UNIVERSITY OF CALIFORNIA

Los Angeles

Evaluation and Competence of Short-term Point Processes Earthquake Models

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Ruoxuan Jia

2019

ABSTRACT OF THE THESIS

Evaluation and Competence of
Short-term Point Processes Earthquake Models

by

Ruoxuan Jia
Master of Science in Statistics
University of California, Los Angeles, 2019
Professor Frederic Paik Schoenberg, Chair

Understanding that large earthquakes can be violent to human beings, a wide variety of seismicity forecasting models are proposed. To perform useful comparisons among these point process models, graphical residual methods are proposed because they visualize and highlight when and where a given model does not agree closely with either the observed seismicity or another model. In this paper, we work with the one-day observational and forecasting earthquake data from Southern California in the entire year of 2017. Our goal is to utilized Voronoi analysis and Super-thinned analysis to evaluate and compare the Epidemic-Type Aftershock Sequence model (ETAS) and the Modified Nonparametric Hawkes Point Process model (Modified MISD). Essentially, our results suggest that the Modified MISD Model returns comparatively better predictions and better goodness of fit.

The thesis of Ruoxuan Jia is approved.

Yingnian Wu

Hongquan Xu

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2019

To my mother . . .

who—among so many other things—

saw to it that I learned to touch-type

while I was still in elementary school

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I truly appreciate that Professor Schoenberg has been so generous in supporting me and giving me so much valuable suggestion and advice. Also, many thanks to Professor Werner and William Savran for providing me with earthquake data from CSEP as well as gaining me useful insight into the big data. Last but not least, I want to thank Professor Xu and Professor Wu for the helpful feedback that enables me to polish my paper.

CHAPTER 1

Introduction

Earthquake refers to those sudden shaking of the ground caused by the passage of seismic waves through Earth rocks. While it can be too weak to be felt, it can also be violent enough to toss people around and destroy whole cities. Therefore, to be better prepared for those traumatic earthquakes, scientists devote to researching effective models that are helpful in anticipating future earthquakes before they occur. A comprehensive description to the development of earthquake and models can be found in the book by Turner et al. ([**?**]).

As a result, a variety of point process models have been proposed to describe and forecast earthquake occurrences in seismically active zones. More specifically, these models are to identify the probability of which a major earthquake would occur within a designated spacial-time window. Although these models are well-tuned through training and testing, different models often result in very different forecasts. As indicated by Bray and Schoenberg in [**?**], it turns out that unlike dealing with regression models or time series models, it can be hard to assess individual model or to justify the competencies across several models. In order to better utilize these models, we need to evaluate the goodness of fit of these models to recognize the appropriate model under different concerns (i.e. regarding different time bins, sub-regions and etc.).

Historically, several numerical tests were proposed to evaluate point process models, which are discussed and compared in [**?**]. For those numerical tests, each model consists of the estimated number of earthquakes in each of the spatial-temporal-magnitude bins, where the number of events in each bin is assumed to follow a Poisson distribution with an intensity parameter equivalent to the forecasted rate. For example, the Likelihood test evaluates the probability of the observed data under the proposed model. However, these number tests

unfortunately have low power and may be misleading for model comparison purposes.

Unlike numerical summaries, 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. As suggested by Baddeley et al. in [1], the most natural way to define residuals in analogy with other types of processes is by examination of the residual process. However, in practice, often such residual analysis is not ideal, especially when the conditional intensity is highly volatile, as in the case for most models for earthquake occurrences. The analysis essentially examines the difference between the observed number of points and the expected number of points of a model where we then inspect the integral of this residual process over each cell. Nonetheless,Bray et al. in [2] pointed out that if cells are small, then residual plots tend to look similar to data itself; and if cells are large, there can be a great loss of information such that the analysis may not be optimally informative. Some alternative residual analysis techniques that were proposed for space-time point process models are Voronoi residuals, deviances, and Super-thinned residuals, which were reviewed by Clements et al. in [3].

In this paper, we work with the one-day observational and forecasting earthquake data from Southern California in the entire year of 2017. We apply the Voronoi likelihood analysis, Voronoi deviance analysis, and Super-thinned analysis to evaluate the Epidemic-Type Aftershock Sequence model (ETAS) implemented by Zhuang et al. in [17] and the Modified Nonparametric Hawkes Point Process model (Modified MISD) implemented by Fox et al. in [8]. We will discuss the strengths and weaknesses of these two point process models in the Collaboratory for the Study of Earthquake Predictability (CSEP) from different perspectives.

CHAPTER 2

Models and Data

In this chapter, we introduce the models and data we use for comparison. The data is collected on daily based and the models are one-day seismicity forecasting models selected from CSEP's rate-based repository that anticipate the expected number of earthquakes in each spatial grid for each day.

2.1 Introduction to Models for Comparison

In this section, we introduce space-time point process, Hawkes process and the two one-day forecasting models that we evaluate and compete in the next Chapter. The two models are the Epidemic-Type Aftershock Sequence Model (ETAS) and the Modified Nonparametric Hawkes Point Process Model (Modified MISD).

The ETAS Model has been one the most extensively studied short-term seismicity forecasting model. Despite it being easy to implement and well-grounded, it relies heavily on assumptions. In [11], Marsan and Lengliné first proposed a flexible nonparametric approach, named the Model Independent Stochastic Declustering (MISD). This approach assumes constant background rates and implements an iterative EM-algorithm to estimate background rates and triggering function. Building upon that, Fox et al. in [8] explores a modified version of the original MISD algorithm.

Space-Time Point Process

Point process is a collection of random points falling in some metric space, such as time and location. Adding to that, a marked point process is defined if each point has some

mark or random variable associated with it. Consider a marked space-time point process $N(t, x, y, m)$ that represents the times, locations, and magnitudes, where $(t_i, x_i, y_i, m_i) : i = 1, N$ suggest earthquake occurrences. Then the "behavior" of this space-time point process can be specified with a conditional intensity function $\lambda(t, x, y, m)$ which is defined as the infinitesimal expected rate at which events occur around (t, x, y) given the history of the process H_t, where $H_t = (t_i, x_i, y_i, m_i) : t_i < t$. That is:

$$\lambda(t, x, y|H_t) = lim_{\delta t, \delta x, \delta y \to 0} \frac{E[N\{(t, t+\delta t) \times (x, x+\delta x) \times (y, y+\delta y)\}|H_t]}{\delta t \delta x \delta y}.$$

In fact, the conditional intensity λ can uniquely determine the distributions of simple point processes because the finite-dimensional distributions of a simple point process is uniquely determined by its condition intensity. This was mentioned by Daley and Vere-Jones in [6]. The conditional intensity function $\lambda(t, x, y)$ can be estimated using either parametric or nonparametric approaches, which we will discuss in models.

Space-Time Hawkes Process

Hawkes Process is also called the self-exciting spatio-temporal point process, which was introduced by Hawkes et al. in [10]. It is oftenly used to anticipate the rate of events as a function of space, time, and the previous history of events. Since these models can naturally capture triggering and clustering behavior, these models have been widely used in fields where space-time clustering of events is observed, including forecasting seismology. The Hawke process is called "self-exciting" because the current conditional intensity is determined by the past history $t' < t$ of the process. Again we associate H_t with the history of all events up to time t, that is $H_t = (t_i, x_i, y_i, m_i) : t_i < t$.

Then a Hawkes process can be modeled as a self-exciting process with the following conditional intensity as follow:

$$\begin{aligned}\lambda(x,y,t|H_t) &= \mu(s) + \int_{t_i:t_i<t} g(x-x_i, y-y_i, t-t_i)dN(x_i,y_i,t_i) \\ &= \mu(x,y) + \sum_{t_i:t_i<t} g(x-x_i, y-y_i, t-t_i),\end{aligned} \tag{2.1}$$

where $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_n\}$ denotes the observed sequence of locations of events and $\{t_1, t_2, ..., t_n\}$ denote the observed times of these events.

In the equation, $\mu(x, y)$ is the background rate of events in a spacial region, which is assumed a nonstationary poisson process in space and stationary process in time. The triggering function is denoted as g, which determines the form of the self-excitation. Generally the triggering function g is nonnegative, and is often a kernel function or power law decay function; often, for simplicity, it is taken to be separable in space and time, so that $g(x - x_i, y - y_i, t - t_i) = f(x - x_i, y - y_i)h(t - t_i)$, similar to covariance functions in other space-time models ([5]).

2.1.1 Epidemic-Type Aftershock Sequence Model by Zhuang et al.

The Epidemic Type Aftershock Sequences Model (ETAS) is a typical type of self-exciting point process model. ETAS models were first introduced by Ogata in [13] for the description of earthquake catalogs. An ETAS process is essentially a marked version of the Hawkes process, where points have different triggering functions depending on their magnitudes. Such models categorize earthquake occurrences into two types: mainshocks and aftershocks. Ogata introduces

$$\lambda(t,x,y) = \mu(x,y) + \sum_{\{t',x',y':t'<t\}} g(t-t', x-x', y-y')h(m'),$$

where $\mu(x, y)$ is the rate of mainshocks occurring over a spatial region, usually estimated by smoothing observed large earthquakes. In this ETAS model, the background intensity $\mu(x, y)$ is assumed to be nonstationary poisson process in space and stationary in time.

One major concern with the Ogata's model is that the there were problems in predicting large earthquakes with clusters of aftershocks. These earthquake clusters complicate the

statistical analysis of the background seismic activities, which prevail over the aftershock activity.While Ogata chosed a conventional declustering algorithm for preliminary estimation of the background rate by fitting bicubic B-spline functions before fitting the space-time ETAS model to all the events, in [17], Zhuang et al. further improves upon the declustering method of the model. This procedure stochastically splits the whole process into background events and offspring events grouped into clusters. Each such splitting can be regarded as one possible realization of how the original catalog could have been produced, taking into account that individual events can be classified as background events or cluster events only with a certain probability. We then can estimate the significance and uncertainty of the features of the catalog not explainable in terms of the assumed model for clustering by running the procedure many times ([17]).

2.1.2 Modified Nonparametric Hawkes Point Process Model by Fox et al.

Traditionally, Hawkes point process models of earthquake seismicity rely heavily on parametric assumptions about the triggering function for the spatial-temporal rate of aftershock activity following an earthquake (i.e, such as the ETAS Model). However, Marsan and Lengliné in [11] proposed the first nonparametric approach that assumes constant background rate and utilizes EM-algorithm to estimate background rates and triggering functions.

While the original MISD model can be helpful in estimating the mean mainshock rate over an observation region, the modified version further enhances the background rate assumption made by the MISD algorithm. This Modified MISD model implemented in [8] is an expansion of the original MISD that incorporates a nonstationary background component, which is an effective improvement by allowing for localized estimates mainshock activity related to variations in the underlying tectonic field and the locations major faults. Moreover, an estimate of a spatially varying background rate can be used to identify regions with a persistent and heightened incidence of large seismic events, independent of aftershock clustering features which diminish over time. Additionally, Fox et al. in [8] improves upon the sampling variability and bias in the estimation of the underlying seismic process. In

this manner, the designed method allows for localized estimated mainshock activity related to variations in the underlying tectonic field and the locations major faults. This model also incorporate a spatially varying background rate into the MISD method by applying the variable kernel estimator, used by [17] for semi-parametric estimation, into the context of MISD model.

2.2 Data Preparation

We proceed by using earthquake observations and forecasting outputs for both the ETAS Model and the Modified MISD Model for the entire year of 2017 (i.e., from Jan 1st, 2017 to Dec 31st, 2017). Our data is obtained from CSEP repository. Both the observational data and the forecasting data are collected on daily basis. In the observation data, for individual earthquake, the observation data indicates the latitude of earthquake, the longitude of earthquake, the exact date and time of the earthquake, the magnitude of the earthquake, and the depth of the earthquake. In the forecasting data, the space is divided into cells of size 0.1° longitude by 0.1° latitude. Both models forecast within latitude range $[-125.4^\circ,\ -113.2^\circ]$ and longitude range $[31.5^\circ,\ 42.9^\circ]$. Then for each forecasted earthquake, the data clarifies the 0.1° longitude cell and 0.1° latitude cell, the range of the magnitude, the range of the depth, and the corresponding conditional intensity for this predicted earthquake.

We address the observational data and forecasting data for residual analysis as follow. From the observational data, we select the latitude of earthquake, the longitude of earthquake and the exact date and time of the earthquake for our consideration. There are a total of 19 observations within the year of 2017 and Table 2.1 below displays a sample of our observational data that we use in following parts.

Table 2.1: Observation Data

	Date and Time	Longitude	Latitude
1	2017.140	-122.74	38.80
2	2017.178	-124.77	40.35
3	2017.221	-115.23	32.23
4	2017.245	-121.61	36.88
5	2017.327	-122.78	38.79
...	...	...	...

The forecasting output for individual model is addressed as follow: first we aggregate the forecasted rate within each 0.1° longitude by 0.1° latitude grid; from there we obtain the expected conditional rate (conditional intensity) for each grid in the region. These aggregated rates within each 0.1° longitude by 0.1° latitude is the expected conditional intensity for individual grid. After this step, we have 7682 grids for each model. In proceeding analysis, the latitude range, the longitude range, and the expected rate are selected from forecasting data for consideration.

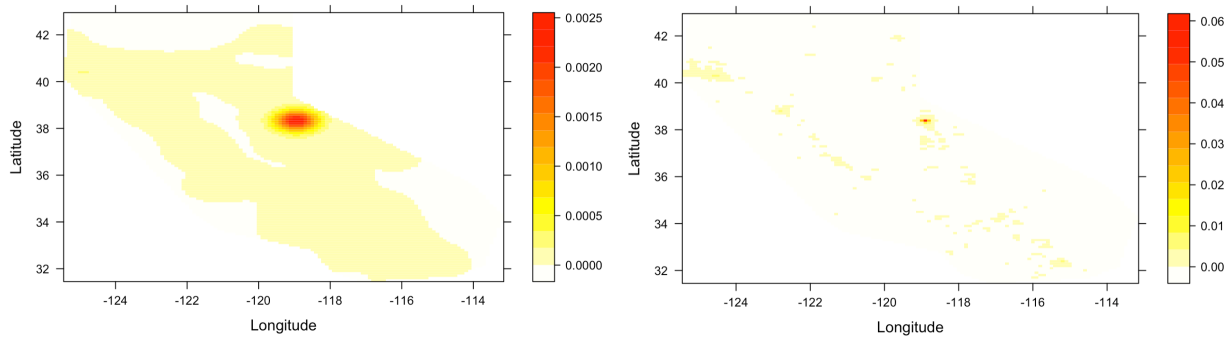
Table 2.2: Expected Conditional Intensities for Each Grid Predicted by ETAS Model (Right) and Modified MISD Model (Left)

Longitude	Latitude	Rate
$[-125.4, -125.3]$	$[40.3, 40.4]$	0.0521
$[-125.4, -125.3]$	$[40.4, 40.5]$	0.0576
$[-125.4, -125.3]$	$[40.5, 40.6]$	0.0596
$[-125.3, -125.2]$	$[40.3, 40.4]$	0.0549
$[-125.3, -125.2]$	$[40.4, 40.5]$	0.0606
...	...	...

Longitude	Latitude	Rate
$[-125.4, -125.3]$	$[40.3, 40.4]$	0.0781
$[-125.4, -125.3]$	$[40.4, 40.5]$	0.215
$[-125.4, -125.3]$	$[40.5, 40.6]$	0.118
$[-125.3, -125.2]$	$[40.3, 40.4]$	0.118
$[-125.3, -125.2]$	$[40.4, 40.5]$	0.330
...	...	...

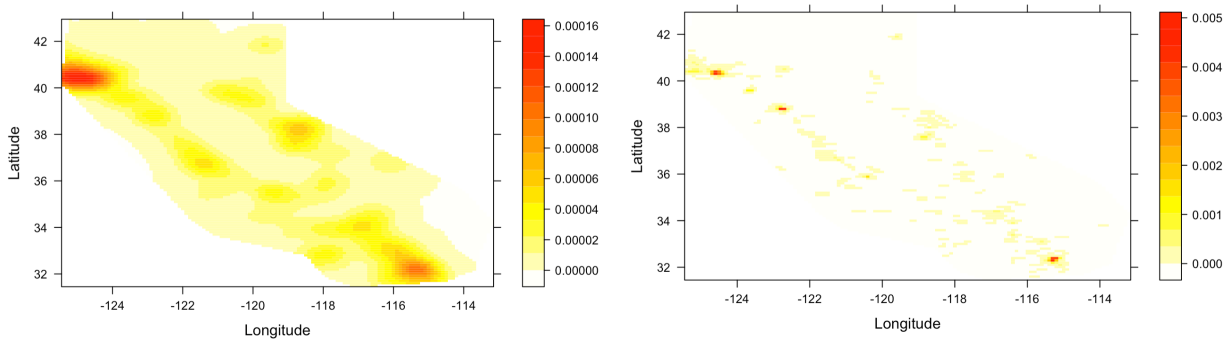
2.3 Forecasting Data Visualization

Before pursuing model analysis and comparison, we visualize the forecasting data from the two models as follow. Figure 2.1, Figure 2.2, and Figure 2.1 display forecasting data from the two models as heat maps on Jan 1st, 2017, Dec 31st, 2017, and the cumulative plot in the year of 2017. Images on the left display output from the ETAS Model while images on the right display outputs from the Modified MISD Model. Each figure exhibits the distribution of forecasting earthquakes in space, where the x-axis and y-axis represent longitude and latitude respectively. Legends on the right demonstrate the color scale corresponding to different values of rates, where darker shades (i.e. red) indicate higher expected conditional intensities and lighter shades (i.e. yellow) suggest low expected conditional intensities.



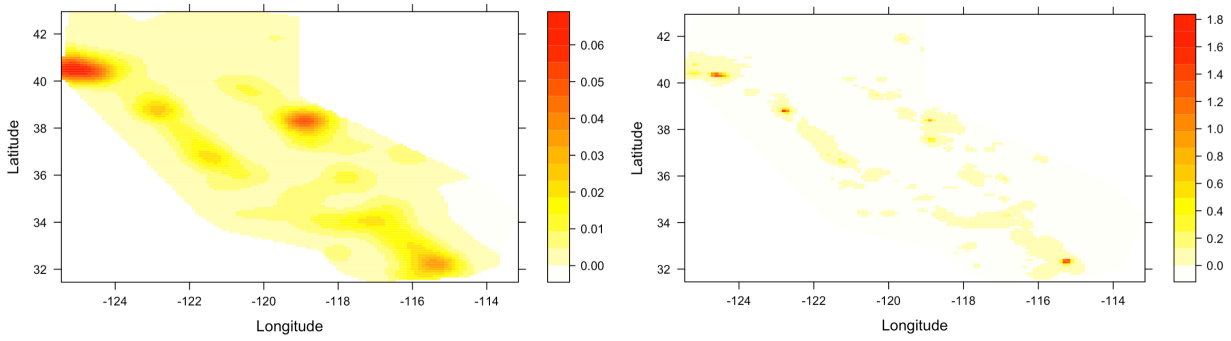
(a) ETAS Model (b) Modified MISD Model

Figure 2.1: One-day Forecast Plot on Jan 1st, 2017



(a) ETAS Model (b) Modified MISD Model

Figure 2.2: One-day Forecast Plot on Dec 31st, 2017



(a) ETAS Model (b) Modified MISD Model

Figure 2.3: Cumulative Forecast Plot in year 2017

Looking at the heat plots from the two point process models above, we can tell that

the two methodologies forecast the occurrence of earthquake distinctively. Under the ETAS Model, the forecasted earthquakes seems to be more evenly distributed along the contour lines. Meanwhile, under the Modified MISD Model, the forecasted earthquakes appears to be more accentuated in designated areas with minimal appearance around. In general, despite the discrepancy in prediction methods, the "most likely" earthquake regions (the darker fields on plots) indicated by both models distribute in similar geographic locations.

In the following chapters, we evaluate and compare the ETAS Model and the Modified MISD Model. A variety of pixel-based residuals for spatial point processes were proposed by Baddeley et al. in [1], discussed by Clements et al. in [3] and Gordon et al. in [9]. Common diagnostics plots showing the standardized differences between the number of points occurring in each pixel and the number expected according to the fitted model, where the standardization may be performed in various ways. For instance, for Pearson residuals, one divides the difference by the estimated standard deviation of the number of points in the pixel, in analogy with Pearson residuals in the context of linear models. Voronoi residuals and Super-thinning can be especially useful at identifying departures from the data.

CHAPTER 3

Voronoi Residual Analysis

In this chapter, our goal is to perform Voronoi residual analysis to compare the two one-day forecasting models. In the following part, we start by introducing the concepts and benefits of Voronoi residual analysis. Then, we prepare our one-year data for Voronoi residual analysis. Finally, we utilize Voronoi likelihood analysis and Voronoi deviance analysis to evaluate both the well-behave of each model individually and the discrepancy between these two models in practice.

3.1 Introduction to Voronoi Residual Analysis

Voronoi residuals are constructed using Voronoi tessellation, which is a partition of the metric space on which a point process is defined to be 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 then the collection N of such Voronoi cells which we assume fills the complete window C_T such that $C_t = \cup C_i$. 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. These were discussed in detail in [14].

Given a model for the conditional intensity of a spatial or space-time point process, one may construct residuals simply by evaluating the residuals over cells rather than rectangular pixels, where the cells comprise the Voronoi tessellation of the observed spatial or spatial-

temporal point pattern. We will refer to such residuals as Voronoi residuals.

The advantage of Voronoi residuals compared to conventional pixel-based methods is that the partition is entirely automatic and spatially adaptive. Therefore the Voronoi residual analysis can be especially helpful in addressing problems induced by skewness in distribution of raw residuals integrated over pixels, when pixels have a small integrated conditional intensity. Indeed, since each Voronoi cell has exactly one point inside it by construction, the raw Voronoi residual for cell i is given by

$$\begin{aligned}\hat{r}_i &:= 1 - \int_{C_i} \hat{\lambda} d\mu \\ &= 1 - |C_i|\overline{\lambda},\end{aligned}$$

where $\overline{\lambda}$ denotes the mean of the proposed model $\hat{\lambda}$ over C_i. These raw residuals can then be scaled in various ways based on the distribution of N.

As with pixel residual, for each Voronoi residual we can plot the raw residual, scaled residual, or the deviances (discussed by Bray and Schoenberg in [**?**]). Voronoi residuals are described in detail by Bray et al. in [2] 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. In [2], Bray et al. 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.

3.2 Voronoi Data Preparation

In this section, we prepare our data for the following Voronoi residual analysis. To start with, as indicated by Voronoi analysis, we divide the space-time window of our observational

data into 19 regions where each observation corresponds to one Voronoi region. Each Voronoi region is consisted of all locations that are closer to the generating point than to any other observation points. In the next step, we divide the forecasting data by allocating every forecasting grid (0.1° longitude by 0.1° latitude pixel) to the "closest" Voronoi region in terms of euclidean distance. In this manner, for individual model, we assign the 7682 grids into the 19 Voronoi cells as described above. Finally, we aggregate the expected conditional intensities of forecasting grids within each Voronoi cell to obtain an expected conditional intensity for each Voronoi cell. The expected conditional intensity for each cell is comparable to the observational data, where the true intensity is 1 for every cell since by design there is exactly one earthquake in each cell.

Table 3.1: Expected Conditional Intensities Predicted by ETAS Model (Right) and Modified MISD Model (Left)

	Date and Time	Aggregated Rate
1	2017.140	1.422
2	2017.178	2.890
3	2017.221	4.614
4	2017.245	0.942
5	2017.327	1.536
6	2017.373	2.716
...	...	...

	Date and Time	Aggregated Rate
1	2017.140	2.537
2	2017.178	7.377
3	2017.221	9.310
4	2017.245	0.910
5	2017.327	3.248
6	2017.373	0.849
...	...	...

Table 3.1 above displays the expected conditional intensities for the first six Voronoi cells (in time order) predicted by ETAS Model and Modified MISD Model. In general, it seems that both models over predict the occurrence of earthquakes. For example, regarding the third observation that happened on Mar 22nd (2017.221), the ETAS Model predicts four times more rate while the Modified MISD Model predicts nine times more rate compare to the actual earthquake occurrence.

3.3 Likelihood Analysis

For Voronoi likelihood analysis, we divide the space-time window into bins as described in the previous section and then calculates residuals within each bin for ETAS Model and Modified MISD Model respectively. Because the bins are based on a tessellation, each bin contains at most one point. Denote bin i as (V_i) and the residual for each bin can be defined as follow (for two models):

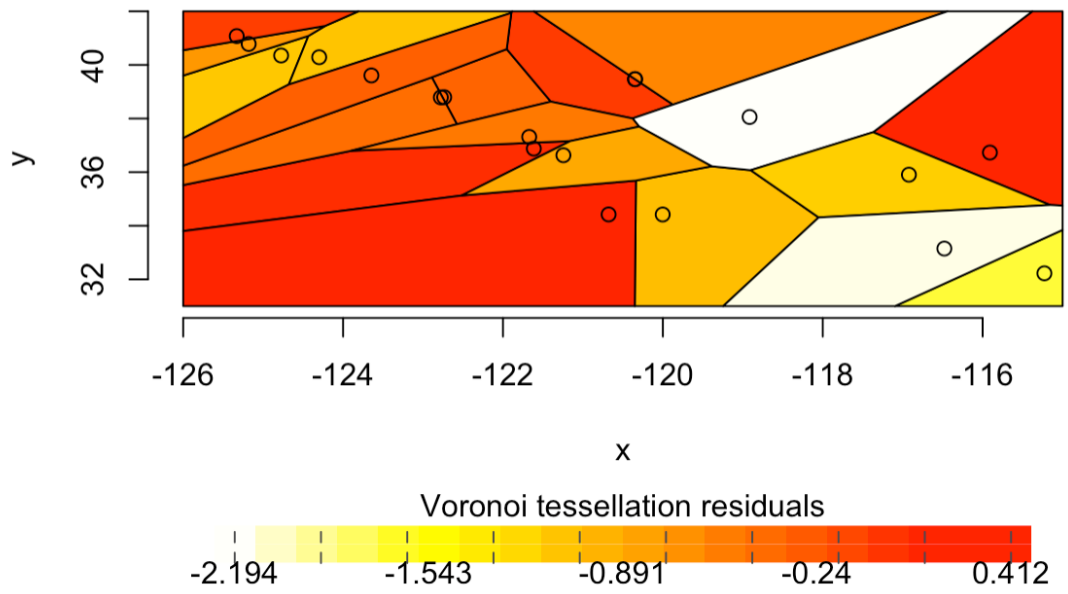
$$R_T(V_i) = \frac{1 - \int_{V_i} \hat{\lambda}_{ETAS}(x)dx}{\sqrt{\int_{V_i} \hat{\lambda}(x)dx}},$$

and

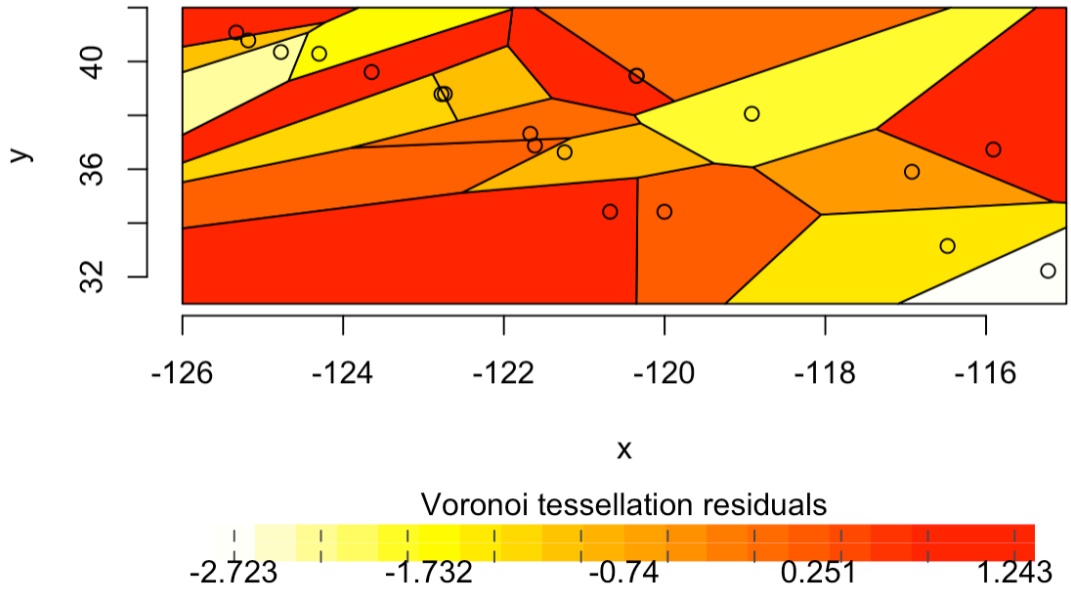
$$R_T(V_i) = \frac{1 - \int_{V_i} \hat{\lambda}_{MMISD}(x)dx}{\sqrt{\int_{V_i} \hat{\lambda}(x)dx}},$$

where $\hat{\lambda}_{ETAS}(x)$ and $\hat{\lambda}_{MMISD}(x)$ each represents the fitted conditional intensity output from the ETAS Model and the Modified MISD Model. These two formulas obtain us likelihood for each Voronoi cell with respect to individual model. Negative values correspond to over-predictions, positive values correspond to under-predictions, and values around 0 indicate well-behave predictions made by a model.

Figure 3.1a and Figure 3.1b display the likelihood plots for ETAS Model and Modified MISD Model respectively, where legends in the bottom indicate the likelihood value with corresponding colors in plots. The lighter shades (i.e. light yellow shades) suggest over-predictions while the darker shades (i.e. bright red shades) suggest under-predictions on earthquake occurrences. In addition, the orange shades indicate better achieved predictions made by the models.



(a) ETAS Model Likelihood Plot over Voronoi Tessellation



(b) Modified MISD Model Likelihood Plot over Voronoi Tessellation

Figure 3.1: Likelihood Plot for the Two Models

To begin with, the overall log-likelihood for the ETAS Model is -13.721 while the over-

all log-likelihood for the Modified MISD Model is -8.865. It follows that the average log-likelihood (information gain per earthquake) for the ETAS Model is -0.722 while the average log-likelihood for the Modified MISD Model is -0.467. These outcomes generally suggest that the Modified MISD Model seems to have a better performance according to Voronoi Analysis.

According to the figures above, it seems that both models do a decent job in forecasting the following events: the earthquake happened at $longitude = -120.35°$, $latitude = 39.47°$ on June 27th around the Sierra area, near the Mohawk Valley fault zone; the earthquake happened at $longitude = -121.67°$, $latitude = 37.31°$ on October 10th around Santa Clara, which is in between Silver Creek fault and San Jose fault; and the earthquake happened at $longitude = -121.24°$, $latitude = 36.63°$ on November 13rd around Paicines in San Benito, which locates on the Paicines fault in Calaveras fault zone.

The Modified MISD model appears to perform well in anticipating earthquakes happened at $longitude = -120.00°$, $latitude = 34.42°$ on May 17th located in the pacific ocean area; the one happened at $longitude = -121.61°$, $latitude = 36.88°$ on March 31st, which is near the San Andreas fault zone; and $longitude = -116.92°$, $latitude = 35.90°$ on August 22nd, around the Death Valley area. Meanwhile, the ETAS Model appeared to forecast the seismicity around the Collayomi fault zone area more accurately. Both earthquakes happened at $longitude = -122.78°$, $latitude = 38.79°$ on April 30th and $longitude = -122.73°$, $latitude = 38.80°$ on May 17th are well-indicated.

Finally, both models present relatively poor performances while anticipating the earthquakes happened on the outer-ring of Southern California; such as the two earthquakes happened at $longitude = -125.33°$, $latitude = 41.08°$ on July 29th and $longitude = -120.68°$, $latitude = 34.42°$ on May 17th around the pacific ocean area; and the one happened at $longitude = -115.23°$, $latitude = 32.23°$ on March 22rd near Mexicali at the Mexico-California border.

To sum up, the log-likelihood outputs suggests that the Modified MISD Model seems to have a better seismicity forecasting power. While both the ETAS Model and the Modified

MISD Model forecast well in the central part of Southern California, it seems that the Modified MISD Model appears to improve upon the ETAS Model near the San Andreas fault zone. On the other hand, in the periphery of California, despite that both models present lack of accuracy, it seems that the ETAS Model makes more reasonable predictions while the Modified MISD Model has a tendency of over-predicting seismicity, such as when forecasting the earthquake happened at $longitude = -124.77°$, $latitude = 40.35°$ on March 6th in the pacific ocean. We further compare the predictions made by the two models in the deviance section bellow.

3.4 Voronoi Deviance Analysis

In this section, we competing point process models compare the ETAS Model to the Modified MISD Model utilizing Voronoi cell based deviance analysis. Similar to the deviances defined for generalized linear models in the regression framework, it measures "the distance", which is the difference between the log-likelihoods of the two point process models over every Voronoi tessellated cells. In our case, we again address our data as above. Then instead of computing deviances over evenly spaced grids, we compute deviances from Voronoi tessellated cells. 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.

The tessellation deviance residuals are the differences in the tessellation residuals of model 1 vs. model 2 within each Voronoi tessellation cell, denoted here as V_i. The tessellation deviance residual is given by

$$R_{TD}(V_i) = \frac{1 - \int_{V_i} \hat{\lambda}_{MMISD}(x)dx}{\sqrt{\int_{V_i} \hat{\lambda}_{MMISD}(x)dx}} - \frac{1 - \int_{V_i} \hat{\lambda}_{ETAS}(x)dx)}{\sqrt{\int_{V_i} \hat{\lambda}_{ETAS}(x)dx}},$$

where $\hat{\lambda}_{ETAS}(x)$ and $\hat{\lambda}_{MMISD}(x)$ are the fitted conditional intensity output from the ETAS Model and the Modified MISD Model. This equation returns the deviance between the two point process models based on Voronoi cells. If $R_{TD}(V_i)$ value is close to 0, then the two models make similar predictions in cell i. On the other hand, if $|R_{TD}(V_i)|$ is large,

then there can be big discrepancies between forecasts made by the two models. Deviances between the ETAS Model and the Modified MISD Model are seen in Figure 3.2.

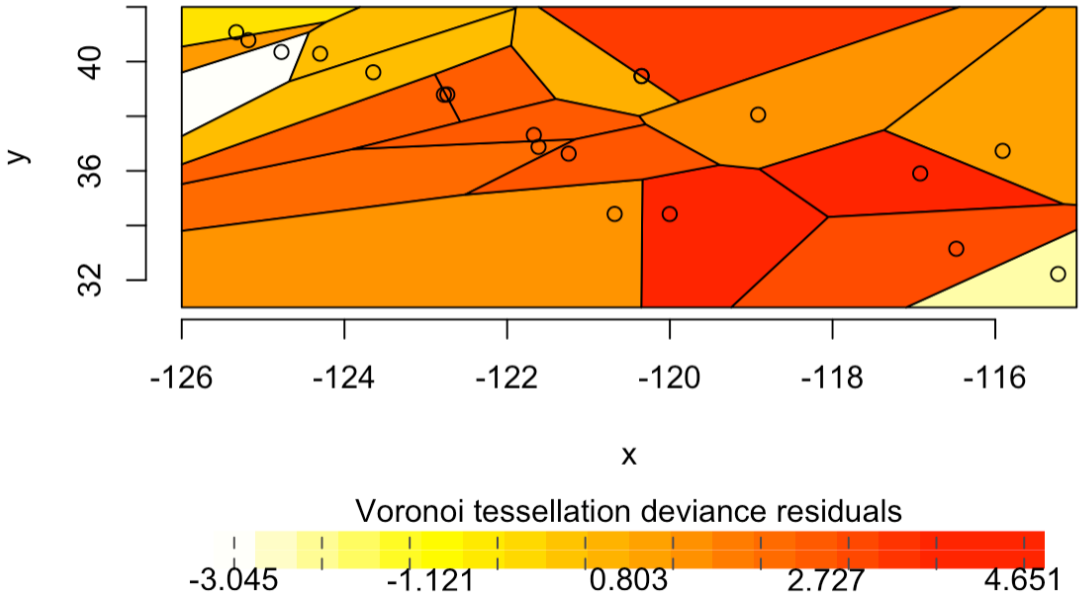


Figure 3.2: Deviance Plot Comparing ETAS Model with Modified MISD Model over Voronoi Tessellation

The legend in the bottom indicates the deviance values with corresponding colors in plots. Lighter shades (i.e. light yellow shades) indicates negative deviance while darker shades (i.e. bright red shades) suggest positive deviances. Also, medium orange shades suggest regions where the two models predict similarly. This plot confirms the previous findings in likelihood section. It suggests that the ETAS Model and the Modified MISD Model make similar predictions in the central part of Southern California. On the other hand, more disparities come up when forecasting the events on the outer-ring of the area. In combine with the previous likelihood plots, we can see that while both models over-predict the earthquake happened at $longitude = -115.23°$, $latitude = 32.23°$ near Mexicali, the Modified MISD is still way off than the ETAS Model because it excessively over-predict earthquakes. Similarly, both models fail to accurately indicate the earthquake happened at $longitude = -116.48°$, $latitude = 33.15°$ on December 7th, yet the ETAS Model is more

imprecise since it six times over-predict seismicity while the Modified MISD Model three times over-predict the result.

CHAPTER 4

Super-Thinning

In this chapter, our goal is to perform Super-thinning analysis to compare the goodness of fit for the two models. In the following part, we will introduce Super-thinning, describe the sudo steps we apply in working with Super-thinning, and utilize Super-thinning analysis to analyze how well the two models fit.

4.1 Introduction to Super-thinning

Super-thinning is a combination of thinning and Superposition methods. Thinning for residual analysis was proposed by Schoenberg in [15]. Using thinning, each point of a point process N, is retained with some independent probability $\frac{b_i}{\hat{\lambda}_i}$ If the model is correct, then the residual process will be homogeneous Poisson with rate b. Each thinning is a distinct process and one may inspect several realizations of thinned residuals and analyze the entire collection to get an overall assessment of goodness of fit, as in [15].

Superposition, meanwhile, was proposed by Brémaud et al. in [12]. He suggested Superposing a simulated point process onto an observed point process realization so as to yield a homogeneous Poisson process, which is essentially an addition operator on point processes (i.e. N_3 is the Superposition of point processes N_1 and N_2). However, as indicated by Clements et al. in [4], tests based on thinned or Superposed residuals tend to have low power when the model $\hat{\lambda}$ for the conditional intensity is volatile, which is typically the case with earthquake forecasts since earthquakes tend to be clustered in particular spatial-temporal regions.

Based upon the two methods, a hybrid approach, Super-thinning, is a more powerful ap-

proach than either thinning or Superposition alone. This combined approach was introduced in [4], suggesting thins in areas of high intensity and Superposes simulated points in areas of low intensity. In such way, we transform the observed points to a homogeneous point process if the point process model we use in the thinning and Superposition is correct. With this in mind, the resulting Super-thinned residuals can be plotted and assessed for homogeneity as a way of evaluating the model. Visualization of Super-thinned residuals allows detection of any clustering or inhibition in the residual points which indicates a lack of fit.

4.2 Super-thinning Steps

This section discusses the Super-thinning method we adopted in addressing our earthquake observation and forecasting data. Let $\hat{\lambda}_{obs}$ be the sequence of log transformed expected conditional intensity computed for each pixel (0.1° longitude by 0.1° latitude forecast grid) that contains an observed earthquake (i.e. every point in the sequence is denoted as $\hat{\lambda}_{obs}[i]$); as a result, the length of $\hat{\lambda}_{obs}$ is 19 where each correspond to an observational point. Further, let $\hat{\lambda}_{fore}$ be the sequence of log transformed expected conditional intensity computed for each pixel of the forecasting data; it turns out that $\hat{\lambda}_{fore}$ is a sequence of length 7682.

Bellow is a summary of the steps we use for individual point process model:

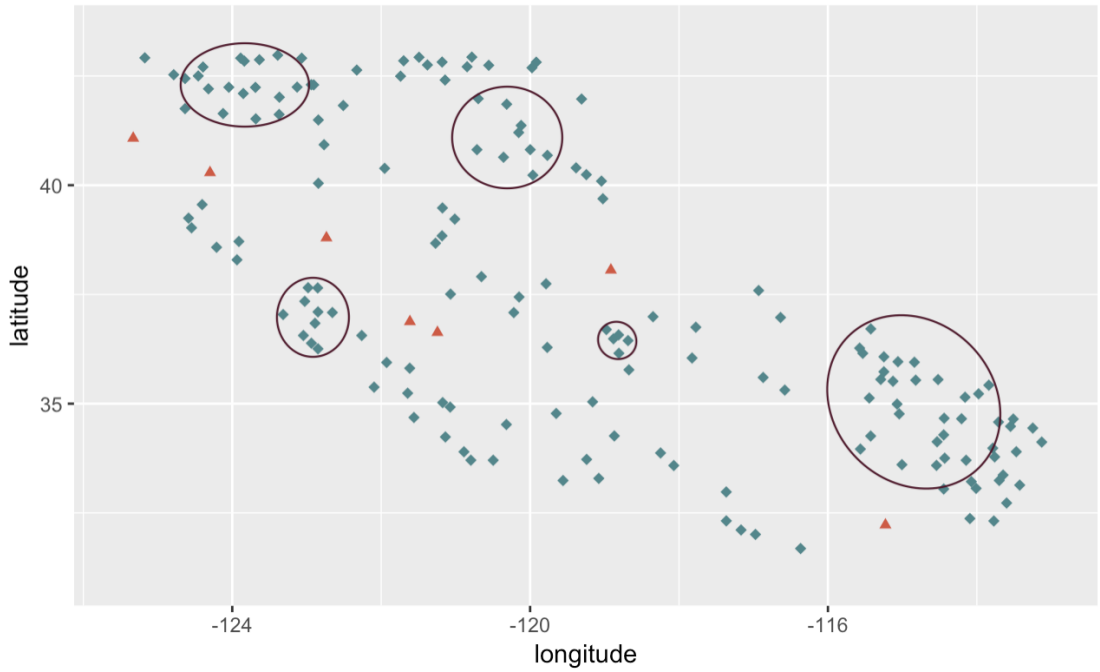
1. First, we take some value b to be the mean of the sequence $\hat{\lambda}_{obs}$ (i.e. $b = mean(\hat{\lambda}_{obs})$).
2. Second, we thin the observation data by comparing every value in $\hat{\lambda}_{obs}$ to the value b:
   - if $\hat{\lambda}_{obs}[i] > b$, then we keep the point i with probability $\frac{b}{\hat{\lambda}_{obs}[i]}$;
3. Third, we Super-position forecasting data by simulating random points as follow:
   - compare every value in the sequence of $\hat{\lambda}_{fore}$ to value b;
   - for any pixel, if $\hat{\lambda}_{fore}[i] < b$, then we simulate poisson points at rate $(b - \hat{\lambda}_{fore}[i])$, where m denotes the number of points generated from the simulation;
   - if $m > 0$, then we add m points randomly uniformly scattered into the corresponding 0.1° longitude by 0.1° latitude designated pixel;

After processing the data, we plot the thinned observation points together with the Super-positioned forecasting points for each model to analyze and compare their distribution.

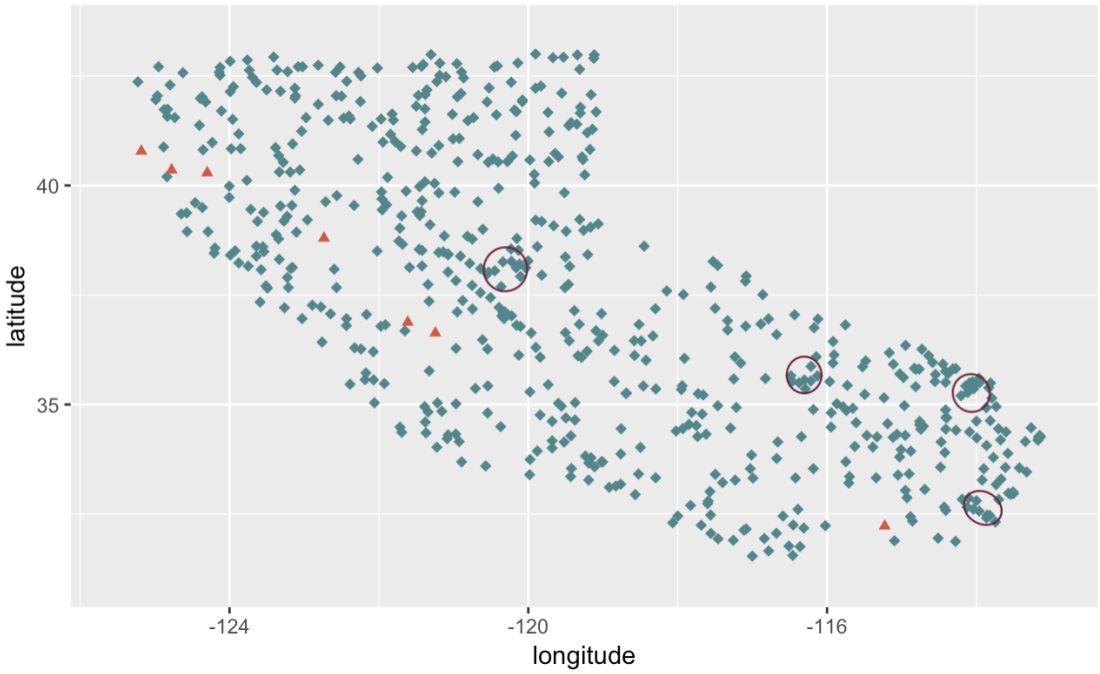
4.3 Super-thinning Analysis

Below, Figure 4.1a and Figure 4.1b display the Super-thinned residuals plots for ETAS Model and Modified MISD Model. On these two plots, red triangles represent the thinned observations while blue squares indicate simulations from forecasting data. The plots suggest that the two models thinned the observational data similarly. Regarding forecasting data, it seems that the ETAS Model plot displays severe clustering and inhibition structures that deviates the homogeneous point process. Meanwhile, the Modified MISD Model appears to fit well in general where the points overall uniformly distributed, despite some minor clustering structures at small radius (i.e. from 0° to 0.2°).

As indicated by Figure 4.1a, some major clusterings happen at radius around 1.2° to 1.5° and other clusters are at radius around 0.2° to 0.5°. The center of the larger cluster in the right corner locates at $longitude \approx -115.5^\circ$, $latitude \approx 34.3^\circ$, which is close to the Calumet fault. Other major clusterings are around $longitude \approx -123^\circ$, $latitude \approx 36.8^\circ$, which is close to the Paleo-subduction zone; $longitude \approx -124.1^\circ$, $latitude \approx 42.3^\circ$ in the mountain areas ; and $longitude \approx -120.1^\circ$, $latitude \approx 40.8^\circ$ locating close to faults on east side Madeline Plains. Also, this plot exhibits greater inhibition than one would expect. Some major inhibitions happen near Willard fault ($longitude \approx -117.2^\circ$, $latitude \approx 33.5^\circ$), near Rinconada fault ($longitude \approx -120.7^\circ$, $latitude \approx 35.6^\circ$), and near Tahoe Valley fault zone ($longitude \approx -120.0^\circ$, $latitude \approx 38.9^\circ$), and near Concord Fault ($longitude \approx -122.0^\circ$, $latitude \approx 38.0^\circ$).



(a) ETAS Model Super-thinning Plot



(b) Modified MISD Model Super-thinning Plot

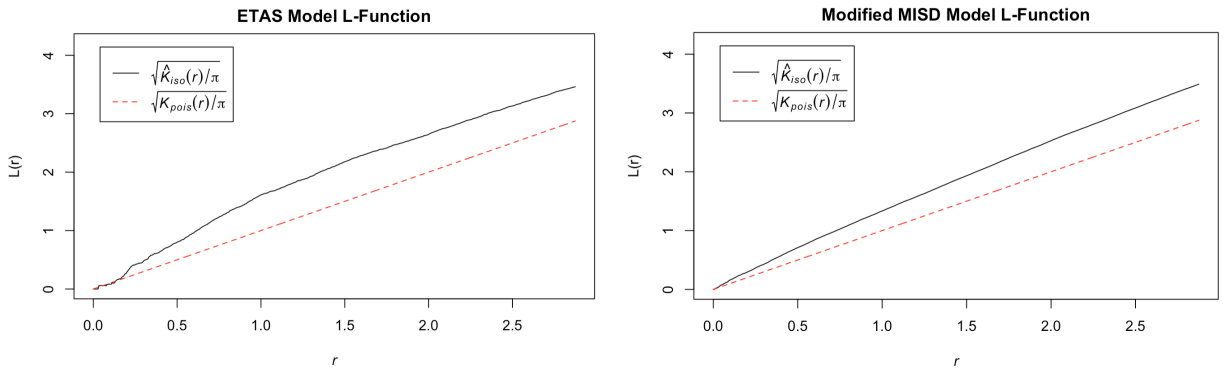
Figure 4.1: Likelihood Plot for the Two Models

Looking at Figure 4.1b, it seems that enough points were simulated such that the points

overall randomly and uniformly distributed. The three circles on the plot recognize some minor clusterings patterns in this plot, which are generally at small distance (from 0° to 0.2°). Some noticeable clusters are close to Red Pass fault ($longitude \approx -116.3°$, $latitude \approx 35.2°$) and in the Homer Mountain area ($longitude \approx -114.9°$, $latitude \approx 35.0°$). Some other ones center at $longitude \approx -114.0°$, $latitude \approx 32.5°$, and $longitude \approx -119.9°$, $latitude \approx 37.7°$.

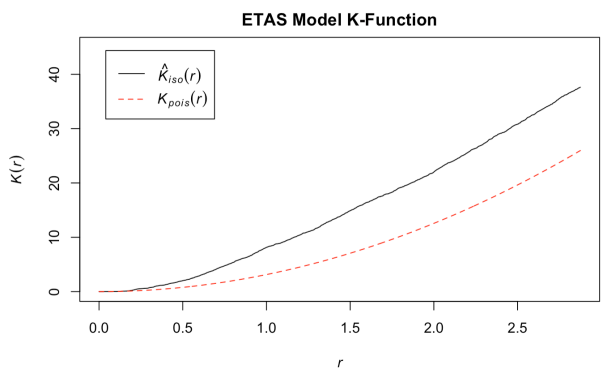
Additionally, there appears to be a small amount of inhibition in this plot as well. For example, there is inhibition near the Cambria fault zone ($longitude \approx -121.0°$, $latitude \approx 35.5°$), La Nacion fault zone ($longitude \approx -117.0°$, $latitude \approx 32.6°$), and near Mission Fault ($longitude \approx -122.0°$, $latitude \approx 37.6°$).

To achieve a more concrete comparison of the Super-thinned distributions presented by the two models, we plot the K-function and L-function for our models in Figure 4.2 and Figure 4.3. In both plots, comparing the plots for the ETAS Model (left) and that for the Modified MISD Model (right), it seems that the estimated K and L of the Modified MISD Model fit the theoretical values better than that of the ETAS Model.

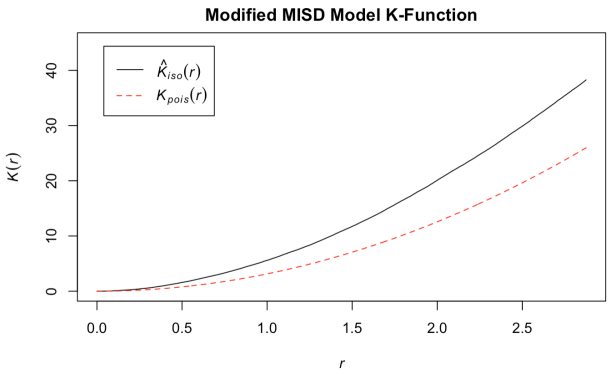


(a) ETAS Model L-Function (b) Modified MISD Model L-Function

Figure 4.2: L-Functions for the Two Models



(a) ETAS Model K-Function



(b) Modified MISD Model K-Function

Figure 4.3: K-Functions for the Two Models

As displayed in above figures, the Modified MISD Model seemingly performs better than the ETAS Model in Super-thinned analysis. While the Modified MISD Model exhibits moderate clustering and inhibition effects, the ETAS Model indeed presents a more severe problem where larger clusters and regions of missing points show up. Such patterns are further affirmed by the K-function plots and L-function plots. As we have discussed above, some major clusterings indicated by the plot for ETAS model are close to the Calumet fault and around the Paleo-subduction zone.

Despite that point densities on the two plots are different, they display similar patterns of inhibitions around some common fault zone areas, which should be more prone to ground-breaking earthquakes and liquefaction (i.e. ground fails) effect. It seems that the two models have a tendency to under-predict seismicity in these regions. This may be due to the fact that based historical seismicity patterns, there is little possibility that earthquake shall happen in these areas in the year of 2017. As we can see from the observational data, there were no actual earthquakes happening in these inhibitions. In the future, if we have the chance to perform Super-thinning on data from a bigger time-span, we may have a better standing point.

CHAPTER 5

Discussion and Conclusion

5.1 Conclusion

While many spatial-time point process models have been proposed for forecasting earthquake occurrences in seismically active zones, the problem of how to assess and compare the goodness of fit of such models remains open. It turns out that a variety of common residual analysis methods for spatial point processes can be implemented to provide powerful summaries that reveal strengths and weaknesses as well as assess goodness of fit in point process models. In this paper, we utilized Voronoi residuals analysis and Super-thinned analysis to the two earthquake forecast models (i.e. the ETAS Model and the Modified MISD Model) in the Collaboratory for the Study of Earthquake Predictability (CSEP) for a catalog spanning from January 1st 2017 to December 31st 2017.

The advantage of Voronoi residuals compared to conventional pixel-based methods is that the partition is entirely automatic and spatially adaptive. Model evaluation using the Voronoi residual analysis visualizes the strength and weakness of the ETAS Model and the Modified MISD Model. Both the likelihood plots and the deviance plot suggest that the ETAS Model and the Modified MISD Model make similar predictions in the central part of Southern California. Additionally, while both models appear to perform well around the central regions, the Modified MISD Model further improves upon the ETAS Model near the San Andreas fault zone. On the other hand, it seems that the two models are disparate in anticipating earthquake events on the outer-ring of the area. However, despite that both models present lack of accuracy in the outer fields, the ETAS Model seems to make more reasonable predictions while the Modified MISD Model displays a tendency to over-

predicting seismicity. As we can tell, in predicting the one event happened at $longitude = -115.23°$, $latitude = 32.23°$ near Mexicali, the Modified MISD is way off by nine times over-predicting earthquakes. Therefore, based on the Voronoi analysis, it can be more effective to utilize the Modified MISD Model in forecasting earthquakes in central California while consider the ETAS Model for the periphery fields of the region.

Super-thinning is a hybrid approach of Thinning and Superposition. The method suggests thinning in areas of high intensity and point simulations in areas of low intensity. In such way, the resulting plot should display a homogeneous point process if the model we use in the thinning and Superposition is correct. Here we use the Super-thinning steps described in Section 4.2 to analyze and compare the goodness of fit for the ETAS Model and the Modified MISD Model. In general, the Super-thinning analysis suggests that the Modified MISD model has a better fit. Despite some minor clusterings and inhibition regions, the plot from the Modified MISD Model displays randomly and uniformly distributed points with clustering at small distances. Meanwhile, the plot for the ETAS Model indicates that there are visible clustering and missing point structures that deviate the homogeneous point process. Some major clusterings indicated by the plot for ETAS model are close to the Calumet fault and around the Paleo-subduction zone. However, while the two plots are disparate, both of which display some pattern of inhibitions around several fault zone areas. It seems that the two models under-predict seismicity in these regions, which lead to inhibition effects. It can be due to the fact that historical seismicity data indicates little possibility in having earthquakes in these areas in the year of 2017. If additional data from a longer time-span is incorporated, we may achieve a more comprehensive analysis on goodness of fit.

5.2 Future Improvement

In this section, we discuss our reflections regarding our analysis as well as future improvements we consider. Regarding the Voronoi residual analysis we pursued, there are disadvantages in that some sampling variabilities are induced by the random cell areas and that the residuals are dependent. Therefore the Voronoi residual analysis relies upon an i.i.d.

(i.e. independent and identically distributed) assumption, which should be used cautiously. As a result, an alternative method to consider in the future is to is to consider residuals based on a model-based centroidal Voronoi tessellation introduced in [7], which provides a partition that creates residuals that are independent of one another if the underlying model is Poisson. Additionally, it is worth to perform Voronoi Analysis base on different time bins, such as one-day bins or one-month bins. These tests may reveal how well the two models behave with respect to different times around the year.

Further, in the Super-thinned analysis section, from plots of both models, we saw a pattern of inhibition around similar fault zone areas. While we make the guess that historical seismicity data leads to such results, we can achieve a more comprehensive analysis if we have the chance to incorporate data from a longer time-span. In the future, we shall analyze whether the two models inheritly under-predict seismicity in these areas.

Finally, numerical tests may be implemented to compete the two point process models. Despite that most numerical tests are not ideal for model comparison or of low power, Schorlemmer et al. in [16] proposed an additional test, named the R-test (i.e. ratio test) that allows for a direct comparison of the performance of two models.

Bibliography

[1] Adrian Baddeley, Rolf Turner, Jesper Møller, and Martin Hazelton. Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666, 2005.

[2] Andrew Bray, Ka Wong, Christopher D. Barr, and Frederic Paik Schoenberg. Voronoi residual analysis of spatial point process models with applications to california earthquake forecasts. *Ann. Appl. Stat.*, 8(4):2247–2267, 12 2014.

[3] Robert Alan Clements, Frederic Paik Schoenberg, and Danijel Schorlemmer. Residual analysis methods for space-time point processes with applications to earthquake forecast models in california. *The Annals of Applied Statistics*, pages 2549–2571, 2011.

[4] Robert Alan Clements, Frederic Paik Schoenberg, and Alejandro Veen. Evaluation of spacetime point process models using super-thinning. *Environmetrics*, 23(7):606–616, 2012.

[5] N Cressie and CK Wikle. Statistics for spatio-temporal data john wiley & sons. *Inc., NJ, USA*, 2011.

[6] Daryl J Daley and David Vere-Jones. An introduction to the theory of point processes. vol. i. probability and its applications, 2003.

[7] Q. Du, V. Faber, and M. Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM Review*, 41(4):637–676, 1999.

[8] Eric Warren Fox, Frederic Paik Schoenberg, and Joshua Seth. A note on nonparametric estimates of space-time hawkes point process models for earthquake occurrences.

[9] Joshua Seth Gordon, Robert Alan Clements, Frederic Paik Schoenberg, and Danijel Schorlemmer. Voronoi residuals and other residual analyses applied to csep earthquake forecasts. *Spatial Statistics*, 14:133–150, 2015.

[10] Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971.

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

[12] K. Matthes. Brmaud, p.: Point processes and queues. martingale dynamics. springer-verlag, berlin heidelberg new york 1981, 373 s., 31 abb., dm 88,. *Biometrical Journal*, 30(2):248–249, 1988.

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

[14] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, Sung Nok Chiu, and D. G. Kendall. *Spatial Tessellations*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition, 2000.

[15] Frederic Paik Schoenberg. Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, 98(464):789–795, 2003.

[16] D. Schorlemmer, M. Gerstenberger, S. Wiemer, and D. Jackson. Earthquake likelihood model testing. *Seismological Research Letters*, pages 17–29, 2007.

[17] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458):369–380, 2002.