draw from the posterior of $\mu_\omega$ and $\sigma_\omega^2$ by the following procedure. Conditioning on the data and the current value of $\sigma_\omega^2$, $\mu_\omega$ has a conjugate posterior:

$$\mu_\omega \mid \{\sigma_i^2\}, \sigma_\omega^2 \sim \text{LogNormal}\left(\frac{1}{D}\left(\frac{M}{V} + \frac{\sum \log \omega_i}{\sigma_\omega^2}\right), \frac{1}{D}\right)$$

where $D$ is defined as before and log is the natural logarithm.

This step is followed by a Metropolis sample for $\sigma_\omega^2$. The Metropolis likelihood ratio here is

$$\begin{aligned}\frac{L(\sigma_\omega^{2*})}{L(\sigma_\omega^2)} &= \prod_{i=1}^{n} \frac{\omega_i\sqrt{2\pi\sigma_\omega^2}}{\omega_i\sqrt{2\pi\sigma_\omega^{2*}}} \exp\left(-\frac{(\log\omega_i - \mu_\omega)^2}{2\sigma_\omega^{2*}} + \frac{(\log\omega_i - \mu_\omega)^2}{2\sigma_\omega^2}\right) \\ &= \left(\frac{\sigma_\omega}{\sigma_\omega^*}\right)^{n/2} \exp\left(\left(-\frac{1}{2\sigma_\omega^{2*}} + \frac{1}{2\sigma_\omega^2}\right)\sum_{i=1}^{n}(\log\omega_i - \mu_\omega)^2\right)\end{aligned}$$

Because of the parametrization of the log-Normal in terms of the mean and variance of the logarithm of the random variable, we can equivalently write the model as

$$\begin{aligned}\log(\omega_i) \mid \sigma_\omega, \mu_\omega &\sim \text{Normal}(\mu_\omega, \sigma_\omega^2) \\ \sigma_\omega &\sim \text{Uniform}(0, A) \\ \mu_\omega &\sim \text{Normal}(M, V).\end{aligned}$$

Provided $A$, $M$, and $V$ are kept on the log scale, the same conjugate and Metropolis updates used in the normal case can be used on $\log(\omega_i)$. This simplifies our model implementation.

### 7.5.4 Next Steps

With the basics of the Bayesian hierarchical model established, several extensions and applications can be explored in future work. It will be necessary to extend the diagnostics tools of Chapter 4 to the Bayesian hierarchical model, and to explore further simulations of model misspecification to understand the role of the model priors and hyperpriors. Other simulations could compare the performance of the Bayesian model against the maximum likelihood estimator presented in this thesis, to determine if pooling across cities improves predictive performance.

Once the model is well-understood, it could be applied to an analysis of crime in several cities at the same time, giving the first ever systematic comparison of crime dynamics between multiple cities. (It would also be possible to compare multiple regions of the same city, such as neighborhoods of New York or Chicago.) The Bayesian model could be extended and used in other ways as well. For example, a

hierarchical multivariate model—modeling multiple types of events which mutually excite each other, such as multiple types of crimes—could pool parameter information between event types, while still allowing their excitation effects to differ. Or city-level covariates could be introduced to account for the differences in parameter values among cities. And, of course, there are many other possible applications, some of which undoubtedly haven't occurred to anyone yet. Self-exciting point process models are flexible and powerful tools whose uses are only just beginning to be discovered.

# *A*

# Raw Data and Source Code

In the interests of reproducibility, data and source code are available for the methods described in this thesis, as well as all plots and analyses presented in this thesis.

All statistical methods were implemented in Python 3. The code was wrapped up into a Python package, available at `https://bitbucket.org/capnrefsmmat/crime-mapping`, containing functions to load data, fit the models described in this thesis, evaluate predictive performance, and produce visualizations and diagnostics. The complete source code revision history is available as a Git repository. Installation instructions are given in the included `README.md` file.

Separately, the analyses and plots produced as a part of this thesis were written as Python scripts using the analysis package, and fully automated with a Makefile to automatically generate the results needed for this thesis. These scripts are available in a separate archive at `https://www.refsmmat.com/files/thesis-files.zip`.

Source code for the package and thesis is licensed under the GNU General Public License, version 2, meaning it may be freely reused and redistributed, under certain terms. The full license is available at `https://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html`, and is provided in both source repositories in a `LICENSE.txt` file.

# Bibliography


Adelfio, G., & Chiodi, M. (2015a). Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. *Stochastic Environmental Research and Risk Assessment*, *29*(2), 443–450. doi:10.1007/s00477-014-0873-8

Adelfio, G., & Chiodi, M. (2015b). FLP estimation of semi-parametric models for space–time point processes and diagnostic tools. *Spatial Statistics*, *14*, 119–132. doi:10.1016/j.spasta.2015.06.004

Bacry, E., Mastromatteo, I., & Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, *01*(01), 1550005. doi:10.1142/s2382626615500057

Baddeley, A., Møller, J., & Pakes, A. G. (2007). Properties of residuals for spatial point processes. *Annals of the Institute of Statistical Mathematics*, *60*(3), 627–649. doi:10.1007/s10463-007-0116-6

Baddeley, A., Rubak, E., & Møller, J. (2011). Score, pseudo-score and residual diagnostics for spatial point process models. *Statistical Science*, *26*(4), 613–646. doi:10.1214/11-sts367

Baddeley, A., Turner, R., Moller, J., & Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B*, *67*(5), 617–666. doi:10.1111/j.1467-9868.2005.00519.x

Bauwens, L., & Hautsch, N. (2009). Modelling financial high frequency data using point processes. In T. Mikosch, J.-P. Kreiß, R. A. Davis, & T. G. Andersen (Eds.), *Handbook of financial time series* (pp. 953–979). doi:10.1007/978-3-540-71297-8_41

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, *18*(9), 509–517. doi:10.1145/361002.361007

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., … Bengio, Y. (2010). Theano: A CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Austin, TX.

Bernasco, W., Johnson, S. D., & Ruiter, S. (2015). Learning where to offend: Effects of past on future burglary locations. *Applied Geography*, *60*, 120–129. doi:10.1016/j.apgeog.2015.03.014

Bowers, K. J., Johnson, S. D., & Pease, K. (2004). Prospective Hot-Spotting: The Future of Crime Mapping? *British Journal of Criminology*, *44*(5), 641–658. doi:10.1093/bjc/azh036

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The Effects of Hot Spots Policing on Crime: An Updated Systematic Review and Meta-Analysis. *Justice Quarterly*, *31*(4), 633–663. doi:10.1080/07418825.2012.673632

Bray, A., & Schoenberg, F. P. (2013). Assessment of Point Process Models for Earthquake Forecasting. *Statistical Science*, *28*(4), 510–520. doi:10.1214/13-STS440

Bray, A., Wong, K., Barr, C. D., & Schoenberg, F. P. (2014). Voronoi residual analysis of spatial point process models with applications to California earthquake forecasts. *Annals of Applied Statistics*, *8*(4), 2247–2267. doi:10.1214/14-AOAS767

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, *439*, 462–465. doi:10.1038/nature04292

Brunsdon, C., Corcoran, J., Higgs, G., & Ware, A. (2009). The influence of weather on local geographical patterns of police calls for service. *Environment and Planning B: Planning and Design*, *36*(5), 906–926. doi:10.1068/b32133

Cerdá, M., Tracy, M., Messner, S. F., Vlahov, D., Tardiff, K., & Galea, S. (2009). Misdemeanor Policing, Physical Disorder, and Gun-related Homicide: A Spatial Analytic Test of "Broken-Windows" Theory. *Epidemiology*, *20*(4), 533–541. doi:10.1097/EDE.0b013e3181a48a99

Chainey, S., Tompson, L., & Uhlig, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, *21*(1-2), 4–28. doi:10.1057/palgrave.sj.8350066

Chen, J., Hawkes, A. G., Scalas, E., & Trinh, M. (2017). Performance of information criteria for selection of hawkes process models of financial data. *Quantitative Finance*, *18*(2), 225–235. doi:10.1080/14697688.2017.1403140

Chen, S., Shojaie, A., Shea-Brown, E., & Witten, D. (2017). The multivariate Hawkes process in high dimensions: Beyond mutual excitation. arXiv. Retrieved from https://arxiv.org/abs/1707.04928

Chiodi, M., & Adelfio, G. (2011). Forward likelihood-based predictive approach for space-time point processes. *Environmetrics*, *22*(6), 749–757. doi:10.1002/env.1121

Clarke, R. V., & Cornish, D. B. (1985). Modeling Offenders' Decisions: A Framework for Research and Policy. *Crime and Justice*, *6*, 147–185. doi:10.1086/449106

Clements, R. A., Schoenberg, F. P., & Veen, A. (2012). Evaluation of space-time point process models using super-thinning. *Environmetrics*, *23*(7), 606–616. doi:10.1002/env.2168

Cohen, J., Gorr, W. L., & Olligschlaeger, A. M. (2007). Leading Indicators and Spatial Interactions: A Crime-Forecasting Model for Proactive Police Deployment. *Geographical Analysis*, *39*(1), 105–127. doi:10.1111/j.1538-4632.2006.00697.x

Cohen, L. E., & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, *44*(4), 588–608.

Cowling, A., & Hall, P. (1996). On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society Series B*, *58*(3), 551–563.

Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Wiley.

Daley, D. J., & Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods* (2nd ed.). Springer.

Daley, D. J., & Vere-Jones, D. (2004). Scoring Probability Forecasts for Point Processes: The Entropy Score and Information Gain. *Journal of Applied Probability*, *41*, 297–312. doi:10.1239/jap/1082552206

Daley, D. J., & Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure* (2nd ed.). Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximimum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

Diggle, P. J. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns* (3rd). CRC Press.

Diggle, P. J., Moraga, P., Rowlingson, B., & Taylor, B. M. (2013). Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm. *Statistical Science*, *28*(4), 542–563. doi:10.1214/13-STS441

Drawve, G. (2016). A Metric Comparison of Predictive Hot Spot Techniques and RTM. *Justice Quarterly*, *33*(3), 369–397. doi:10.1080/07418825.2014.904393

Drawve, G., Moak, S. C., & Berthelot, E. R. (2014). Predictability of gun crimes: a comparison of hot spot and risk terrain modelling techniques. *Policing and Society*, 1–20. doi:10.1080/10439463.2014.942851

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi:10.1016/j.patrec.2005.10.010

FBI. (2004). *Uniform Crime Reporting Handbook*. Department of Justice. Retrieved from https://ucr.fbi.gov/additional-ucr-publications/ucr_handbook.pdf

Field, S. (1992). The effect of temperature on crime. *British Journal of Criminology*, *32*(3), 340–351.

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, *23*(7), 1025–1044. doi:10.1068/a231025

Fox, E. W., Schoenberg, F. P., & Gordon, J. S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *Annals of Applied Statistics*, *10*(3), 1725–1756. doi:10.1214/16-AOAS957

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D., & Bertozzi, A. L. (2016). Modeling E-mail Networks and Inferring Leadership Using Self-Exciting Point Processes. *Journal of the American Statistical Association*, *111*(514), 564–584. doi:10.1080/01621459.2015.1135802

Freed, A. M. (2005). Earthquake triggering by static, dynamic, and postseismic stress transfer. *Annual Review of Earth and Planetary Sciences*, *33*(1), 335–367. doi:10.1146/annurev.earth.33.092203.122505

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. doi:10.1198/016214506000001437

González, J. A., Rodríguez-Cortés, F. J., Cronie, O., & Mateu, J. (2016). Spatio-temporal point process statistics: A review. *Spatial Statistics*, *18*, 505–544. doi:10.1016/j.spasta.2016.10.002

Gorr, W. L. (2009). Forecast accuracy measures for exception reporting using receiver operating characteristic curves. *International Journal of Forecasting*, *25*(1), 48–61. doi:10.1016/j.ijforecast.2008.11.013

Gorr, W. L., & Lee, Y. (2015). Early Warning System for Temporary Crime Hot Spots. *Journal of Quantitative Criminology*, *31*(1), 25–47. doi:10.1007/s10940-014-9223-8

Gray, A. G., & Moore, A. W. (2003). Nonparametric Density Estimation: Toward Computational Tractability. In *SIAM International Conference on Data Mining* (pp. 203–211). Philadelphia, PA: Society for Industrial and Applied Mathematics.

Green, B., Horel, T., & Papachristos, A. V. (2017). Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago, 2006 to 2014. *JAMA Internal Medicine*, *177*(3), 326–333. doi:10.1001/jamainternmed.2016.8245

Guttman, A. (1984). R-trees: A Dynamic Index Structure for Spatial Searching. In *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data* (pp. 47–57). SIGMOD '84. doi:10.1145/602259.602266

Haberman, C. P., & Ratcliffe, J. H. (2012). The Predictive Policing Challenges of Near Repeat Armed Street Robberies. *Policing*, *6*(2), 151–166. doi:10.1093/police/pas012

Hart, T., & Zandbergen, P. (2014). Kernel density estimation and hotspot mapping. *Policing*, *37*(2), 305–323. doi:10.1108/PIJPSM-04-2013-0039

Harte, D. (2012). Bias in fitting the ETAS model: a case study based on New Zealand seismicity. *Geophysical Journal International*, *192*(1), 390–412. doi:10.1093/gji/ggs026

Harte, D., & Vere-Jones, D. (2005). The Entropy Score and its Uses in Earthquake Forecasting. *Pure and Applied Geophysics*, *162*(6), 1229–1253. doi:10.1007/s00024-004-2667-2

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *51*(1), 83–90. doi:10.1093/biomet/58.1.83

Hawkes, A. G., & Oakes, D. (1974). A Cluster Process Representation of a Self-Exciting Process. *Journal of Applied Probability*, *11*(3), 493–503. doi:10.1017/S0021900200096273

Hodges, J. S., & Reich, B. J. (2010). Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. *The American Statistician*, *64*(4), 325–334. doi:10.1198/tast.2010.10052

Hunt, P., Saunders, J., & Hollywood, J. S. (2014). *Evaluation of the Shreveport Predictive Policing Experiment*. RAND.

Johnson, D. H. (1996). Point process models of single-neuron discharges. *Journal of Computational Neuroscience*, *3*(4), 275–299. doi:10.1007/bf00161089

Johnson, S. D., Bernasco, W., Bowers, K. J., Elffers, H., Ratcliffe, J., Rengert, G., & Townsley, M. (2007). Space–Time Patterns of Risk: A Cross National Assessment of Residential Burglary Victimization. *Journal of Quantitative Criminology*, *23*(3), 201–219. doi:10.1007/s10940-007-9025-3

Kennedy, L. W., Caplan, J. M., & Piza, E. L. (2010). Risk Clusters, Hotspots, and Spatial Intelligence: Risk Terrain Modeling as an Algorithm for Police Resource Allocation Strategies. *Journal of Quantitative Criminology*, *27*(3), 339–362. doi:10.1007/s10940-010-9126-2

Kennedy, L. W., Caplan, J. M., Piza, E. L., & Buccine-Schraeder, H. (2015). Vulnerability and Exposure to Crime: Applying Risk Terrain Modeling to the Study of Assault in Chicago. *Applied Spatial Analysis and Policy*. doi:10.1007/s12061-015-9165-z

Knox, E. G. (1964). The Detection of Space-Time Interactions. *Applied Statistics*, *13*(1), 25–30. doi:10.2307/2985220

Kumazawa, T., & Ogata, Y. (2014). Nonstationary ETAS models for nonstandard earthquakes. *The Annals of Applied Statistics*, *8*(3), 1825–1852. doi:10.1214/14-aoas759

Lang, D. (2004). *Fast Methods for Inference in Graphical Models* (Doctoral dissertation, University of British Columbia). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.9783&rep=rep1&type=pdf

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, *38*(5), 404–415. doi:10.1016/j.jbi.2005.02.008

Levine, N. (2008). The "Hottest" Part of a Hotspot: Comments on "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime". *Security Journal*, *21*(4), 295–302. doi:10.1057/sj.2008.5

Lewis, E., Mohler, G., Brantingham, P. J., & Bertozzi, A. L. (2011). Self-exciting point process models of civilian deaths in Iraq. *Security Journal*, *25*(3), 244–264. doi:10.1057/sj.2011.21

Lewis, P. A. W., & Shedler, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly*, *26*(3), 403–413. doi:10.1002/nav.3800260304

Lippiello, E., Giacco, F., Arcangelis, L. d., Marzocchi, W., & Godano, C. (2014). Parameter estimation in the ETAS model: Approximations and novel methods. *Bulletin of the Seismological Society of America*, *104*(2), 985–994. doi:10.1785/0120130148

Loeffler, C., & Flaxman, S. (2017). Is Gun Violence Contagious? *Journal of Quantitative Criminology*. doi:10.1007/s10940-017-9363-8

Mares, D. (2013). Climate change and crime: monthly temperature and precipitation anomalies and crime rates in St. Louis, MO 1990–2009. *Crime, Law and Social Change*, *59*(2), 185–208. doi:10.1007/s10611-013-9411-8

Marsan, D., & Lengliné, O. (2008). Extending earthquakes' reach through cascading. *Science*, *319*(5866), 1076–1079. doi:10.1126/science.1148783

Marsan, D., & Lengliné, O. (2010). A new estimation of the decay of aftershock density with distance to the mainshock. *Journal of Geophysical Research*, *115*(B9), B09302. doi:10.1029/2009JB007119

McLachlan, G. J., & Krishnan, T. (2008). *The EM Algorithm and Extensions* (2nd). Wiley.

Meyer, S. (2010). *Spatio-temporal infectious disease epidemiology based on point processes*. (Master's thesis, Ludwig-Maximilians-Universität München).

Meyer, S., Elias, J., & Höhle, M. (2012). A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence. *Biometrics*, *68*(2), 607–616. doi:10.1111/j.1541-0420.2011.01684.x

Meyer, S., & Held, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics*, *8*(3), 1612–1639. doi:10.1214/14-AOAS743

Meyer, S., Warnke, I., Rössler, W., & Held, L. (2016). Model-based testing for space–time interaction using point processes: An application to psychiatric hospital admissions in an urban area. *Spatial and Spatio-temporal Epidemiology*, *17*, 15–25. doi:10.1016/j.sste.2016.03.002

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

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association*, *106*(493), 100–108. doi:10.1198/jasa.2011.ap09546

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*, *110*(512), 1399–1411. doi:10.1080/01621459.2015.1077710

Møller, J., & Rasmussen, J. G. (2005). Perfect Simulation of Hawkes Processes. *Advances in Applied Probability*, *37*(3), 629–646. doi:10.1239/aap/1127483739

Musmeci, F., & Vere-Jones, D. (1992). A space-time clustering model for historical earthquakes. *Annals of the Institute of Statistical Mathematics*, *44*(1), 1–11. doi:10.1007/bf00048666

Nandan, S., Ouillon, G., Wiemer, S., & Sornette, D. (2017). Objective estimation of spatially variable parameters of epidemic type aftershock sequence model: Application to California. *Journal of Geophysical Research: Solid Earth*, *122*(7), 5118–5143. doi:10.1002/2016jb013266

Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N., & Marathe, M. V. (2013). A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and Other Respiratory Viruses*, *8*(3), 309–316. doi:10.1111/irv.12226

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, *30*(1), 243–261. doi:10.1007/BF02480216

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, *83*(401), 9–27. doi:10.1080/01621459.1988.10478560

Ogata, Y. (1998). Space-Time Point-Process Models for Earthquake Occurrences. *Annals of the Institute of Statistical Mathematics*, *50*(2), 379–402. doi:10.1023/A:1003403601725

Ogata, Y. (1999). Seismicity Analysis through Point-process Modeling: A Review. *Pure and Applied Geophysics*, *155*(2-4), 471–507. doi:10.1007/s000240050275

Ogata, Y., & Katsura, K. (1988). Likelihood analysis of spatial inhomogeneity for marked point patterns. *Annals of the Institute of Statistical Mathematics*, *40*(1), 29–39.

Ogata, Y., Katsura, K., & Tanemura, M. (2003). Modelling heterogeneous space-time occurrences of earthquakes and its residual analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*(4), 499–509. doi:10.1111/1467-9876.00420

Ogata, Y., & Zhuang, J. (2006). Space-time ETAS models and an improved extension. *Tectonophysics*, *413*, 13–23. doi:10.1016/j.tecto.2005.10.016

Ornstein, J. T., & Hammond, R. A. (2017). The Burglary Boost: A Note on Detecting Contagion Using the Knox Test. *Journal of Quantitative Criminology*, *33*(1), 65–75. doi:10.1007/s10940-016-9281-1

Papangelou, F. (1972). Integrability of expected increments of point processes and a related random change of scale. *Transactions of the American Mathematical Society*, *165*, 483–483. doi:10.1090/s0002-9947-1972-0314102-9

Peng, R. D., Schoenberg, F. P., & Woods, J. A. (2005). A Space–Time Conditional Intensity Model for Evaluating a Wildfire Hazard Index. *Journal of the American Statistical Association*, *100*(469), 26–35. doi:10.1198/016214504000001763

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation.

Porter, M. D., & White, G. (2012). Self-exciting hurdle models for terrorist activity. *Annals of Applied Statistics*, *6*(1), 106–124. doi:10.1214/11-AOAS513

Rasmussen, J. G. (2013). Bayesian inference for Hawkes processes. *Methodology and Computing in Applied Probability*, *15*(3), 623–642. doi:10.1007/s11009-011-9272-5

Ratcliffe, J. H. (2009). *Near Repeat Calculator*. Temple University and National Institute of Justice. Retrieved from http://www.cla.temple.edu/cj/resources/near-repeat-calculator/

Ratcliffe, J. H., & Rengert, G. F. (2008). Near-Repeat Patterns in Philadelphia Shootings. *Security Journal*, *21*(1-2), 58–76. doi:10.1057/palgrave.sj.8350068

Ratcliffe, J. H., Taniguchi, T., Groff, E. R., & Wood, J. D. (2011). The Philadelphia foot patrol experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots. *Criminology*, *49*(3), 795–831. doi:10.1111/j.1745-9125.2011.00240.x

Rathbun, S. L. (1996). Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference*, *51*(1), 55–74. doi:10.1016/0378-3758(95)00070-4

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B*, *39*(2), 172–212.

Ross, G. J. (2016). *Bayesian Estimation of the ETAS Model for Earthquake Occurrences*. Preprint. Retrieved from http://www.gordonjross.co.uk/bayesianetas.pdf

Rosser, G., Davies, T., Bowers, K. J., Johnson, S. D., & Cheng, T. (2017). Predictive crime mapping: Arbitrary grids or street networks? *Journal of Quantitative Criminology*, *33*(3), 569–594. doi:10.1007/s10940-016-9321-x

Sarma, S. V., Nguyen, D. P., Czanner, G., Wirth, S., Wilson, M. A., Suzuki, W., & Brown, E. N. (2011). Computing confidence intervals for point process models. *Neural Computation*, *23*(11), 2731–2745. doi:10.1162/NECO_a_00198

Schoenberg, F. P. (2003). Multidimensional Residual Analysis of Point Process Models for Earthquake Occurrences. *Journal of the American Statistical Association*, *98*(464), 789–795. doi:10.1198/016214503000000710

Schoenberg, F. P. (2013). Facilitated estimation of ETAS. *Bulletin of the Seismological Society of America*, *103*(1), 601–605. doi:10.1785/0120120146

Schoenberg, F. P. (2016). A note on the consistent estimation of spatial-temporal point process parameters. *Statistica Sinica*, *26*, 861–879. doi:10.5705/ss.2014.150

Schoenberg, F. P., Hoffman, M., & Harrigan, R. (2017). *A recursive point process model for infectious diseases*. https://arxiv.org/abs/1703.08202.

Short, M. B., D'Orsogna, M. R., Brantingham, P. J., & Tita, G. E. (2009). Measuring and Modeling Repeat and Near-Repeat Burglary Effects. *Journal of Quantitative Criminology*, *25*(3), 325–339. doi:10.1007/s10940-009-9068-8

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall.

Stan Development Team. (2016). Stan Modeling Language Users Guide and Reference Manual. http://mc-stan.org.

Tanemura, M. (2003). Statistical distributions of Poisson Voronoi cells in two and three dimensions. *Forma*, *18*, 221–247.

Taylor, B., Koper, C. S., & Woods, D. J. (2011). A randomized controlled trial of different policing strategies at hot spots of violent crime. *Journal of Experimental Criminology*, *7*(2), 149–181. doi:10.1007/s11292-010-9120-6

Townsley, M., Homel, R., & Chaseling, J. (2003). Infectious burglaries: A test of the near repeat hypothesis. *British Journal of Criminology*, *43*(3), 615–633. doi:10.1093/bjc/43.3.615

Van Patten, I. T., McKeldin-Coner, J., & Cox, D. (2009). A Microspatial Analysis of Robbery: Prospective Hot Spotting in a Small City. *Crime Mapping*, *1*(1), 7–32.

Veen, A., & Schoenberg, F. P. (2008). Estimation of Space–Time Branching Process Models in Seismology Using an EM–Type Algorithm. *Journal of the American Statistical Association*, *103*(482), 614–624. doi:10.1198/016214508000000148

Vere-Jones, D. (1998). Probabilities and Information Gain for Earthquake Forecasting. *Computational Seismology*, *30*, 248–263.

Vere-Jones, D. (2009). Some models and procedures for space-time point processes. *Environmental and Ecological Statistics*, *16*(2), 173–195. doi:10.1007/s10651-007-0086-0

Wang, Q., Schoenberg, F. P., & Jackson, D. D. (2010). Standard Errors of Parameter Estimates in the ETAS Model. *Bulletin of the Seismological Society of America*, *100*(5A), 1989–2001. doi:10.1785/0120100001

Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology*, *53*(2), 133–157. doi:10.1111/1745-9125.12070

Youstin, T. J., Nobles, M. R., Ward, J. T., & Cook, C. L. (2011). Assessing the Generalizability of the Near Repeat Phenomenon. *Criminal Justice and Behavior*, *38*(10), 1042–1063. doi:10.1177/0093854811417551

Zhuang, J. (2006). Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *Journal of the Royal Statistical Society: Series B*, *68*(4), 635–653. doi:10.1111/j.1467-9868.2006.00559.x

Zhuang, J. (2011). Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth, Planets and Space*, *63*(3), 207–216. doi:10.5047/eps.2010.12.010

Zhuang, J., Ogata, Y., & Vere-Jones, D. (2002). Stochastic Declustering of Space-Time Earthquake Occurrences. *Journal of the American Statistical Association*, *97*(458), 369–380. doi:10.1198/016214502760046925

Zhuang, J., Ogata, Y., & Vere-Jones, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research*, *109*, B05301. doi:10.1029/2003JB002879

Zipkin, J. R., Schoenberg, F. P., Coronges, K., & Bertozzi, A. L. (2015). Point-process models of social network interactions: Parameter estimation and missing data recovery. *European Journal of Applied Mathematics*, *27*(03), 502–529. doi:10.1017/S0956792515000492