Nonparametric Estimation, Forecasts, and Model Evaluation of Spatial Temporal Point Process Models for California Seismicity

Gordon, Joshua Seth

2017

UNIVERSITY OF CALIFORNIA
Los Angeles

Nonparametric Estimation, Forecasts, and Model Evaluation of Spatial Temporal Point Process Models for California Seismicity

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Joshua Seth Gordon

2017

ABSTRACT OF THE DISSERTATION

Nonparametric Estimation, Forecasts, and Model Evaluation of Spatial Temporal Point
Process Models for California Seismicity

by

Joshua Seth Gordon
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2017
Professor Frederic R Paik Schoenberg, Chair

Point processes models describe random sequences of events. One key model is the self-exciting point process model where the covariance between events is positive. Models of this type have many applications including seismology, epidemiology, and crime. While model estimation is a primary focus, quantification and assessment of model performance are also useful at identifying departures of model fit from the data. This dissertation discusses two applications of self-exciting point process models. The proposed models are fit to data sets from California seismology and plague data. Performance is verified using simulation studies. We also introduce various model evaluation techniques and conduct detailed model evaluations.

This dissertation is organized as follows: Chapter 1 of this dissertation provides preliminary background on point process models. Chapter 2 introduces a nonparametric technique for estimating self-exciting point process models and proposes an extension which seeks to capture anisotropy in the spatial distribution of aftershocks. A forecasting approach is developed and model performance is evaluated retrospectively in comparison to Helmstetter et al. (2007). Chapter 3 describes a mathematical curiosity that allows one to compute exact maximum likelihood estimates of the triggering function in a direct and extremely rapid manner when the number p of intervals on which the nonparametric estimate is sought equals the number n of observed points. Chapter 4 uses Voronoi residuals, super-thinning, and some other residual analysis methods to evaluate a selection of earthquake forecast models in the Collaboratory for the Study of Earthquake Predictability (CSEP). Chapter 5 introduces `NonParametricHawkes`, an R package for estimation, forecast, and evaluation of spatial-temporal models, created to increase the accessibility of the methods introduced here.

The dissertation of Joshua Seth Gordon is approved.

Lingsen Meng

Jingyi Li

Hongquan Xu

Frederic R Paik Schoenberg, Committee Chair

University of California, Los Angeles

2017

To my mother for her ongoing love and support and to my father who could not see this thesis completed. To my sister whose input I value greatly, to Larry for his years of encouragement, and to Megan for her enduring faith in my journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge my advisor Frederic P. Schoenberg for his advice and support during my pursuit of this doctorate. Rick has be an inspiration from day one, and my research abilities have improved immensely under his guidance. I would also like to acknowledge Eric Fox for his effort in development of the algorithms used as a foundation for so much of this work. Eric has always been a mentor and I am lucky to have had to the opportunity to work with him. In addition, I benefited greatly from reading the prior work of Andrew Bray, Robert, Clements, Ka Wong, and Eric Fox. I am also very grateful for the advice of James Molynuex and Bart Blackburt.

Chapter 2 of this dissertation is in review with the Journal of the American Statistical Association. Chapter 3 of this dissertation is in review with the Journal of Nonparametric Statistics. Chapter 4 of this dissertation appeared as Gordon et al. (2015).

VITA

2009	B.A. (Economics), Willamette Univeristy.
2010	AmeriCorps VISTA, Portland, Oregon.
2013	M.S. (Statistics), UCLA, Los Angeles, California.
2013-2017	Sr. Data Scientist, Pacific Life Insurance.
2015	Instructor, Statistics Department, UCLA.
2015	Teaching Assistant Consultant, Statistics Department, UCLA.
2015-2017	Teaching Assistant, Statistics Department, UCLA.
2015-2017	Engineer, Seagate Technology.

PUBLICATIONS

Gordon, J.S., Clements, R.A., Schoenberg, F.P., and Schorlemmer, D. (2015). Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. *Spatial Statistics*, 14b, 133-150.

Fox, E.W., Schoenberg, F.P., and Gordon, J.S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *A*nnals of Applied Statistics, **10**(3), 1725-1756.

Schoenberg, F.P., Gordon, J.S., and Harrigan, R. (2016) Analytic computation of nonparametric Marsan-Lengliné estimates for Hawkes point processes.

Gordon, J.S., and Schoenberg, F.P. A nonparametric Hawkes model for forecasting California seismicity (2017). *JASA*.

CHAPTER 1

Introduction

1.1 Scientific Context

Stochastic point process models describe random sequences of events and provide tools which help improve our understanding of the physical world. One application of such models is in seismology, where we have only indirect observations of the processes and the underlying physics are extremely complex. The main focus of this dissertation is on a class of branching point process models used in statistical seismology called Hawkes or epidemic-type aftershock sequence (ETAS) models. Model estimation is a primary focus along with quantification of model performance. While statistical seismologists have often used retrospective analysis to train models, there has been increasing interest in improving reproducibility. As such, seismologists have created a framework which focuses on the assessments of earthquake forecast models.

In 2006, the Collaborative Study of Earthquake Predictability (CSEP) was created. CSEP is a multinational effort to organize seismological forecast models and facilitate prospective testing (Jordan, 2006). Not only are forecasts generated independently of user's direct inputs, all stochastic elements are reproducible, allowing for investigations into forecast successes and failures. Hundreds of models of various types such as 1-day, 3-month, and 5-year models are submitted to CSEP. While not all forecasts entered are purely statistical in nature, their standardized nature makes it quite simple to compare them.

We take advantage of the unique opportunities provided by CSEP in two ways:

1. We develop and introduce two models which have been accepted into CSEP and

2. We evaluate existing long-term forecast models for California Seismicity

1.2 Point Process Models

Point processes are a useful tool for modeling events which occur randomly in time and in space. A point process is a random collection of points falling in some space where each point represents the time and/or location of an event. Examples of events include incidence of disease, sightings or births of a species, or the occurrences of fires, earthquakes, lightning strikes, tsunamis, or volcanic eruptions. A point process can be mathematically defined where N is a random measure on a complete separable metric space S taking values in the non-negative integers. This measure represents the number of points falling in the subset A of S.

A point process (Daley and Vere-Jones, 2003, 2007) is a collection of points $\{\tau_1, \tau_2, ...\}$ occurring in some metric space. Frequently in applications the points occur in time, or in space and time. Such processes are typically modeled via their conditional rate (also called conditional intensity), $\lambda(t)$ or $\lambda(s,t)$, which represents the infinitesimal rate at which points are accumulating at time t or at location (s,t) of space-time, given information on all points occurring prior to time t.

Given the history of the process the conditional intensity is defined:

$$\lambda(t,x,y|\mathcal{H}_{t_i}) = \lim_{\Delta t, \Delta x, \Delta y \downarrow 0} \frac{E[N\{(t,t+\Delta t) \times (x, x+\Delta x) \times (y, y+\Delta y)\}|H_t]}{\Delta t \Delta x \Delta y}$$

1.2.1 Poisson Process

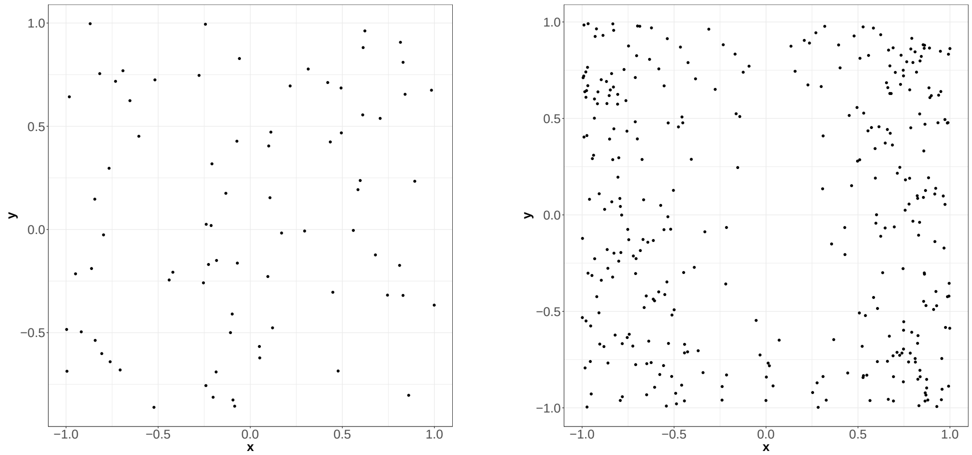
Perhaps the most important type of point process is the Poisson process. The Poisson process is defined as a simple point process N such that the number of points in any set follows a Poisson distribution and the numbers of points in disjoint sets are independent. That is, N is a Poisson process if its subsets are independent Poisson random variables, for any disjoint, measurable subsets.

When a Poisson process has a constant conditional intensity rate it is called a stationary Poisson process. This model implies risk of an event is the same at all times, regardless of where and how frequently such disturbances have occurred previously. A homogeneous Poisson process, such as Figure 1.1(a), is a Poisson point process whose intensity is a non-negative constant. That is, for a finite region A, the number of events in the region $N(A)$ follows a Poisson distribution with mean $\lambda|A|$. For any n events in region A, the events are an independent random sample from a uniform distribution in A.

Alternatively, shown in Figure 1.1(b), an inhomogeneous is a Poisson process with rate parameter $\lambda(s)$ such that the rate of the process is a function of time and location. That is, they are a class of Poisson Process with a varying intensity function. A point process can be called simple if all points are distinct, self-exciting if covariance between points is positive, and self correcting if the covariance is negative. Traditionally the points of a point process are thought to be indistinguishable, other than by their times and/or locations. Often, however, there is other important information to be stored along with each point such as the magnitudes of earthquakes and such processes may be viewed as marked point processes. We can mathematically define X as a simple point process in $\mathbb{R}^2$ where a mark $m_i \in M$ is associated with each point $x_i \in X$. Then a point process Y is called a marked point process in $\mathbb{R}^2$ with mark space M. The space of marks M can be very general. It may be a finite set, a continuous interval of real numbers, or a more complicated space such as the set of all convex polygons. Marked point processes are also used in the formal description of operations like thinning and clustering. For example, thinning a point process X is formalized by constructing a marked point process with marks in $\{0, 1\}$. The mark $\mathbf{1_x}$ attached to each point x indicates whether the point is to be retained (1) or deleted (0).

1.2.2 Hawkes Process

*H*awkes or *s*elf-exciting point processes (Hawkes, 1971) are a type of branching point process model that has become widely used in modeling seismicity (Ogata, 1988, 1998). For a purely temporal Hawkes process, the conditional rate of events at time t, given information $\mathcal{H}_t$ on



$\lambda(x, y) = 75$ $\lambda(x, y) = 300(x^2 + y^2)$

Figure 1.1: Poisson Processes

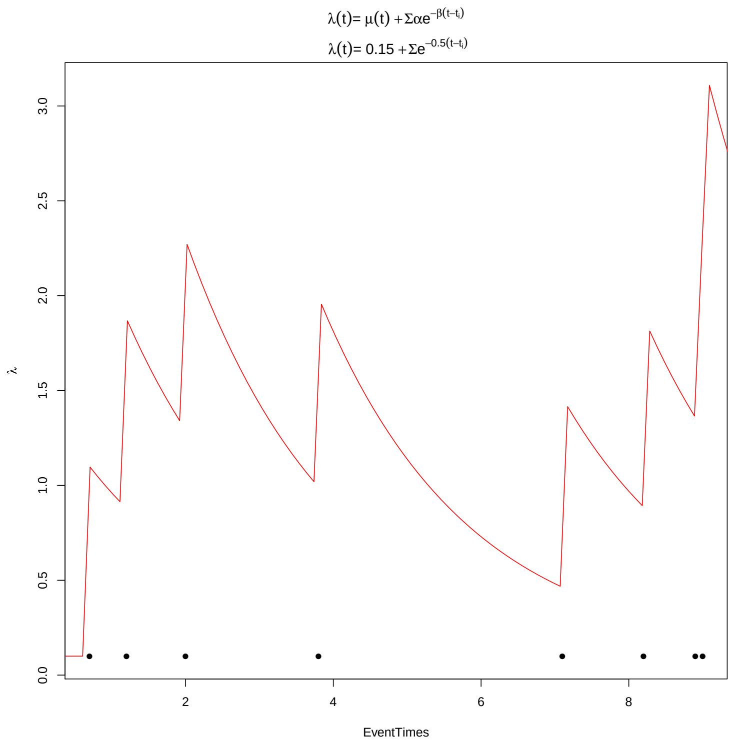
all events prior to time t, can be written

$$\lambda(t|\mathcal{H}_t) = \mu + K \sum_{i:t_i<t} g(t - t_i), \tag{1.1}$$

where $\mu > 0$, is the background rate, $g(u) \geq 0$ is the triggering density satisfying $\int_0^\infty g(u)du = 1$, which describes the secondary activity induced by a prior event, and the constant K is the productivity, which is typically required to satisfy $0 \leq K < 1$ in order to ensure stationarity (Hawkes, 1971). An example of a Hawkes process is seen in Figure 1.2.

Several forms of the triggering function g have been proposed for modeling aftershock activity, such as $g(u_i; m_i) = \frac{1}{(u_i+c)^p} e^{a(m_i - M_0)}$, where $u_i = t - t_i$ is the time elapsed since event i, and M_0 is the lower cutoff magnitude for the earthquake catalog (Ogata, 1988). Extensions to the spatial-temporal case have been posited, and typically one assumes the spatial triggering to be isotropic, meaning the rate of aftershock activity following an earthquake only depends on the distance r from the earthquake's epicenter. For instance, a spatial-temporal-magnitude version of (1.1) suggested by Ogata (1998) uses circular aftershock regions where the sqaured distance between an aftershock and its triggering event follows a Pareto distribution. The

Figure 1.2: Example of Hawkes Process



model may be written

$$\lambda(t, \mathbf{x}|\mathcal{H}_t) = \mu(\mathbf{x}) + K \sum_{i:t_i<t} g(t - t_i, \mathbf{x} - \mathbf{x}_i, m_i), \tag{1.2}$$

with triggering function

$$g(t - t_i, \mathbf{x} - \mathbf{x}_i, m_i) = \exp\{a(m_i - M_0)\}(t - t_i + c)^{-p}(||\mathbf{x} - \mathbf{x}_i||^2 + d)^{-q}, \tag{1.3}$$

where $||\mathbf{x} - \mathbf{x}_i||^2$ represents the squared distance between the epicenter or hypocenter $\mathbf{x}_i$ of earthquake i, and $d > 0$ and $q > 0$ are parameters describing the spatial distribution of triggered earthquakes about their respective mainshocks. An example of the spatial distribution of a Hawkes model specified by (1.2) is shown in Figure 1.3 where $q = 2.721$, $d = 0.0501$, and 300 background events. Triggered events are shown in red, background events are simulated from a homogenous poisson process and appear in black.

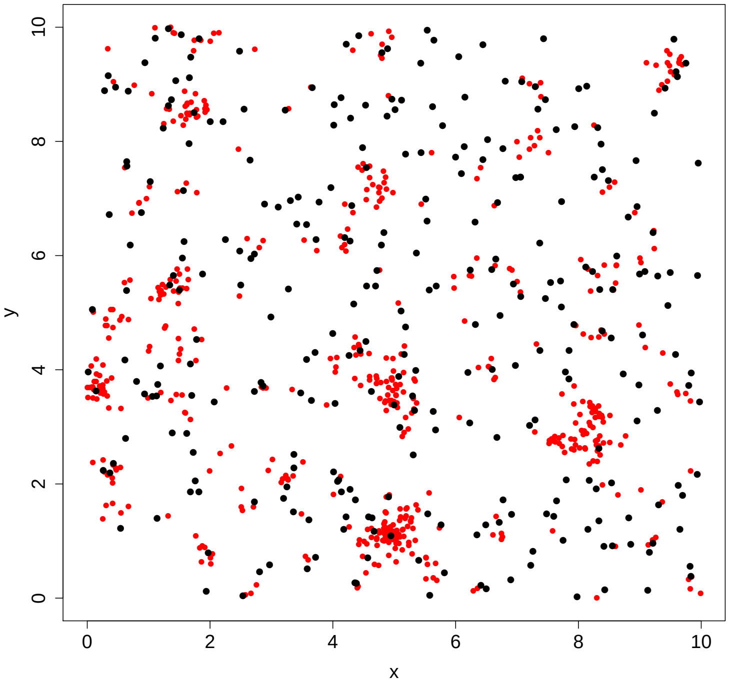


Figure 1.3: Hawkes model with $q = 2.721$, $d = 0.0501$, and 300 background events. Triggered events are shown in red, background events are simulated from a homogenous poisson process and appear in black.

The parameters in models such as (1.3) have traditionally been estimated by maximizing the log-likelihood function

$$\ell(\theta) = \sum_i \log\left(\lambda(t_i, \mathbf{x}_i|\mathcal{H}_{t_i})\right) - \int_0^T \int_S \lambda(t, \mathbf{x}|\mathcal{H}_t)\, d\mathbf{x}dt \tag{1.4}$$

where $\theta = (a, c, d, K_0, p, q, \mu)$ is the parameter vector to be estimated and $S \times [0, T]$ is the space-time window in which the data $(t_i, \mathbf{x}_i, m_i)$ are observed (Daley and Vere-Jones, 2003). Under rather general conditions, MLEs are consistent, asymptotically normal, and efficient (Ogata, 1978), and estimates of their variance can be derived from the negative of the diagonal elements of the Hessian of the likelihood function (Ogata, 1978; Rathbun, S.L. and Cressie, 1994). Functional forms for the background rate $\mu(\mathbf{x})$ are not typically given; instead μ is assumed constant or estimated by smoothing the largest events in the catalog, e.g. using bi-cubic B-splines or kernel smoothing (Ogata 1998; Zhuang et al. 2002).

CHAPTER 2

A nonparametric Hawkes model for forecasting California seismicity

2.1 Introduction

The long history of failed attempts to predict earthquakes has rendered the task of rigorous assessment of earthquake forecasts the primary task for statistical seismology (Schorlemmer et. al, 2010). As a result of the wide variety of models that have been proposed for forecasting future seismicity, earthquake forecasting centers have been established recently, such as the Regional Earthquake Likelihood Model (RELM) working group and the Collaboratory for the Study of Earthquake Predictability (CSEP). Such centers, which require forecasts to be fully automatic, with no subjective or retrospective adjustments made by the modelers, are essential tools to assess the fit of models to observed seismicity and to determine which models seem best suited to earthquake forecasting.

In this study, we construct long-term earthquake forecasts for California seismicity using a Hawkes model with a nonparametrically estimated triggering function. The method for nonparametric estimation was first proposed in Marsan and Lengliné (2008), and in the implementation discussed here, this estimation technique is merged with a flexible, local fault estimation idea similar to that implemented by Helmstetter et al. (2007) and Fox et al. (2016), which allows detailed and locally varying estimates of the spatial distribution of aftershocks. The proposed forecasting model has been submitted to and accepted by CSEP. Here, we discuss this model and assess its performance retrospectively, especially relative to other leading models in CSEP such as Helmstetter et al. (2007).

Hawkes models and their slight variants such as the Epidemic-Type Aftershock Sequence Models of Ogata (1988, 1998) have thus far outperformed their competitors in RELM and CSEP (Schorlemmer et. al, 2010; Zechar, 2013). While such models have traditionally been estimated by maximum likelihood (Ogata, 1978), concerns about model misspecification have focused recent attention on nonparametric estimation methods for Hawkes processes, using a variety of approaches in various different applications. Lewis and Mohler (2011) considered an application of the spread of violence whereas Bacry et al. (2012) developed a nonparametric estimation technique for modeling high frequency fluctuations in financial prices. Building on Fox et al. (2016), we investigate a more detailed formulation of the spatial triggering function which varies with direction, magnitude, and region. First, we estimate a primary fault direction for each earthquake based on prior local seismicity using weighted least squares, and estimate the triggering density of its aftershocks relative to this estimated primary direction. Second, we allow the triggering density to incorporate magnitude scaling, so that the spatial distribution of aftershock activity is dependent not only on distance and direction, but also on the magnitude of the triggering event. Lastly, we allow the spatial triggering function to vary based on subregions within California.

The idea to use spatially varying triggering kernels in a Hawkes model is not new. Indeed, Ogata (1998) noted the benefits of anisotropic spatial response functions in ETAS models, in agreement with the observed elliptical nature of aftershock activity. Additional parametric forms of an anisotropic spatial response function have been proposed, such as the tapered Pareto distribution proposed in Wong (2009). Ogata (1998) also considered the inclusion of a scale factor which allows the spatial distribution of aftershocks to depend on magnitude. Marsan and Lengliné (2008) investigated the inclusion of local fault structure in model estimation but provided little detail on implementation. Fox et al. (2016) used similar methods to show that measuring an aftershock's distance from the triggering event relative to the local faulting of the triggering event, as opposed to epicentral distance, tended to result in a lower estimated background rate. Here we further explore the use of local fault geometry and provide details on implementation for both MISD estimation and forecasting, as well as model evaluation.

The structure of this chapter is as follows. Section 2.2 outlines MISD nonparametric estimation of Hawkes models incorporating local fault directions, evaluates three formulations of the spatial response, and describes forecasting methods. Earthquake data are explained in Section 2.2.6 and in Section 2.3, MISD is applied to two earthquake catalogues. Section 2.4 discusses forecast selection and evaluation. Important subjects for future research are discussed in Section 2.5.

2.2 Methods

2.2.1 MISD Estimation Incorporating Local Fault Direction

As an alternative to fitting a parametric triggering function in a Hawkes process, one may instead estimate the triggering function nonparametrically. In the nonparametric method of Marsan and Lengliné (2008), estimation of a Hawkes model is conducted by maximizing the expectation of the complete data log-likelihood and assigning probabilities that a child event i is caused by an ancestor event j. Given N events in a catalog, these probabilities can be written in matrix form:

$$P = \begin{bmatrix} p_{11} & & & & \\ p_{21} & p_{22} & & & \\ p_{31} & p_{32} & p_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ p_{N1} & p_{N2} & p_{N3} & \cdots & p_{NN} \end{bmatrix}.$$

Here the probabilities for each row i and column j are correspond with triggering probabilities as follows:

$$p_{ij} = \begin{cases} \text{probability earthquake } i \text{ is an aftershock of } j, & i > j \\ \text{probability earthquake } i \text{ is a mainshock,} & i = j \\ 0, & i < j. \end{cases}$$

The rows must sum to 1 ($\sum_{j=1}^{i} p_{ij} = 1$) since each event i must either be caused by one of the previous events $j = 1, ..., i-1$ or be a background event ($i = j$).

Marsan and Lengliné (2008) allowed the spatial-temporal triggering function g to depend on magnitude of the mainshock and on distance from the mainshock relative to local faulting. Here, we allow the triggering function further to depend on the subregion where the events took place, and we assign to each potential mainshock an estimate of its local primary direction of faulting based on weighted regression applied to locally recorded earthquakes, as described in Section 3.2. That is, we consider the spatial-temporal triggering function to depend not only on the magnitude m_j of the triggering earthquake but also on the distance and angular separation from the location (x, y) in question to the triggering event, relative to the estimated local strike angle ϕ_j associated with the triggering event. Thus, we will write the model as

$$\lambda(t, m, x, y|\mathcal{H}_t) = \mu(x, y) + \sum_{j:t_j<t} \kappa(m_j) g(t - t_j) f(x - x_j, y - y_j; \phi_j, m_j), \tag{2.1}$$

and the triggering probabilities can be estimated using essentially the method of Marsan and Lengliné (2008), but where

$$\begin{aligned} p_{ij} &= \frac{g(u) f(x, y; \phi, m)}{\mu(x, y) + \sum g(u) f(x, y; \phi, m)}, \\ p_{ii} &= \frac{\mu(x, y)}{\mu(x, y) + \sum g(u) f(x, y; \phi, m)}. \end{aligned}$$

When the triggering is isotropic, for example, the angle ϕ_j is irrelevant, and one may write the spatial-magnitude portion of the triggering function as

$$f(x - x_j, y - y_j; \phi_j, m_j) = h(r; m_j).$$

More generally, in the case where the triggering may depend on the distance to the triggering event as well as the angle relative to the local fault direction of the triggering event, one may write this portion of the triggering function as

$$f(x - x_j, y - y_j; \phi_j, m_j) = h(r, \theta; m_j),$$

where θ is the difference between the angle of the segment connecting (x, y) and (x_j, y_j) and the angle ϕ_j, and $\int\int_A f(x, y)dA = \int_0^\infty \int_0^{2\pi} f(r\cos\theta, r\sin\theta)r\ drd\theta = \int_0^\infty \int_0^{2\pi} h(r, \theta)drd\theta = 1$.

2.2.2 Local Fault Estimation Using Weighted Least Squares

A Hawkes model specified according to (2.1) requires estimation of each event's local strike angle. Because focal mechanism estimates are not available for all earthquakes and are frought with large errors (Kagan and Jackson 2014), we instead propose the estimation, for each event j in the catalog, of the corresponding primary slope ϕ_j associated with its aftershocks, using observed local seismicity.

Specifically, in order to estimate the primary angle of faulting for earthquake j, we use the entire catalog of events, including very small earthquakes, occurring both before and after earthquake j. One may then calculate the Euclidean distance r_{ij} between the estimated epicenter of earthquake j and that of any other earthquake i, and fit a weighted least squares regression, constrained to pass through the epicenter of earthquake j, and with weights $w_i = 1/r_{ij}$, to produce an estimate ϕ_j of the slope of the anticipated primary direction of aftershock activity associated with earthquake j. Thus, events located further away from earthquake j are given proportionally less weight in determining the local angle of faulting associated with earthquake j. If any event i has exactly the same recorded epicenter estimate as the mainshock j, then we remove the event i from this regression.

One may then define, for mainshock j and any subsequent potential aftershock, i, the angular separation between i and j relative to the slope ϕ_j. That is, one may calculate θ_{ij}, the angle formed by the line through (x_j, y_j) with slope ϕ_j and the segment connecting mainshock epicenter (x_j, y_j) to aftershock epicenter (x_i, y_i). Letting ϕ_{ij} denote the slope of this latter segment, the angle θ_{ij} is given simply by

$$\theta_{ij} = \arctan(\phi_i - \phi_j).$$

As in Wong (2009), we may consider the *m*inimal angle separating earthquakes i and j, and thus the estimated angle θ_{ij} can be constrained to the quarter circle, $\left[0, \dfrac{\pi}{2}\right]$. Details of incorporating these estimates of θ_{ij} into the MISD algorithm are given in Appendix A.

The degree to which estimated epicenters of California earthquakes follow locally linear faults can vary substantially. For instance, three examples of fault estimates with varying degrees of scatter are shown in Figure 2.1. Here the weighted regression R^2 values vary, descending from 95% to 50% to 10%.

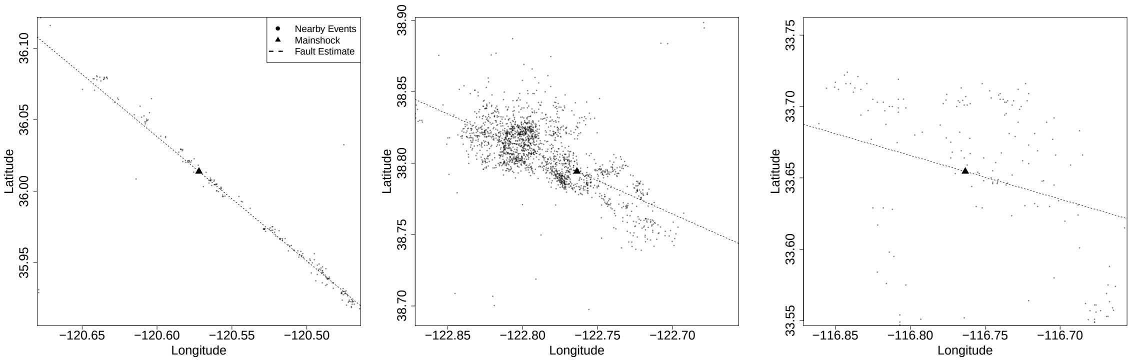


Figure 2.1: Three examples of fault estimates with R^2 values of (a) 95%, (b) 50%, and (c) 10%. Mainshock strike angles are estimated using events that occur both before and after the mainshock, as well as all events $m \geq 2.5$. Events used to estimate strike angles were in a window of $0.1°(121\text{km}^2)$ around the mainshock.

2.2.3 Spatial Triggering Performance Verification

As a proof of concept and to illustrate the performance of the MISD estimation method specified by (2.1), even in highly nonstandard conditions, we repeatedly simulate a parametric Hawkes model under various formulations of the model, and verify that the estimates recover the appropriate functional form of the triggering density. We proceed by first assessing how well the method is able to recover a spatially anisotropic triggering function. Second, we investigate how well MISD is able to recover a magnitude-dependent spatial triggering density. Third, we partition the space into subregions and verify the ability of MISD to recover different spatial triggering densities, each associated with a different subregion of the observation window.

To confirm the ability of MISD to estimate an anisotropic triggering density, earthquake occurrences were simulated from Hawkes models with a parametric triggering function. One example of an anistropic triggering function is a bivariate normal density with rotation $\theta = 5\pi/12$, as shown in Figure 2.2(a). Here the spatial triggering function $h(r, \theta)$ is evaluated at bin centers for both distance (degrees) and direction (radians). For convenience of visualization, one realization of a process corresponding to this triggering density is shown in Figure 2.2(b). Background points ($\circ$) are drawn uniformly in the spatial region $[0, 10] \times [0, 10]$ over time $[0, 5000]$ and triggered events are overlaid ($+$). In the simulations, earthquake magnitudes are generated independently of other model components according to an exponential (Gutenberg-Richter) distribution with b-value equal to 1 (see e.g. Utsu 1999 for details and a history of the investigation into the distribution of earthquake sizes). The process is repeated 200 times using the above specifications. The means of the 200 corresponding isotropic estimates based on (1.2) are shown in Figure 2.2(c). Here the estimates $\hat{h}(r)$ are constrained to have equal probability of triggering in all directions, and thus cannot accurately recover the true triggering function, whereas the mean of the 200 corresponding anisotropic estimates of $\hat{h}(r, \theta)$ using (2.1) quite accurately recovers the shape of the true triggering density shown in Figure 2.2(a).

Next we assess how well the MISD algorithm recovers a spatial density that depends on the magnitude of the preceding event. Specifically, earthquake occurrences are simulated from a Hawkes model with a single parametric triggering function, specified by (1.3), where for any earthquake of magnitude in $(0, 1]$, the parameters governing its aftershocks are (d, q) = (0.00171, 1.960), and otherwise the parameters governing aftershock activity are (d, q) = (0.05010, 2.721). 200 realizations are generated in the spatial region $[0, 6] \times [0, 4]$ over time $[0, 5000]$, and one such simulation is shown in Figure 2.3(a). 200 estimates of $\hat{h}(r|m)$

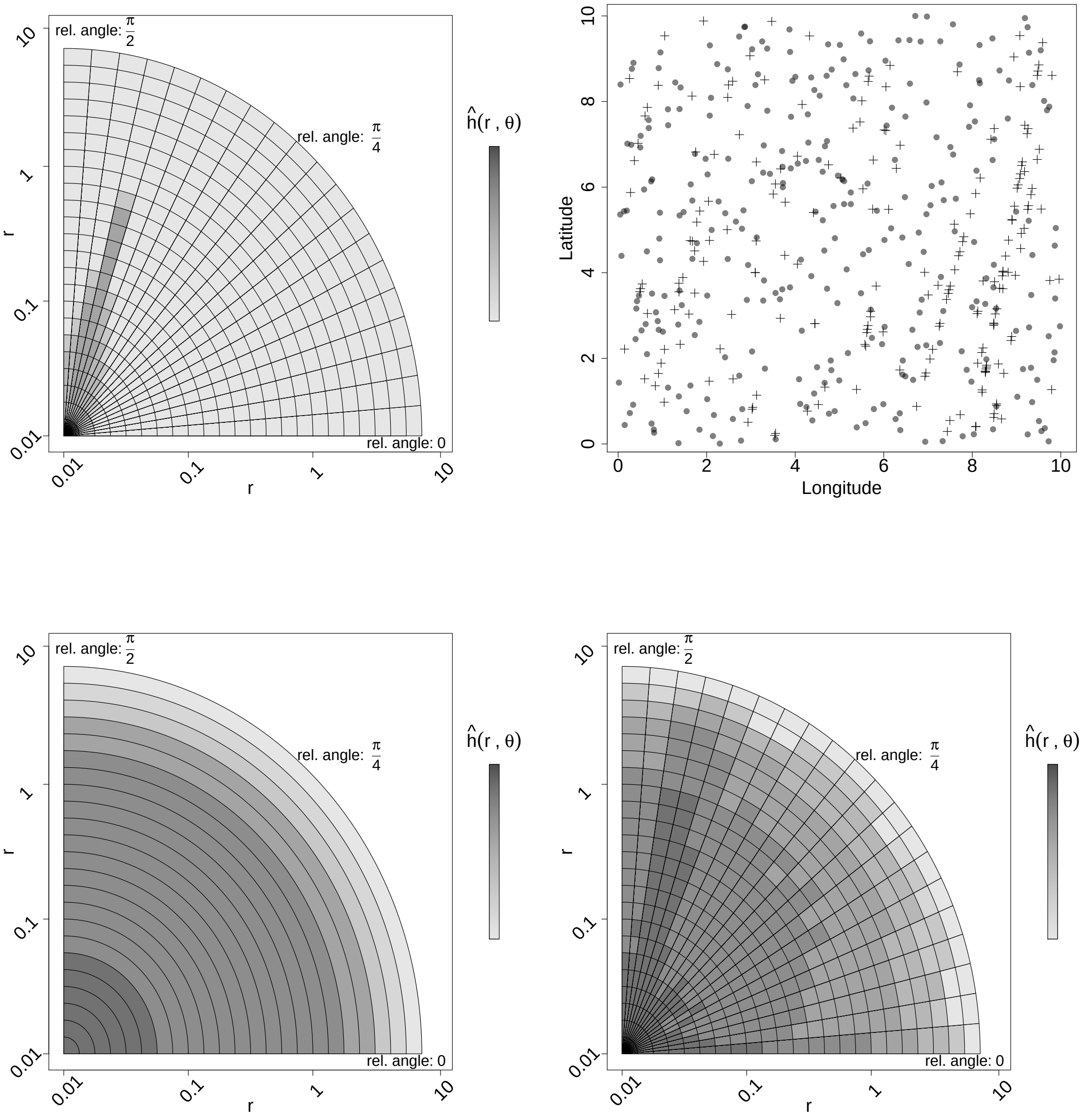


Figure 2.2: (a) The true spatial triggering function $h(r, \theta)$ used in simulation, evaluated at bin centers. The distribution is a bivariate normal with rotation $\theta = 5\pi/12$. (b) One realization of a simulated Hawkes process. Background events ($\circ$) are distributed uniformly in space and time. Triggered events ($+$) are simulated from the bivariate normal distribution with rotation $\theta = 5\pi/12$ shown in (a). (c) Mean estimated triggering function $\hat{h}(r)$ over 200 simulations using MISD with isotropic triggering as specified in (1.2). (d) Mean estimated spatial triggering function $\hat{h}(r, \theta)$ over 200 simulations using MISD with anisotropic triggering as specified in (2.1).

constructed using the MISD algorithm described in Section 3.1 are shown in Figure 2.3(b,c). The dashed(b) and dotted(c) curves in Figures 2.3(b) and 2.3(c) represent the true spatial triggering functions used in the simulations. Despite the fact that the triggering densities are mixed together in the simulation, the MISD algorithm appears to be able to recover the two spatial triggering densities quite accurately.

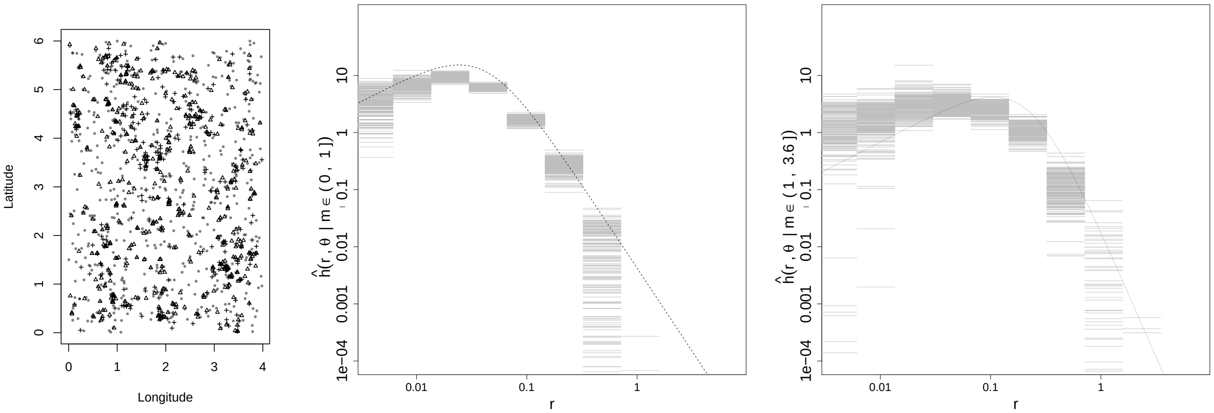


Figure 2.3: (a) One instance of a simulated Hawkes process used to evaluate the performance of the isotropic (1.2) and anisotropic (2.1) MISD algorithms under magnitude scaling conditions. Background events ($\circ$) are distributed uniformly in space and time. Aftershocks of events with magnitude in $(0, 1]$ are labeled $+$ and are generated according to (1.3) with parameters $(d; q) = (0.00171;\ 1.960)$, and aftershocks of events with magnitude greater than 1 are labeled $\triangle$ and are generated according to (1.3) with parameters $(d; q) = (0.05010;\ 2.721)$. (b,c) Estimated magnitude-dependent spatial triggering functions $\hat{h}(r|m)$ resulting from 200 simulations with triggering function specified by (1.3), for mainshocks with magnitude in $(0, 1]$ and for mainshocks with magnitude greater than 1, shown in panel (b) and (c), respectively. The dashed curve (b) represents the true triggering function for mainshocks with magnitudes in $(0, 1]$, and the dotted curve (c) represents the true triggering function for mainshocks with magnitudes greater than 1. In these simulations $n_\theta^{bins} = 1$.

A third proof of concept involves the situation where the spatial-temporal triggering function varies spatially from region to region. The spatial region is partitioned into two regions, North and South, and events are simulated according to (1.3). Figures 2.4(b) and 2.4(c) reveal and that both spatial response functions appear to be recovered quite accurately by the MISD algorithm. Even under such nonstandard conditions, the method described in Section 2.2.1 appeared to perform well in recovering the triggering densities used to generate the synthetic catalogues in all three proofs of concept.

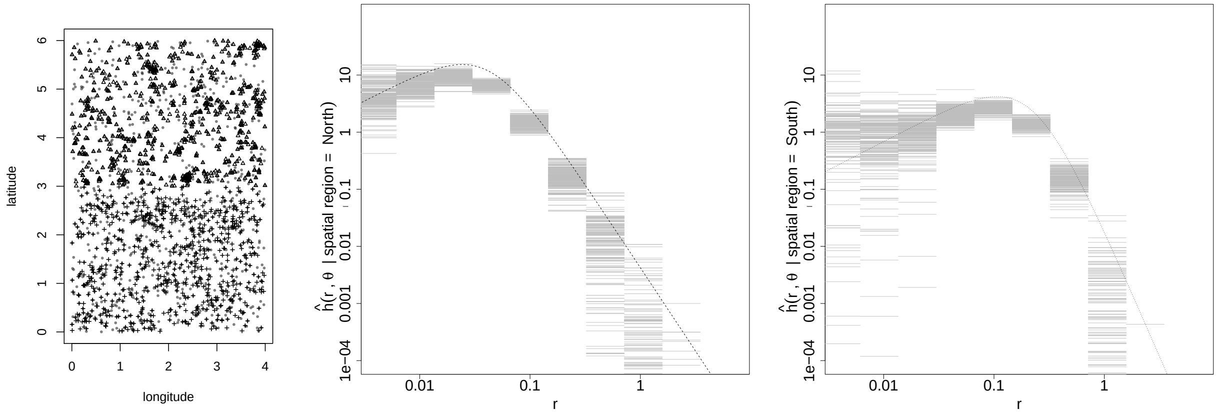


Figure 2.4: (a) One instance of a simulated Hawkes process used to evaluate MISD with anisotropic triggering function according to (2.1). Background events ($\circ$) are distributed uniformly in space and time. Triggered events in the Northern region are labeled $\triangle$ and those in the Southern region are labeled $+$. (b,c) Estimated region-dependent spatial triggering function $\hat{h}(r|Z)$ for 200 simulations with two sets of parameters, $(d;q) = (0.00171; 1.960)$ and $(d;q) = (0.05010; 2.721)$ corresponding to the Northern and Southern regions, respectively. The dashed(b) line represents the true triggering function from which aftershocks were simulated in the Northern region while the dotted(c) line represents the true triggering function from which aftershocks were generated in the Southern region. In this simulation $n_{\theta}^{bins} = 1$.

2.2.4 Forecasting Methods

The fitted model (2.1), using the methods of Section 3.1 and 3.2, can be used to forecast future seismicity. Under paradigms such as CSEP, the goal is to predict the number of earthquakes occurring within spatial-temporal-magnitude bins $(t, t+\Delta t] \times S_i \times (m_j, m_{j+1}))$, for each i and j, given only information on seismicity up to time t. A common approach for computing the conditional expectation of the number of events in such a bin, given only information up to time t, is to use the integral of the conditional intensity:

$$\Lambda([t, t+\Delta t] \times S_i \times (m_j, m_{j+1})) \;\; = \int\limits_{m_j}^{m_{j+1}} \int\limits_{x,y\in} \int\limits_{S_i} \int_t^{t+\Delta t} \lambda(t, x, y|\mathcal{H}_t)\ \mathrm{d}t\ \mathrm{d}x\ \mathrm{d}y\ \mathrm{d}m,$$

and approximating $\lambda(t, x, y|\mathcal{H}_t)$ using the value it would take if no additional events occurred in the time interval $(t, t + \Delta t)$. While this approximation may be satisfactory in some situations such as when Δt is extremely small, it is unlikely to be a close approximation over a multi-year period, such as a 5 year CSEP forecast. The result of using such an approach would be an underestimation in total seismicity, especially if the forecast was generated during a period of active triggering. Given this limitation we believe a simulation approach may be preferred. Following Zhuang (2011), we construct forecasts of seismicity by first simulating events from the estimated background rate using the thinning procedure of Lewis and Shedler (1979) and subsequently adding in triggered offspring events as in Fox et al. (2016). This process is repeated $K = 10,000$ times in order for the simulations to sample a wide range of possibilities for the seismicity after time t (Zhuang, 2011; Werner et. al, 2011). Implementation details are provided in Appendix B.

Once synthetic catalogs are generated for $[t, t + \Delta t]$, we smooth events according to local fault directions while simultaneously smoothing their magnitude distribution. Since CSEP requires the spatial distribution be estimated over equally sized cells, we choose 150×150 $0.1^\circ \times 0.1^\circ$ cells which are identical to those used in the background rate estimation. Magnitude is smoothed in bins of 0.1 magnitudes units ranging from 3.95 to 8.95, plus an additional bin for magnitudes 8.95 - 10.

Anisotropic smoothing potentially has the advantage of an improvement in forecast accuracy, especially when applied to a catalog generated by (2.1). Anisotropic smoothing can be achieved by kernel smoothing the spatial (vector) distances between each earthquake and each subsequent earthquake, relative to the estimated principal strike angle of the prior event. These principal directions associated with each event are estimated by the weighted least squares approach discussed in Section 2.2.1. That is, we define an anisotropic smoothing kernel estimate via

$$\begin{gathered} P_m(\vec{r}, m) = \sum_{i=1}^{N} K_d(\vec{r} - \vec{r}_i, \theta_i) G_h(m - m_i) \\ \theta_i = \arctan\left(|\phi_i|\right), \quad \theta_i \in [0, \pi/2] \end{gathered} \tag{2.2}$$

with

$$K_{d_i}(\vec{r}, \theta) = \frac{1}{2\pi d_1 d_2 \sqrt{1-\rho^2}} \exp\left(-\frac{z}{2(1-\rho)^2}\right),$$

and

$$G_h(m) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(m-m_i)^2}{2h^2}\right). \tag{2.3}$$

In (2.2), the rotation angle θ_i is simply the angle between the estimated strike angle ϕ_i of earthquake i and the principal horizontal axis. (2.3) is an isotropic Gaussian function with mean m and standard deviation h. As in Helmstetter et al. (2007), h was fixed to 0.15 magnitude units.

2.2.5 Model Evaluation

To evaluate model performance we use both Voronoi residuals (Bray et al., 2014) and deviance residuals (Clements et al. 2011). Voronoi residuals and Voronoi deviances are useful for evaluating gridded forecasts especially when a substantial proportion of pixels have very small integrated conditional intensities. Furthermore, Voronoi based residual methods offer advantages over grid based residuals in that with the former type of residuals, the spatial partition is data-driven and spatially adaptive, and the resulting distribution of residuals is usually far less skewed in such situations than residuals integrals over fixed rectangular grid cells (Bray et al. 2014).

Voronoi residuals are constructed by partitioning the space using a Voronoi tessellation. For any point in a point pattern of N observed events in the forecast evaluation period, one may define its corresponding Voronoi cell as the region consisting of all locations that are closer to the observed event than to any of the other $N-1$ points. A Voronoi tessellation is the collection of such Voronoi cells. Details on Voronoi tessellations and their properties are found in Okabe (2000).

Each Voronoi cell C_i has only one point inside it by construction. Hence, a raw Voronoi residual for each cell C_i is given simply by

$$\begin{aligned}\hat{R}_i &:= 1 - \int\limits_{C_i} \hat{\lambda}\,\mathrm{d}\mu \\ &= 1 - |C_i|\bar{\lambda}, \end{aligned} \tag{2.4}$$

where $\bar{\lambda}$ denotes the mean of the proposed conditional intensity estimate, $\hat{\lambda}$, over C_i. Since standardizing residuals often improves their usefulness (Baddeley et al., 2005), Bray et al. (2014) proposed rescaled Voronoi residuals of the form

$$\hat{R}_{\mathrm{V}}(C_i) = \frac{1 - \int \hat{\lambda}(t,x,y)\ \mathrm{d}t\ \mathrm{d}x\ \mathrm{d}y}{\sqrt{\int \hat{\lambda}(t,x,y)\ \mathrm{d}t\ \mathrm{d}x\ \mathrm{d}y}}.$$

In order to compare two competing models, a useful criterion is the deviance, which is the ratio of the two corresponding local loglikelihoods, evaluated over spatial-temporal bins. The fit of competing point process models can be readily compared using pixel based deviances proposed by Wong and Schoenberg (2009), which share similarities with deviances defined for generalized linear models in the regression framework. Deviances may be computed over evenly spaced pixels, but instead of simply comparing the observed to the forecasted seismicity within each pixel, the difference between the log-likelihoods of two competing models is examined. That is, the deviances in a given bin, B_i, given two models for the conditional intensity, $\hat{\lambda}_1$ and $\hat{\lambda}_2$, is calculated as follows:

$$\begin{aligned} R_{\mathrm{D}}(B_i) &= \sum_{i:(t_i,x_i,y_i)\in B_i} \log\left(\hat{\lambda}_1(t_i,x_i,y_i)\right) - \int_{B_i} \hat{\lambda}_1(t,x,y)\ \mathrm{d}t\ \mathrm{d}x\ \mathrm{d}y \\ &- \left(\sum_{i:(t_i,x_i,y_i)\in B_i} \log\left(\hat{\lambda}_2(t_i,x_i,y_i)\right) - \int_{B_i} \hat{\lambda}_2(t,x,y)\ \mathrm{d}t\ \mathrm{d}x\ \mathrm{d}y \right). \end{aligned}$$

A positive residual implies the model $\hat{\lambda}_1$ fits better in the given pixel and negative residuals imply that $\hat{\lambda}_2$ provides a better fit. Of course, to get an overall view of which model fits better we can take sum of the deviances, $\sum_i R_{\mathrm{D}}(B_i)$ and obtain a log-likelihood ratio score.

2.2.6 Earthquake Data

California is populated by a large number of faults, and these faults tend to be highly locally fractured and heterogeneous (Bolt, 2006), which makes the region an ideal location for the assessment of Hawkes models with anisotropic spatial triggering as in model (2.1). To estimate the anisotropic spatial triggering function for California seismicity, construct forecasts for $m \geq 4.95$ seismicity, and to compare our forecasts with e.g. those of Helmstetter et al. (2007), we use the same catalog as Helmstetter et al. (2007), made available by the Advanced National Seismic System (ANSS) at http://www.ncedc.org/anss/catalog-search.html. The ANSS comprehensive catalog contains information on estimated origin times, magnitudes, and origin locations of many thousands of seismic events dating back several decades; for details on the catalog and its completeness see Kagan (2003) and Kagan and Jackson (2014). We consider shallow (depth ¡ 30km) events in the period from 1 January 1981 until 23 Aug 2005 with a lower magnitude cutoff of $m \geq 2.5$ and a spatial window of 15° by 15° region covering California, from Lon −127.0°W to Lon −112.0°W and Lat 29.0°N to Lat 45.0°N. Locally linear features representing faults are clearly visible, such as those at Lat 36.0°N, Lon −121.0°W. The estimated epicenters of these 71,662 recorded events are shown in Figure 2.5(a) and the distribution of magnitudes for the time window is shown in Figure 2.5(b). The catalog is considered complete for magnitudes above 3.0 (Werner et. al, 2011) and while we use the entire catalog of earthquakes with $m > 3.0$ in fitting (2.1), we also use smaller events, including those of $m \leq 3.0$, for the local fault estimation described in Section 2.2.2.

16,569 of the 71,662 recorded events have magnitude greater than 3.0 and 5,159 fall outside the window specified as the official Collaboratory for the Study of Earthquake Predictability (CSEP, www.cseptesting.org) testing region. Of the events within the CSEP testing region, 184 events have magnitude $\geq$ 4.95. Unlike Helmstetter et al. (2007), when forecasting seismicity we do not remove potential explosions from the learning catalog, to avoid any subjective data processing issues, and this results in a slightly higher overall predicted number of events during the forecast period. We also make no parametric assumption regarding

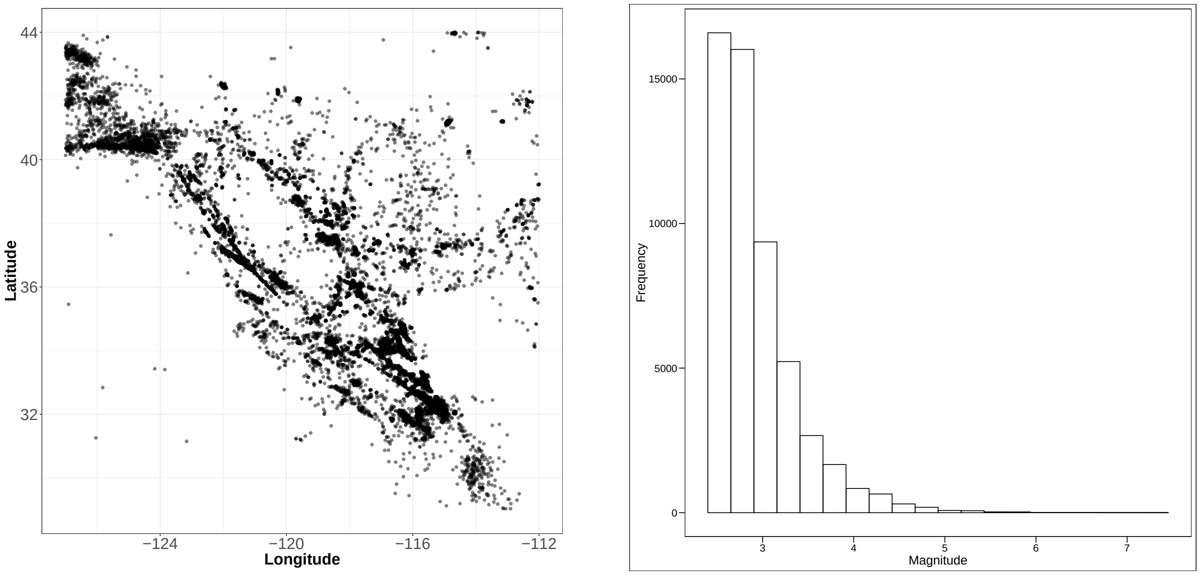


Figure 2.5: (a) 71,662 events from the ANSS catalog falling within the the spatial window, a 15 degree by 15 degree area surrounding California from Lon −127.0°W to Lon −112.0°W and Lat 29.0°N to Lat 44.0°N. (b) The density of magnitudes in the ANSS catalog for the temporal window 1 January 1981 until 23 Aug 2005 with a lower magnitude cutoff of $m \geq 2.5$.

the magnitude distribution, such a tapered Gutenberg-Richter law, and instead use of the observed empirical magnitude distribution shown in Figure 2.5(b). Hence the largest event possible in our forecasts is equivalent to the largest in the dataset, which is magnitude 7.31. In assessment of the spatial triggering function in Section 2.3, we also make use of the U.S. Geological Survey database (http://earthquake.usgs.gov/hazards/qfaults/google.php) on faults in the United States that are believed to be sources of $m > 6$ earthquakes during the Quaternary period (the past 1,600,000 years).

2.3 Application to California seismicity

2.3.1 Estimation of Isotropic and Anisotropic Triggering Functions

The estimated isotropic spatial triggering function of (1.2) applied to all magnitude greater than 3.0 events in the ANSS catalog. The resulting triggering function is shown in Figure 2.6(a), on log-log scale for ease of visualization. The fit resembles a power-law (Pareto) distribution. The isotropic, parametric power-law triggering function specified by 1.3, with parameters $(d; q) = (1.238\text{e-}05; 1.244)$, estimated simply by nonlinear least squares, is overlaid for reference. The dot within each bin represents the estimate $\hat{h}(r)$ while the error bars approximate 95% confidence intervals as derived in Fox et al. (2016).

Using this isotropic estimate in (1.2) to decluster the ANSS catalog, we estimate 4,132.44 of the 16,569 events to be background events, yielding an overall background rate of 24.93%. The remaining components of the triggering function and background rate are shown in Figure 2.6(b-d). Figure 2.6(b) shows the estimate of the spatial background rate $\mu(\mathbf{x}) = \mu(x, y)$ for model (1.2). Figure 2.6(c,d) shows the estimated magnitude productivity and temporal triggering components. The estimated magnitude productivity $\hat{\kappa}(m)$ appears approximately exponential while the estimated temporal component of the triggering function resembles a power law.

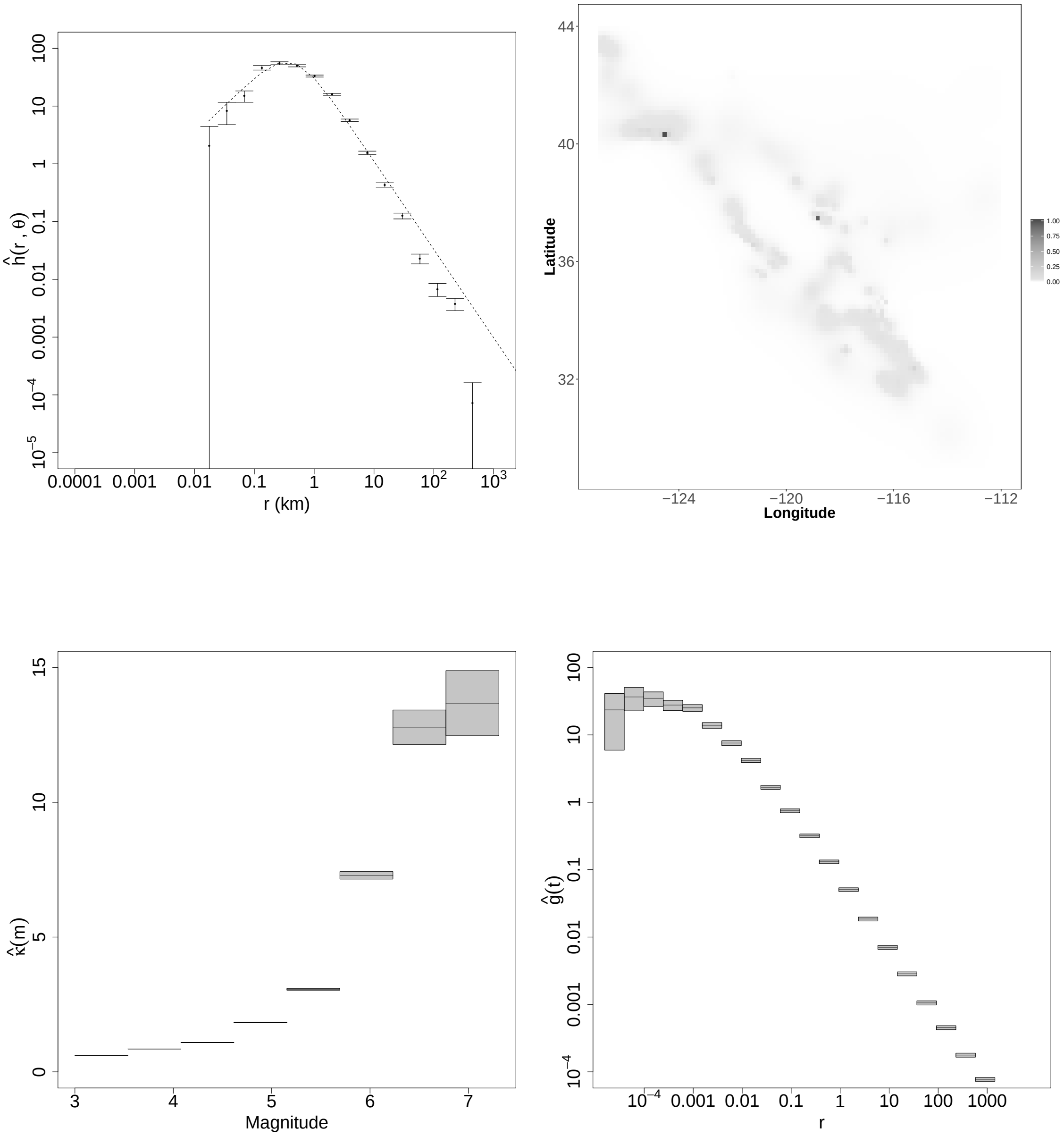


Figure 2.6: (a) Estimated spatial triggering function $\hat{h}(r)$ for ANSS data using MISD assuming isotropic triggering (1.2), and using 25 distance bins. The black dot in each bin represents the estimate $\hat{h}(r)$ while the error bars show approximate 95% confidence intervals as derived in Fox et al. (2016), based on the sampling variation for the histogram estimates in each bin. Parametric density (1.3) is overlaid for reference with parameters $(d; q) =$ (1.238e-05; 1.244) estimated by nonlinear least squares through bin centers. (b) Spatial distribution of estimated background rate for (1.2) using a 150 × 150 grid of $0.1^\circ \times 0.1^\circ$ cells. (c) Estimated magnitude productivity and (d) temporal triggering components using MISD, with 95% errors bars.

In order to investigate spatial inhomogeneity in the triggering function, an anisotropic MISD estimate fitting equation (2.1) was fit to the catalog of ANSS events with $m \geq 3.0$ and the results are shown in Figure 2.7.

The estimate of the spatially anisotropic triggering, shown in Figure 2.7(a), reveals few noticeable departures from isotropy. However, for many of the bins, the estimate in Figure 2.7(a) is based on very few pairs of points, and thus has large sampling error. A more powerful test for anisotropy is obtained by fitting an anisotropic estimate of the spatial triggering function using only 2 directional bins, as shown in Figures 2.7(b) and (c).

Figure 2.7(b) indicates anisotropy, as the estimated spatial density of aftershocks is more concentrated at lower distances for large relative angles θ, measured relative to the estimated strike angles of the mainshocks. Although the shapes of the estimated densities $\hat{h}(r, \theta)$ are similar and roughly power-law shaped for $\theta \in (0, \pi/2]$ degrees compared to $\theta \in (\pi/2, \pi/4]$ degrees, as shown in Figure 2.7(c), the anisotropy is indeed statistically significant, particularly for distances r between 5.0 and 50.0 km. A χ^2 test comparing the two densities reveals the differences are highly significant in fact ($\chi^2_{13} = 85.67, p = 9.773\text{e}{-13}$).

When stochastically declustering the catalog based on the estimated spatial triggering density in Figure 2.7(b), an estimated 3,874.59 of the 16,569 events are background events, corresponding to an overall background rate of 23.39%. Figure 2.7(d) shows the spatial distribution of this estimated background rate for (2.1). The corresponding estimated magnitude productivity and temporal component of the triggering function in Figures 2.7(e,f) are similar to those estimated assuming isotropy.

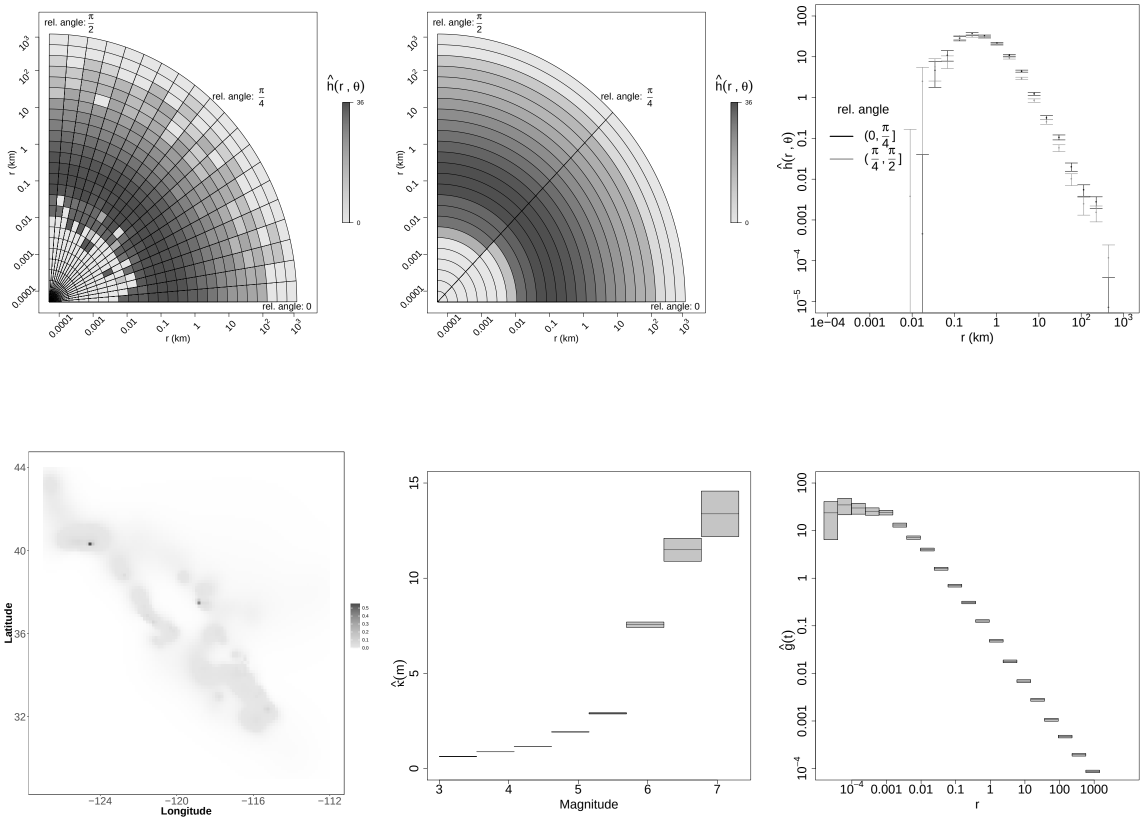


Figure 2.7: (a) Estimated spatial triggering density $\hat{h}(r, \theta)$ using MISD assuming anisotropy (2.1), fit to the ANSS catalog and estimated using 25 logarithmically spaced distance bins and 20 direction bins. (b) Estimated anisotropic spatial triggering density $\hat{h}(r, \theta)$ fit to the ANSS catalog using 25 logarithmically spaced distance bins and 2 direction bins. (c) Estimated anisotropic spatial triggering density $\hat{h}(r, \theta)$ corresponding to (b), along with 95% confidence bounds. The density for θ in $(0, \pi/4]$ and the density for θ in $(\pi/4, \pi/2]$ are significantly different ($\chi^2_{13} = 85.568, p = 9.773\text{e}{-13}$). (d) Estimated spatial background rate for (2.1) using a 150 × 150 grid of 0.1° × 0.1° cells. (e) Estimated magnitude productivity and (f) estimated temporal triggering components with 95% errors bars.

2.3.2 Estimation of Spatial Triggering Functions

In order to investigate the dependence of the spatial triggering function on the magnitude of the mainshock, we fit an anisotropic triggering function to the ANSS earthquake catalog using MISD, where now the spatial distribution was permitted to depend on the magnitude of the triggering earthquake. The result is shown in Figure 2.8(a). There is a statistically significant difference ($\chi^2_{12} = 1732.7, p < 1\text{e}{-32}$) between the spatial triggering functions for smaller mainshock magnitudes versus larger mainshock magnitudes, with the larger earthquakes having significantly more diffuse aftershock regions. The result is not surprising and is in agreement with previous observations of magnitude scaling (see e.g. Kanamori and Anderson 1975 or Kagan 2002).

We similarly fit the spatial distribution after partitioning the space into two regions, North and South, based on latitude. The estimated spatial triggering function $\hat{h}(r, \theta|\text{Region})$ is shown in Figure 2.8(b). There is a relatively small but statistically significant difference between the estimated spatial triggering function for Northern California mainshocks compared to Southern California mainshocks, with Southern California aftershock zones significantly more diffuse ($\chi^2_{11} = 151.51, p < 2.805\text{e}{-26}$). Next, we fit an anisotropic spatial triggering function (2.1) after partitioning mainshocks into two categories based on their positions relative to known faults. Any event within 27km of a known fault was classified as occurring along a fault, and Figure 2.8(c) shows the corresponding estimated spatial triggering function $\hat{h}(r, \theta|\text{Region})$ for such events and their complement. The mainshocks occurring near known faults had significantly less diffuse aftershock regions ($\chi^2_{13} = 1613.9, p < 1\text{e}{-32}$). However, for mainshocks along faults and those further from known faults, the estimated spatial triggering function resembles a power law. Combining the above approaches, we allow the spatial triggering to be impacted by both region and magnitude of the mainshock, and the resulting four spatial triggering estimates are depicted in Figure 2.8(d). Smaller mainshocks have larger aftershock densities at small distances, whether the mainshocks occur along a fault or not. However, the estimated aftershock zone is noticeably more diffuse for large mainshocks away from known faults compared to large mainshocks along known faults. Overall, the differences among the four estimated distributions are highly significant ($\chi^2_{30} = 3360.1, p < 1\text{e}{-32}$).

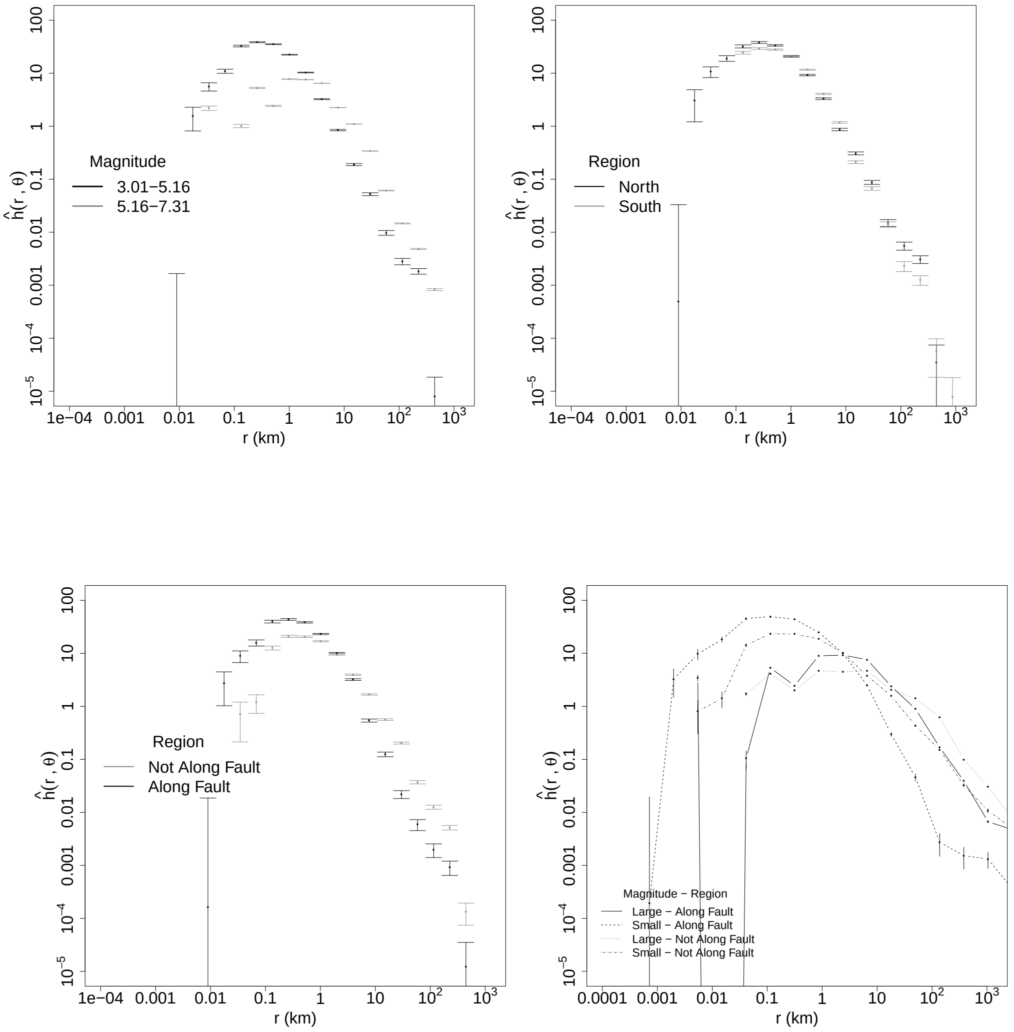


Figure 2.8: (a) Estimated anisotropic spatial triggering density as a function of magnitude, $\hat{h}(r|m \in (\text{Large}, \text{Small}))$, fit to the ANSS catalog by MISD and estimated using 25 logarithmically spaced distance bins and two magnitude bins. Approximate 95% confidence intervals are also shown. The estimated spatial density for smaller mainshocks and the estimated spatial density for larger mainshocks are significantly different ($\chi^2_{12} = 1732.7, p < 1\text{e}{-32}$). (b) Estimated anisotropic spatial triggering density as a function of magnitude and region, $\hat{h}(r|m, \text{Region} \in (\text{North}, \text{South}))$, with approximate 95% confidence bands. The estimated densities for North and South are significantly different from one another ($\chi^2_{11} = 151.51, p < 2.805\text{e}{-26}$. (c) Estimated anisotropic spatial triggering density as a function of magnitude and proximity to a known fault, $\hat{h}(r, \theta|\text{Region} \in (\text{Fault}, \text{No Fault}))$, with approximate 95% confidence bands. The estimated spatial densities for mainshocks near known faults and for mainshocks far from known faults are significantly different ($\chi^2_{13} = 1613.9, p < 1\text{e}{-32}$). Estimated spatial triggering function (d) $\hat{h}(r, \theta|m, \text{Region} \in (\text{Fault}, \text{No Fault}))$ with approximate 95% confidence bands. The 4 estimated densities are significantly distinct ($\chi^2_{30} = 3360.1, p < 1\text{e}{-32}$).

2.4 Model Selection and Forecast Evaluation

A primary goal of this study is to determine if a nonparametric Hawkes model estimated by MISD, coupled with a nonparametric forecast, can outperform the CSEP model introduced in Helmstetter et al. (2007). We chose Helmstetter et al. (2007) since it performed best among those submitted during that time period (Zechar, 2013). Indeed, Schorlemmer et. al (2010) indicated that both the mainshock and mainshock-aftershock models of Helmstetter et al. (2007) performed best in their classes by the L- and N-tests, and in comparisons with every other model using the R-test, each time the other model was rejected in favor of the Helmstetter et al. (2007) model.

To compare how well the Hawkes model estimated by MISD performs retrospectively compared with Helmstetter et al. (2007), we applied both (1.2) and (2.1) to the ANSS catalog specified by CSEP. Unlike Fox et al. (2016), we chose to use a fixed bandwidth, as opposed to an adaptive bandwidth, while smoothing the background rate during model estimation. Further detail regarding the tuning of inputs for these models is covered in Appendix B.

Example forecasts using the selected model inputs in Table 2.1 are shown in Figure 2.9, which depicts the mean forecasts generated from isotropic MISD, anisotropic MISD, and from Helmstetter et al. (2007), respectively. The anisotropic and isotropic MISD forecasts are very similar to one another, and clearly much smoother than that of Helmstetter et al. (2007). Figures 2.9(d-f) show the corresponding distributions of forecasted rates in the 7682 cells of the CSEP testing region. Helmstetter et al. (2007) has a higher concentration of cells with medium rate forecasts where the nonparametric models proposed here have more cells with low forecasted rates. In terms of the number of events forecasted, both our forecast and that of Helmstetter et al. (2007) overestimated total seismicity, as only 31 $m \geq 4.95$ events were actually recorded in the temporal window of 24 August 2005 to 24 Aug 2010, compared to the forecasted estimates of 37.33 using MISD and 35.40 from Helmstetter et al. (2007).

Table 2.1: Selected Model and Forecast Inputs for (1.2) and (2.1). Full descriptions for variables are provided in Appendix Section 7.

Model	n_r^{bins}	n_θ^{bins}	n_m^{bins}	n_{m2}^{bins}	n_t^{bins}	n_x^{bins}	n_y^{bins}	background bandwidth	forecast bandwidth	forecast bandwidth ratio	mag_cut$_{\text{fit}}$	mag_cut$_{\text{fault}}$	Deviance
(1.2)	25 bins		8 bins		25 bins	150 bins	150 bins	0.08 degrees	0.08 degrees		3.50		4.99
(2.1)	25 bins	7 bins	8 bins	2 bins	25 bins	150 bins	150 bins	0.08 degrees	0.08 degrees	3/4	3.50	2.50	6.38

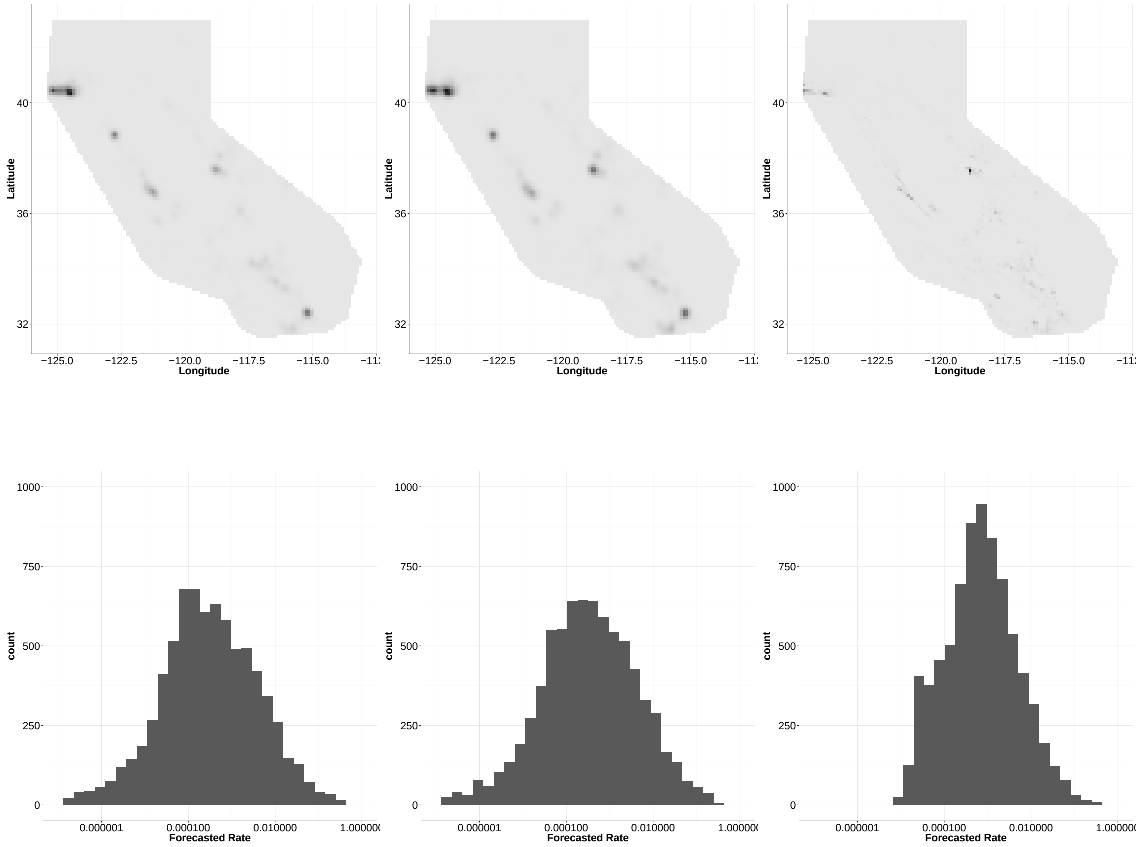


Figure 2.9: (a,b) Retrospective forecast for 24 August 2005 to 24 Aug 2010 using isotropic MISD (1.2) and anisotropic MISD (2.1), respectively, with $N_{pred} = 37.33$ events. (c) CSEP forecast for Helmstetter et al. (2007) with $N_{pred} = 35.4$ events. (d,e,f) Histograms of forecasted rates in the 7682 cells of the CSEP testing region for isotropic MISD (1.2), anisotropic MISD (2.1), and Helmstetter et al. (2007), respectively. Rates on the x-axis are shown on a logarithmic scale.

2.4.1 Forecast Evaluation

Figure 2.10 shows the Voronoi residuals of the isotropic and anisotropic MISD models, as well as the Voronoi residuals for Helmstetter et al. (2007). The Voronoi residuals for a null homogeneous Poisson model are shown for comparison in Figure 2.10(a). For the null model, the rate is constant, so small cells have large positive residuals (horizontal lines) which indicate underprediction of seismicity, and large cells have large negative residuals, indicating overprediction of seismicity (Gordon et al., 2015). In Figure 2.10 one sees the MISD and Helmstetter et al. (2007) forecasts performed similarly to the null model in the largest cell. The MISD and Helmstetter et al. (2007) models underpredicted seismicity in the southernmost part of the forecast area, where a cluster of large events occurred on 4 April 2010 near Baja California along the Laguna Salada fault. In the largest cell in the Northern region of the forecast area near Eureka, CA, Helmstetter et al. (2007) has a relatively large negative residual compared with the forecasts (1.2) and (2.1), indicating a relative overprediction of seismicity in this region.

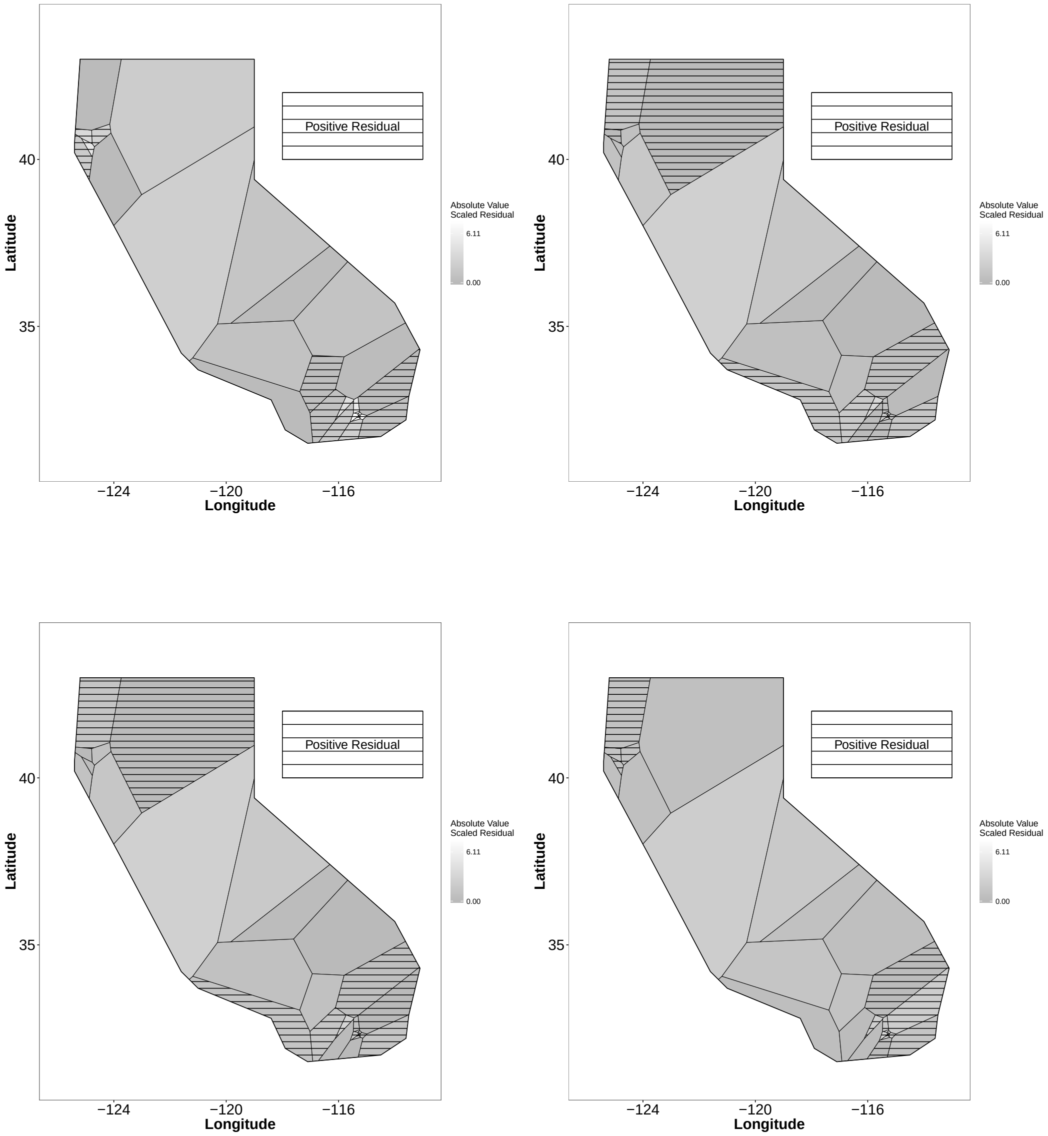


Figure 2.10: Voronoi residuals for (a) Null Poisson model, (b) Hawkes model estimated using isotropic MISD (1.2), (c) Hawkes model estimated using anisotropic MISD (2.1), and (d) Helmstetter et al. (2007). Striped shells indicate positive residuals and solid cells indicate negative residuals, with lighter shading indicating larger absolute values of the residuals.

Figure 2.11 shows the pixel deviances for the isotropic MISD (1.2), anisotropic MISD (2.1), and Helmstetter et al. (2007) models relative to one another.

Figure 2.11 shows pixel deviance while Figure 2.12 shows Voronoi deviances for the isotropic MISD (1.2), anisotropic MISD (2.1), and Helmstetter et al. (2007) models relative to one another. Helmstetter et al. (2007) outperforms (1.2) in the largest cell, while (1.2) performs better in the areas surrounding the cluster of events near the Laguna Salada fault at Latitude -115.25°, Longitude 32.30°. The Northern portion of the forecasts near Eureka and Trinidad fault look similar since both (1.2) and Helmstetter et al. (2007) perform well in only part of the region. Indeed, neither model clearly outperforms the other overall. (1.2) performs slightly better than (2.1), near both Eureka and near Baja California. Helmstetter et al. (2007) has large positive residuals near Alum Rock California. However, both (1.2) and (2.1) perform similarly over most of the spatial window, as indicated by the relatively small Voronoi deviances. A large residual is observed for the event on 2005 October 16 in the SouthWest near San Clemente Island where (2.1) outperforms both Helmstetter et al. (2007) and (2.1).

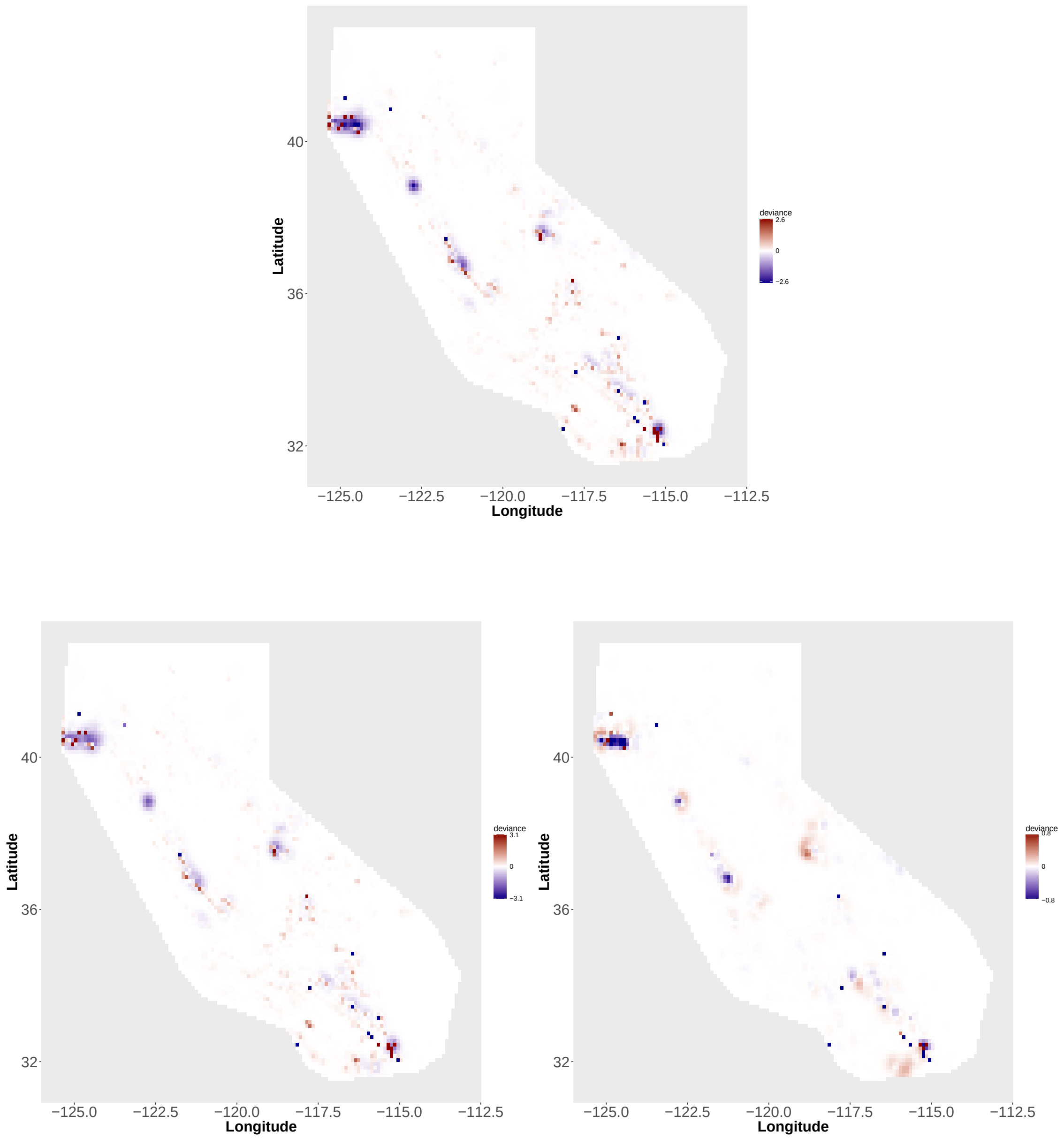


Figure 2.11: (a) Pixel deviances for (1.2) vs. Helmstetter et al. (2007), (b) (2.1) vs. Helmstetter et al. (2007) and (c) (1.2) vs. (2.1), respectively. Red cells shells indicate positive deviance residuals and blue cells indicate negative deviance residuals, with lighter shading indicating larger values of the deviance residuals.

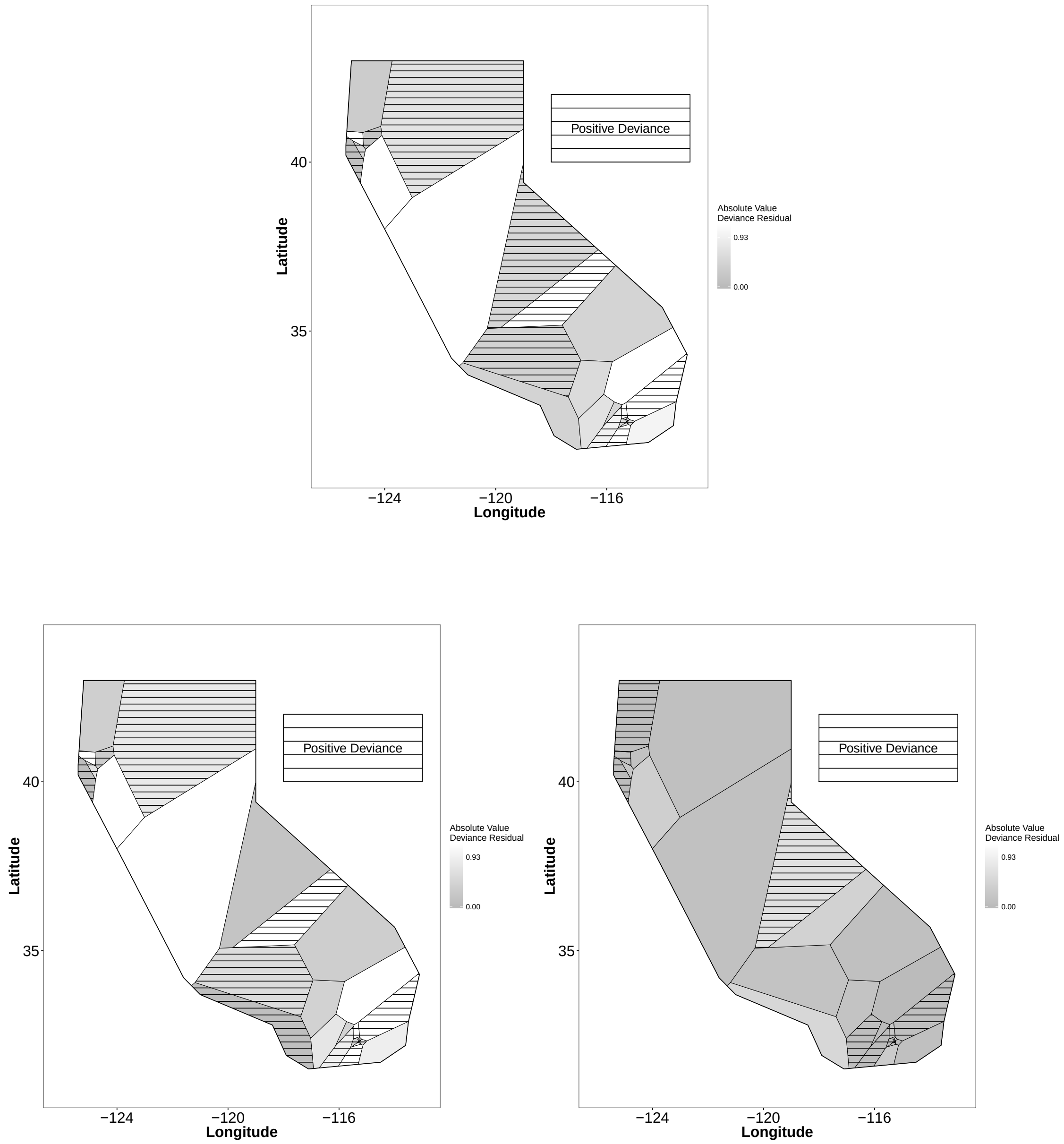


Figure 2.12: (a) Voronoi deviances for (1.2) vs. Helmstetter et al. (2007), (b) (2.1) vs. Helmstetter et al. (2007) and (c) (1.2) vs. (2.1), respectively. Striped shells indicate positive deviance residuals and solid cells indicate negative deviance residuals, with lighter shading indicating larger absolute values of the deviance residuals.

2.5 Discussion

We have constructed long-term earthquake forecasts for California seismicity using a nonparametrically estimated Hawkes model. The MISD method, merged with a flexible, local fault estimation approach, performs well in both describing and forecasting earthquake occurrences. Hawkes models estimated using MISD, with spatial distributions computed relative to local strike angle estimates, are able adequately to summarize the spatial distribution of aftershocks and dependence upon region, distance from known faults, and mainshock magnitude. Simulation studies suggest the method performs well, even under highly nonstandard conditions.

Alternative approaches for estimating fault planes used in (2.1) can be made. Under the current approach, each estimate of ϕ is given equal weight in the estimation of the spatial distribution of aftershocks even though their associated mean squared errors are significantly different. In future work, one might loosen restrictions in order to allow ϕ not to pass through the epicenter of the corresponding mainshock. In addition, estimated strike angles ϕ may be smoothed or constrained to allow block estimates, since it may be unrealistic for strike angles to differ substantially among earthquakes only a few meters apart. Another area for future work is to combine information from seismic moment tensors into the estimation of ϕ. Because such estimates may have large errors, robust regression techniques could also be applied to reduce the impact of outliers. Similarly, using Deming regression to estimate strike angles based on previous seismicity may improve estimation. Further research is needed to evaluate the sensitivity of estimates of the spatial distribution of aftershocks to strike angle estimates in general.

Additional opportunity exists to test and improve our earthquake forecasts. While we considered the ANSS catalog for estimation and forecasting, the size of the catalog used is relatively small. While we show slightly improved performance compared to Helmstetter et al. (2007) in various regions by using a nonparametric estimation and forecasting approach, more alternative models should be submitted to and tested in CSEP and we intend to continually improve our forecasts in the long term (5 year) model group. We hope that these contributions will lead to improved probabilistic earthquake forecasts.

CHAPTER 3

Analytic computation of nonparametric Marsan-Lengliné estimates

3.1 Introduction.

The spatial-temporal spread of infectious disease has traditionally been described via compartmental SIR models or their variants. Such models involve dividing populations according to disease status, and then modeling the changes in numbers of infected, susceptible, and recovered individuals in the population using systems of simple differential equation models (e.g. Meyers 2007, Grassly and Fraser 2008, Vynnycky and White 2016). Similarly, reaction-diffusion or regression-based methods have also been used with infectious disease or invasive species data to describe the amount of area being infected over time or the spatial-temporal spread of an infestation (e.g. Thompson 1991, Lonsdale 1993, Perrins et al. 1993, Pysek and Prach 1993, Higgins and Richardson 1996, Delisle et al. 2003, Peters 2004, Riley 2007, Vynnycky and White 2016). The processes by which humans spread contagious diseases and plants spread seeds naturally lend themselves to spatial-temporal Hawkes point process analysis, however, and it is these Hawkes point process models that are the subject of investigation here.

Purely temporal self-exciting point process models were proposed to describe the temporal spread of smallpox in Brazil by Becker (1977), and by Farrington et al. (2003) to describe the effect of vaccinations on the spread of measles in the United States, but the use of spatial-temporal Hawkes models for describing infectious diseases has so far remained under-utlized. As an alternative to compartmental SIR models and their variants, Hawkes models can provide different insights into the spread of epidemics and invasive species, including a description of the spread via an estimated spatial-temporal triggering kernel. As noted by Law et al. (2009), unlike grid-based studies on area occupation, where the surface of study is divided into an array of pixels on a grid, spatial-temporal point processes can enable greater precision of forecasts in space and time, and can offer a more detailed and precise account of spatial heterogeneity and clustering. To this end, Diggle (2006) investigated inhomogeneity in foot-and-mouth disease using spatial-temporal point process models estimated by partial likelihood methods, and Diggle (2014) surveyed successful uses of spatial-temporal point process modeling in describing in detail ecological phenomena such as the locations of Japanese black pine saplings as well as public health data such as liver cirrhosis in Northeastern England. However, the use of Hawkes processes for spatial-temporal epidemic data has been sparse. Exceptions are Becker (1977), who proposed purely temporal self-exciting point process models to describe the temporal spread of smallpox in Brazil, Farrington et al. (2003), who describe the effect of vaccinations on the spread of measles in the United States using self-exciting point process models, and Balderama et al. (2012), who model red banana plant locations and times using a parametric space-time Hawkes point process model, whose components were assumed to follow simple exponential laws. Such Hawkes models have long been used in seismology to describe the rate of aftershock activity following an earthquake (Ogata 1988, Ogata 1998) and have outperformed alternatives for earthquake forecasting (Zechar et al. 2013, Gordon et al. 2015).

Traditionally, the functional form of the triggering function in a Hawkes process must be specified by the researcher, and can then be estimated parametrically, using maximum likelihood estimation (Ogata 1978, Schoenberg 2013, Schoenberg 2016). One of the most exciting recent advances in this area was the discovery by Marsan and Lengliné (2008) of a method for estimating the triggering function of a Hawkes process nonparametrically. Their method, which uses a variant of the E-M algorithm, writes the triggering function as a step function and then estimates the steps by approximate maximum likelihood. The procedure thus does not rely on a parametric form for the triggering function, and is extremely useful as a tool for a variety of purposes including suggesting the functional form of a triggering function, assessing the goodness of fit of a particular proposed functional form, and simulating or forecasting the process without relying on a particular and possibly mis-specified functional form for the triggering function.

Unfortunately the method proposed by Marsan and Lengliné (2008) requires an iterative and computationally intensive procedure. In addition, the method ultimately produces approximate maximum likelihood estimates whose asymptotic properties are not well understood. Here, we describe a mathematical curiosity that allows one to compute exact maximum likelihood estimates of the triggering function in a direct and extremely rapid manner. One of the key ideas is to let the number p of intervals on which the nonparametric estimate is sought equals the number n of observed points, and we also discuss extensions to the more standard case where n is much larger than p. The computation times for our proposed method are many times smaller than with the iterative method of Marsan and Lengliné (2008). We evaluate the performance of this newly developed approach to estimating triggering functions in a variety of simulations. We then apply the method to two real-world datasets involving Loma Prieta earthquakes from 1989-2016 and plague occurrences in the United States from 1900-2012, in order to produce estimates of the triggering function and accompanying 95%-confidence bands. Such confidence bands, obtained via repeated simulation and re-estimation, would be very difficult to obtain using prior methods, and are useful for quantifying the uncertainty in estimates of the triggering function and corresponding rates of spread of plague as well as aftershock activity.

The structure of this paper is as follows. Following a brief review of the algorithm of Marsan and Lengliné (2008) in Section 2, the technique proposed here for the case $p = n$ is described in Section 3, followed by simulations in Section 4. Applications to seismological an epidemic data are shown in Sections 5 and 6, respectively. Concluding remarks are given in Section 8. Extensions to the case where $n >> p$ and other details regarding implementation are discussed in Appendix E.

3.2 Existing methods of nonparametric estimation.

Rather than rely on a particular parametric form of the Hawkes triggering function such as (1.3), one may instead use the iterative method of Marsan and Lengliné (2008) to estimate the background and triggering function of a spatial-temporal Hawkes process.

One tactic used by Marsan and Lengliné (2008) is to assume the estimated triggering function to be a step function with p steps of fixed widths and unknown heights, and then to consider estimating the p step heights via approximate maximum likelihood.

Specifically, for the case where $g(u)$ is a step function with value β_k for u in the range U_k, setting the partial derivatives with respect to each β_k of ℓ in equation (1.4) to zero, for $k = 1, ..., p$, reduces to

$$0 = \partial\ell(\theta)/\partial\beta_k = \sum_{(i,j):\tau_j - \tau_i \in U_k} K/\lambda(\tau_i) - Kn|U_k|, \tag{3.1}$$

resulting in p equations with p unknowns. However, the equations are nonlinear, depending in particular on $1/\lambda(\tau_i)$, which in turn depends on the p unknowns. One may attempt to resort to gradient descent or other optimization routines to find approximate solutions, but these are typically prohibitively slow, since they rely on computing or approximating ℓ and its derivatives many times, and each computation of ℓ requires $O(n^2)$ computations. Thus, Marsan and Lengliné (2008) and Marsan and Lengliné (2010) resort to finding approximate maximum likelihood estimates using the E-M based algorithm of Veen and Schoenberg (2008). Unfortunately, this algorithm also requires an iterative procedure and considerable computation time.

The Marsan-Lengliné algorithm works by first starting with a guess at the background rate

μ and triggering function g, and then repeating the following two steps until convergence is reached.

(a) First, one computes estimates of p_{ij} = the probability that occurrence i triggered occurrence j, for each i and j, using the previous (or initial) estimates of μ and g. One similarly obtains estimates p_{0j} for each j as the probability that observation j was a background event, i.e. was not triggered by any previous events.

(b) Second, one updates the estimate of $g(u)$ for any u by using a histogram or kernel smoothing of the pairs of observed events separated by a spatial-temporal distance of approximately u, where each such pair of observations (i, j) is given a weight of p_{ij}. Similarly, one can simultaneously obtain an updated estimate of the background rate μ simply by smoothing the observed points and weighting each of them by its probability p_{0j}, as estimated in step (a), of being a background event.

In Marsan and Lengliné (2008), the function quantities $g(u)$ are estimated for a discrete set of values u using a weighted histogram type estimator. Other nonparametric estimates of Hawkes triggering functions have been suggested, such as a maximum penalized likelihood estimator proposed by Lewis and Mohler (2011), and an estimator based on the Laplace transform of the sample covariance function proposed by Bacry et al. (2012). Adelfio and Chiodi (2015) suggest a semi-parametric estimator for Hawkes processes where the background rate is nonparametrically estimated while the triggering function is estimated parametrically.

Nonparametric estimates not only enable robust estimation of the triggering function without relying on a particular parametric form, but in addition, these methods may also be used to solve the problem of declustering observed sequences of invasive species or infectious diseases. Indeed, the stochastic declustering algorithm of Zhuang et al. (2002) uses a realization of an ETAS model to calculate the probability that an event is a background event, and its complement gives the probability of being a triggered event. One then obtains a complete probabilistic reconstruction of the branching structure. Similarly, Marsan and Lengliné (2008) decluster earthquake sequences using their nonparametric estimates.

3.3 Proposed nonparametric estimates.

As mentioned in the preceding Section, the Marsan and Lengliné (2008) algorithm for estimating the triggering function of a Hawkes process is an iterative procedure requiring substantial computation time. A relatively simple mathematical curiosity enables us to reduce the computation time dramatically, however, for the special case when $p = n$, as described below.

Evaluating the conditional rate in equation (1.1) at the n observed points, we have the n equations

$$\lambda(\tau_j) = \mu + K \sum_{i<j} g(\tau_j - \tau_i), \tag{3.2}$$

for $j = 1, 2, ..., n$, where $\lambda(\tau_j)$ is the conditional intensity at point τ_j. For simplicity, assume for the moment that μ and K are known. The case where μ and K are unknown, and other issues arising in practical implementation, are discussed in Appendix E.

As with Marsan and Lengliné (2008), we assume our estimate $\hat{g}(u)$ will be a step function, so that for u in a given interval U_k, we have $\hat{g}(u) = \beta_k$. Again, for simplicity, let us assume that the process is observed over a spatial region S and time interval $[0, T]$, and that the intervals U_k are time intervals of equal size Δ, so that the size $|U_k|$ of each interval is simply $\Delta|S|$. Setting the derivative of the loglikelihood L with respect to β_k to zero yields the p

equations

$$0 = \partial L/\partial \beta_k = \sum_{(i,j):\tau_j - \tau_i \in U_k} K/\lambda(\tau_j) - Kn|U_k|. \tag{3.3}$$

Unfortunately these terms on the right hand side of (3.3) are highly nonlinear functions of the desired parameters, $\beta = \{\beta_1, \beta_2, ..., \beta_p\}$. However, (3.3) consists of p *l*inear equations in terms of the n constants $1/\lambda(\tau_j)$. Thus, if $p = n$, we may readily use equation (3.3) to solve for $1/\lambda(\tau_j)$, which therefore yields $\lambda(\tau_j)$, and subsequently use the n linear equations (3.2) to solve for β.

Indeed, with some additional notation we may write the resulting estimator in a very simple and condensed form as follows. First, let us use the following notation: for any vector $z = \{z_1, z_2, ..., z_k\}$, we will let $1/z$ represent the vector $\{1/z_1, 1/z_2, ..., 1/z_k\}$. Next, consider the $p \times p$ adjacency matrix A defined so that, for $j, k \in \{1, 2, ..., p\}$, $A[k, j]$ is the number of points τ_i such that $\tau_j - \tau_i \in U_k$. Let S denote the spatial observation region. Let $\boldsymbol{\lambda}$ represent the n-vector $\{\lambda(\tau_1), \lambda(\tau_2), ..., \lambda(\tau_n)\}$ and let $\mathbf{1}$ denote the n-vector $\{1, 1, ..., 1\}$.

With this notation, equation (3.3) can be rewritten as

$$0 = KA(1/\boldsymbol{\lambda}) - Kb, \tag{3.4}$$

where $b = n|S|\Delta\mathbf{1}$. The estimate of $1/\boldsymbol{\lambda}$ satisfying equation (3.4) is thus $A^{-1}b$, assuming A is invertible. Similarly, equation (3.2) may be rewritten

$$\boldsymbol{\lambda} = \mu + KA^T\beta, \tag{3.5}$$

whose solution is $\beta = (KA^T)^{-1}(\boldsymbol{\lambda} - \mu)$. Combining these two formulas, the resulting vector $\hat{\beta}$ of estimates may be written

$$\hat{\beta} = (KA^T)^{-1}[1/(A^{-1}b) - \mu]. \tag{3.6}$$

Equation (3.6) is remarkably simple, trivial to program, and extremely rapid to compute. The speed with which these nonparametric estimates can be obtained facilitates certain computations that might otherwise be prohibitively time consuming. For instance, one may obtain approximate standard errors for these estimates by repeated simulation and estimation of the triggering function.

When n is large, the estimates in (3.6) will have extremely high variance. However, when it is reasonable to assume that g is smooth, these estimates in (3.6) may be smoothed to provide more stable estimates of g. Appendix E contains discussion of how this smoothing may be achieved, and other computational issues. In particular, the focus here is exclusively on the estimation of the parameters β governing the triggering density. The cases where μ and K are not known and where the binwidths Δ are not equal are discussed in Appendix E.

3.4 Performance in simulations.

As a proof of concept, Figure 3.1 shows the estimates in (3.6) applied to simulated temporal Hawkes processes with triggering function g given by the exponential, truncated normal, uniform, and Pareto densities. One can see that the estimate (3.6) is reasonably accurate in each case.

Figure 3.2 shows the computation times and mean squared errors of the estimates obtained via (3.6) and those obtained using the approximate MLE method of Marsan and Lengliné (2008), using simulations of a Hawkes process with exponential triggering function with mean 300 and productivity $K = 0.3$ on a square region $[0, 1] \times [0, 1]$. The circles in Figure 3.2a and b correspond to (3.6), and the triangles correspond to Marsan and Lengliné's estimates. The mean squared error is typically somewhat lower for the estimates obtained via equation (3.6), though the results are somewhat mixed. However, one sees from Figure 3.2a the dramatic improvement in computation time for the algorithm based on equation (3.6) compared with the Marsan and Lengliné algorithm. These differences in computation time, which are shown on logarithmic scale, become substantial as n grows. For a simulation with $T = 100,000$, $n = 1,378$, and the Marsan and Lengliné algorithm took 10.82 min, while the estimator proposed here took 9.39 sec. A simulation with $T = 200,000$ produced $n = 2,778$ events and the Marsan and Lengliné algorithm and the one proposed here took 3.37 hours and 41.78 sec, respectively.

Note that the code for both methods was written in R for comparability; certainly the Marsan and Lengliné code, which relies on iteration via loops, could be considerably faster in C or Fortran. Also, the computation time for the Marsan and Lengliné algorithm depends on a stopping criterion. Here for each estimate we had a stopping criterion of $\epsilon = 10^{-5}$, i.e. if the estimate of $g(u)$ did not change by at least ϵ for any u during a given iteration, then the procedure terminated, and a maximum of 100 iterations was used for each estimate in case this convergence threshold ϵ was not reached. For the estimator proposed here, the computation times reflect the time not only to compute g but also to smooth the g-values, using a Gaussian moving average filter with bandwidth 1100.

3.5 Application to Loma Prieta seismicity.

The estimation method of equation (3.6) and the method of Marsan and Lengliné (2008) were used to estimate the triggering function for the moment magnitude 6.9 earthquake occurring in Loma Prieta, California on October 17, 1989, and its aftershocks over the subsequent 27 years. Specifically, 5,567 shallow (depth $<$ 75km) earthquakes of magnitude at least 2.1, occurring between Oct 16, 1989 and Feb 24, 2016, with longitude between -121.0 and -124.0 and latitude between 36.0 and 39.0, were recorded and catalogued by the U.S. Geological Survey (USGS), and made publicly available on USGS.gov. A Hawkes model was fitted to this Loma Prieta catalog, and the resulting estimates of the triggering function are shown in Figure 3.3.

When examined over a long time scale, the estimates using equation (3.6) are indistinguishable from those using the method of Marsan and Lengliné (2008) in Figure Figure 3.3a. However, as shown in Figure Figure 3.3b, there do appear to be substantial differences between the two triggering density estimates for shorter time scales, with the estimate proposed here placing considerably more mass at time intervals of 8 hours to 1 day. The estimation method proposed here using equation (3.6) is sufficiently computationally fast to allow construction of standard errors and confidence intervals via simulation. Specifically, 100 simulations of a Hawkes process with triggering function equal to that estimated from the Loma Prieta data using equation (3.6) were generated. For each simulation, estimates of the triggering function were obtained, again using equation (3.6), and for each of 50 values of u, the standard deviation of the resulting estimates of $g(u)$ was taken as the estimated standard error. The estimate in (3.6) ± 1 and ± 2 standard errors are depicted with the dotted curves in Figure 3b, and the results indicate that the apparent differences between the two estimated triggering functions are not statistically significant. The only time interval where the estimate based on Marsan and Lengliné (2008) is outside the estimate using (3.6) ± 1 standard error is around 0.4 days, where the Marsan and Lengliné (2008) estimate assigns less mass to the estimated triggering density than the estimate based on (3.6).

3.6 Application to United States Plague Data.

Figure 3.4a shows the reported times of confirmed human plague cases in the continental United States from 1900-2012. These data represent 1006 confirmed cases, combined from various sources including the United States Public Health Service and later the Center for Disease Control (CDC); see Kugeler et al. (2015) for details. As noted in Kugeler et al. (2015), there was a period of apparently high frequency of plague in the United States from 1900-1925, followed by a period of relatively few cases until 1965, and subsequently sporadic cases annually occurring. For 138 cases, most of which occurred between 1900-1907, no specific estimated onset date was recorded. We focus our analysis, therefore, on the 868 cases with estimated onset dates. For these cases, time of day was not reported. Ignoring the time of day for each occurrence is problematic in this case, as if one neglects to consider this or treats each case as if it occurred at the same time of day, then there will be multiple pairs of points with exactly the same time interval, and the resulting paucity of unique interevent times causes the adjacency matrix A to be singular. Hence times of days for each event were randomly simulated uniformly and independently for these 868 cases.

The estimated triggering density $\hat{g}$, using both the Marsan and Lengliné method and the analytic method based on equation (3.6), is shown in Figures 4b and 4c. The results show how highly concentrated the cases are in time, with the bulk of the mass in the estimated triggering function occurring at very short time intervals of one week or less. More precisely, the estimated triggering density using equation (3.6) has 80% of its density corresponding to time intervals of 4.5 days or less, and 99% of its density corresponding to time intervals of 7.5 days or less. In other words, the fitted model suggests that given that subject B contracted the plague from subject A, the delay between the onset time for subject A and the onset time for subject B was most likely 4.5 days or less, and highly unlikely to be more than 7.5 days. The productivity constant K and background rate μ were estimated by maximum likelihood as 0.3267 and 0.01391 events/day, respectively, with corresponding standard errors of 0.002205 and 6.433×10^{-6} events/day, respectively, constructed using the inverse of the diagonal of the Hessian of the loglikelihood as is standard for maximum likelihood estimates (Ogata 1978).

We performed 1000 simulations of Hawkes processes each with $\hat{g}$ set to that estimated from the data using equation (3.6), and with K and μ equal to their MLEs. For each of these 1000 simulations, estimates of the triggering function were performed using equation (3.6), and the results were used to obtain the 95% confidence bands shown in Figure 3.4c. Little should be inferred from the noticeably higher density in the equation (3.6) estimates compared to the Marsan-Lengliné estimates in the range of 3-4 days, since these differences are well within the 95% confidence bands. Moreover, it is evident that even with more than 100 years of data on plague in the United States, there is still substantial variability in the estimates up to time intervals of 20 days. However, there is also a clear decrease in the triggering density after 1-2 days, and much of the triggering density corresponds to these very short time intervals of 4.5 days or less. Note that the 95% confidence bands are obtained *i*ndividually for each time interval, u; while it is not uncommon to obtain a large estimated triggering density of 0.2 or more for a particular time interval of 10-20 days, for example, it is extremely unlikely to see such high density estimates for many time intervals in this range.

3.7 Concluding remarks

Prescribing the number p of parameters to be estimated to equal the number n of observations allows one to obtain analytic MLEs for those parameters, by setting the derivatives of the log-likelihood to zero, solving for $1/\lambda_i$ and therefore obtaining λ_i, and then solving for β which defines the estimated triggering function. We observe major computation time savings from this method. For datasets of only 100-300 points, the savings are negligible. However, for 5,000 points, whereas the Marsan and LenglinÃĺ (2008) algorithm in R with a stopping criterion of 100 iterations maximum takes about 7 hours on a 2.5 GHz Intel Core i5 processor, the analytic method proposed here and also implemented in R takes only 1.3 minutes on the same machine. This speed facilitates computations such as simulation based confidence intervals.

According to the fitted Hawkes model for United States plague data from 1900-2012, the disease appears to have spread rapidly, with 99% of contagion occurring within 7.5 days. Note that this estimate of contagion is based on the estimated onset days of the reported plague symptoms; the actual times of day when the disease was contracted were unknown. The modeled contagion is not necessarily direct: the model does not discriminate between the situations where subject A transmits the disease directly to subject B and the situation where subjects A and B are both infected by some other source (human or animal). Hence the estimated contagion time of 0-7.5 days as suggested by the fitted model includes both direct and indirect transmission, and actual direct transmission times might be shorter. As noted in Kugeler et al. (2015), plague is a zoonotic disease, and in the United States, for only 30% of cases is the specific source of exposure known. Indeed, of the 305 human cases in the United States for which such source information is available, 16.1% of the cases were attributed to human to human transmission, while the remainder were attributed to exposure from fleas, rodents or other animals (Kugeler et al. 2015). The estimated triggering function reported here should be viewed as a description of the spatial-temporal spread of the disease, rather than an estimate of the direct human to human transmission rate. Furthermore, missing observations (unreported or undiagnosed plague cases) as well as the confirmed plague cases prior without assigned onset dates were removed from the analysis, and these omissions might also have led to an overestimate of the transmission times. The application to Loma Prieta earthquake data suggests a similarly rapid contagion time between earthquakes and their aftershocks, with most aftershock activity occurring within 0.5 days of the triggering event according to the fitted model. Of course, in the case of the Loma Prieta seismicity, most of the events in the catalog may be attributed to the magnitude 6.9 Loma Prieta mainshock itself. The fitted model here instead attributes many of the subsequent events to the cascading effect of aftershocks of the M6.9 Loma Prieta event triggering future aftershocks, and so on, and it is these triggerings that are estimated to occur primarily over a time span of only several hours. In the case of earthquakes as well as plague, there may be numerous other covariates, such as climate, geographical and geological variables for instance, that are omitted here yet may influence the relationship observed here between previously observed points and the rate of future points. The conditional intensity may nevertheless be consistently estimated in the absence of such information provided the impact of the missing covariates is suitable small, as shown in Schoenberg (2016).

While the method proposed here extends readily to the space-time-magnitude case and to the case where both μ and g are estimated simultaneously, as described in Appendix E, more work is needed to see what the limits are on this method. An important topic for future research is to investigate whether the method proposed here can be extended to other types of point process models as well, such as Cox processes, inhibition processes, Gibbs point processes, or other models. In addition, more work is needed to adapt the current method to the case where the adjacency matrix is singular or nearly singular, and to determine ideal means and bandwidths for smoothing the resulting estimates or perhaps to incorporate this smoothing into the estimation procedure. As mentioned in the Introduction, while Hawkes point process models are widely used in seismology, they have been sparsely used in epidemiology, and we hope the nonparametric estimation methods described here might facilitate the use of Hawkes processes to describe such datasets in the future.

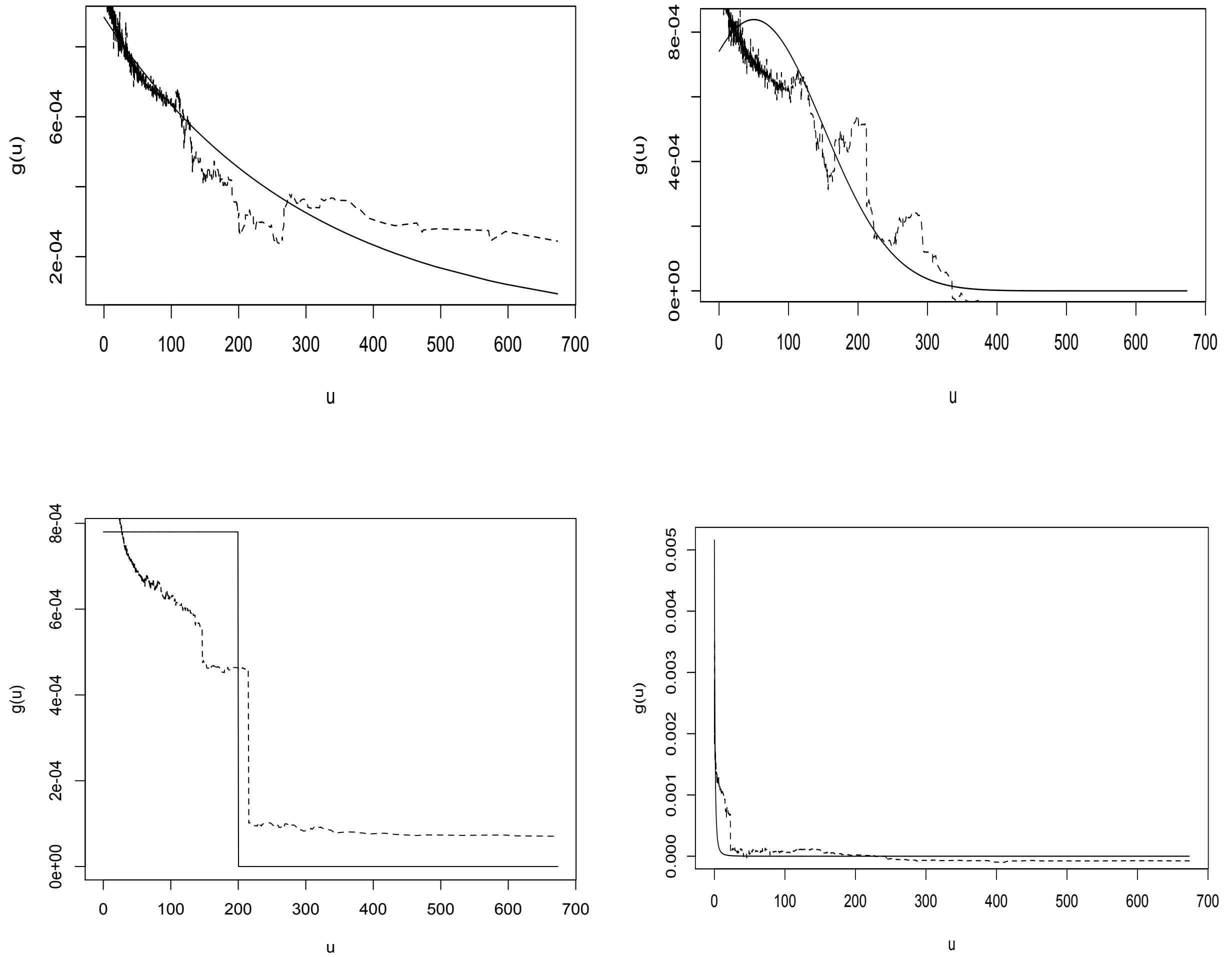
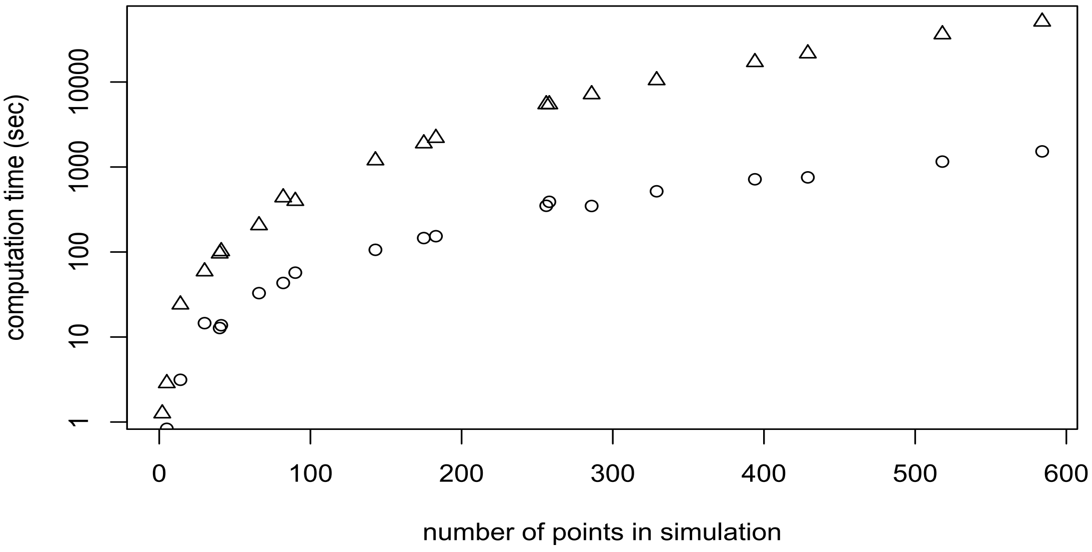


Figure 3.1: Estimated triggering density (dashed curve) using (3.6) and true triggering density (solid curve), for a) exponential, b) truncated normal, c) uniform, and d) Pareto densities. The exponential density used had mean 300, the truncated normal was the restriction to positive values of the normal density with mean 50 and SD 100, the uniform density was on the interval (0,200), and the Pareto triggering function had density $g(u) = (p-1)c^{p-1}/(u+c)^p$, with $c = 2$ and $p = 2.5$.



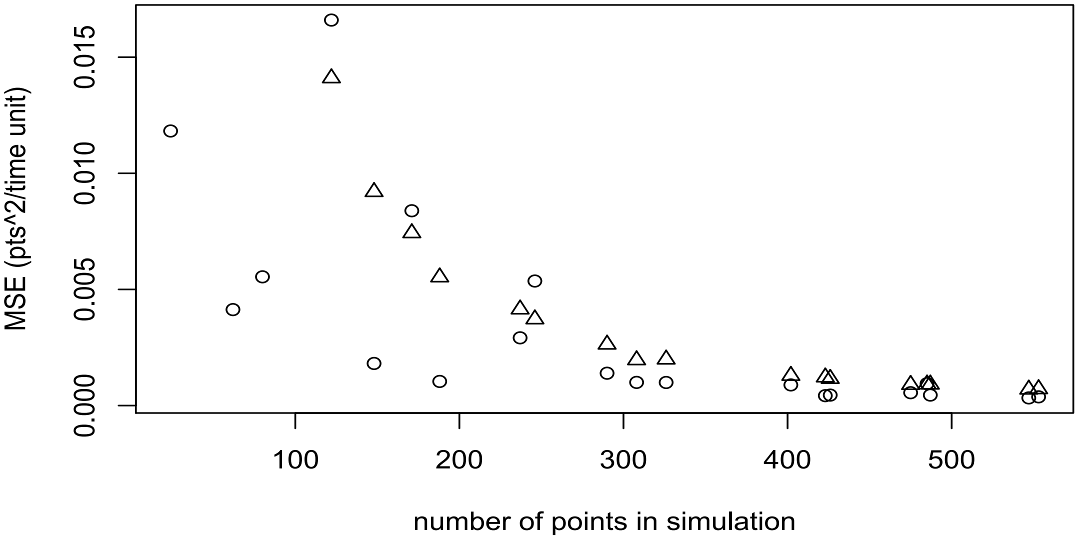


Figure 3.2: Computation times and MSE of triggering functions estimated using (3.6), represented by circles, or the method of Marsan and Lengliné (2008), represented by triangles. Computations were performed in *R* on a MacBook Pro with 2.5 GHz Intel Core i5.

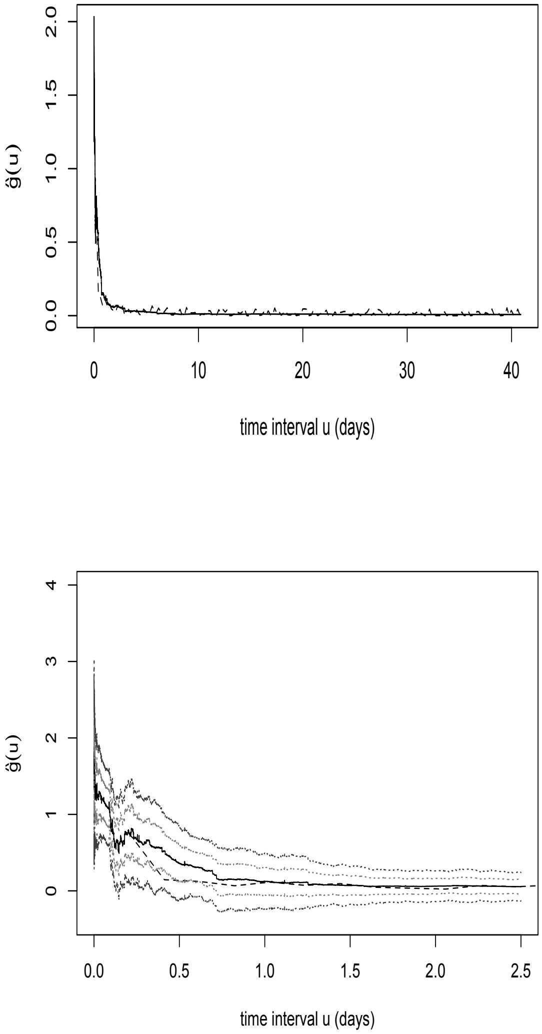


Figure 3.3: Estimated triggering function for Loma Prieta seismicity of $M \geq 2.1$ from 10/16/1989 to 1/17/1990. Solid curves represent the method proposed here in equation (3.6), and dashed curves represent the iterative method of Marsan and Lengliné (2008). Grey, dotted curves are estimates based on equation (3.6) ± 1 or 2 SEs, for light grey and dark grey, respectively. SEs were computed using the SD of equation (3.6) estimates in 100 simulations of Hawkes processes with triggering functions equal to that estimated using equation (3.6).

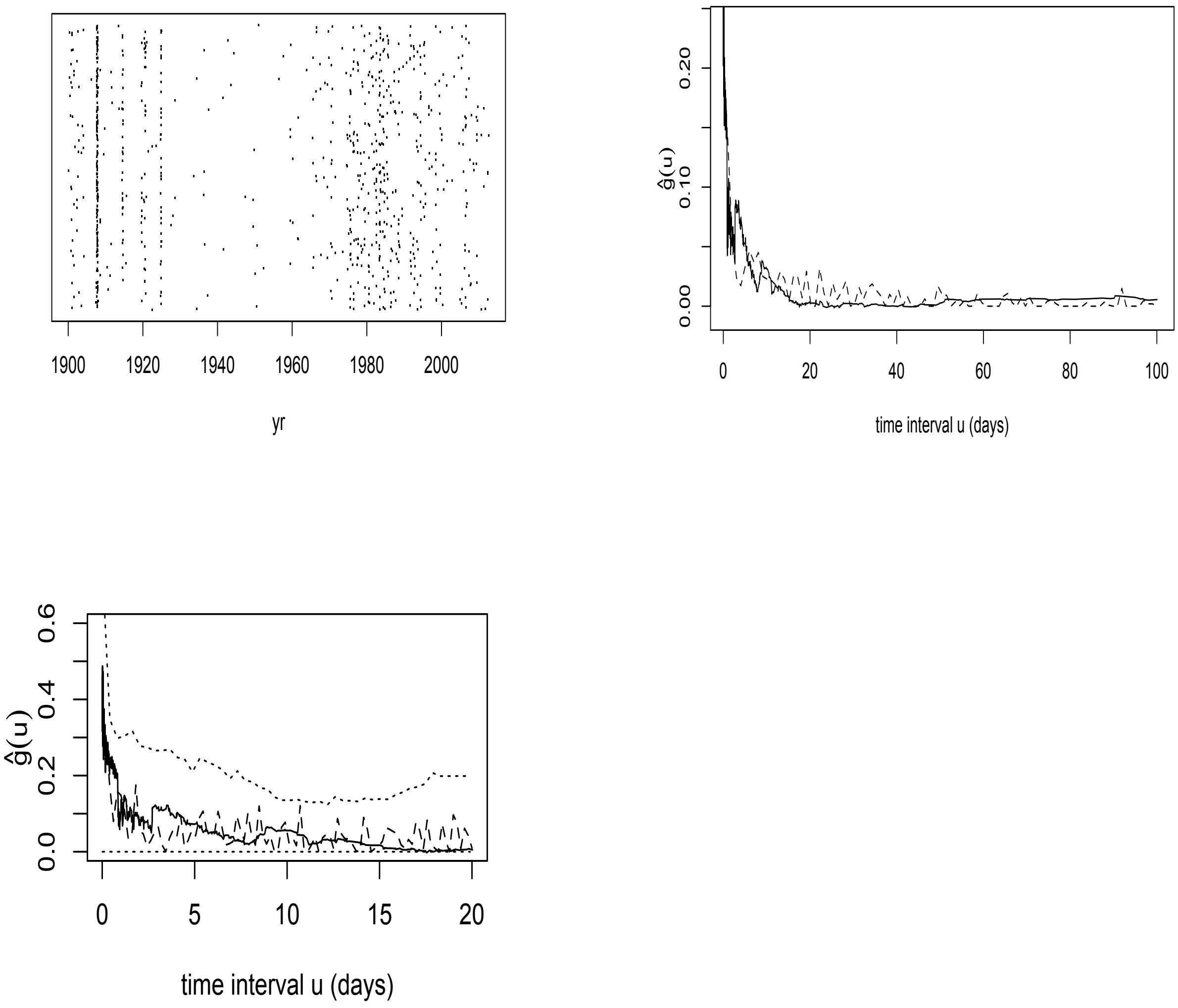


Figure 3.4: (a) Onset dates of reported and confirmed occurrences of plague in the United States from 1900-2012, according to data from the CDC. The y-coordinates are scattered uniformly at random on the y-axis for ease of visualization. (b) Estimated triggering function, $\hat{g}$, for the reported onset times of U.S. plague cases. (c) Estimated triggering function $\hat{g}$, for U.S. plague data, for intervals up to 20 days. In (b) and (c), the solid curves correspond to equation (3.6), the dashed curves result from the method of Marsan and Lengliné (2008), and the dotted curves are the middle 95% range for $\hat{g}$ from equation (3.6) resulting from simulating Hawkes models where the true triggering function is that estimated from the data using equation (3.6).

CHAPTER 4

Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts.

4.1 Introduction

Voronoi residuals, super-thinning, and some other residual analysis methods are applied to a selection of earthquake forecast models in the Collaboratory for the Study of Earthquake Predictability (CSEP). Unlike simple numerical summaries such as the N-test, L-test, or R-test, graphical residual methods are proposed which can be useful for comparing multiple models and for highlighting when and where a given model does not agree closely with the observed seismicity. For gridded forecasts Voronoi residuals seem preferable for assessing one model individually. For models outputting an estimated conditional rate at any particular space-time location, Voronoi residuals and super-thinning can be especially useful at identifying departures from the data.

Some alternative residual analysis techniques for spatial-temporal point process models have been proposed recently, such as Voronoi residuals and super-thinned residuals. These residual methods were reviewed in Clements et al. (2011) and applied to earthquake forecast models from the Regional Earthquake Likelihood Models (RELM) project (Field 2007, Schorlemmer and Gerstenberger 2007) in the framework of the Collaboratory for the Study of Earthquake Predictability (CSEP) described in Jordan (2006). The evaluations in Clements et al. (2011) were made using earthquake occurrence data from January 2006 to September 2009. Here, we review the strengths and weaknesses of these various model evaluation techniques and apply the residual analysis methods to a longer catalog, extending from January 2006 to

September 2014.

We also suggest an improvement with respect to color scaling of Voronoi residual diagrams. Previous research has struggled with this issue. Meijering (1953) demonstrated that the expected area of a Voronoi cell is equal to the reciprocal of its intensity for a homogeneous Poisson process and Hinde and Miles (1980) that the area of Voronoi cells is approximately gamma distributed. Tanemura (2003) showed that, for an inhomogeneous Poisson process, the reduced Voronoi cell area X is well approximated by a gamma distribution with a rate and shape of 3.569. However, as noted in Bray et al. (2014), this model is a poor approximation to the distribution of Voronoi residuals for earthquake data, since earthquakes are so highly clustered. Bray et al. (2014) proposed using a probability integral transformation (PIT) to scale the Voronoi residuals uniformly. The method was shown to work well, in terms of providing useful graphics when applied to earthquake data, but was computationally intensive, requiring repeated simulation of the model under consideration, and the resulting values of the PIT transformed Voronoi residuals are not so easy to interpret. Here, we propose simply using the Voronoi residuals applied to a homogeneous Poisson process model, with rate fit by maximum likelihood, as a scale by which to judge the Voronoi residuals of alternative models. The resulting scaling is trivial computationally, provides useful graphics when applied to earthquake data from Southern California, and results are very easy to interpret. Bright red cells and bright blue cells indicate areas where the proposed model performed alarmingly poorly, as these are areas where the residual was similar in size to that of the cells of maximum under-prediction or over-prediction of seismicity, respectively, of the homogeneous Poisson model.

Note that, for the purpose of evaluation and comparison of earthquake forecast models, CSEP currently implements several numerical summary tests, such as the Likelihood-test (L-test) (Schorlemmer et al. 2007) and the Number-test (N-test) (Zechar et al. 2013) to measure the consistency of a forecast with the observation and comparative tests, such as the T- and W-tests (Rhoades et al. 2013). For instance, the L-test works by first simulating some fixed number s of realizations from the forecast model and comparing the log-likelihood (ℓ) for the observed earthquake catalog (ℓ_{obs}) with that of the simulations (ℓ_j, for $j = 1, 2, \ldots, s$). The quantile score, γ, is defined as the fraction of simulated likelihoods that are less than the observed catalog likelihood:

$$\gamma = \frac{\sum_{j=1}^{s} \mathbf{1}_{\{\ell_j < \ell_{\text{obs}}\}}}{s},$$

where $\mathbf{1}$ denotes the indicator function. A value of γ close to zero is considered strong evidence of inconsistency between the model and the observed seismicity. The N-test is similar to the L-test, except that the quantile score examined is instead the fraction of simulations that contain fewer points than the actual observed number of points in the catalog, N_{obs}. That is,

$$\delta = \frac{\sum_{j=1}^{s} \mathbf{1}_{\{N_j < N_{obs}\}}}{s},$$

where N_j is the number of points in the jth simulation of the model. With the N-test, the model is rejected if δ is close to 0 or 1. Application of these test statistics and others to the RELM and CSEP models studied here are shown in Clements et al. (2011).

Such tests provide a score for the overall fit of the model but fail to indicate where a model may be fitting poorly. In addition, as noted in Clements et al. (2011), in practice both statistics γ and δ test essentially the same thing, namely the agreement between the observed and modeled $t$otal number of points. Indeed, for a typical model, the likelihood for a given simulated earthquake catalog depends critically on the number of points in the simulation. Furthermore, both the L-test and N-test have very low power, as shown via simulations in Clements et al. (2011). Instead of numerical tests and functional summaries such as weighted K-functions (see Adelfio and Schoenberg 2009), here we focus on graphical residual analysis methods that can be useful to suggest areas where one model outperforms another, or where one given model may need improvement.

This chapter is organized as follows: Section 4.2 presents the observed earthquake occurrences and the forecasted models for comparison, in Section 4.3 we apply pixel-based Pearson residuals and identify specific locations where each forecast performs well and/or poorly, Section 4.4 presents super-thinned residuals, and Section 4.5 describes Voronoi residuals, demonstrates their utility and introduces a new method of coloring scaling the cells. Section 4.6 summarizes the results and discusses some of the strengths and weaknesses of these methods.

4.2 Data and models for comparison

The data and models explored here are similar to those used in Clements et al. (2011), the main difference being the extension of the earthquake catalog to include 8.7 years instead of 3.7 years. Figure 4.1 shows estimated earthquake hypocenter locations for 510 shallow earthquakes (M 3.95+), which occurred in RELM's testing spatial-temporal window in Southern California between 1 January 2006 and 2 September 2014, and were obtained for this study from the Advanced National Seismic System (ANSS). 43 earthquakes were of magnitude $\geq M$ 4.95 and 10 were events $\geq M$ 5.5. The largest event, El Mayor M 7.2, occurred in Baja California on 4 April 2010.

The RELM models selected for comparison come from CSEP's rate-based repository, and are the same as those discussed in Clements et al. (2011). They are the following 3 models (named here A, B, and C as in Clements et al. 2011):

A: Helmstetter et al. (2007).

B: Kagan et al. (2007).

C: Shen et al. (2007).

All of these models except Model C forecast based exclusively on previous seismicity, whereas Model C incorporates geodetic and geological data as well. Models A, B, and C are five-year models, producing just one forecast of seismicity in each spatial-magnitude grid. Not included are STEP and ETAS Models which are one-day models, producing a forecasted expected number of events in each spatial grid for each day. Note that since not all RELM models in CSEP produce estimates in every pixel within this space-time

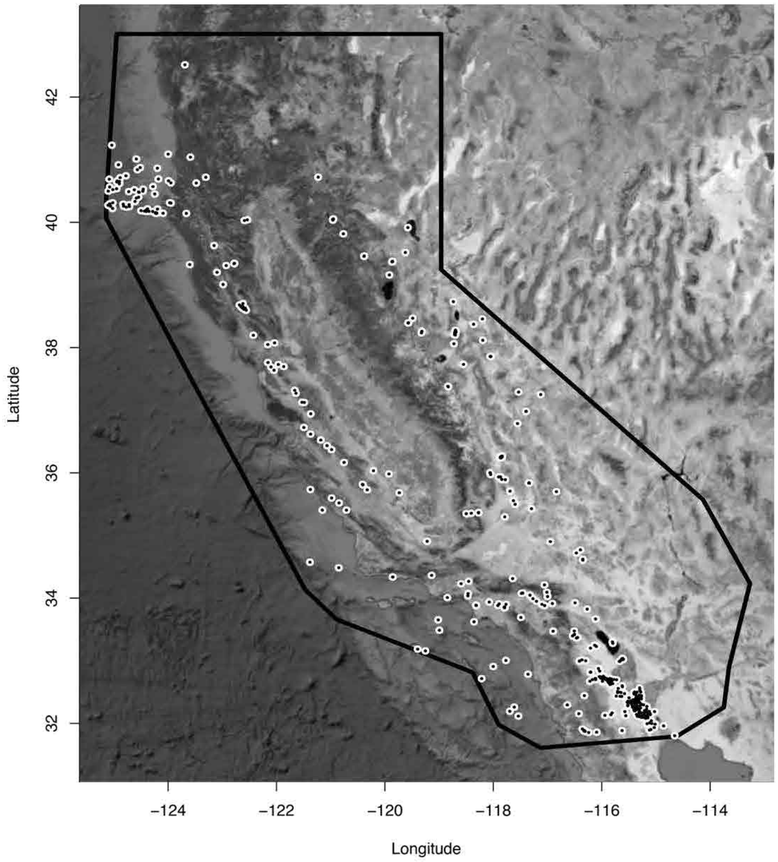


Figure 4.1: Locations of 510 earthquakes with magnitude $M \geq 3.95$ in the RELM testing region from 1 January 2006 to 2 September 2014.

window, there are different numbers of observed earthquakes occurring in the relevant forecast regions corresponding to different models. As such, Model A had 509 earthquakes observed within its corresponding space-time window, Model B had 343, and Model C had 356. All 5 year models were scaled proportionally to adjust to the temporal window used here.

The RELM testing region was designed to include all earthquakes in California and $\sim 1°$ around it. The space is divided into cells size of $0.1°$ longitude by $0.1°$ latitude. The magnitude dimension is also divided into $0.1°$ bins for earthquake magnitudes ranging from 3.95 to 8.95. For magnitudes 8.95–10, there is a single bin of size $0.1°$ by $0.1°$ by 1.05 units of magnitude. Note that forecast bins can be masked, meaning that the bins should be ignored when evaluating the forecast. For the RELM Models A, B and C, a lower magnitude bound of 4.95 was imposed, but as in Clements et al. (2011), for purposes of model evaluation and comparison, we extrapolate down to magnitude 3.95 using the model's fitted magnitude distribution. Specifically, Models A and B assume the magnitude distribution follows a tapered Gutenberg-Richter law (Gutenberg and Richter, 1944) with a b-value of 0.95 and a corner magnitude of 8.0. Model C uses a b-value of 0.975 and the same corner magnitude. Model A adjusts the magnitude distribution in a small region in northern California influenced by geothermal activity ($-122.9° < \text{lon} < -122.7°$ and $38.7° < \text{lat} < 38.9°$) by using a b-value of 1.94 instead of 0.95.

4.3 Raw and Pearson residuals

Raw and Pearson residuals can be useful for detecting a model's lack of fit as expressed by large differences between the number of points occurring in each pixel and the number expected according to the fitted model. Consider a space-time point process with conditional intensity $\hat{\lambda}(t, x, y)$ at any time t and location (x, y). The raw residual process r may be defined following Baddeley et al. (2005) as the difference between the point process and its conditional intensity process, i.e. a process with integral

$$\begin{aligned} R(B) &= \int_B r(t,x,y)\ dtdxdy \\ &= \int_B dN - \int_B \hat{\lambda}\ (t,x,y)\ dtdxdy = N(B) - \int_B \hat{\lambda}\ (t,x,y)\ dtdxdy, \end{aligned} \tag{4.1}$$

for any measurable set B. One may then observe these integrated raw residuals over a sequence of pixels B_i. Note that Baddeley et al. (2005) only consider the case of purely spatial point processes characterized by their Papangelou intensities; Zhuang (2006) showed one may nevertheless extend the definition to the spatial-temporal case using the conventional conditional intensity.

One may rescale the raw residuals such that they have mean 0 and variance approximately equal to 1. The resulting Pearson residual process r_p are defined so that, for measurable B,

$$\begin{aligned} R_{\mathrm{P}}(B) &= \int_B r_{\mathrm{P}}(t,x,y)\ dtdxdy = \int\limits_B \frac{1}{\sqrt{\hat{\lambda}}} dN - \int_B \sqrt{\hat{\lambda}}\ dtdxdy \\ &= \sum_{(t_j,x_j,y_j)\in B} \frac{1}{\sqrt{\hat{\lambda}\ (t_j,x_j,y_j)}} - \int_B \sqrt{\hat{\lambda}\ (t,x,y)}\ dtdxdy, \end{aligned}$$

provided $\hat{\lambda}(t_i,x_i,y_i) > 0$. This form is analogous to the Pearson residuals in Poisson log-linear regression.

In practice, with Pearson residuals, standardization is problematic when one or more events occur in spatio-temporal locations with forecasted conditional intensity of 0. Often a minor adjustment may be made to the estimated conditional intensity in each of these locations such that the forecasted conditional intensity is slightly greater than 0. Model C needs such an adjustment in a few bins whereas Models A and B did not have this constraint and we were able to obtain Pearson residuals for each pixel. However, when several models assign very low conditional intensity in one area of space-time, the Pearson residuals in these locations of very low conditional intensity tend to overwhelm the others in a visual inspection, resulting in a Pearson residual plot that is of limited use in terms of evaluating the quality of the fit of the models either in absolute or relative terms.

Figure 4.2(b) shows the Pearson residuals for Model B with the largest residual (4.47) located on the pixel that spans the California border with Mexico. This region is just east of Mexicali and the Imperial Valley fault zone (lon ≈ -115.8°W and lat ≈ 32.7°N), and is the location just East of a large cluster of earthquakes. Nearby there are many other large Pearson residuals. A residual of 3.95 occurs just West, and a large cluster of large residuals for Model B all occur south of this region (lon ≈ -115.2°W and lat ≈ 32.2°N). Outside of this region, other notable residuals for Model B (3.36) are located above the San Bernardino and Inyo county border near the Panamint Valley fault zone (lon ≈ -117.9°W and lat ≈ 36.4°N), and (1.68) just north (lon ≈ -117.4°W and lat ≈ 36.0°N). Since the Pearson residuals should be approximately standardized, values much greater than 2 (in absolute value) suggest a significant lack of fit.

Model A, shown in Figure 4.2(a), shares many similarities to Model B in terms of model fit. The largest Pearson residual for Model A (4.57) is located near Mexicali at at the Mexico-California border (lon ≈ -115.2°W and lat ≈ 32.2°N). In fact, this region contains most of the largest residuals for Model A. Large residuals (such as 3.51) also occur just north of California in the Battle Rock fault zone (lon ≈ -123.8°W and lat ≈ 42.6°N) and (2.59) near the Bare Mountain fault zone (lon ≈ -117.8°W and lat ≈ 36.4°N).

The largest residual for Model C (4.44) occurs in close proximity to the Bare Mountain fault zone (lon ≈ -117.8°W and lat ≈ 36.4°N) as seen in Figure 4.2(c). There were 10 earthquakes ranging from magnitude 4.2 to 5.19 in this pixel during the temporal window. A cluster of large negative residuals, as seen in dark blue, occurred near the creeping section of the San Andreas fault zone (lon ≈ -120.4°W and lat ≈ 36.1°N). Seismicity was accurately forecasted in the low intensity area in the Eastern most portion of the forecast.

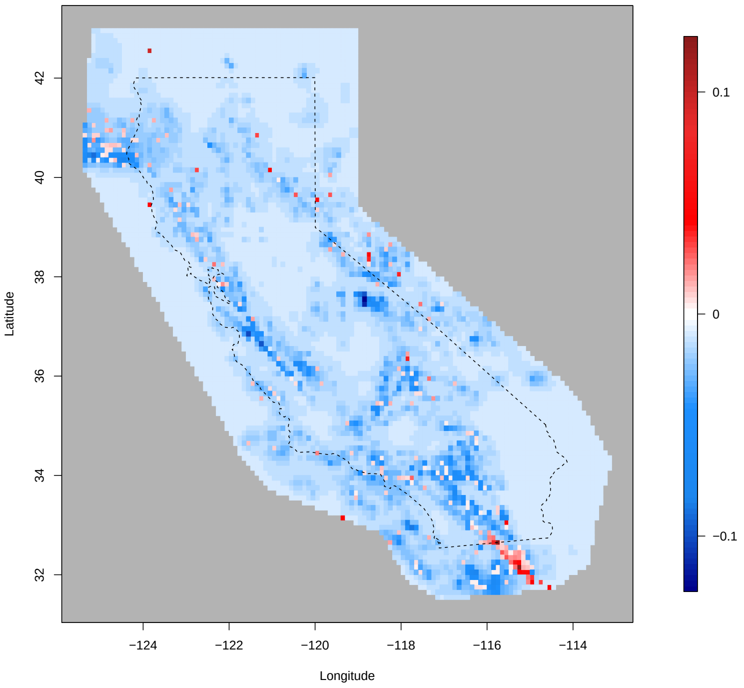
Yet as with Models A and B, the majority of the largest Pearson residuals for Model C occurred near the Imperial Fault region.

While raw and Pearson residuals can be an effective tool for evaluating a model's lack of fit, these residuals may be highly skewed and hence yield potentially misleading results when spatial-temporal bins are small and/or the estimated conditional intensity in some bins is very low, as in many locations in the models described above. Indeed, when this occurs, plots of the pixellated raw or Pearson residuals tend to resemble plots of the points themselves, and thus reveal little about the goodness-of-fit of the models in question. Pearson residuals and raw residuals are effective at identifying areas where the model fit should be adjusted and can be a good starting point for analysis, and alternative approaches, such as deviances, can be useful to evaluate how poorly a model fits in locations where it underpredicted or overpredicted seismicity.

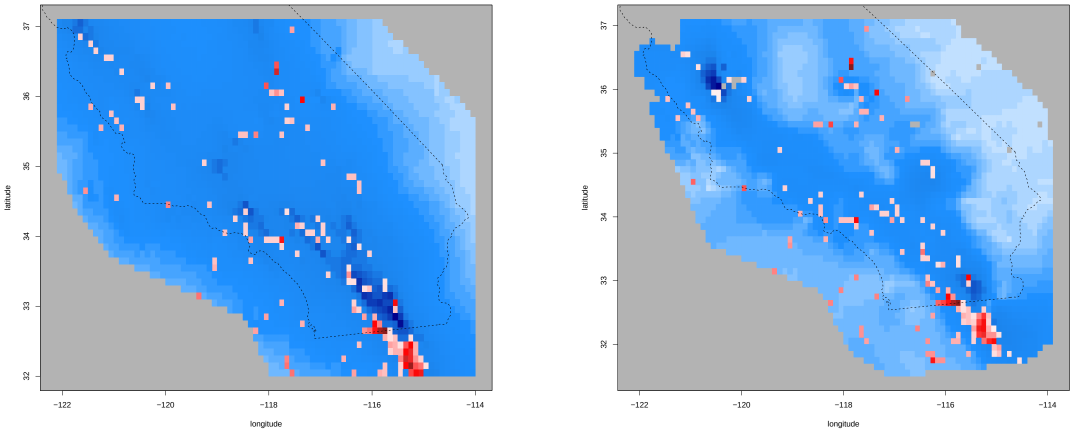
4.4 Super-thinning

Thinning for residual analysis was proposed in Schoenberg (2003) and superposition by Brémaud (1981). Using thinning, each point τ_i, of a point process N, is retained with some probability p_i. Superposition, meanwhile, is essentially an addition operator on point processes, i.e. N_3 is the superposition of point processes N_1 and N_2. While both are individually effective, a more powerful approach than either thinning or superposition alone is super-thinning. This combined approach, introduced by Clements et al. (2012), thins in areas of high intensity and superposes simulated points in areas of low intensity, resulting in a homogeneous point process if the model for λ used in the thinning and superposition is correct.

Using the super-thinning method, N can be transformed into Z, using the following algorithm. First thin N, keeping each point (t_i, x_i, y_i) independently with probability



(a) Pearson residuals for Model A



(b) Pearson residuals for Model B

(c) Pearson residuals for Model C

Figure 4.2: The maximum observed Pearson residual is 4.57 for Model A and is located near Mexicali at at the Mexico-California border (lon ≈ $-115.2°$W and lat ≈ $32.2°$N). The maximum observed Pearson residual is 4.47 for Model B and is located on the pixel that spans the California border with Mexico. The largest residuals for Model C occur in close proximity to the Bare Mountain fault zone (lon ≈ $-117.8°$W and lat ≈ $36.4°$N).

$p_i = \min\{1, k/\hat{\lambda}(t,x,y)\}$, to obtain a thinned residual process Z_1. Next, simulate points according to a Cox process Z_2 directed by $\max\{0, k - \hat{\lambda}(t,x,y)\}$. The points of the residual point process $Z = Z_1 + Z_2$, obtained by superposing the thinned residuals and the simulated Poisson process, are called super-thinned residuals. The procedure results in a homogeneous Poisson process Z with rate k if and only if the thinning and superposition are performed using an estimate $\hat{\lambda}$ that is equal to the true conditional intensity λ almost everywhere (Clements et al. 2012).

An advantage of this method is that the user may specify the overall rate of the resulting residual point process, Z, so that it contains neither too few or too many points. In this case, for Models A, B, and C, k was chosen to be the total number of expected earthquakes according to each forecast. The resulting super-thinned residuals can be plotted and assessed for homogeneity as a way of evaluating the model. Visualization allows detection of any clustering or inhibition in the residual points which indicates a lack of fit. One can also use the L-function applied to the residual super-thinned process Z to assess the uniformity of the residuals.

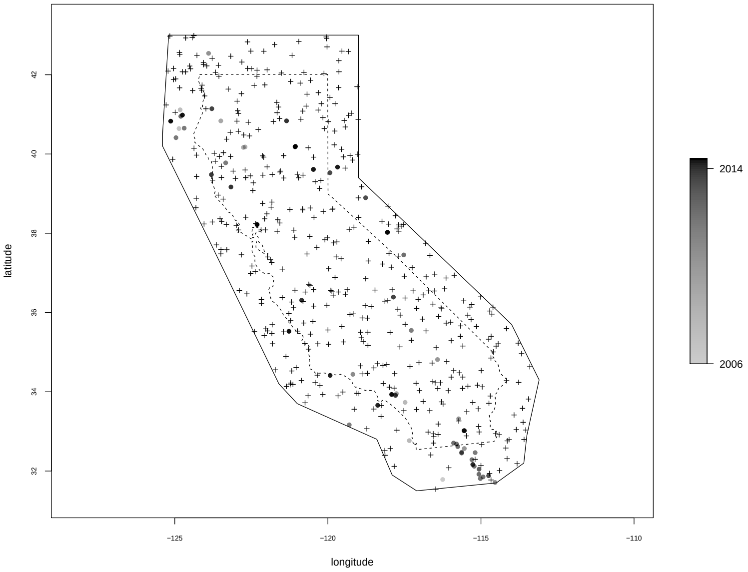
Figure 4.3(a) shows Model A fits well overall despite some clustering in the residuals at very small distances (from $0°$ to $0.1°$) but not much otherwise. Circles indicated observed earthquakes and plus signs indicate simulated points, while lighter points indicate events that occurred earlier in the time and darker points occurred later. There is a significant cluster near the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N) and in the Trinidad fault zone (lon $\approx -124.5°$W and lat $\approx 40.5°$N). However, the centered weighted L-function for the corresponding residuals for Model A, shown in Figure 4.5(a), reveals that the model performs well is most areas. The L-function is best interpreted along with 95%-confidence bands which are plotted as dash lines. While there is a small amout of inhibition in the residual process but Model A seems to accurately predict the rate of seismicity outside of the interfault zones.

The super-thinned residuals for Model B, shown in Figure 4.3(b), contain a few significant clusters near the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N), Laguna Salada (lon $\approx -117.1°$W and lat $\approx 32.4°$N), and La Habra (lon $\approx -118.0°$W and lat $\approx 34.0°$N). Indeed, there is significant clustering for Model B up to distances of 0.3°. As with Model A, the weighted L-function for Model B, shown in Figure 4.5(b), indicates little inhibition outside of this range. There is also little overprediction as evident by a consistent covering of residual points. This result means enough points were simulated and the residual process is close to what is expected.

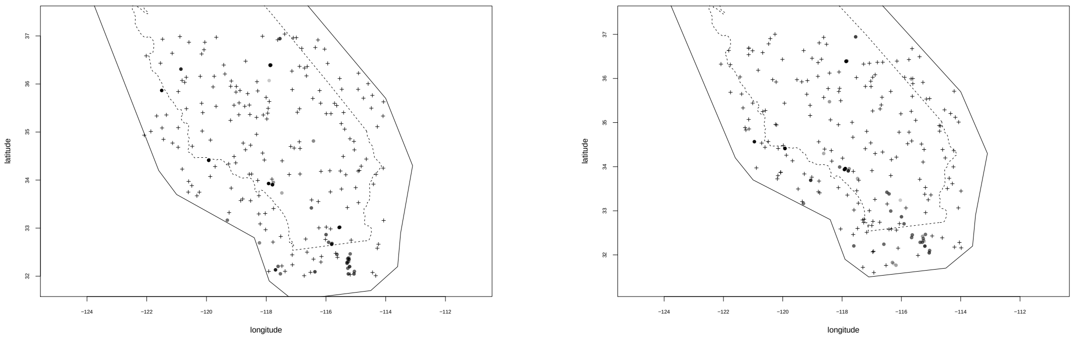
As seen in Figure 4.3(c), there is also significant clustering for Model C up to distances of 0.2°, which occur in similar regions to Models A and B. Clustering for Model C occurs in the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N), Laguna Salada (lon $\approx -117.1°$W and lat $\approx 32.4°$N), and off the coast of Baja California (lon $\approx -117.7°$W and lat $\approx 32.0°$N). Several of the nearly linear patterns in the data also appear in the residuals, such as in the Imperial Cluster, which indicates that the rates in these locations may be misspecified. Indeed, as seen Figure 4.5(c), at short distances the centered weighted L-function indicates the Model C is under-predicting but the L-functions are within the Poisson bounds for large distances.

4.4.1 Spatial Temporal Super-thinning

Spatial-temporal super-thinned residuals allow the evaluation of a model's performance over various time domains. The forecast period was split into four equal temporal windows and super-thinned residuals were visualized for Model B to assess the spatial temporal performance of the model. Figure 4.4 shows these results for Model B. The super-thinned residuals contain clusters near the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N), specifically in temporal window 2 (2 March 2008 to 3 May 2010) and temporal window 3 (3 May 2010 to July 7 2012). Otherwise, there is little over-prediction as evident by a consistent covering of residual points. This result implies enough points were simulated



(a) Super-thinned residuals Model A



(b) Super-thinned residuals Model B

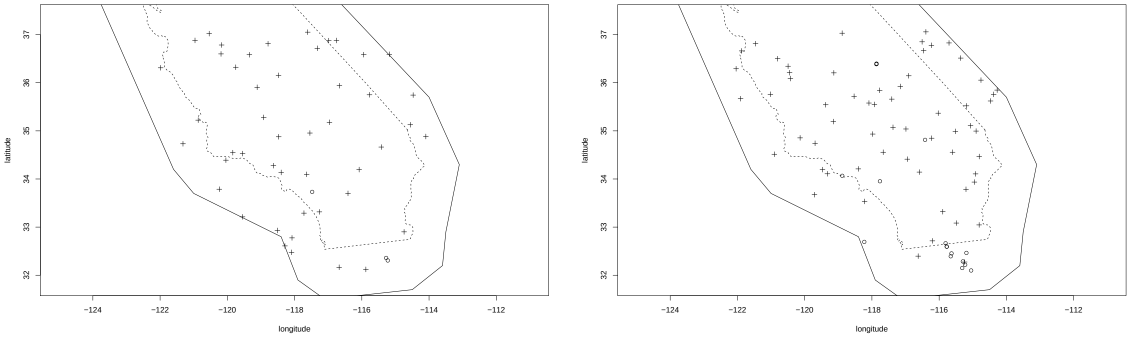
(c) Super-thinned residuals Model C

Figure 4.3: One realization of super-thinned residuals for each Model. Circles indicated observed earthquakes and plus signs indicate simulated points. Lighter points indicate events that occurred earlier in the time, darker points occurred later. There is a significant cluster near the Imperial Fault (lon ≈ −115.2°W and lat ≈ 32.2°N) for Model A. The super-thinned residuals for Model B contain a few significant clusters near the Imperial Fault (lon ≈ −115.2°W and lat ≈ 32.2°N) and Laguna Salada (lon ≈ −117.1°W and lat ≈ 32.4°N). Clustering for Model C occurs in Laguna Salada (lon ≈ −117.1°W and lat ≈ 32.4°N) and off the coast of Baja California (lon ≈ −117.7°W and lat ≈ 32.0°N).

and the residual process in each temporal window is close to what is expected. Figure 4.6 shows the weighted L-function, corresponding to the four temporal windows. There appears to be significant clustering in temporal window 2 up to distances of 0.3° and in temporal window 3 up to distance of 0.1°. In addition, at distance less than 0.1° in temporal window 4 (2 July 2012 to 2 September 2014), the observed data exhibit greater inhibition than one would expect according to Model B.

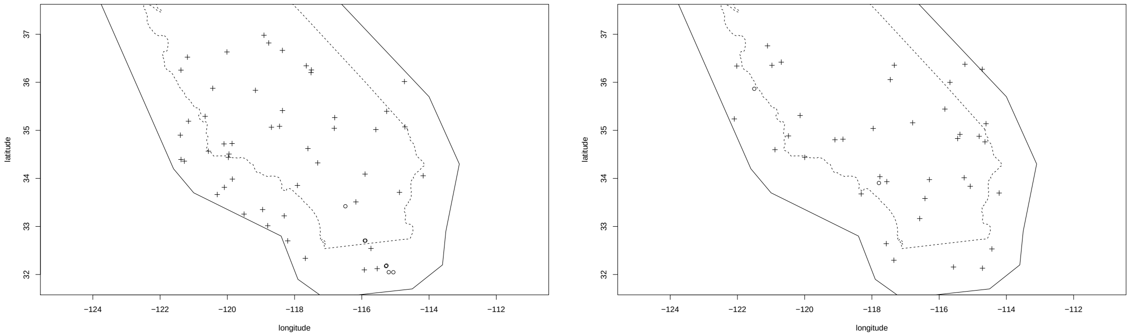
To further evaluate each model's temporal performance, the forecast period was split into 12 equal temporal windows. Given the results from the L-function described above, investigating distances of 0.2° to 0.3° is of interest. During each temporal window, the Z-statistic with the largest absolute value associated with the L function in $r \in [.2, .3)$, was identified. The temporal assessment of the L-function, shown in Figure 4.7, indicates that Model A appears to have larger positive Z-statistics in the first half of the temporal window and has its largest negative Z-statistics in the 8th temporal window. In this window the L-function crosses the 95% confidence bounds shown in Figure 4.5(a). Model A generally performed well, so this result confirms the previous finding. Model B tended to have positive Z-statistics although 3 of the last 4 temporal windows had large negative Z-statistics. Similar to Model B, Model C had large positive Z-statistics in the 5th, 6th, and 7th temporal windows while 3 of the last 4 temporal windows had large negative Z-statistics. The temporal trend in the extremes of the residual process over time appears to be minimal.

Spatial-temporal super-thinned residuals are a valuable tool for assessing a model's performance over varying space-time windows. The temporal windows must typically be defined quite arbitrarily, however. For example, the M 7.2 El Mayor earthquake at the end of the second temporal window, as seen in Figure 4.6(c), and its aftershocks were visible in window 3. Smaller temporal windows can be used, as shown in Figure 4.7 for instance, but as the size of the temporal window decreases, the number of observed events in each window also decreases resulting in less data available within each window for model evaluation.



(a) T_1 : 1 Jan 2006 - 2 Mar 2008

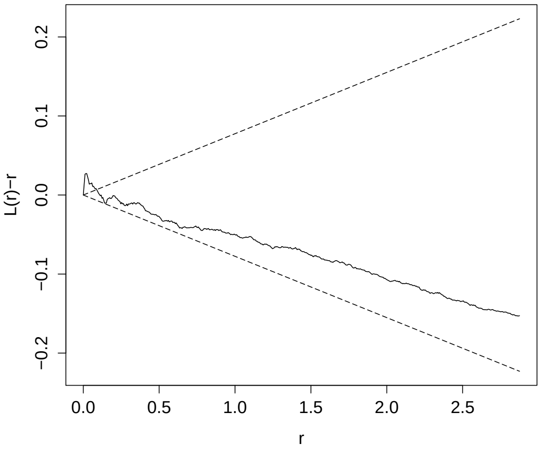
(b) T_2 : 2 Mar 2008 - 3 May 2010



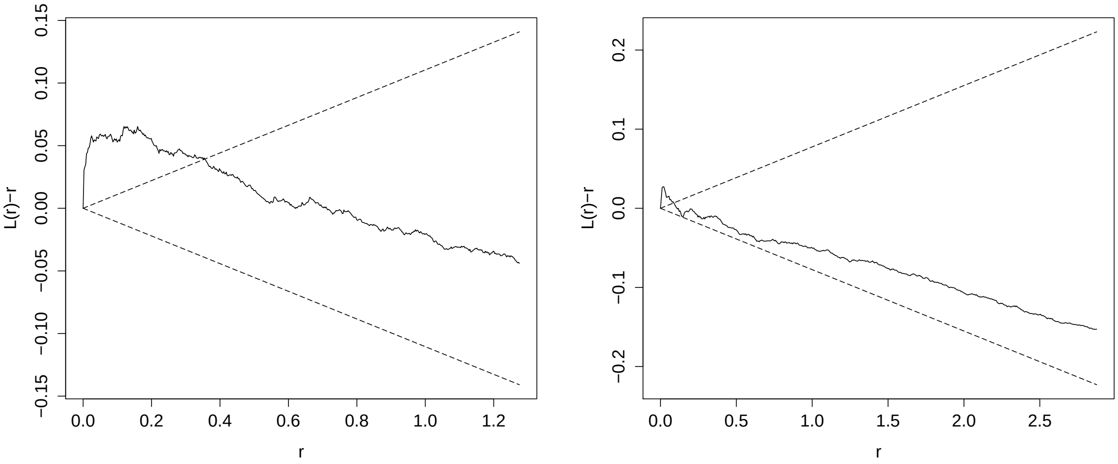
(c) T_3 : 3 May 2010 - 2 Jul 2012

(d) T_4 : 2 Jul 2012 - 2 Sep 2014

Figure 4.4: One realization of temporal Super-thinned residuals for Model B. Circles indicated observed earthquakes and plus signs indicate simulated points. The super-thinned residuals for Model B contain a few significant clusters near the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N) in temporal window 2 (2 March 2008 to 3 May 2010) and temporal window 3 (3 May 2010 to July 7 2012). Otherwise, the covering of residual points appears to be quite uniform.

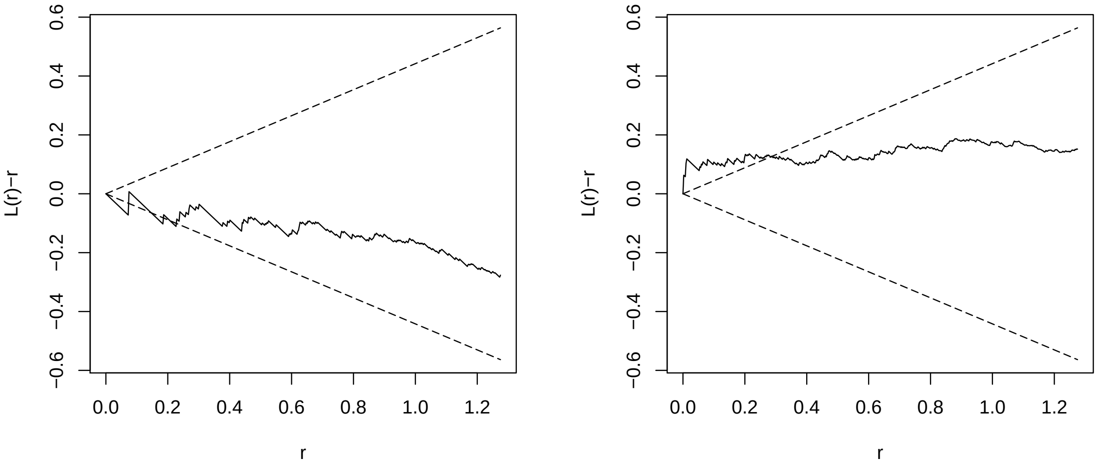


(a) Centered weighted L-function for Model A



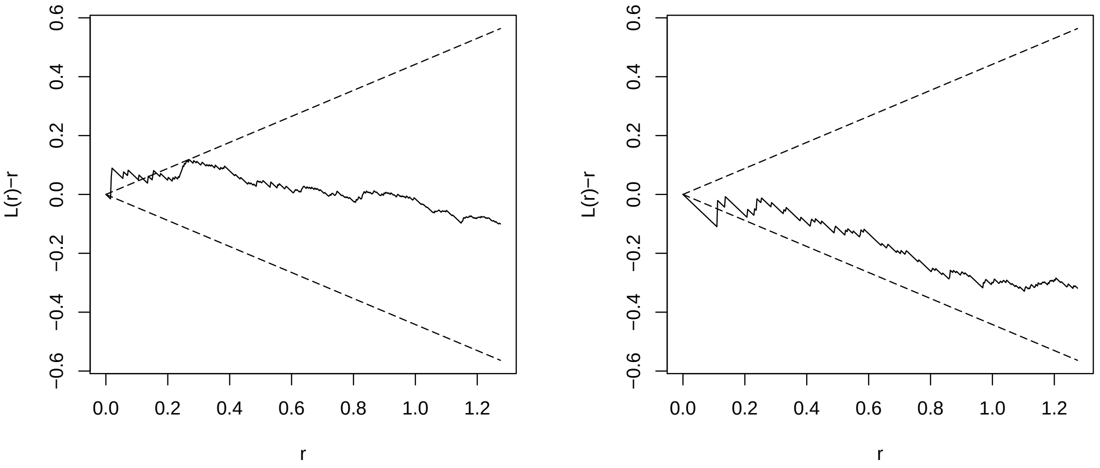
(b) Centered weighted L-function for Model B (c) Centered weighted L-function for Model C

Figure 4.5: Centered weighted L-function for Model A along with 95%-confidence bands. Model A fits well overall despite some clustering in the residuals at very small distances (from 0° to 0.1°) but not much otherwise. There is significant clustering for Model B up to distances of 0.3°. There is also significant clustering for Model C up to distances of 0.2°, which occurs in similar regions to Models A and B.



(a) T_1 : 1 Jan 2006 - 2 Mar 2008

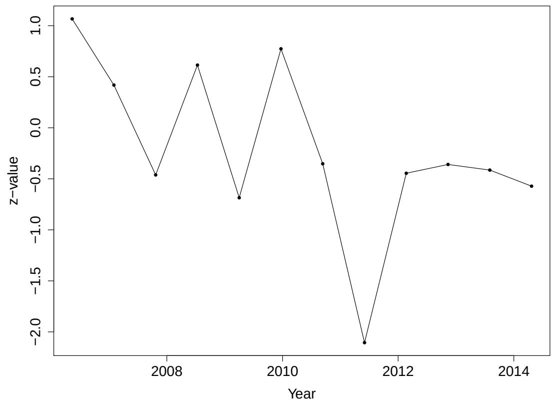
(b) T_2 : 2 Mar 2008 - 3 May 2010



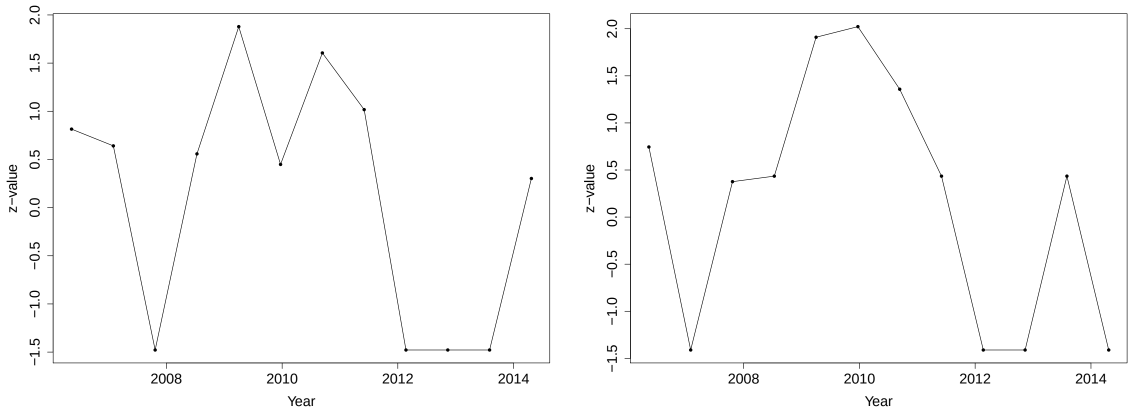
(c) T_3 : 3 May 2010 - 2 Jul 2012

(d) T_4 : 2 Jul 2012 - 2 Sep 2014

Figure 4.6: Centered weighted L-function for Model B along with 95%-confidence bands. There appears to be significant clustering in temporal window 2 up to distances of 0.3° and in temporal window 3 up to distance of 0.1°. However at distance less than 0.1° in temporal window 4 (2 July 2012 to 2 September 2014), the observed data exhibit greater inhibition than one would expect according to Model B.



(a) Model A



(b) Model B

(c) Model C

Figure 4.7: Temporal Assessment of L-function at $r \in [.2, .3)$ enabling an assessment of the stationarity of the residual process over time. Model A appears to have larger positive Z-statistics in the first half of the the temporal windows and has its largest negative Z-statistics in the 8th temporal window. In this window the L function would have crossed the 95% confidence bounds shown in Figure 4.5. Model A generally performed well so this result is not surprising. Model B has its largest Z-statistics in temporal windows 5 and 7. Model B tended to have positive Z-statistics although 3 of the last 4 temporal windows had large negative Z-statistics. Similar to Model B, Model C had large positive Z-statistics in the 5th, 6th, and 7th temporal windows while 3 of the last 4 temporal windows had large negative Z-statistics.

4.5 Voronoi residuals

Voronoi residuals are also useful for overcoming the problems caused by skewness in the distribution of the raw or Pearson residuals integrated over pixels, when pixels have a small integrated conditional intensity. Voronoi residuals are constructed using a Voronoi tessellation, which is a partition of the metric space on which a point process is defined into convex polygons, or Voronoi cells. Specifically, for a point pattern of N events, one may define its corresponding Voronoi tessellation as follows: for each observed point τ_i of the point process, its corresponding cell C_i is defined as the region consisting of all locations that are closer to the generating point τ_i than to any other point of N. The tessellation is the collection of such Voronoi cells which we assume fills the complete window C_T such that $C_\mathrm{T} = \bigcup\limits_{i=1}^{N} C_i$. Voronoi cells are necessarily convex polygons and have many well understood properties; for instance, the mean number of edges in the Voronoi cells induced by a stationary planar Poisson process is six. Okabe et al. (2000) provide a thorough treatment of Voronoi tessellations and their properties.

Given an observed spatial or spatial-temporal point pattern and its corresponding Voronoi tessellation, one may construct residuals for a conditional intensity model simply by evaluating the integral of the raw residual process over the Voronoi cells rather than over rectangular pixels. We will refer to such residuals as Voronoi residuals. Alternatively, one may integrate Pearson residuals or examine deviances over the Voronoi residuals as well.

A key advantage of Voronoi residuals compared to conventional pixel-based residual methods is that the partition is entirely automatic, data-driven, and spatially adaptive. Barr et al. (2010) showed Voronoi estimates can have less variability than kernel estimates in locations of low intensity surrounded by locations of high intensity. Moreover, the resulting distribution of residuals tends to be far less skewed than when one integrates the raw or Pearson residuals over fixed rectangular pixels. Each Voronoi cell has exactly one point inside it by construction, i.e. $N(\tau_i) = 1$ for each Voronoi cell C_i.

Voronoi residuals are described in detail in Bray et al. (2014) and are shown to be considerably less skewed than pixel residuals. One difficulty when plotting Voronoi residuals is the determination of an appropriate color scale, with appropriate limits. Bray et al. (2014) proposed using a probability integral transformation (PIT) to scale the Voronoi residuals uniformly. While the PIT method was shown to work well, in terms of providing useful graphics when applied to earthquake data, it was computationally intensive since it required repeated simulation of the model under consideration. Here, we propose a much simpler alternative. We simply fit a homogeneous Poisson process model, with rate fit by maximum likelihood, and use the standardized Voronoi residuals for this null model as a scale by which to judge the residuals of alternative models.

The resulting scaling is trivial computationally, provides useful graphics when applied to earthquake data from Southern California, and the results are very easy to interpret. Figure 4.8(a) shows the Voronoi cells corresponding to the standardized residuals of the null homogeneous Poisson model, as such the color in each Voronoi cell is defined by the size of the cell. Cells of highly clustered events are colored red and blue cells indicate areas of sparse seismicity. Note that to construct the reference model, all Voronoi cells were truncated by the border of the RELM region, and in computing Voronoi residuals for Models A, B, and C, the Voronoi cells were constructed spatially and each such cell spans the entire temporal and magnitude window.

Figure 4.8(b) shows the standardized Voronoi residuals for Model A. Areas towards the edges of the color scale are where Model A performed alarmingly poorly as compared to the null homogeneous Poisson model. Indeed, these are areas where the residual was similar in size to that of the cells of maximum under-prediction or over-prediction of seismicity for the null model. Model A under-predicts seismicity near the Imperial Fault, as indicated by the tight cluster of red cells, and generally tends to over-predict seismicity compared to the reference model. Model A appears to perform well in the Trinidad fault zone (lon $\approx -124.5°$W and lat $\approx 40.5°$N). Model A also substantially outperforms the reference model near Hawthorne, Nevada (lon ≈ −118.70°W and lat ≈ 38.40°N).

Figure 4.8(c) shows Model B similarly under-predicts seismicity in the Imperial Fault region. Model B appears to forecast seismicity accurately near the Campo Indian Reservation (lon ≈ −116.30°W and lat ≈ 32.40°N) but generally tends to under-predict seismicity in the Southwestern region of the forecasted area. Model B tends to over-predict seismicity in the Eastern region of the forecast and generally in the Northwestern region. However, within a vast region of over-prediction Model B performs well off the coast of Paso Robles (lon ≈ −121.50°W and lat ≈ 35.70°N). The relatively large amount of lighter shades indicates that Model B generally performs well compared to the null model.

Model C (Figure 4.8(d)) similarly under-predicts in the Imperial Valley, though the vast white shade in this region indicates areas where Model C forecast seismicity accurately compared to the homogeneous Poisson model. For example, Model C appears to perform well near the Channel Islands of California (lon ≈ −119.3°W and lat ≈ 33.1°N) and also performs well further South near the Coronado Islands (lon ≈ −117.1°W and lat ≈ 32.3°N). Like Models A and B, Model C appears to under-predict seismicity in the Southwestern area of the forecast region. Alternatively, over-prediction is evident by cluster of large negative residuals located just west of the Sequoia National Forest (lon ≈ −119.5°W and lat ≈ 36.0°N) in the Northwestern area of the forecast.

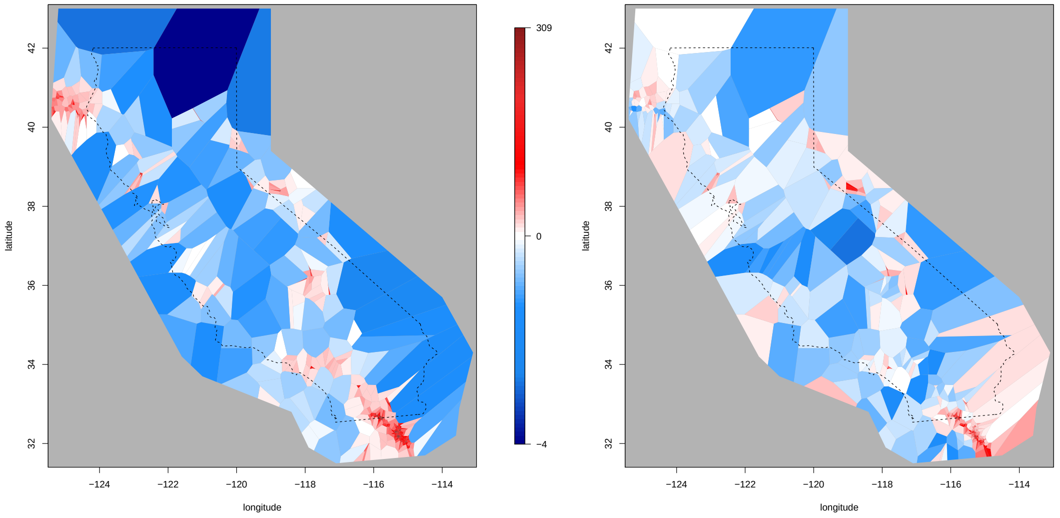
Using the color scaling defined by the null homogeneous Poisson model enables easy comparisons between models. We can see in Figure 4.8(a) and Figure 4.8(b) that Model B under-predicts in the eastern portion of the forecast region, whereas Model A tends to over-predict seismicity in this area. In comparing Model A to Model C, one sees that Models A and B both under-predict seismicity near the Channel Islands of California (lon ≈ −119.3°W and lat ≈ 33.1°N) whereas Model C appears to forecast seismicity accurately in this region. In addition, Model A tends to over-predict just south (lon ≈ −118.0°W and lat ≈ 32.0°N) whereas Models B and C under-predicted seismicity. Compared to Model C, Model B tended to over-predict in areas such as Death Valley (lon ≈ −116.5°W and lat ≈ 36.3°N) and near Pomona (lon ≈ −117.8°W and lat ≈ 34.0°N). Overall, the Voronoi residuals suggest that Model C forecast seismicity the most accurately of the three models.

Voronoi residual plots are able to clearly highlight the regions where Models A, B and C under-predict and over-predict the intensity of the process. Model evaluation using Voronoi residuals has the advantage of offering an adaptive, data-driven grid that requires no input from the user regarding tuning parameters. They are ideal for evaluating when a particular model appears to over-predict or under-predict seismicity, especially with the ease of comparison to a null homogeneous Poisson model. When combined with Pearson and super-thinned residuals, one can conduct a complete evaluation of a group of competing models.

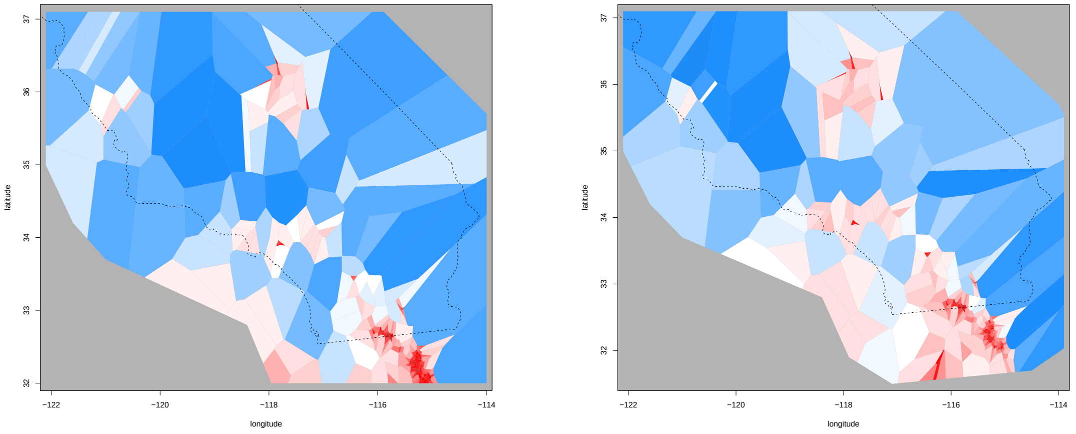
4.6 Summary

A variety of common residual analysis methods for spatial point processes can be implemented to assess fit and reveal strengths and weaknesses in point process models. Voronoi residuals and super-thinned residuals for spatial-temporal point process models appear to provide powerful summaries of model fit. These residual methods were applied to the 5-year earthquake forecast models in the Regional Earthquake Likelihood Models (RELM) project for a catalog spanning from January 2006 to September 2014, including shallow earthquakes of magnitude at least 3.95. Extending the temporal window beyond that in Clements et al. (2011) has allowed a larger number of observed earthquakes to test and led to more detailed and more meaningful results.

Pixel-based Pearson residuals may be valuable for their ease of interpretation and simplicity of calculation. However, problems can arise due to forecasted conditional intensities of 0 and extreme skew in the standardized residuals rendering them difficult to interpret. Model evaluation using the partitions derived from Voronoi tessellation of the observed events



(a) Voronoi residuals for Null Model (b) Voronoi residuals for Model A



(c) Voronoi residuals for Model B (d) Voronoi residuals for Model C

Figure 4.8: Voronoi residuals for each Model. When computing the integrated rate over a Voronoi cell, we treat the forecasted rate over each pixel as constant within each pixel. Model A appears to perform well in the Trinidad fault zone (lon ≈ −124.5°W and lat ≈ 40.5°N) while Model B appears to forecast seismicity accurately near the Campo Indian Reservation (lon ≈ −116.30°W and lat ≈ 32.40°N). Model C appears to perform well near the Channel Islands of California (lon ≈ −119.3°W and lat ≈ 33.1°N).

has the advantage of offering an adaptive, data-driven grid that requires no input from the user regarding tuning parameters. While some sampling variability is induced by the random cell areas, the resulting Voronoi residuals are substantially less skewed than their counterparts over typical rectangular grid cells, and the scaling proposed here for Voronoi residuals using a fitted homogeneous Poisson process is trivial computationally and appears to result in useful graphics for model evaluation and comparison. In particular, Voronoi residuals appear to be ideal for evaluating when a particular model appears to over-predict or under-predict seismicity, with the null homogeneous Poisson model offering a useful and easily interpretable model for comparison. All of these methods may be useful in the CSEP paradigm, and hopefully insight gained during one prediction experiment can inform the building of models for subsequent experiments.

While the N-test and the L-test provide easy to understand statistics that can be used for hypothesis testing, they have very low power and fail to indicate where a model may be fitting poorly. The L-test does not differentiate between over-prediction and under-prediction and the N-test contains no information on the spatial performance of the model. Formulation of these models of space-time point processes allows for additional evaluations to be applied that are not constrained by assumptions such as the independence of bins [Schneider et al. 2014]. When applying point process evaluation techniques, a model's performance can depend on the diagnostic tool. Since Voronoi residuals rely on fewer assumptions due to the spatially adaptive partitioning, as a result, models that predict well in areas of low seismicity may outperform models with more accurate forecasts where events actually occurred. Super-thinning appears to be a promising alternative, but may have low power if the forecasted intensity is volatile. Weighted 2nd-order statistics appear to be quite powerful, especially for comparisons of competing models in space-time [Clements et al. 2011].

Zechar et al. (2013) recommend using all tests in combination since each provides insight into model performance. Surveying results from all evaluation methods applied here, Model C generally appears to perform the best. Model A, B and C appear to over-predict seismicity in many locations but under-predict seismicity in the Imperial Fault region. Voronoi residuals suggest that Model C forecast seismicity more accurately in comparison to Models A and B and this is supported by the results from deviances. Model C outperforms both Models A and B in areas of high seismicity and has less extreme Pearson residuals than Models A and B. Indeed, the spatial distribution of intensity according to Model C appears to be quite accurate in areas of low seismicity and Model C tended not to over-predict seismicity locally as much as Models A and B.

Acknowledgements

We thank the Advanced National Seismic System for the earthquake catalog data, and the Collaboratory for the Study of Earthquake Predictability and the Southern California Earthquake Center, especially Masha Liukis for supplying the earthquake forecasts.

CHAPTER 5

Software

5.1 Introduction to `NonParametricHawkes`

`NonParametricHawkes` is an R package for nonparametric model estimation, forecasting, and evaluation of spatial point process models, created to increase the accessibility of the methods introduced here. The package was primarily designed for submission to CSEP, and thus is most useful for applications related to seismology, however, all of the functionality robust and can be applied to any data. This chapter will introduce a sample of the functionality available in the package. Specifically, the package i illustrated on both a simulated Hawkes process as well as an earthquake catalog.

5.2 Using `NonParametricHawkes`

Included within `NonParametricHawkes` are a variety of example datasets. First, a small ETAS simulation which is useful for demonstrating the methods. The data can be loaded in the usual way. It contains four variables: time (t), magnitude (hm), longitude (lon), and latitude(lat).

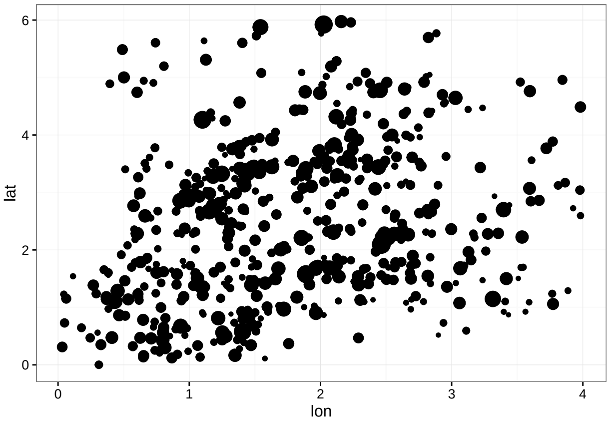
```
# Load
require(NonParametricHawkes)


# Load example data
data(ETASsim)
str(ETASsim)
```

```
## List of 4
##  $ t  : num [1:570] 10.2 11.7 11.7 23.2 35.6 ...
##  $ hm : num [1:570] 0.3811 0.08 0.0293 0.0336 0.5472 ...
```

```
##  $ lon: num [1:570] 1.41 2.84 3.03 1.46 1.87 ...
##  $ lat: num [1:570] 0.8 1.414 1.423 0.866 3.076 ...
```

It is important that data be transformed (sorted) in such a way that the oldest events are first the variables are named appropriately. The `nphData` function provides this functionality.

```
# Cast Data to nphData format
newdata <- nphData(data = ETASsim, time_var = 't',
                   x_var = 'lon', y_var = 'lat', mag = 'hm')

p <- ggplot(newdata) + geom_point(aes(x = lon, y = lat, size = hm))
p + theme_bw() + theme(legend.position = 'none')
```



Model Estimation

The package contains functions for estimating multiple variations of Hawkes point process models. All model estimation procedures start with the function name `nphawkes`:

- `nphawkesT`: Temporal
- `nphawkesTNS`: Temporal Non Stationary
- `nphawkesST`: Spatial Temporal
- `nphawkesMSTH`: Marked Spatial Temporal with Homogenous Background Rate
- `nphawkesMSTNH`: Marked Spatial Temporal with Nonhomogenous Background Rate
- `nphawkesMSTNHA`: Marked Spatial Temporal with Anisotropic Spatial Response and Nonhomogenous Background Rate

- `nphawkesMSTNHAC`: Marked Spatial Temporal with Anisotropic Spatial Response Conditional on Magnitude and Nonhomogenous Background Rate

Each function includes a help file, for example, `help(nphawkesMSTNH)` returns:

Description

Fits a Marked Spatial Temporal Nonparametric Hawkes Point Process Model with a nonhomogenous and stationary background rate

Usage

```
nphawkesMSTNH(data, nbins_m = 8, nbins_t = 25, nbins_r = 30, np = 50, num_iter
= 1000, nbins_x = 100, nbins_y = 100, xrange = c(0, 4), yrange = c(0, 6),
eps = 0.001, verbose = T, warn = T)
```

Arguments

`data` A list or data frame with the times, locations, and magnitudes of interest

`nbins_m` The number of bins on which to estimate the magnitude productivity function for Hawkes Point Process Model

`nbins_t` The number of bins on which to estimate the spatial triggering of the Hawkes Point Process Model

`nbins_r` The number of bins on which to estimate the temporal triggering of the Hawkes Point Process Model

`np` The nearest neighbor used for bandwidth selection

`num_iter` The maximum number of iterations to run

`nbins_x` The number of x bins on which to estimate the inhomogenous background rate

`nbins_y` The number of y bins on which to estimate the inhomogenous background rate

`xrange` The x range on which estimate the spatial triggering function

`yrange` The y range on which estimate the spatial triggering function

`eps` The convergence criteria

`verbose` A logical to control the amount of information printed to the console during estimation

Value

`mu` A single number representing the stationary background rate

`g` A vector of length nbins_t containing the nonparametrically estimated temporal triggering function

`f` A vector of length nbins_r containing the nonparametrically estimated spatial triggering function

`kp` A vector of length nbins_m containing the nonparametrically estimated magnitude productivity function

`delta_t` The bin widths used in the temporal triggering function estimation

`delta_r` The bin widths used in the spatial triggering function estimation

`delta_m` The bin widths used in the magnitude productivity function

An example is shown below. The resulting object contains the estimate model, specifically, the background rate, the magnitude productivity, the temporal response, and the spatial response. A plot method is included for the `nphawkesMSTNH` class. See `help(plot.nphawkesMSTNH)` for a full description of the all available options.

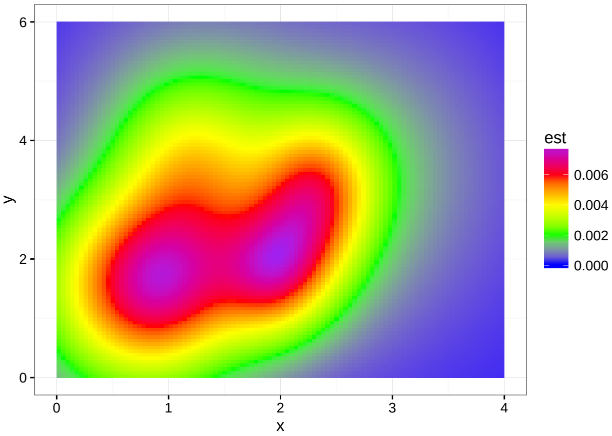
```
res <- nphawkesMSTNH(data = newdata,
                     nbins_m = 8,
                     nbins_t = 25,
                     nbins_r = 25,
                     eps = 1e-3,
                     xrange = c(0,4),
                     yrange = c(0,6),
                     verbose = F
                    )
str(res)
```

```
## List of 14
##  $ mu     : num [1:10000] 0.00142 0.00152 0.00162 0.00173 0.00183 ...
##  $ g      : num [1:25] 4.92 5.29 4.34 5.02 2.61 ...
##  $ f      : num [1:25] 8.45 0 0 0 0 ...
##  $ kp     : num [1:8] 0.253 0.275 0.649 0.875 1.36 ...
##  $ delta_t : num [1:26] 0.001 0.00185 0.00343 0.00636 0.01179 ...
##  $ delta_r : num [1:26] 0.001 0.00143 0.00205 0.00293 0.00419 ...
##  $ delta_m : num [1:9] 0 0.313 0.626 0.939 1.252 ...
##  $ nbins_x : num 100
```

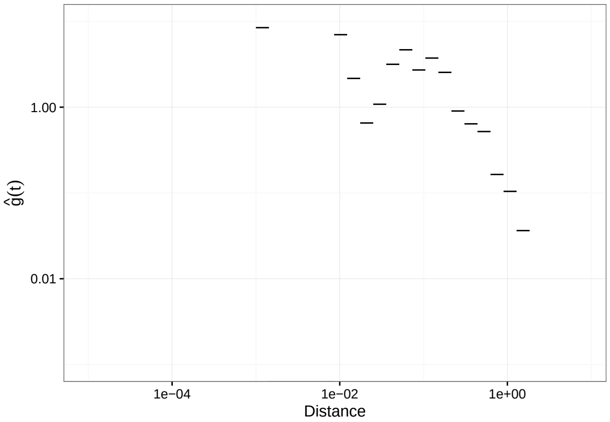
```
##  $ nbins_y : num 100
##  $ tot_time: num 4986
##  $ xrange  : num [1:2] 0 4
##  $ yrange  : num [1:2] 0 6
##  $ np      : num 50
##  $ pbNew   : num [1:570] 1 0.96928 0.00172 0.37037 1 ...
##  - attr(*, "class")= chr "nphawkesMSTNH"
```

```
p <- plot(fit = res, data = newdata, type = 'background')
# Add more layers if needed
p + labs(x = 'x', y = 'y')
```

The plot return a `ggplot2` object, which is convenient since we can add to the plot in layers. For example:



```
p1 <- plot(fit = res, data = newdata, type = 'space')
p1 + scale_y_log10(limits = c(1e-3,1e1))
```



Similarly, we use the `nphawkesMSTNHAC` function which provides model estimation for Marked Spatial Temporal with Anisotropic Spatial Response Conditional on Magnitude and Nonhomogenous Background Rate. For a user, only two additional inputs need to be specified `nbins_a` and `nbins_m2`.

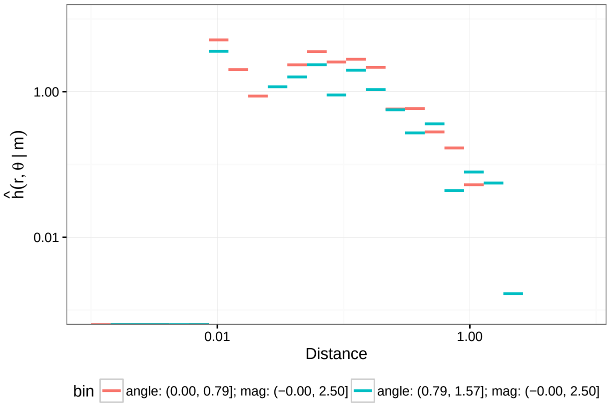
- `nbins_a` The number of direction bins on which to estimate the spatial triggering of the Hawkes Point Process Model
- `nbins_m2` The number of bins on which to estimate the magnitude component of the spatial triggering function.

```
res <- nphawkesMSTNHAC(data = newdata,
                     nbins_m = 8,
                     nbins_t = 25,
                     nbins_r = 25,
                     nbins_a = 2,
                     nbins_m2 = 1,
                     eps = 1e-3,
                     xrange = c(0,4),
                     yrange = c(0,6),
                     verbose = F
                    )
# Note: est_space include angle
str(res)
```

```
## List of 19
##  $ mu       : num [1:10000] 0.00143 0.00153 0.00163 0.00174 0.00184 ...
##  $ g        : num [1:25] 4.85 5.22 4.31 4.73 2.7 ...
##  $ f        : num [1:50] 0.00 0.00 0.00 0.00 2.67e-303 ...
##  $ kp       : num [1:8] 0.254 0.269 0.654 0.895 1.473 ...
##  $ phi      : num [1:570] 0.367 0 0 -0.0717 0 ...
##  $ delta_t  : num [1:26] 0.001 0.00185 0.00343 0.00636 0.01179 ...
##  $ delta_r  : num [1:26] 0.001 0.00143 0.00205 0.00293 0.00419 ...
##  $ delta_a  : num [1:3] 0 0.785 1.571
##  $ delta_arm: num [1:50] 0.000846 0.001211 0.001732 0.002478 0.003545 ...
##  $ delta_m  : num [1:9] -0.000001 0.312952 0.625904 0.938857 1.251809 ...
##  $ delta_m2 : num [1:2] -1.0e-06 2.5
##  $ nbins_x  : num 100
##  $ nbins_y  : num 100
##  $ mag_cut  : num 0
##  $ tot_time : num 4986
##  $ xrange   : num [1:2] 0 4
##  $ yrange   : num [1:2] 0 6
##  $ np       : num 50
##  $ pbNew    : num [1:570] 1 0.9997 0.00134 0.51946 0.99989 ...
##  - attr(*, "class")= chr "nphawkesMSTNHAC"
```

Calling the plot function for spatial distribution on the `nphawkesMSTNHAC` object shows the conditional distribution of triggering distance, given some mainshock magnitude.

```
p1 <- plot(fit = res, data = newdata, type = 'space', print = FALSE)
p1 + scale_y_log10(limits = c(1e-3,1e1))
```



5.3 California Earthquake Catalog Example

Included in the package is an earthquake catalog with over 150,000 events from Jan 1990 to June 2016. The spatial windows is the area surrounding California. Once again we can use the `nphData` function to transform the catalog into the correct format. Below an Isotropic version of MISD if fit using two different background rate smoothing bandwidths. In the first, a variable with bandwidth of 100th nearest neighbors is used. In the second, 20th nearest neighbor is used. We can see a clear distinction the estimated background rates of the two bandwidths. The `nphawkesMSTNHAC` unction also supports fixed bandwidth background state smoothing. See `help(nphawkesMSTNHAC)` for more options. It is worth noting that the estimation procedure in `nphawkesMSTNH` is special case of `nphawkesMSTNHAC` where the number of anisotropy and conditional bins are both equal to 1. See Gordon et. al. 2017 for a full explanation.

```
# Load Included California Catalog
data(catalog)
# Cast to nphData class
newdata <- nphData(data = catalog[catalog$Magnitude > 3.95,],
                   time_var = 'tdiff', x_var = 'Longitude',
                   y_var = 'Latitude', mag = 'Magnitude')
str(newdata)
```

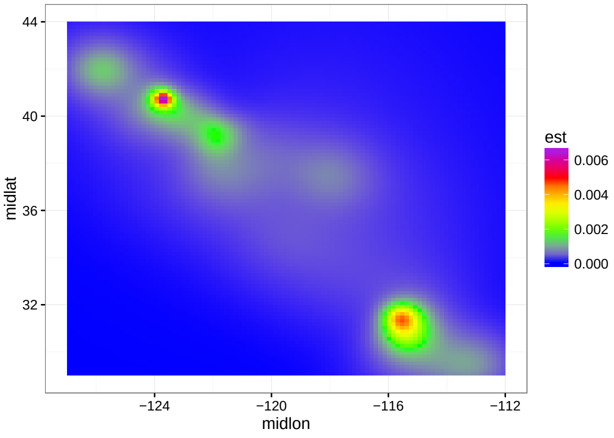
```
## Classes 'nphData' and 'data.frame':  2170 obs. of  6 variables:
##  $ DateTime: POSIXlt, format: "1990-01-05 06:42:10" ...
##  $ lon     : num  -127 -127 -125 -118 -121 ...
##  $ lat     : num  41.7 41.8 40.4 35.2 36.4 ...
##  $ hm      : num  4.4 4.7 4.3 4.01 4.3 4.6 4.53 4.3 5.4 4.9 ...
```

```
##  $ Depth   : num  10 10 4.3 2.59 12.93 ...
##  $ t       : num  4.28 4.3 5.23 10.06 11.38 ...
##  - attr(*, "na.action")=Class 'omit'  Named int 85495
##   .. ..- attr(*, "names")= chr "NA"
```

```
# Smoothing with 100 nearest neighbors
fit <- nphawkesMSTNH(data = newdata, np = 100,
                     eps = 1e-2, xrange = c(-127, -112),
                     yrange = c(29,44))
```

```
## start convergence algorithm...
## Iteration:  1 SupDist:    0.99948986 Nb:  37.22205
...
## Iteration:  14 SupDist:    0.01376935 Nb:  542.8207
```

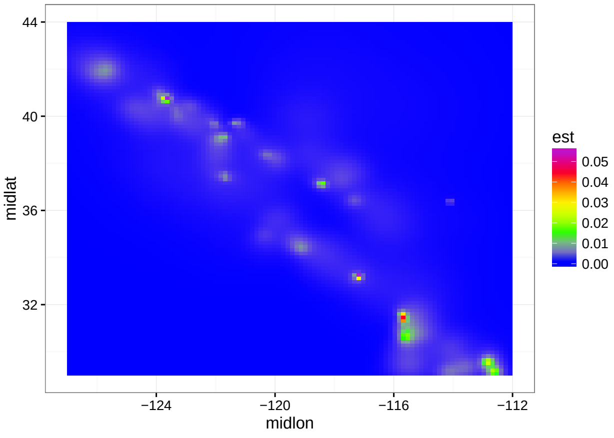
```
plot(fit, newdata, type = 'background')
```



```
# Smoothing with 20 nearest neighbors
fit <- nphawkesMSTNH(data = newdata, np = 20,
                     eps = 1e-2, xrange = c(-127, -112),
                     yrange = c(29,44))
```

```
## start convergence algorithm...
## Iteration:  1 SupDist:    0.99948986 Nb:  73.64299
...
## Iteration:  14 SupDist:    0.01394689 Nb:  832.0025
```

```
plot(fit, newdata, type = 'background')
```



```
# choose a forecast data
forecast_date <- as.Date(max(catalog$DateTime)) + 1
forecast_date
```

5.3.1 Forecasting Seismicity

What if we would like to forecast seismicity? The `forecastMSTH` function provides such ability. For the most accurate forecast of tomorrow, we need the most recent events in the catalog. Since 9 June 2016 is the last available date in this catalog, a retrospective daily forecast will be demonstrated.

```
## [1] "2016-06-10"
```

```
# Try it out, generate one forecasted point pattern
set.seed(54)
myForecast <- forecastMSTH(fit = fit, data = newdata,
                           forecast_date = forecast_date)
str(myForecast)
```

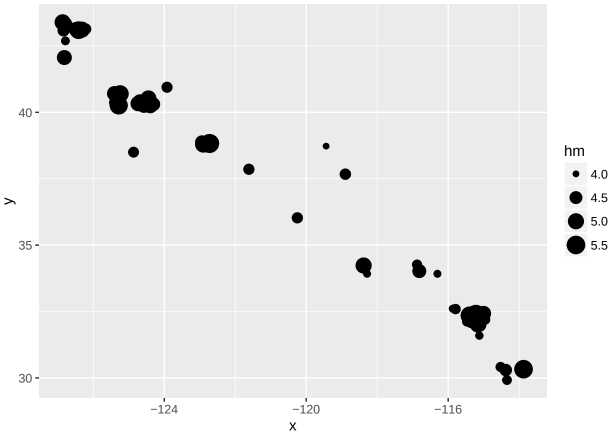
```
## List of 1
##  $ new:List of 4
##   ..$ t  : num 9657
##   ..$ hm : num 3.98
##   ..$ lon: num -123
##   ..$ lat: num 38.9
```

Instead of simply forecasting one time, it is required to forecast many times, say 1000. The `repForecastMSTH` provides this functionality.

```
# Generate Multiple Forecasts
myRepForecast <- repForecastMSTH(fit = fit,
                                 data = newdata,
                                 num_sim = 1000,
                                 forecast_date = forecast_date,
                                 verbose = T)
```

```
## starting repeated forecasts...
## 25 % complete
## 50 % complete
## 75 % complete
## 100 % complete
```

```
# Put all the result in one data.frame
forecastResults <- data.frame(
  x = unlist(myRepForecast$lons),
   y = unlist(myRepForecast$lats),
   hm = unlist(myRepForecast$mags),
   time = unlist(myRepForecast$times)
                              )
p <- ggplot(forecastResults) + geom_point(aes(x = x, y = y, size = hm))
p
```

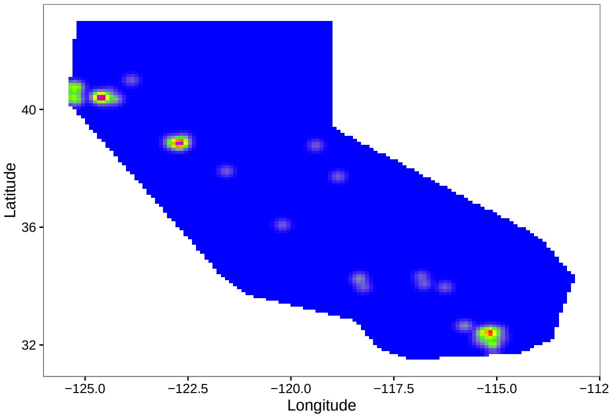


Once forecasts are generated we want to smooth them over the spatial window. Included in the package are many smoothing functions, two of which are demonstrated here. For isotropic smoothing we can use `forecastSmoothing` which enables smoothing of not only the spatial distribution but also the magnitude distribution. If we are interested in only the spatial distribution of forecasted events we can use `quickForecastSmoothing` which ignores the magnitude distribution. Both methods rely on `Rcpp` for high performance. Note: fdata in the function below is generated by the `forecastDataPrep` function. Included is a plotting method for the `nphIsoForecast` class, see `help(plot.nphIsoForecast)` for additional information.

```
# includes magnitude smoothing
res <- forecastSmoothing( x = forecastResults$x,
                          y = forecastResults$y,
                          mags = forecastResults$hm,
                          magbins = csep_mag_grid,
                          lims = c(fdata$xrange, fdata$yrange),
                          nbins = 150,
                          mh = .15,
                          h = rep(.1, nrow(forecastResults)),
                          verbose = FALSE)


# ignores magnitude smoothing
res <- quickForecastSmoothing(x = forecastResults$x,
                              y = forecastResults$y,
                              lims = c(fdata$xrange, fdata$yrange),
                              nbins = 150,
                              mags = forecastResults$hm,
                              h = rep(.1, nrow(forecastResults)),
                              verbose = FALSE)
class(res)
```

```
## [1] "nphIsoForecast"
```

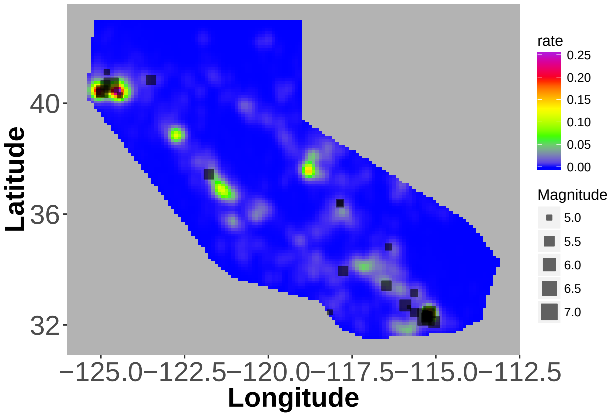
```
plot(res)
```



Also included as a data object is an example CSEP five year forecast. We can use the `plot.csepForecast` method to show the spatial distribution.

```
List of 8
 $ minlon  : num [1:7682] -117 -117 -117 -117 -117 ...
 $ maxlat  : num [1:7682] 31.6 31.6 31.6 31.6 31.6 31.6 31.6 31.6 31.7 31.7 ...
 $ maxlon  : num [1:7682] -117 -117 -117 -117 -117 ...
 $ minlat  : num [1:7682] 31.5 31.5 31.5 31.5 31.5 31.5 31.5 31.5 31.6 31.6 ...
 $ mindepth: num [1:7682] 0 0 0 0 0 0 0 0 0 0 ...
 $ maxdepth: num [1:7682] 30 30 30 30 30 30 30 30 30 30 ...
 $ rate    : num [1:7682] 1.85e-04 1.04e-04 8.17e-05 1.06e-04 9.80e-05 ...
 $ mask    : num [1:7682] 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "row.names")= int [1:7682] 1 2 3 4 5 6 7 8 9 10 ...
 - attr(*, "class")= chr "csepForecast"
```

```
data(csepFiveYearForecast)
data("catalog495")
p <- plot(csepFiveYearForecast, legend = T)
p + geom_point(aes(x = Longitude, y = Latitude, size = Magnitude),
               data = catalog495,
               alpha = .6,
               shape = 22,
               colour = "black",
               fill = "black")
```



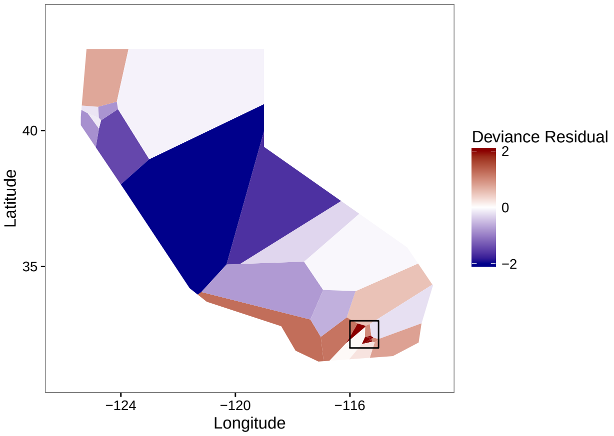
Once forecasts are generated, we can apply various residual analysis techniques to them such as Voronoi Residuals. Voronoi tessellation is available in the `deldir` package, which is a vital dependency for the `NonParametricHawkes` package. A simple workflow is demonstrated: first we tesselate the observed point pattern, second we extract the vertices of the tiles of the Voronoi tessellation. Next, if needed, we account for any boundary issues by augmenting the tessellation to only include areas inside the CSEP spatial window (known as Regional Earthquake Likelihood Models or RELM). Lastly we transform the `list` into a `data.frame` which is easier to work with. Lastly we call the voronoiResiduals function to calculate the raw and scaled residuals for the forecast in question. The function also return an object called total rate which is the integrated conditional intensity over each cell. Once the data is joined to the original tiles, we can plot the resulting using `ggplot2`, specifically the `geom_map` functional allows us the draw and fill the tessellation.

```
deldf <- deldir(data.frame(x = catalog495$Longitude, y = catalog495$Latitude), rw=c(-127, -112, 29, 44))
tl <- tile.list(deldf)
tl <- relmTileFix(tl)
tiles <- extractTiles(tl)

vor4 <- voronoiResiduals(forecast = data.frame(unclass(csepFiveYearForecast)), tl = tl)
```

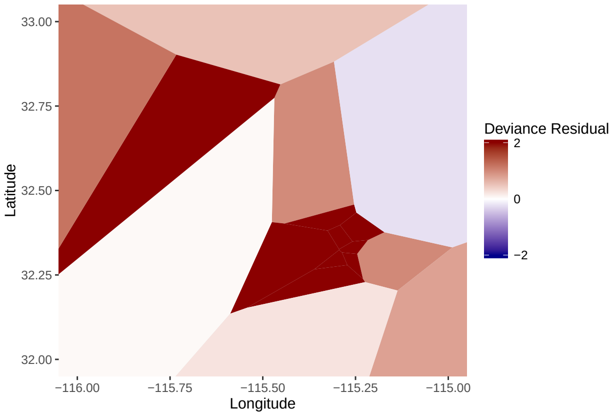
```
## [1] 5
## [1] 10
## [1] 15
## [1] 20
## [1] 25
## [1] 30
```

```
res_df4 <- data.frame(id = 1:length(vor4$scaled.res),
                      scaled.res = vor4$scaled.res,
                      raw.res = vor4$raw.res,
                      total.rate = vor4$total.rate)
tiles_res4 <- left_join(tiles, res_df4, by = ’id’)


p1 <- ggplot(tiles_res4) +
  theme_bw() +
  xlim(c(-126,-113)) +
  ylim(c(31,44)) +
  geom_map(map = tiles,
           aes(x = x, y = y, map_id=id, fill = scaled.res),
           size=0.25)  +
  scale_fill_gradientn(
    colours = c(’darkblue’, ’white’, ’darkred’), values = NULL ,
    breaks = c(-2, 0, 2),
    limits = c(-2,2),
    oob = scales::squish,
    guide   = guide_colourbar(title = ’Deviance Residual’,
                              title.vjust = ’1’)) +
  theme( panel.grid.major = element_blank(),
         panel.grid.minor = element_blank()) +
  labs(x = ’Longitude’, y = ’Latitude’) +
  geom_polygon(data = data.frame(x1 = c(-116,-116, -115,-115),
                                 y1 = c(32,33,33,32)),
```

```
                                      mapping = aes(x1,y1),
                                      fill = NA,
                                      colour = 'black')
p1
```



```
p2 <- ggplot(tiles_res4) +
  xlim(c(-116,-115)) +
  ylim(c(32,33)) +
  geom_map(map = tiles,
    aes(x = x, y = y, map_id=id, fill = scaled.res), size=0.25)  +
  scale_fill_gradientn(
    colours = c('darkblue', 'white', 'darkred'), values = NULL ,
    breaks = c(-2, 0, 2), limits = c(-2,2), oob = scales::squish,
    guide   = guide_colourbar(title = 'Deviance Residual', title.vjust = '1')) +
  theme( panel.grid.major = element_blank(),
         panel.grid.minor = element_blank()) +
  labs(title = '', x = 'Longitude', y = 'Latitude')

p2
```



CHAPTER 6

Discussion

Development of both short and long-term earthquake forecasts are still areas that require much work. The methods demonstrated in this Dissertation provide some fundamentals for estimation, prediction, and evaluation of these seismological models. The MISD methods demonstrated in Chapter 2 performs well in both describing and forecasting earthquake occurrences. Yet, since the algorithm requires an iterative approach, fast estimation can better be accomplished using the methods demonstrated in Chapter 3. Additional opportunities exist to test and improve our earthquake forecasts using the methods outlined in Chapters 2 and 4. Voronoi residuals, as well as Deviance residuals, and super-thinned residuals provide some great insights into where models perform well or poorly relative to one another. The `NonParametricHawkes` R package provide easy to use, well documented and reproducible examples which may prove vital to making further improvements to these methods in the future.

Further research is needed to assess and evaluate the two only day forecast models we have submitted to CSEP. The models began running in Oct 2016 and have been generating daily forecasts ever since. Evaluation of daily forecasts may require innovation in the visualization of temporal residuals. The methods introduced earlier are primarily focused on static representation which allows the comparison of data features without a temporal dimension. Yet, in the case of daily forecasts, it may make sense to focus on dynamic representation. Advantages of dynamic representation are that all data is shown and temporal trends can be observed.

Recently, fault angles based on focal mechanisms, as opposed to historic seismicity, have been used to describe the anisotropic spatial distribution of events following a mainshock.

These moment tensor estimates can be incorporated into the MISD algorithm demonstrated in Chapter 2 or perhaps used in conjunction with MISD through a semi parametric approach. While local fault estimation was shown to work quiet well, the goal of including the best available geophysical measurements should remain a key research goal.

The primarily application of this research was an application for California seismicity. Perhaps another avenue of research would be to apply these methods to global tectonic zones to see what can be learned about the seismicity in those regions. Similarly, more work is needed to better generalize the `NonParametricHawkes` R package, thus increasing its relevance to statisticians and scientists around the world.

CHAPTER 7

Appendix

7.1 Appendix A: Algorithm A

1. Initialize $P(0)$, set iteration index $v = 0$.

2. Estimate inhomogenous background rate $\mu(x, y)$

$$\mu_{k,\ell} = \frac{1}{T\Delta x\Delta y}\sum_{D_{k,\ell}} p_{ii}^{(v)}, \quad k = 0, ..., n_x^{bins} - 1, \ \ell = 0, ..., n_y^{bins} - 1.$$

3. Estimate triggering components $k(m), g(t)$ and $h(r, \theta|m)$:

$$\begin{aligned}
\kappa(m)_k^{(v)} &= \frac{\sum_{A_k} p_{ij}^{(v)}}{N_k^{mag}}, && k = 0, ..., n_m^{bins} - 1;\\
g(t)_k^{(v)} &= \frac{\sum_{B_k} p_{ij}^{(v)}}{\Delta t_k \sum_{i=1}^{N}\sum_{j=1}^{i-1} p_{ij}^{(v)}}, && k = 0, ..., n_t^{bins} - 1;\\
h(r, \theta, m)_{k,\ell,q}^{(v)} &= \frac{\sum_{C_{k,\ell,q}} p_{ij}^{(v)}}{\Delta r_k \Delta\theta_\ell \Delta m_q \sum_{i=1}^{N}\sum_{j=1}^{i-1} p_{ij}^{(v)}}, && k = 0, ..., n_r^{bins} - 1,\\
& && \ell = 0, ..., n_\theta^{bins} - 1, \quad q = 0, ..., n_{m2}^{bins} - 1;\\
h(m)_q^{(v)} &= \frac{\sum_{E_q} p_{ij}^{(v)}}{\Delta m_q \sum_{i=1}^{N}\sum_{j=1}^{i-1} p_{ij}^{(v)}}, && q = 0, ..., n_{m2}^{bins} - 1;
\end{aligned}$$

4. Update probabilities $P^{(v+1)}$ after transforming polar $h^{(v)}$ to Cartesian $f^{(v)}$.

$$\begin{aligned}
p_{ij}^{(v+1)} &= \frac{\kappa^{(v)}(m_j)g^{(v)}(t_i - t_j)f^{(v)}(r_{ij}, \theta_{ij}|m_j)}{\mu^{(v)}(x_i, y_i) + \sum_{j=1}^{i-1}\kappa^{(v)}(m_j)g^{(v)}(t_i - t_j)f^{(v)}(r_{ij}, \theta_{ij}|m_j)},\\
p_{ii}^{(v+1)} &= \frac{\mu^{(v)}(x_i, y_i)}{\mu^{(v)}(x_i, y_i) + \sum_{j=1}^{i-1}\kappa^{(v)}(m_j)g^{(v)}(t_i - t_j)f^{(v)}(r_{ij}, \theta_{ij}|m_j)}.
\end{aligned}$$

5. Run the algorithm until convergence.

The notation for the above is defined as follows:

- n_x^{bins} and n_y^{bins} are the number of bins used for estimating the background rate.
- Δx and Δy are bin widths and $D_{k,\ell}$ is defined as the pairs of points that fall in the k and ℓth bins respectively such that: $D_{k,\ell} = \{i : (k-1)\Delta x < x_i \leq k\Delta x, (\ell - 1)\Delta y < y_i \leq \ell \Delta y\}$.
- $n_m^{bins}, n_t^{bins}, n_r^{bins}, n_\theta^{bins}$ are the number of bins used in estimating the triggering function components for magnitude, time, and space.
- n_{m2}^{bins} is the number of bins used in estimating the spatial-magnitude distribution.
- N_k^{mag} is the number of earthquakes in whose magnitude fall in the k^{th} interval: $\sum_{j=1}^{N} \mathbb{1}_{\delta m_k < m_j \leq m_{k+1}}$.
- $A_k = \left\{(i,j) \middle| \delta m_k \leq m_j \leq \delta m_{k+1}, i > j\right\}$ is the set of indices of all pairs of earthquakes whose mainshock magnitudes fall within the k^{th} bin of the histogram estimator $\kappa(m)$.
- $B_k = \left\{(i,j) \middle| \delta t_k \leq t_i - t_j \leq \delta t_{k+1}, i > j\right\}$ is the set of indices of all pairs of earthquakes whose inter event times fall within the k^{th} bin of the histogram estimator $g(t)$.
- $C_{k,\ell,q} = \left\{(i,j) \middle| \delta r_k \leq r_{ij} \leq \delta r_{k+1},\ \delta\theta_\ell \leq \theta_{ij} \leq \theta_{\ell+1},\ \delta m_q \leq m_j \leq \delta m_{q+1}, i > j\right\}$ is the set of indices of all pairs of earthquakes whose epicentral distances fall within the k^{th} bin and whose angular separation fall in the ℓ^{th} bin, and mainshock magnitude fall in the q^{th} bin of the histogram estimator $h(r, \theta, m)$.
- $E_q = \left\{(i,j) \middle| \delta m_q \leq m_j \leq \delta m_{q+1}, i > j\right\}$ is the set of indices of all pairs of earthquakes whose mainshock magnitude fall in the q^{th} bin of the histogram estimator $h(m)$.
- $\Delta m_q = \delta m_{q+1} - \delta m_q$ is the size of the q^{th} bin shared by both the joint and marginal probability distributions for magnitude in h.

The approach shares some similarities to the fault distance approach described in Marsan and Lengliné (2008). However, in their approach, distance was calculated between an earthquake j and the fault plane of earthquake i, as opposed to epicentral distance. The fault plane was defined as the least squares line passing through the centroid of events within a rupture plane, as opposed to the mainshock epicenter itself.

7.2 Appendix B: Algorithm B

Algorithm B1: Generating the background catalog of Algorithm for Equations 1.2 and 2.1

1. Let N_{days} be the number of days in the forecast window $t, t + \Delta$, and T be the length of time (days) of the fitting catalog. Let γ be the estimated number of background events in the forecast window, such that $\gamma = N_{days} \cdot \frac{\sum_{i=1}^{N} p_{ii}}{T}$.
2. Let μ^{max} be the maximum of the background density $\hat{\mu}(x, y)$.
3. Generate a Poisson random variable N with mean $\lambda = \mu^{max} \cdot |S|$.
4. Uniformly distribute the locations of N events over the spatial-temporal window.
5. Generate U_i, a uniformly distributed random variable.
6. Retain each event if $U_i < \gamma \cdot \frac{\hat{\mu}(x, y)}{\mu^{max}}$.
7. Let N(0) be the number of retained events.
8. Draw a sample of size N(0) magnitudes from the magnitudes of events in the catalog $\{m_i : i = 1, ..., N\}$
9. Let generation 0, $G'(0) = \{(t_i^{(0)}, x_i^{(0)}, y_i^{(0)}, m_i^{(0)})\}$, be the set of times, locations, and magnitudes of the background events.

Algorithm B2: Forecast the triggering process for (1.2)

1. Set generation index $v \leftarrow 0$

2. Let the initial catalog $G(0)$ be the union of all observed events in the catalog and the simulated background events in $G'(0)$.

3. For each event i in $G^{(0)}$, generate $k = 1, \ldots, N^{(i)}$ offspring where $N^{(i)}$ is a Poisson random variable with mean $\kappa(m_i)$. Generate offspring t_k, m_k from the temporal response function $g(t - t_i)$ and empirical magnitude distribution $s(m_i)$.

4. Locations x_k, y_k are generated by sampling a distance from spatial response function $f(x - x_i, y - y_i)$ and a direction by sampling a Uniform random variable in $(0, 2\pi)$.

5. Let offspring $O_i'(v) \leftarrow \{(t_k, x_k, y_k, m_k) : t_k \in [t, t + \Delta t]\}$.

6. Keep only the offspring which occurred in $[t, t + \Delta t]$, thus set $G^{(v+1)} \leftarrow \cup_{i \in G^{(\ell)}} O_i'(v)$.

7. If $G^{(\ell)}$ is not empty (there are new simulated events in $[t, t+\Delta t]$), set generation index $v \leftarrow v + 1$, else return all events in $t, t + \Delta t : G'(0) \cup \left(\cup_{j=1}^{v} G^{(j)} \right)$.

Simulating from the anisotropic triggering process of (2.1) once again requires the use of a lower magnitude cut-off for fault estimation. Although the entire catalog up to time t is used to estimate faults for forecasted events, the lower magnitude catalog is not used to generate new events. At each iteration of the simulation we re-estimate fault directionality of new events, using the entire catalog, as opposed to assigning the fault direction of the preceding fault. Algorithm B3 is designed such that the if number of direction bins and magnitude bins in the spatial response function are both to set to 1, Algorithm B3 reduces to the Algorithm B2.

Algorithm B3: Forecast of the triggering process for Algorithm A

1. Set the lower magnitude cut-off c_1, upper magnitude cut-off c_2, and let the fault estimation catalog H be all events where $m \geq c_1$.

2. Set generation index $v \leftarrow 0$.

3. Let the initial catalog $G(0)$ be the union of all observed events greater or equal to $m \geq c_2$ and the simulated background events in $G'(0)$ ((2.1)).

4. For each event i in $G^{(0)}$, estimate a fault orientation ϕ_i associated with each event using catalog H.

5. Set $\theta_i = \arctan(|\phi_i|)$.

6. For each event i in $G^{(0)}$, generate $k = 1, \ldots, N^{(i)}$ offspring where $N^{(i)}$ is a Poisson random variable with mean $\kappa(m_i)$. t_k is generated from the temporal response function $g(t - t_i)$ and m_k from the magnitude distribution $s(m_i)$.

7. Locations x_k, y_k are generated from spatial response function $f(x - x_i, y - y_i, \theta_i | m_i)$ with distance r_k and direction θ_k.

8. Define R as a 4 directional rotation distribution:

R_θ	R_π
1	0
1	1
-1	0
-1	1

9. Since the spatial response is estimated on a quarter circle, sample $k = 1, \ldots, N^{(i)}$ rotations uniformly from rotation R.

10. Let direction $\theta'_k = \theta_i + \theta_k \cdot R_\theta^{(k)} + \pi \cdot R_\pi^{(k)}$.

11. Set $x_k = x_i + \cos(\theta'_k) \cdot r_k$ and $y_k = y_i + \sin(\theta'_k) \cdot r_k$.

12. Let $O'_i(\ell) \leftarrow \{(t_k, x_k, y_k, m_k) : t_k \in [t, t + \Delta t]\}$.

13. Keep only the offspring which occurred in $[t, t + \Delta t]$, thus set $G^{(v+1)} \leftarrow \cup_{i \in G^{(\ell)}} O'_i(v)$.

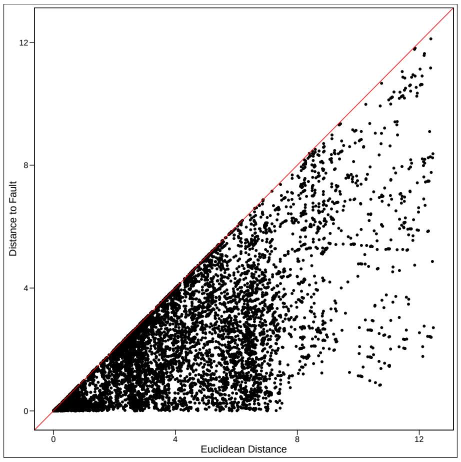
14. If $G^{(\ell)}$ is not empty (there are new simulated events in $[t, t+\Delta t]$), set generation index $v \leftarrow v + 1$, else return all events in $t, t + \Delta t : G'(0) \cup \left(\cup_{j=1}^{v} G^{(j)}\right)$.

7.3 Appendix C1: MISD (1.2) Distance to Fault

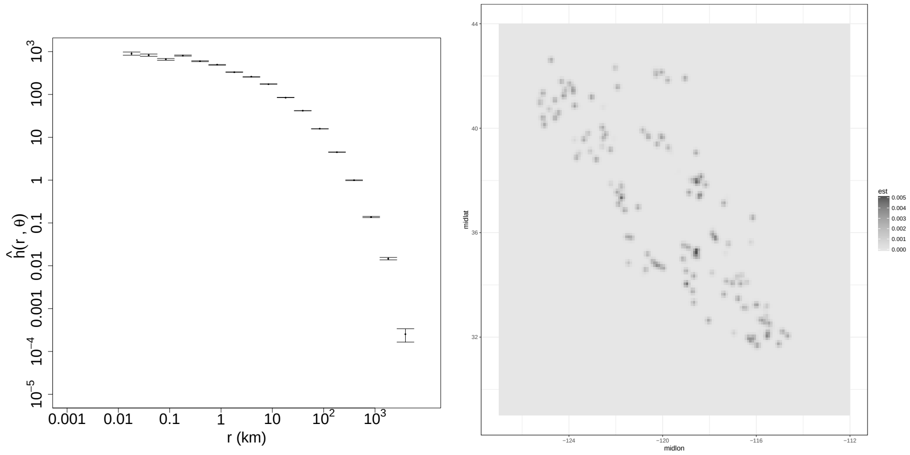
The orientation of a fault on which a mainshock occurs provides a reasonable approximation to an events rupture mechanics. By approximating a fault we are able to estimate the spatial distribution of aftershocks using distance from the fault rather than epicentral distance. To estimate a fault we used the location of all nearby events, both before and after, and performed a weighted least squares regression.

Fault orientation was estimated by setting the magnitude cut offs to $c_1 = c_2 = 4.0$ and a surrounding area of .045 degrees by .045 degrees (roughly 100km^2) for each event in the catalog. All input parameters were set identically to (1.2) in the epicentral approach outlined of Section 2.3.

Figure 7.1(b) shows a scatter plot comparing the epicentral distance to fault distance for a sample of earthquakes in the catalog. We can see that epicentral distance is strictly greater than distance from fault. The method (provided in Appendix C2) estimated a larger proportion of aftershocks then the model using using epicentral distance (see Fox et al. (2016) supplement for more details). Figure 7.1)(b) shows the estimate of the underlying spatial inhomogeneous Poisson processes $\mu(x, y)$ for mainshock activity. Although the background rate is much lower, its spatial distribution is somewhat similar to the background rate described in Section 2.3. Histogram estimates of the components of the spatial triggering function are shown in Figure 7.1)(b). Although not shown, compared to the estimate derived using epicentral distance in Figure 2.6, the density estimate of $g(t)$ and the magnitude productivity function $\kappa(m)$ are quite similar, however, the estimates (shown) for the spatial distribution $h(r)$ shows a much higher rate of triggering at shorter distances. The results for both the background rate and triggering function suggest that epicentral distance outperforms distance to fault.



Epicentral vs. Fault Distance



$\hat{h}(r)$ $\mu(x, y)$

Figure 7.1: Fault Distance

7.4 Appendix C2: Fault Distance Methods

Using the following steps we can approximate a fault for each mainshock:

1. Let earthquake j be called the mainshock. Using the entire catalog of events, both before and after the mainshock, identify all $m_i \geq c_1$ events within some area surrounding a $m_j \geq c_2$ event, where $c_1 \leq c_2$. Here c is a magnitude cut off and m is the magnitude of an earthquake.

2. Calculate the Euclidean distance between a m_j earthquakes epicenter and all other m_i

events denoted r_{ij}.

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

3. Fit a weighted ordinary least squares regression for each event j constrained to pass through the m_j epicenter with weights $w_j = 1/r_{ij}$.
4. Store the coefficient of the weighted regression β_j as the estimated fault plane for each m_j event.

If any event is in the same exact location as the mainshock as a result of a location error, we remove the event since it adds no information to the fault estimate.

Distance to fault was calculated by computing the perpendicular distance for each pair (i, j) of earthquakes, where $t_i > t_j$, in the catalog using the fault approximated by the method described above. Since the regression line must pass through the mainshock location we can rewrite the regression equation as:

$$-\beta_j \text{ Longitude}_i + \text{Latitude}_i = 0,$$

and apply the formula:

$$d_{ij} = \frac{|-\beta_j \text{ Longitude}_i + \text{Latitude}_i|}{\sqrt{\beta_j^2 + 1^2}}.$$

7.5 Appendix D: Model Selection

Following an approach similar to Helmstetter et al. (2007), to estimate the expected number of events within the forecast period we counted the number of events per year in the learning catalog. In this case, 184 events fell within the CSEP testing region over a time of 24.64 years, giving us $N_{pred} = 37.33$ events in the forecast window. Helmstetter et al. (2007) conducted a procedure to remove potential explosions from the catalog, and as such estimated $N_{pred} = 35.40$ during the same period.

Selection of model inputs was conducted using loglikelihood ratios of our models with that of Helmstetter et al. (2007), to identify the best inputs for the nonparametric estimation and forecasting. We find that larger bandwidths result in poorer performance while including smaller events in the catalog improves our results. In addition, the forecasts appears to improve when more directional bins are used in conjunction with slight directional smoothing while magnitude scaling in the spatial triggering function does not substantially improve forecast performance. Table 2.1 shows selected model inputs for (1.2,2.1) as determined by the overall deviance of the model compared with that of Helmstetter et al. (2007).

7.6 Appendix E: Computational issues.

E1. Extension to the case where the binwidths Δ are not equal.

Formulas (7-9) are simplified somewhat by the assumption that the bins U_k are all of size Δ. However, these formulas can easily be amended when the bins are of unequal sizes. Indeed, equation (7) would remain unchanged but instead of the definition $b = n|S|\Delta\mathbf{1}$, we would have $b = n|S||U|\mathbf{1}$, where $|U|$ is the vector of bin sizes $\{U_1, U_2, ..., U_n\}$. Formulas 8 and 9 would remain exactly the same, but again with $b = n|S||U|\mathbf{1}$ in equation (9). The method for smoothing $\hat{\beta}$ should be adjusted to account for the modification to unequal bin widths as well. For instance, in smoothing $\hat{\beta}$ via simple moving average (MA) filtering as is done in Figures 3 and 4, one must be careful to account for the varying binwidths when smoothing. In these cases, the values of $\hat{\beta}$ were smoothed using MA filtering with a Gaussian kernel applied to the distances from the bin centers. One could also weight each bin according to its binwidth when smoothing by MA filtering, though simple unweighted MA filtering was implemented for Figures 3 and 4.

E2. Extension to the case where μ and/or K are unknown.

When the parameters μ and/or K are unknown, one can estimate g and subsequently estimate μ and/or K by maximum likelihood using e.g. a gradient descent search algorithm. Alternatively, one may estimate μ and K simultaneously with the β_k. Suppose both μ and K are unknown. Consider augmenting the vector β, so that $\beta_{p+1} = \mu$ and $\beta_{p+2} = K$, so that now β has length $p + 2$. Since $\partial L/\partial\mu = \sum_j 1/\lambda(\tau_j) - ST$ and $\partial L/\partial K \approx \sum_j A[k, j]\beta_k - n$, where in the approximation above we are ignoring mass of the triggering function g outside the observed spatial-temporal window, triggered by events inside the spatial-temporal window. This approximation would be exact if we were observing the process over infinite time and space. Setting these partial derivatives to zero, we obtain 2 additional linear equations, $\sum_{j=1}^{n} 1/\lambda(\tau_j) = ST$ and $\sum_{j=1}^{n} \sum_{k=1}^{n} A[k,j]\beta_k/\lambda(\tau_j) = n$. One can thus amend equations (7-9) accordingly to solve simultaneously for the $p+2$ unknowns in the augmented vector β, using the p equations in (7-8) and the two equations above.

E3. The case where A is not invertible.

A problem with the method proposed here is that in many cases the adjacency matrix A will not be of rank n, and thus the solution in (9), which requires inversion of A, cannot be obtained. In particular, if there is any interval U_k such that no observed pairs of points τ_i and τ_j satisfy $\tau_j - \tau_i \in U_k$, then $A[k,m] = 0$ for all m, and this column of all zeros will render A singular.

One way of dealing with such situations is to let the intervals U_k depend on the observations, choosing the intervals U_k carefully to ensure that each such interval corresponds to at least one time interval $\tau_j - \tau_i$. For instance, one may choose some very small constant $\epsilon > 0$, and let the time windows be $[u_i + \epsilon, u_{i+1} + \epsilon)$, where $u_i = \tau_{i+1} - \tau_i$ are the observed times between successive events. This ensures that each row and each column of A has at least one nonzero element. This alone seems to be sufficient in most cases to render A invertible. Little is known about the theoretical properties of MLEs when the parameters are defined within random intervals rather than fixed intervals, but this method was employed here throughout and seems to yield good results.

Even with these intervals carefully selected this way, it is still possible for A to have rank less than n. In such cases, one may opt to choose elements of A at random and augment them by one until A is invertible; the result seems to have relatively little effect on the resulting estimates. An alternative we explored briefly is to perturb all the elements of A very slightly by adding independent normal random variables with mean zero and very small variance, but this resulted in drastic changes to the resulting estimates and added substantial variability to the results. Better solutions to the problem of singularity of the adjacency matrix remains an important area for future work in order to make the method proposed here more stable and robust.

E4. Smoothing the estimates $\hat{\beta}$.

Determining the proper amount of smoothing for the estimates generated using equation (9) seems to be an extremely difficult problem. In the simulations in Section 5, we simply applied a Gaussian moving average filter to the estimates $\hat{\beta}$. The bandwidths selected for this filter were 1100 for the exponential and truncated normal, 880 for the uniform, and 500 for the Pareto triggering density. In general, a bandwidth of roughly $n/4$ or $n/3$ seems reasonable, but further study is needed in order to determine what bandwidth works optimally with the estimates proposed here using equation (9). As mentioned in item 1 of this Appendix, when unequal binwidths are used in defining the intervals U_k corresponding to the parameters β_k being estimated, one may modify the simple MA filtering in order to account for the varying binwidths, e.g. by applying the Gaussian kernel to the time difference between the midpoints of the corresponding intervals, and/or by weighting each parameter estimate $\hat{\beta}$ by the size of its corresponding interval. Alternative, more elaborate smoothing methods could also be used, including techniques involving more complex filters or splines. Future research should focus on the possibility of smoothing and inverting the matrix A in tandem, to address not only the problem of variability in the estimates $\hat{\beta}$ but also the singularity problem described in item 3 of this Appendix.

Bibliography

Adelfio, G. and Schoenberg, F.P. (2009). Point process diagnostics based on weighted second-order statistics and their asymptotic properties. *A*nnals of the Institute of Statistical Mathematics, **61**(4), 929-948.

Bacry, E. Dayri, K. and Muzy, J.F (2012).. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *Eur. Phys. J.* **B** **85**(157)

Baddeley, A., Turner, R., Möller, J., and Hazelton, M. (2005). Residual analysis for spatial point processes (with discussion). *J*ournal of the Royal Statistical Society, series B, **67**(5), 617-666.

Barr, C.D., and Schoenberg, F. P. (2010). On the Voronoi estimator for the intensity of an inhomogeneous planar Poisson process. *B*iometrika, **97**(4), 977-984.

Bray, A., and Schoenberg, F.P. (2013). Assessment of point process models for earthquake forecasting. *S*tatistical Science, **28**(4), 510-520.

Bray, A., Wong, K., Barr, C.D., and Schoenberg, F.P. (2014). Voronoi cell based residual analysis of spatial point process models with applications to Southern California earthquake forecasts. *A*nnals of Applied Statistics, **8**(4), 2247-2267.

Bolt, B. (2006). Earthquakes, 5th ed. W.H. Freeman, New York.

Brémaud, P., 1981. Point Processes and Queues. Springer, NY.

Clements, R.A., Schoenberg, F.P., and Schorlemmer, D. (2011). Residual analysis for space-time point processes with applications to earthquake forecast models in California. *A*nnals of Applied Statistics, **5**(4), 2549-2571.

Clements, R.A., Schoenberg, F.P., and Veen, A. (2012). Evaluation of space-time point process models using super-thinning. *E*nvironmetrics, **23**(7), 606-616.

Daley, D., and Vere-Jones, D. (1988). *A*n Introduction to the Theory of Point Processes. Springer, New York.

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

Daley, D., and Vere-Jones, D. (2007), An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure, New York: Springer.

Field, E. H. (2007). Overview of the working group for the development of Regional Earthquake Likelihood Models (RELM). *S*eismological Research Letters **78** 7-16.

Fox, E.W., Schoenberg, F.P., and Gordon, J.S. (2016). Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. *A*nnals of Applied Statistics, **10**(3), 1725-1756.

Gordon, J.S., Clements, R.A., Schoenberg, F.P., and Schorlemmer, D. (2015). Voronoi residuals and other residual analyses applied to CSEP earthquake forecasts. *S*patial Statistics, 14b, 133-150.

Gutenberg, B. and Richter, C.F. (1944). Frequency of earthquakes in California. *B*ull. Seismol. Soc. Amer. **142**, 185-188.

Hawkes, A. G.(1971). Spectra of some self-exciting and mutually exciting point processes. *B*iometrika **58** 83-90.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2006). Comparison of short-term and time-independent earthquake forecast models for southern California. *B*ulletin of the Seismological Society of America, 96 (1), 90-106.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2007), High-resolution time-independent grid-based forecast for M = 5 earthquakes in *C*alifornia, Seismol. Res. Lett., 78 (1), 78 - 86.

Hinde, A.L. and Miles, R.E. (1980). Monte Carlo estimates of the distributions of the random polygons of the Voronoi tessellation with respect to a Poisson process. *J*. Statist. Comput. Simul., **10**, 205-223.

Jordan, T. (2006). Earthquake predictability, brick by brick. *S*eismological Research Letters **77** 3-6.

Kagan, Y.Y. (2002). Aftershock zone scaling. *B*ull. Seism. Soc. Am. **92**(2), 641–655.

Kagan, Y.Y. (2003). Accuracy of modern global earthquake catalogs. *P*hysics of the Earth and Planetary Interiors **135**, 173-209.

Kagan, Y. Y., Jackson, D. D. and Rong, Y. (2007). A testable five-year forecast of moderate and large earthquakes in southern California based on smoothed seismicity. *S*eismological Research Letters **78**, 94-98.

Kagan, Y.Y., and Jackson, D.D. (2014). Statistical earthquake focal mechanism forecasts. *G*eophysical J. Int. 197(1), 620-629.

Kanamori, H., and D. Anderson (1975). Theoretical basis of some empirical relations in seismology. *B*ull. Seism. Soc. Am. **65**, 1073–1095.

Lewis, E. and Mohler, G (2011). A Nonparametric EM algorithm for Multiscale Hawkes Processes. *J*ournal of Nonparametric Statistics.

Lewis, P. A. and Shedler, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. Naval Research Logistics Quarterly 26 403âĂŞ413.

Marsan, D. and Lengliné, O. (2008). Extending earthquakes' reach through cascading. *S*cience **319** 1076-1079.

Marsan, D. and Lengliné, O. (2010). A New Estimation of the Decay of Aftershock Density With Distance to the Mainshock. *J*. Geophys. Res. **115**

Meijering, J.L. (1953). Interface area, edge length, and number of vertices in crystal aggregation with random nucleation. *P*hilips Research Reports, **8**, 270-290.

NCEDC (2014), Northern California Earthquake Data Center. UC Berkeley Seismological Laboratory. Dataset. doi:10.7932/NCEDC.

Okabe, A., Boots, B. Sugihara, K., and Chiu, S. (2000) Spatial Tessellations, 2nd ed. Wiley, Chichester.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *A*nn. Inst. Statist. Math., **30**, 243–261.

Ogata, Y. 1988. Statistical models for earthquake occurrences and residual analysis for point processes.. *J*. Amer. Statist. Assoc. **83** 9-27.

Ogata, Y. 1998. Space-Time Point-Process Models for Earthquake Occurrences. *A*nnals of the Institute of Statistical Mathematics **50** 379-402.

Rathbun, S.L., Cressie, N. 1994. Asymptotic properties of estimators for the parameters of spatial inhomogeneous Poisson point processes. Adv. Appl. Probab. 26, 122–154.

Rhoades, D.A., Schorlemmer, D., Gerstenberger, M.C., Christophersen, A., Zechar, J.D., and Imoto, M. (2011). Efficient testing of earthquake forecasting models. *A*cta Geophysica, **59**(4), 728-747.

Schneider, M., Clements, R., Rhoades, D., Schorlemmer, D. (2014). Likelihood- and residual-based evaluation of medium-term earthquake forecast models for California. *G*eophysical Journal International, **198**(3), 1307-1318.

Schoenberg, F.P. (2003). Multi-dimensional residual analysis of point process models for earthquake occurrences. *J*. Amer. Statist. Assoc., **98**(464), 789-795.

Schorlemmer, D., and M. Gerstenberger (2007), RELM testing center, *S*eismol. Res. Lett., 78(1), 30.

Schorlemmer, D., Gerstenberger, M., Wiemer, S., Jackson, D. and Rhoades, D. (2007), Earthquake likelihood model testing. *S*eismological Research Letters, **78**, 17-27.

Schorlemmer, D., Zechar, M. Werner, D. D. Jackson, E. H. Field, T. H. Jordan, and the RELM Working Group (2010). First results of the Regional Earthquake Likelihood Models experiment. Pure and Applied Geophysics, **167**, 859–876.

Shen, Z., Jackson, D. D. and Kagan, Y. Y. (2007). Implications of geodetic strain rate for future earthquakes, with a five-year forecast of M5 earthquakes in southern California. *S*eismological Research Letters **78**, 116-120.

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

Werner, M. J., A. Helmstetter, D. D. Jackson, and Y. Y. Kagan (2011), High resolution long- and short-term earthquake forecasts for California, *B*ulletin of the Seismological Society of America 101(4), 1630–1648.

Utsu, T. (1999). Representation and analysis of the earthquake size distribution: a review and some new approaches. *P*ure Appl. Geophys. **155**, 509-535.

Wong, K. and Schoenberg, F. P. (2009). On mainshock focal mechanisms and the spatial distribution of aftershocks. *Bulletin of the Seismological Society of America* **99(6)** 3402-3412.

Zhuang, J., Y. Ogata, and D. Vere-Jones, Analyzing earthquake clustering features by using stochastic reconstruction, *J*. Geophys. Res., B5(3), B05301, 2004.

Zhuang, J. (2006). Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *J*. R. Stat. Soc. Ser. B Stat. Methodol. **68**, 635-653.

Zhuang, J. (2011), Next-day earthquake forecasts for the Japan region generated by the ETAS model, *E*arth Planets Space. (63) 207–216.

Zechar, J. D., D. Schorlemmer, M. J. Werner, M. C. Gerstenberger, D. A. Rhoades, and T. H. Jordan (2013). Regional Earthquake Likelihood Models I: First-order results, *B*ull. Seismol. Soc. Am. (103) 787–798.