Title
Evaluating One-day Earthquake Forecast Models Using Numerical Tests and Residual Analysis Methods

Author
Zhang, Zhe

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Evaluating One-day Earthquake Forecast Models Using Numerical Tests and Residual Analysis Methods

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Zhe Zhang

2022

ABSTRACT OF THE THESIS

Evaluating One-day Earthquake Forecast Models Using Numerical Tests and Residual Analysis Methods

by

Zhe Zhang
Master of Science in Statistics
University of California, Los Angeles, 2022
Professor Frederic Paik Schoenberg, Chair

Residual analysis has long been an effective tool for evaluating earthquake forecast models. In this thesis, we not only present simple numerical summaries, but also focus on graphical residual methods including pixel-based approaches, Voronoi methods and super-thinned residuals. Competing models in the Collaboratory for the Study of Earthquake Predictability (CSEP) are evaluated and compared using the residual analysis techniques, and the goodness-of-fit and deficiencies are assessed for each model.

The thesis of Zhe Zhang is approved.

Qing Zhou

Yingnian Wu

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2022

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 would like to express my sincere gratitude to my advisor Professor Rick Schoenberg for his support and advice. His patience, encouragement, vision and motivation have deeply inspired me. I am also very grateful for the discussion with Josh Ward. In addition, I thank Professor Werner and William Savran for supplying the earthquake forecasts. Finally, I am exceedingly thankful to the members of my committee, Professor Qing Zhou and Professor Yingnian Wu.

CHAPTER 1

Introduction

A fundamental issue in seismology is the need to accurately anticipate where and when earthquakes are most likely to occur (Bolt, 2003). Forecasts are significant and necessary for both short-term allocations of emergency resources and long-term earthquake preparedness and building codes (Jordan and Jones, 2010). To this end, different space-time models for earthquake occurrences are developed, and prospective analyses of forecasting models become the focus of seismologists in comparison with retrospective assessments involving lack of reproducibility, overfitting and data selection (Bray and Schoenberg, 2013). The efforts to assess and compare their goodness-of-fit have led to developments such as the Regional Earthquake Likelihood Models (RELM) project and its successor, the Collaboratory for the Study of Earthquake Predictability (CSEP). CSEP inherited the community consensus reached in RELM that all models will be tested by tests including several numerical summary tests like the Number or N-test and the Likelihood or L-test to measure the consistency of the data with the model. However, the N-test and L-test are unable to discern a significant lack of fit unless the overall rate of the model fits extremely poorly. In addition, they fail to indicate where a model fits poorly and may be highly misleading in model comparison.

The statistical developments in the assessment of space-time point process models have resulted in new, powerful model evaluation tools including comparative quadrant methods such as Pearson residuals, and residual point process methods such as Voronoi residuals. The general form of these residual measures is a standardized difference between the number of points observed and the number expected of a fitted model. For the Pearson residuals, one may examine the integral of the residual process over regular, rectangular grid cells. However, a problem expressed in Bray et al. (2014) is that if pixels are large, then the method may

overlook local inconsistencies between the observation and a forecast and result in even gross misspecification; if pixels are small, residuals that correspond to pixels with an earthquake may overwhelm the others in a visual inspection such that a plot of the Pearson residuals may not be informative. To address the problem of pixel size specification, Voronoi residuals and Voronoi deviances are proposed, which have the advantage of being less skewed. Some other powerful alternative residual analysis techniques like super-thinned residuals have also been proposed, which were reviewed in Clements et al. (2011).

The purpose of the current paper is to review various model evaluation techniques and to demonstrate their use in one-day earthquake forecasting models for California. In Chapter 2, we describe the observed earthquake occurrences and the forecasted model, the Epidemic-Type Aftershock Sequence (ETAS) model (Ogata and Zhuang, 2006) and the Short-Term Earthquake Probability (STEP) model (Gerstenberger et al., 2005), for comparison. Chapter 3 discusses the methods currently used by CSEP for model evaluation. Pixel-based Pearson residuals and deviances are applied in Chapter 4, and Chapter 5 investigates Voronoi residuals and Voronoi deviances. Chapter 6 introduces and applies the method of super-thinning. Chapter 7 summarizes the results as well as some of the benefits and weaknesses of these methods.

CHAPTER 2

Models and Data for Comparison

In this chapter, we lay out the fundamental frameworks for point processes and describe two models which have been accepted into CSEP and the data we use for comparison in detail.

2.1 Point Processes

A point process is a collection of points representing the time and/or location of an event which falls in some space. A spatial-temporal point process is a stochastic process that generates a countable set of points $\{(s_i, t_i) : i = 1, 2, ..., N\}$ in $\mathbb{R}^2 \times \mathbb{R}$. It is often characterized by its associated conditional intensity function where, given the history of the process, H, up to time t,

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

which represents the infinitesimal expected rate by which points occur in space s and time t, given information on all of the occurrences of points prior to time t. Several marks along with each observation may also be recorded to help forecast future events. In seismology, the magnitude of the event is an important mark. In addition, the finite-dimensional point processes are considered to be simple, i.e. $(s_i, t_i) \neq (s_j, t_j)$ for all $i \neq j$, so the point process can be uniquely determined by the conditional intensity (Daley and Vere-Jones, 2007).

2.2 Self-exciting Point Processes

Self-exciting or Hawkes point processes are a class of branching point processes in which there is a positive correlation between points. With the ability to capture triggering and clustering behaviors, the model has become widely used in modeling seismicity since the occurrence of an earthquake could trigger nearby aftershocks. In the spatial and temporal context, the conditional intensity of a self-exciting point process depends on the past history H_t and is of the form

$$\lambda(x, y, t|H_t) = \mu(x, y, t) + K \sum_{i:t_i<t} g(x - x_i, y - y_i, t - t_i),$$

where K describes the productivity of the triggering function, $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_n\}$ denote the sequence of locations and $\{t_1, t_2, ..., t_n\}$ the times of the events within an earthquake catalog.

The events are classified into two types: background and triggered. The background rate of the process as a function of location and time is modeled by the Poisson process $\mu(x, y, t)$, and the triggering function g models additional events induced by a prior event. The triggering function is typically nonnegative and often a power law decay function or kernel function. Meanwhile, it can be separated into space and time for simplicity such that $g(x-x_i, y-y_i, t-t_i) = f(x-x_i, y-y_i)h(t-t_i)$. The summation term gives the contribution of all previous events to the overall intensity at location (x, y) and time t.

2.2.1 Epidemic-Type Aftershock Sequence Model by Ogata and Zhuang (2006)

The Epidemic-Type Aftershock Sequence (ETAS) model is one of the most important seismicity forecasting models. The initial ETAS model was proposed by Ogata (1988), modeling the temporal background seismic activities and their aftershocks. Ogata (1998) soon extended it to a spatial-temporal model, attempted different parameterizations of the model and considered the separable form in space and time clustering effects.

The model is essentially a marked self-exciting point process, expressed as

$$\lambda(x, y, t) = \mu(x, y) + \sum_{i:t_i<t} g(x - x_i, y - y_i; m_i)v(t - t_i),$$

where m is the magnitudes of the observed events and the aftershock decay is represented by the Modified Omori function $v(t) = K(t+c)^{-p}$. The common standard form of the triggering functions is typical the multiplication of normalized time and space density distributions besides the multiplication of size function $\kappa(m) = \text{const.} \times \sigma(m) \propto e^{\alpha m}$ such that

$$g(x, y; m) = \kappa(m) \times \frac{(p-1)c^{p-1}}{(t+c)^p} \times [\frac{1}{\pi\sigma(m)} h\{\frac{(x, y)S(x, y)^t}{\sigma(m)}\}].$$

Zhuang et al. (2004) made a diagnostic analysis of space-time features of clusters by stochastic declustering, and the expected number of offspring reveals a significant bias in the spatial scaling factor. Hence, a better-fitted model that reduces the bias was proposed by Ogata and Zhuang (2006). The constraint between κ and σ is removed, leading to a new model

$$g(x - x_i, y - y_i; m_i - m_c) = e^{(\alpha-\gamma)(m_i-m_c)} \times [\frac{(x - x_i, y - y_i)S_i(x - x_i, y - y_i)^t}{e^{\gamma(m_i-m_c)}} + d]^{-q},$$

which requires the eight parameters $\theta = (\mu, K, c, \alpha, \gamma, p, d, q)$. A further term γ is introduced, taking into account the correlation between the aftershock area and the aftershock's magnitude.

2.2.2 Short-Term Earthquake Probability model by Gerstenberger et al. (2005)

The STEP model introduced by Gerstenberger et al. (2005) to assess seismic hazards is an extension of the marked temporal earthquake clustering model proposed by Reasenberg and Jones (1989, 1990 1994). They assumed that both the aftershock magnitudes and the variation in aftershock rate with the mainshock magnitude follow the Gutenberg-Richter relationship, and the rate of aftershocks varies with time obeying the modified Omori Law.

The aftershock occurrence is described as a nonhomogeneous Poisson process occurring at time t with a rate greater than a magnitude threshold M_c given as

$$\lambda(t, M) = \frac{10^{a+b(M_m-M)}}{(t+c)^p}.$$

The Reasenberg and Jones model was then expanded by Gerstenberger et al. (2005) in several ways. The new model evaluates the spatial distribution of the aftershock sequence and computes the overall hazard from multiple aftershocks. With sufficient observations, sequence-specific aftershock parameters are determined using maximum likelihood techniques. All possible magnitudes of seismic shaking are included in a hazard assessment, and the hazard computation is based on an automated decision for real-time processing. In this manner, the STEP model helps establish an appropriate null hypothesis for testing short-term conditional probabilistic seismic hazards.

2.3 The Data

The ETAS and STEP models explored here come from CSEP's rate-based repository, and both of the models produce a forecasted expected number of events in each spatial-magnitude bin for each day. Each bin is of size 0.1° longitude (lon) by 0.1° latitude (lat) by 0.1 units magnitude for earthquake magnitudes ranging from 3.95 to 8.95. There is a single bin of size 0.1° by 0.1° by 1.05 units of magnitude for magnitudes 8.95-10. The RELM testing region was designed to incorporate all earthquake occurrences in California and $\sim 1°$ around it.

The observational and forecasting data selected is for a period of one year from October 1, 2016 to September 30, 2017. Within the space-time window, there are 22 earthquakes observed. In addition, for residual analysis, the conditional intensity is aggregated for each spatial bin by summing the forecasted rate over all magnitude bins and the forecast period. As a result, the overall intensity is the expected conditional rate for each of the 7,682 grids for each model, shown in Table 2.1.

Longitude	Latitude	Rate
[125.4, 125.3]	[40.1, 40.2]	0.0367
[125.4, 125.3]	[40.2, 40.3]	0.0472
[124.4, 124.3]	[40.1, 40.2]	0.0355
[124.4, 124.3]	[40.2, 40.3]	0.0422
...	...	...

Longitude	Latitude	Rate
[125.4, 125.3]	[40.1, 40.2]	0.0000
[125.4, 125.3]	[40.2, 40.3]	0.0000
[124.4, 124.3]	[40.1, 40.2]	0.0072
[124.4, 124.3]	[40.2, 40.3]	0.1362
...	...	...

Table 2.1: Expected Conditional Rate for Each Grid Forecasted by the ETAS model (Left) and the STEP model (Right). STEP has several pixels with forecasted conditional intensities of 0.

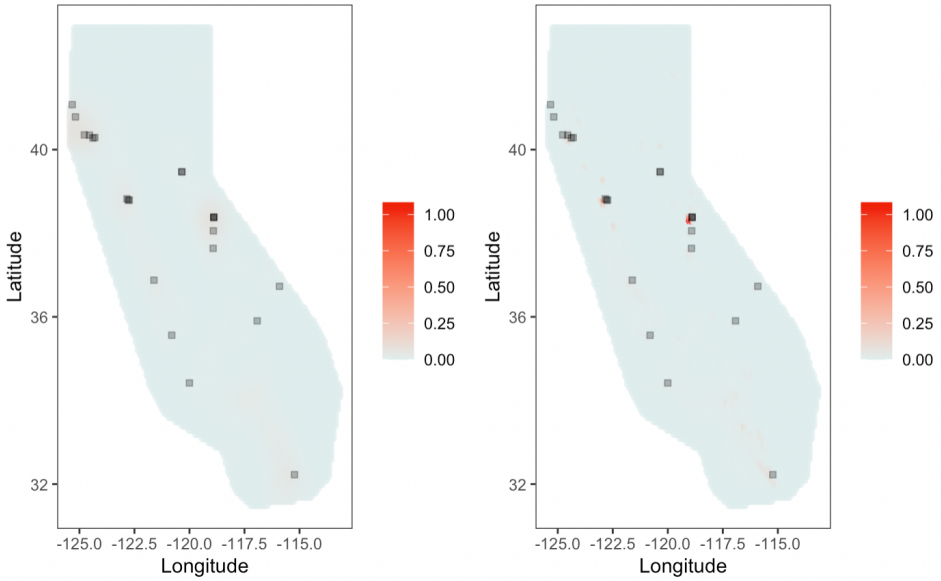

Figure 2.1: Cumulative forecast plot from October 1, 2016 to September 30, 2017 and locations of earthquakes with magnitude $M \geq 3.95$ in the RELM testing region.

Figure 2.1 visualizes the cumulative intensity for the one year period from the two models as well as the estimated earthquake hypocenter locations indicated by the black square for the 22 earthquake occurrences in space. The color scale corresponding to the expected conditional intensities is presented through the legends on the right, where darker red indicates higher expected conditional rates. Meanwhile, the color scale is set to be the same for the two models in order to make comparisons compatible. It can be seen that the two plots share some similarities in forecasting earthquake regions, which are also aligned with most events. Meanwhile, there are some differences between the forecasts made by the two methodolo-

gies. The spatial distribution of forecasting earthquakes of the ETAS model tends to be more evenly and on a larger scale, with smaller expected conditional rates. The discrepancies are demonstrated especially in the Trinidad fault zone (lon $\approx -124.5°$W and lat $\approx 40.5°$N) and near Hawthorne, Neveda (lon $\approx -118.7°$W and lat $\approx 38.4°$N). In contrast, the STEP model tends to make accentuated forecasts such as near the Geysers (lon $\approx -122.7°$W and lat $\approx 38.8°$N) and Hawthorne, Neveda.

CHAPTER 3

Numerical Tests

This chapter reviews the suite currently implemented by CSEP to assess seismological models. Two types of numerical tests will be introduced and used to evaluate the considered models. Then, the limitations are discussed.

3.1 Introduction to numerical tests

Two numerical summary tests, the Likelihood-test (L-test) and the Number-test (N-test), were initially implemented by CSEP to evaluate and compare earthquake forecast models. For the numerical tests, each model consists of the estimated number of earthquakes in each of the spatial-temporal-magnitude bins.

The L-test, described by Schorlemmer et al. (2007), assesses the quality of a forecast in the likelihood space, evaluating the probability of the observed events under the fitted model. It simulates some fixed number s of realizations from the forecast model and compares the log-likelihood (ℓ) of the observed earthquake catalog (ℓ_{obs}) with that of each simulation (ℓ_j, for $j = 1, 2, \ldots, s$). The quantile score, γ , is given by the fraction of simulated likelihoods that are less than the observed catalog likelihood in such a way that

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

where $\mathbf{1}$ denotes the indicator function. The forecast is considered to be inconsistent with the observed seismicity if the value of γ is close to zero and can be rejected.

The N-test focuses on the number of events, and the quantile score δ is defined as the

fraction of simulations that contain fewer points than the actual observed number of points in the catalog, N_{obs}:

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

where N_j is the number of points in the jth simulation of the model. If δ is close to 0 or 1, then the model is under-predicting or over-predicting the total number of earthquakes respectively, which would be rejected.

3.2 Analysis and limitations of numerical tests

Table 3.1 shows results for the L and N-test for the two models. From the γ score alone, the ETAS and STEP model would not be rejected, so the N-test should be applied for a final decision. At the 5% level of significance, the δ scores demonstrate that the ETAS model is underpredicting earthquake rates, while the actual observed number of events falls inside the 95% interval for the number of events in the forecast of the STEP model, passing the N-test.

Model	ℓ_{obs}	γ	N_{obs}	δ
ETAS	-184.92	1.00	22	**1.00**
STEP	-173.73	0.81	22	0.87

Table 3.1: Results of the L and N-test including the observed log-likelihoods, ℓ_{obs}, the L-test γ scores, the observed number of events, N_{obs} and the N-test δ scores. The bold-faced score implies significance at the 5% level resulting in rejection of the forecast.

Overall, the L-test evaluates the forecast based on magnitude, spatial location and number of events, so it is considered more comprehensive than the N-test which ignores the first two components. However, only the overall goodness-of-fit of the model can be evaluated by the two tests, and they fail to suggest areas where one given model may behave poorly and where one model performs better than the other. As Clements et al. (2011) mentioned, in practice, both statistics γ and δ test essentially the same thing, namely, the agreement between the observed and modeled total 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.

Additionally, when used to compare competing models, the results of likelihood-based tests can be uninformative and misleading due to the problem of the variable null hypothesis (Bray and Schoenberg, 2013). For example, if model A has a higher likelihood than model B, it would be quite likely for model A to be rejected while model B not to be rejected based on the L-test. It is mainly the result of different null hypotheses of the two tests. When L-test is utilized to test model A, the null hypothesis is model A, and when model B is tested, the null hypothesis is model B. Hence, under different hypotheses, the test statistic may distribute very differently, leading to misleading or even seemingly contradictory results.

CHAPTER 4

Pixel-based Residual Methods

The residual analysis method for spatial-temporal point processes provides graphical displays of model performance. In this chapter, we introduce three pixel-based residuals and use the methods to analyze how well the two models fit.

4.1 Raw residuals

Baddeley et al. (2005) proposed methods for the residual analysis of purely spatial processes, according to differences between the number of points occurring within prespecified bins and the number expected by the model. Zhuang (2006) extended such methods to the spatial-temporal case which is very appropriate for assessing the CSEP models as the models are restricted to have a constant conditional rate within predetermined bins. Consider a model $\hat{\lambda}(t, x, y)$ for the conditional intensity at any location s and time t. Raw residuals may be defined following Baddeley et al. (2005) as the number of points observed minus the number of expected points in each bin B_i, that is,

$$R(B_i) = \int_{B_i} r(t, x, y)\, dtdxdy = \int_{B_i} dN - \int_{B_i} \hat{\lambda}(t, x, y)\, dtdxdy = N(B_i) - \int_{B_i} \hat{\lambda}(t, x, y)\, dtdxdy.$$

The adequacy of the fitted model can be checked by inspecting whether $R(B_i) \approx 0$.

Figure 4.1(a) shows the raw residuals for the ETAS model with the largest residual (2.93) located on the pixel that is near Hawthorne, Neveda (lon $\approx -118.7°$W and lat $\approx 38.4°$N) which is also the location of a small cluster of earthquakes. Another notable residuals (1.99) occurs in the Sierra area in close proximity to the Mohawk Valley fault zone (lon $\approx$ -120.4°W

and lat ≈ 39.4°N). Besides that, a residual of 1.96 occurs around the Collayomi fault zone area (lon ≈ -122.8°W and lat ≈ 38.7°N). The earthquake occurrences were more accurately forecasted in other areas such as the Trinidad fault zone (lon ≈ -124.5°W and lat ≈ 40.5°N) which contains another small cluster of points.

The raw residuals of STEP model, visualized in Figure 4.1(b) share some similarities to the ETAS model in terms of model fit, while differences also exist. The largest raw residuals (1.95) for the STEP model occurs near the Mohawk Valley fault zone (lon ≈ -120.4°W and lat ≈ 39.4°N). Large residual (1.91) is also located close to Hawthorne, Neveda (lon ≈ -118.7°W and lat ≈ 38.4°N) and (1.85) around the Collayomi fault zone area (lon ≈ -122.8°W and lat ≈ 38.7°N).

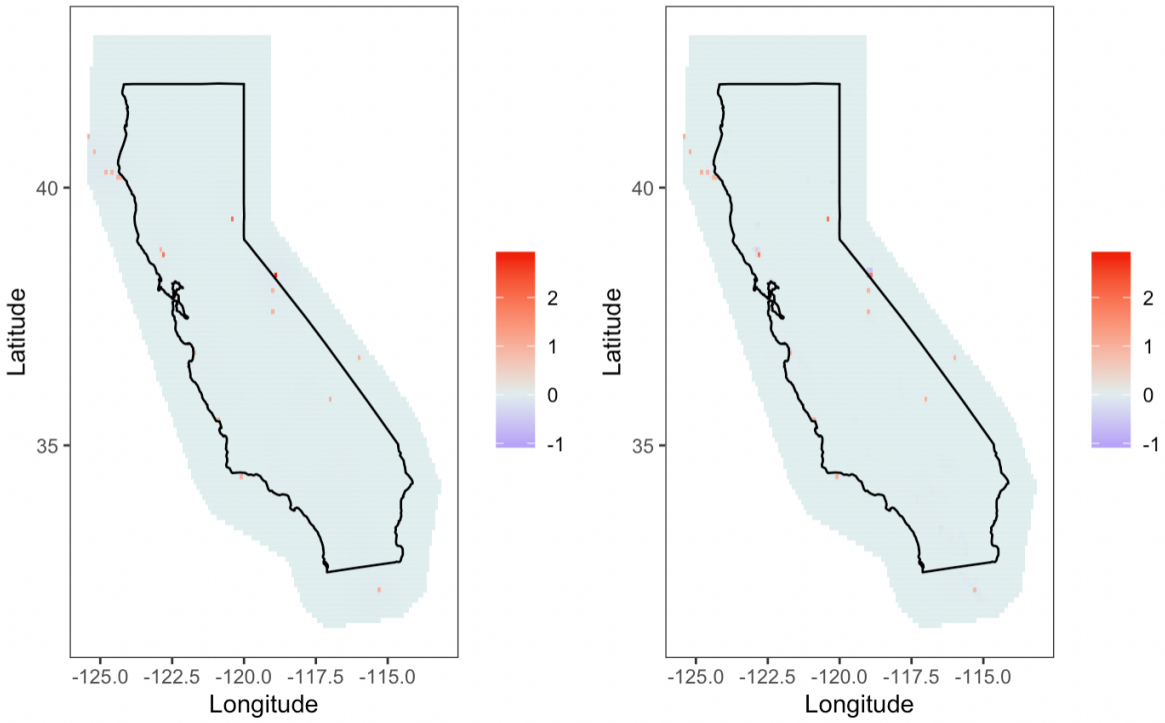

Figure 4.1: Left panel (a): pixel-based raw residuals for the ETAS model. Right panel (b): pixel-based raw residuals for the STEP model. The maximum observed pixel-based raw residual is 2.93 for ETAS and is located near Hawthorne, Neveda (lon ≈ −118.7°W and lat ≈ 38.4°N). The maximum observed pixel-based raw residual is 1.95 for STEP and is close to the Mohawk Valley fault zone (lon ≈ -120.4°W and lat ≈ 39.4°N).

4.2 Pearson residuals

Baddeley et al. (2005) also introduced various ways of standardizing the difference between the number of events and expected points such as the Pearson residuals which are defined as

$$R_P(B_i) = \int_{B_i} r_P(t,x,y)\,dtdxdy = \int_{B_i} \frac{1}{\sqrt{\hat{\lambda}}}\,dN - \int_{B_i} \sqrt{\hat{\lambda}}\,dtdxdy$$

$$= \sum_{(t_j,x_j,y_j)\in B_i} \frac{1}{\sqrt{\hat{\lambda}(t_j,x_j,y_j)}} - \int_{B_i} \sqrt{\hat{\lambda}(t,x,y)}\,dtdxdy,$$

for all $\hat{\lambda}(t_i,x_i,y_i) > 0$. In this case, the raw residuals are rescaled with mean 0 and variance approximately equal to 1.

In our case, since the STEP model has several pixels with an estimated conditional intensity of 0 which complicates the standardization, a minor adjustment would be made so that the forecasted conditional intensity in each of these locations is slightly greater than 0. However, the resulting Pearson residuals of a few pixels are tens or hundreds of times larger than others, rendering a plot of Pearson residuals providing limited information. Indeed, when the aggregated conditional intensity in some pixels is very small or the spatial-temporal pixels are very small, the distribution of raw residuals becomes highly skewed and the standardization makes such residuals more problematic. The residuals in such pixels may dominate the others in a visual setting. Hence, Pearson residuals emphasize more like the locations of the observations themselves and fail to identify the quality of the fit in the other pixels where earthquakes did not occur.

In this case, one possible solution would be to enlarge the pixel size such that the integrated conditional rate in each pixel is higher. However, while this would be effective in the case of a homogeneous process, a different problem could be induced in a nonhomogeneous setting. Since positive and negative values of the residual process within a particular pixel can cancel each other out, a large pixel may overlook gross misspecification such that the residuals will have low power. Therefore, although raw and Pearson residuals could identify individual pixels containing earthquakes which need an improvement in their forecasts, it is

typically hard to avoid the problem of highly skewed or low power residuals such that the residuals cannot figure out other regions where the models may fit relatively well or poorly.

4.3 Deviances

Wong and Schoenberg (2009) introduced pixel-based deviance residuals which are useful for comparing the fit of models. Similar to raw and Pearson residuals, deviances are computed over evenly spaced pixels. The differences between log-likelihoods for the two models would then be examined. Given the conditional intensity $\hat{\lambda}_1$ and $\hat{\lambda}_2$ of two models, the deviances in each pixel, B_i, are expressed as

$$R_D(B_i) = \sum_{i:(t_i,x_i,y_i)\in B_i} log(\hat{\lambda}_1(t_i,x_i,y_i)) - \int_{B_i} \hat{\lambda}_1(t,x,y)\,dtdxdy$$

$$-\Bigg(\sum_{i:(t_i,x_i,y_i)\in B_i} log(\hat{\lambda}_2(t_i,x_i,y_i)) - \int_{B_i} \hat{\lambda}_2(t,x,y)\,dtdxdy \Bigg).$$

Positive residuals imply that the model $\hat{\lambda}_1$ fits better in the given pixel and negative residuals imply that $\hat{\lambda}_2$ provides a better fit. In addition, an overall log-likelihood ratio score could be obtained by summing the deviance residuals up, $\sum_i R_D(B_i)$, to get an overview of the quality of fit and improvements.

The deviances for the ETAS model versus the STEP model can be seen in Figure 4.2. Although the discrepancies between the two models may not be large, the ETAS model outperforms the STEP model in about two-thirds of all locations where earthquakes were observed. Specifically, the ETAS model forecast the seismicity accurately around the Collayomi fault zone (lon $\approx$ -122.8°W and lat $\approx$ 38.8°N) and also South in the forecast region. The STEP model appears to improve upon the ETAS model in the Pacific Ocean (lon $\approx$ -124°W and lat $\approx$ 40.0°N) and also performs well around Hawthorne, Neveda (lon $\approx$ -118.9°W and lat $\approx$ 38.3°N) where clusters of earthquakes are located. Additionally, the ETAS model demonstrates a much better fit in a few pixels on the northwest side as indicated by the dark

red, but the STEP model does a better job in most locations including the vast majority of locations far from seismicity, while the ETAS model tends to over-predict events in these locations. Overall, the log-likelihood ratio score is -6.89, suggesting that the STEP model offers some improvements over the ETAS model.

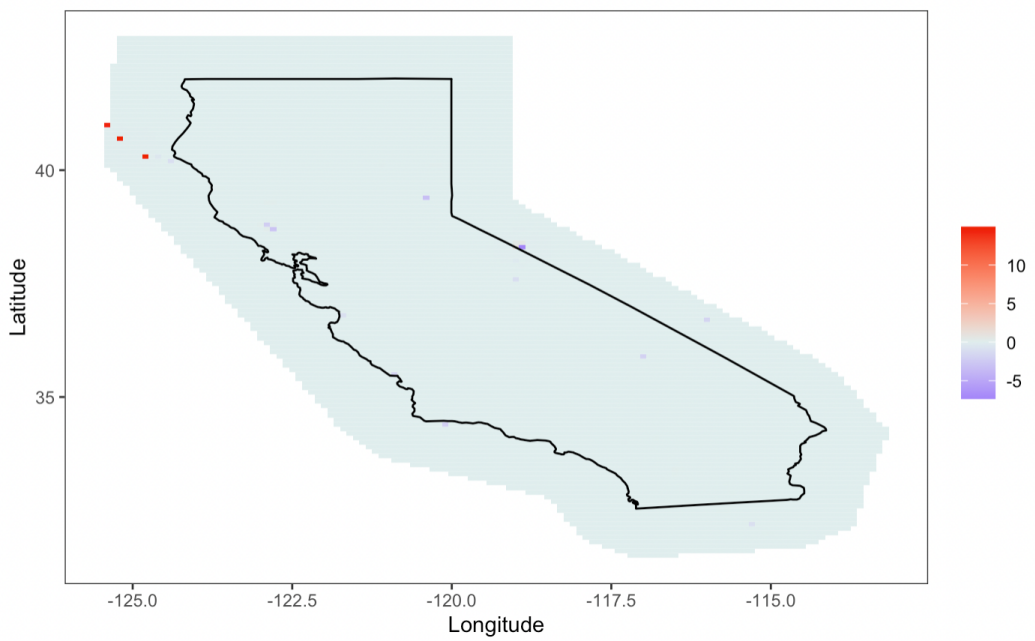

Figure 4.2: Deviance residuals for the ETAS model versus STEP model. The ETAS model performs well in a few pixels on the northwest side shown by the dark red and in two-thirds of locations where earthquakes occurred. The STEP model forecasts well in most locations including the vast majority of locations far from seismicity.

One thing that should be noted is that the result can depend heavily on the 0 replacement values of the STEP model. In our case, the 0 replacement values are chosen to be half of the minimum non-zero conditional intensity which is around 1e-8. However, if the 0 replacement values chosen get closer to 0, it would likely lead to a different result which implies the ETAS model outperforms the STEP.

Overall, deviance residuals help make a comparison between two competing models of similar times and regions. It can identify places where one appears to perform better than the other in alignment with the observed earthquakes, and a log-likelihood ratio score can be computed for an insight into the improvement in fit from one model to another, helping

the building of subsequent models and experiments.

CHAPTER 5

Voronoi Methods

While pixel-based residuals like Pearson residuals and deviances are useful for gridded forecasts, departures from the events can be identified by Voronoi residuals and Voronoi Deviances. They are also valuable for overcoming the challenges which are the results of small aggregated conditional intensities in each pixel, leading to high skewness in the distribution of the raw and Pearson residuals. This chapter presents concepts of Voronoi methods, introduces the data preparation for Voronoi residual analysis and analyzes the results of model evaluation using Voronoi methods.

5.1 Voronoi Residuals

Voronoi residuals are constructed using Voronoi tessellation which is a partition of the metric space on which a point process is defined into convex polygons, or Voronoi cells, C_i. Specifically, given a point pattern of N events, for each observed point x_i of the point process, its corresponding cell is the region including all locations that are closer to x_i than to any other point of N. The Voronoi tessellation is the collection of such cells which we assume fills the complete window C_T such that $C_T = \cup_{i=1}^{N} C_i$. Then, given a model for the conditional intensity of a spatial or spatial-temporal point process, the raw residual process may be aggregated over the Voronoi cells rather than over rectangular pixels to construct residuals for a conditional intensity model, which is referred to as Voronoi residuals.

In comparison with conventional pixel-based methods, the partition of Voronoi residuals is entirely spatially adaptive, data-driven and automatic, so the distribution of residuals tends to be far less skewed than pixel-based methods. Since each Voronoi cell has exactly

one point inside it by construction, the raw Voronoi residual for cell i is given by

$$\hat{R}_i := 1 - \int_{C_i} \hat{\lambda}\, d\mu$$

$$= 1 - |C_i|\bar{\lambda}$$

where $\bar{\lambda}$ denotes the mean of the proposed model $\hat{\lambda}$ over C_i. Following Baddeley et al. (2005), the raw Voronoi residuals can be recalled of the form

$$R_V(C_i) = \frac{1 - \int \hat{\lambda}(t,x,y)\, dtdxdy}{\sqrt{\int \hat{\lambda}(t,x,y)\, dtdxdy}}.$$

The determination of an appropriate color scale with appropriate limits is one difficulty when plotting Voronoi residuals. Bray et al. (2014) suggested a probability integral transformation (PIT) to scale the Voronoi transformation uniformly. However, it was computationally intensive since repeated simulations of the fitted model are required. Thus, although the PIT method performs well in terms of visualizing the residual process in seismology, simpler alternatives may be proposed. For instance, 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 (Gordon et al., 2015).

To prepare for the Voronoi residual analysis, the space-time window of the observational data is divided into 22 cells where each of the 22 observations corresponds to one Voronoi cell. Each Voronoi cell consists of all locations which are closer to the generating point than to any other observation points. Hence, each pixel of size 0.1° longitude by 0.1° latitude is assigned to the "closest" Voronoi cell in terms of Euclidean distance. The expected conditional intensity is then aggregated over each Voronoi cell which will be compared to the observational data to evaluate the quality of the fit of the model. Table 5.1 shows the integrated expected conditional intensities for the first four Voronoi cells forecasted by ETAS and STEP model. It can be seen that the ETAS model forecasts more earthquakes than the

STEP model in general and the STEP model seems to perform poorly in predicting a few events with approximate 0 expected conditional rate .

Date	Integrated Rate
2016.820	0.8939
2016.928	0.7089
2016.952	2.2161
2016.952	1.8414
2016.953	2.0795
...	...

Date	Integrated Rate
2016.820	0.4124
2016.928	0.4087
2016.952	0.6340
2016.952	1.8745
2016.953	1.4659
...	...

Table 5.1: Expected conditional intensities forecasted by ETAS model (Left) and STEP model (Right). The ETAS model tends to forecast more earthquakes than STEP.

For the Voronoi residuals, positive values represent under-predictions, negative values represent over-predictions, and values around 0 indicate well-behaved predictions made by a model. Figures 5.1(a) and 5.1(b) show the standardized Voronoi residuals for the ETAS model and the STEP model respectively. It can be seen that both models tend to over-predict seismicity as indicated by a vast region of red cells. Specifically, both models perform poorly in forecasting earthquakes on the outer ring of Southern California such as the seismicity around the Pacific Ocean area (lon $\approx$ -120.00°W and lat $\approx$ 34.42°N) and near Mexicali at the Mexico-California border (lon $\approx$ -115.23°W and lat $\approx$ 32.23°N). Meanwhile, both models under-predict seismicity near Hawthorne, Nevada (lon $\approx$ -118.89°W and lat $\approx$ 38.37°N). The ETAS model seems to do a decent job while anticipating the earthquakes that happened in the Trinidad fault zone (lon $\approx$ -124.54°W and lat $\approx$ 40.34°N), while the STEP model appears to under-predict seismicity in this region. Furthermore, the STEP model performs well forecasting seismicity in clusters nearby (lon $\approx$ -118.91°W and lat $\approx$ 40.78°N) and (lon $\approx$ -124.38°W and lat $\approx$ 40.27°N). In addition, the STEP model appears to forecast earthquakes accurately around the Collayomi fault zone area (lon $\approx$ -122.77°W and lat $\approx$ 38.78°N).

Using the same color scale enables simple comparisons between models. Overall, the Voronoi residuals indicate that the STEP model appears to demonstrate a better seismicity forecasting ability. It forecasts the earthquake near the Trinidad fault zone and around the Collayomi fault zone accurately, with a Voronoi residual very close to 0. In addition, although

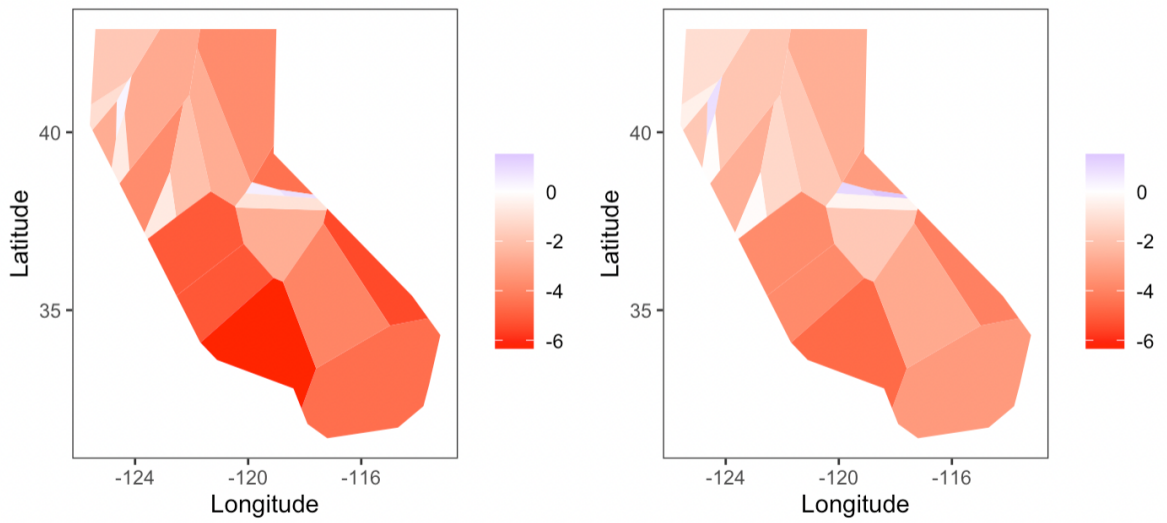

Figure 5.1: Left panel (a): Voronoi residuals for the ETAS model. Right panel (b): Voronoi residuals for the STEP model. Both models perform well in the Trinidad fault zone (lon ≈ -124.5°W and lat ≈ 40.5°N). Moreover, the STEP model appears to forecast earthquakes accurately around the Collayomi fault zone area (lon ≈ -122.77°W and lat ≈ 38.78°N).

the STEP model under-predicts seismicity in a few regions, it seems to improve upon the ETAS model by decreasing the tendency of over-predicting seismicity in most areas.

5.2 Voronoi Deviances

To further assess the predictions made by the two models, Voronoi deviances are utilized to make a comparison. Similar to pixel-based deviances, the log-likelihoods of the ETAS model and the STEP models are examined. In this case, instead of computing deviances over evenly spaced pixels, we compute deviances from Voronoi tessellated cells. The Voronoi deviance residual of a Voronoi cell C_i is then defined as

$$R_V(C_i) = \sum_{i:(t_i,x_i,y_i)\in C_i} log(\hat{\lambda}_1(t_i,x_i,y_i)) - \int_{C_i} \hat{\lambda}_1(t,x,y)\, dtdxdy$$

$$-\left(\sum_{i:(t_i,x_i,y_i)\in C_i} log(\hat{\lambda}_2(t_i,x_i,y_i)) - \int_{C_i} \hat{\lambda}_2(t,x,y)\, dtdxdy \right),$$

where $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the fitted conditional intensity given two models. A $R_V(C_i)$ value close to 0 implies that the two models provide a similar fit in the Voronoi cell C_i, and larger $R_V(C_i)$ values suggest bigger discrepancies between forecasts made by the two models.

Voronoi deviances between the ETAS Model and the STEP model are shown in Figure 5.2. Combing it with the Voronoi residual plots, it can be seen that the STEP model outperforms the ETAS model in most of the regions since the ETAS model excessively over-predict earthquakes. The ETAS model accurately forecasts seismicity in the northwest in the Pacific Ocean and at longitude ≈ -122.73°N and latitude ≈ 38.79°W. Overall, the total log-likelihood ratio is -6.89, confirming the result of pixel-based deviances that the STEP model provides a better fit than the ETAS model, but it would be easier to compare the forecasts with the observational data in each area using the visualization of Voronoi deviances.

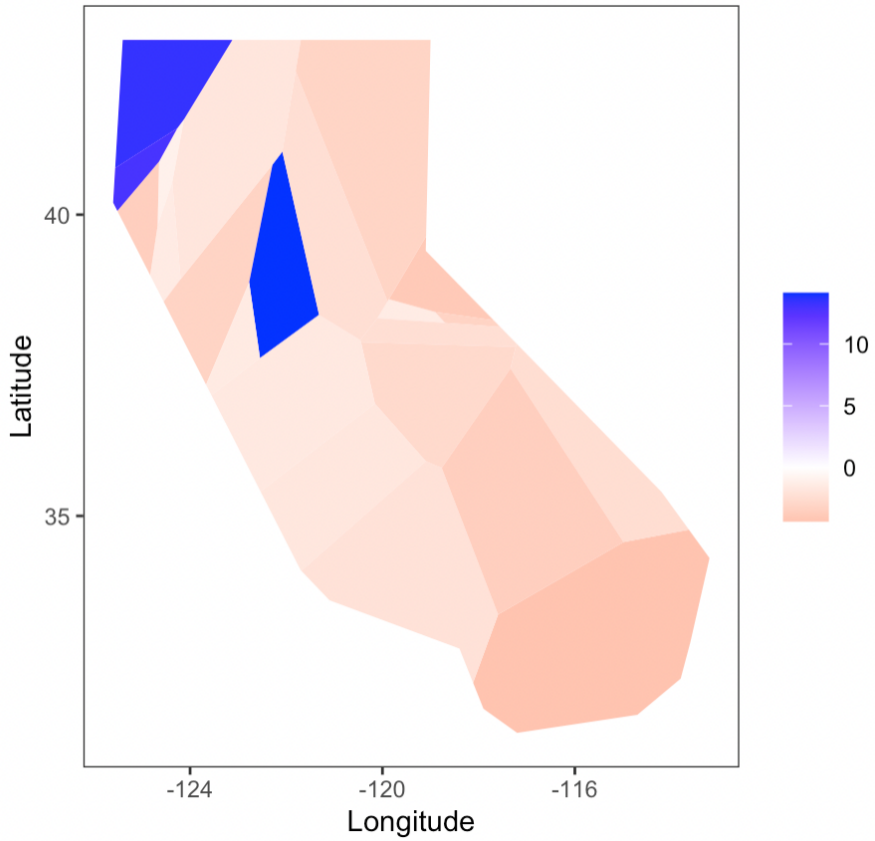

Figure 5.2: Voronoi deviances for the ETAS model versus STEP model. The STEP model outperforms ETAS in most regions.

CHAPTER 6

Super-thinning

In this chapter, an introduction to super-thinning is presented at the beginning. Then, we give steps of applying super-thinning and discuss the resulting super-thinning residuals to compare the goodness-of-fit of the two models.

6.1 Introduction to super-thinning

Super-thinning for residual analysis, proposed by Clements et al. (2012), is a combined approach of thinning introduced by Schoenberg (2003) and superposition by Brémaud (1981). By thinning, each point x_i of a point process N is kept with some probability p_i. If the estimate $\hat{\lambda}$ is equal to the true conditional intensity λ almost everywhere, then the residual process will be homogeneous Poisson with rate k. In residual superposition, it transforms the point process N by simulation, which creates a residual point process superposed onto N, generating a homogeneous Poisson process. It is essentially an addition operator on the point process, i.e. N_3 is the superposition of point process N_1 and N_2. However, when the model $\hat{\lambda}$ for the conditional intensity of the observed point process is volatile, the two residual analysis methods tend to have low power. Therefore, super-thinning which could control the relative amount of thinning and superposition is a more powerful approach. As noted in Clements et al. (2012), it thins in regions of high intensity and superposes simulated points in regions of low intensity, which results in a homogeneous Poisson point process if the model for λ used in the thinning and superposition is correct. Hence, homogeneity is one way of assessing the super-thinned residuals of the model, detecting any clustering or inhibition indicating a lack of fit.

In super-thinning, to transform the point process N into a residual process Z, one first thin N by keeping each observed point (t_i, x_i, y_i) independently with probability $p_i = min\{1, k/\hat{\lambda}(t, x, y)\}$ to obtain a thinned residual Z_1. Subsequently, a Cox process Z_2 may be simulated directly by $max\{0, k - \hat{\lambda}(t, x, y)\}$, which can be easily achieved by simulating a homogeneous Poisson process with rate k and independently retaining each simulated point $(\tilde{t_i}, \tilde{x_i}, \tilde{y_i})$ with probability $max\{0, (k - \hat{\lambda}(\tilde{t_i}, \tilde{x_i}, \tilde{y_i}))/k\}$. The super-thinned residuals are then the points of the residual point process $Z = Z_1 + Z_2$.

6.2 Super-thinning analysis

As Clements et al. (2012) suggested, k was chosen to be the median of $\hat{\lambda}$ for the ETAS and STEP model to optimize the power of the resulting residual test. Figure 6.1 shows the super-thinned residuals for the two models. The dark green circles indicate the thinned events, and plus signs indicate simulated points. It can be seen that the two models thinned the observed earthquakes similarly. Figure 6.1(a) suggests that there contains a few significant clusters in the right southeastern corner (lon ≈ -114.8°W, lat ≈ 33.7°N) and close to the faults on east side Madeline Plains (lon ≈ -120.1°W, lat ≈ 40.8°N). There is another notable cluster at longitude ≈ -119.9°W, latitude ≈ 36.5°N. Besides, some inhibitions exist in the residual process. For example, there are inhibitions near Tahoe Valley fault zone (lon ≈ -120.0°W, lat ≈ 38.9°N) and close to Concord fault (lon ≈ -122.0°W, lat ≈ 37.5°N).

The super-thinned residuals for the STEP model are displayed in Figure 6.1(b). Compared with the ETAS model, the points seem to be more uniformly distributed without major inhibitions. There are some small clusters center at longitude ≈ -122.5°W, latitude ≈ 38.0°N, and longitude ≈ -119.9°W, latitude ≈ 41.5°N. Another cluster is near the Paleo-subduction zone (lon ≈ -124.1°W, lat ≈ 42.0°N). There is also little over-prediction indicated by a consistent covering of residual points, so the residual process is close to what is expected as enough points were simulated. In addition, there is a small number of inhibitions close to Mission fault (lon ≈ -122.0°W, lat ≈ 37.6°N) and near the Cambria fault zone (lon ≈ -121.0°W, lat ≈ 35.5°N).

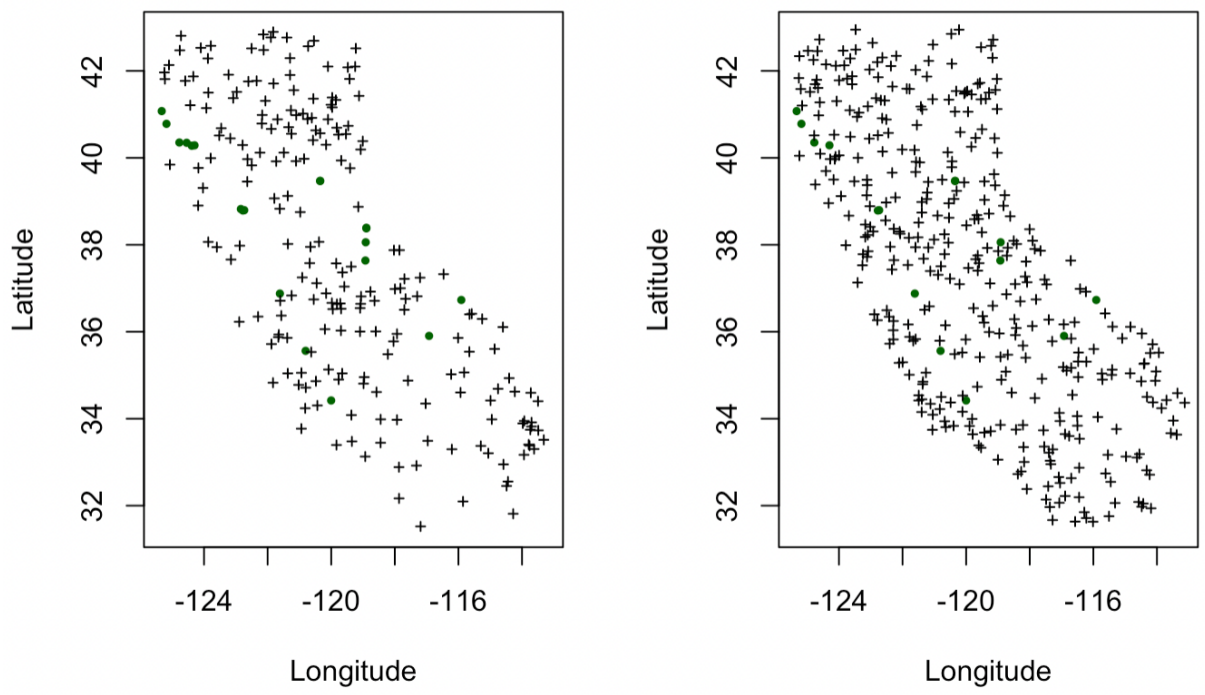

Figure 6.1: One realization of super-thinned residuals. Left panel (a): ETAS. Right panel (b): STEP. The dark green circles indicate the thinned events, and plus signs indicate simulated points. There is a significant cluster at longitude $\approx$ -114.8°W, latitude $\approx$ 33.7°N for ETAS. Clustering for the STEP model occurs near the Paleo-subduction zone (lon $\approx$ -124.1°W, lat $\approx$ 42.0°N).

Additionally, the Ripley's K-function and the variance stabilized version of the K-function called the L-function are utilized to help the super-thinning analysis. Figure 6.2 shows that at almost all distances the actual observed value of $\hat{K}$ is greater than the theoretically derived expected value indicating clustering for both models. Nevertheless, comparing the plots for the ETAS model with those for the STEP model, it can be seen that the estimated K and L of the STEP model fit the theoretical values better. Combining the result with the visualization of super-thinned residuals suggests that the STEP model seemingly performs better in super-thinning analysis.

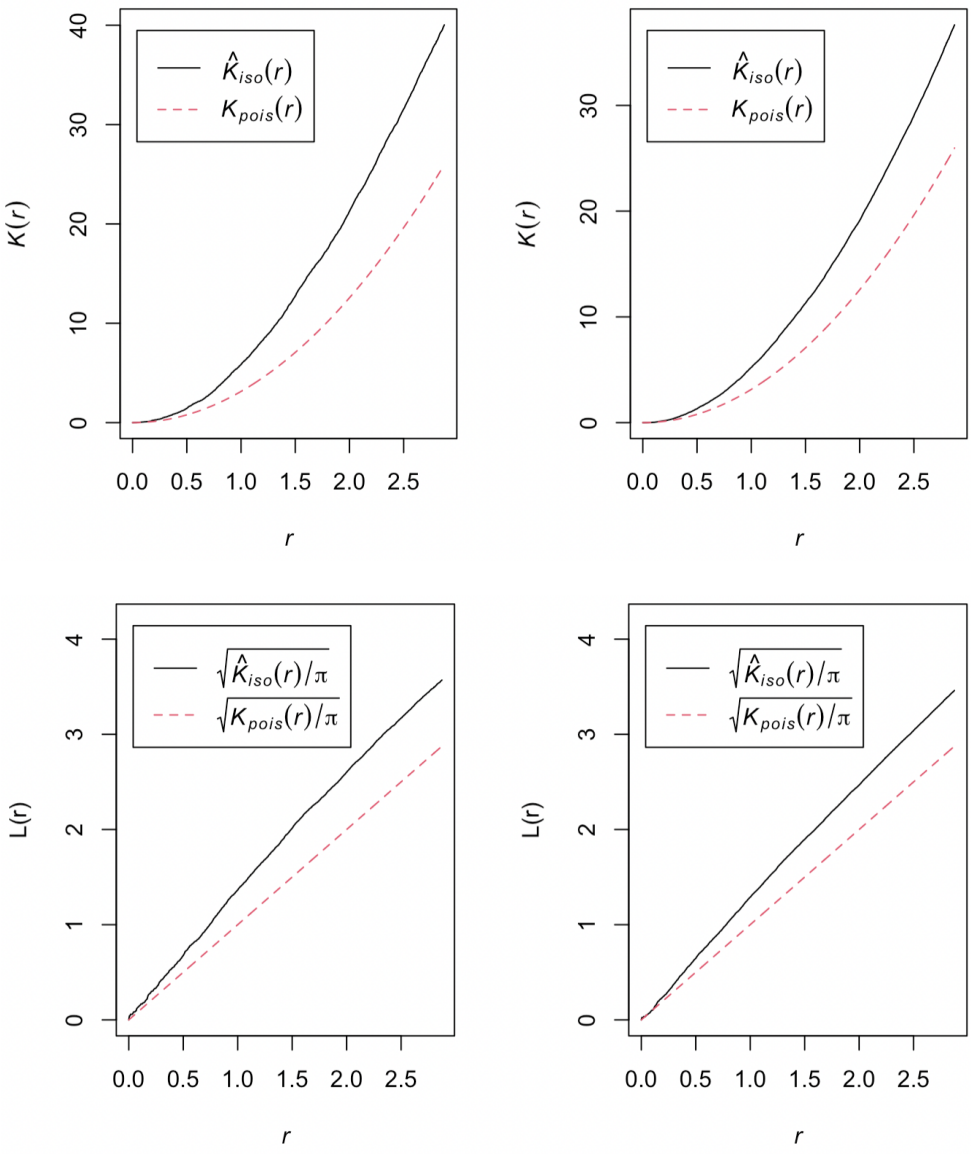

Figure 6.2: Upper left panel (a): estimated K-function for ETAS. Upper right panel (b): estimated K-function for STEP. Bottom left panel (c): estimated L-function for ETAS. Bottom right panel (d): estimated L-function for STEP. The estimated K and L of the STEP model fit the theoretical values better.

CHAPTER 7

Summary

While a variety of spatial-temporal point process models has been proposed to forecast seismicity, assessing point processes poses unique challenges. As a result, many common residual analysis methods for spatial point process models can be implemented to evaluate the quality of fit and assess strengths and weaknesses in point processes. In this paper, residual methods including pixel-based residuals, deviances, Voronoi residuals and super-thinned residuals were applied to two one-day earthquake forecast models, the Epidemic-Type Aftershock Sequence (ETAS) model and the Short-Term Earthquake Probability (STEP) model, for a catalog spanning from October 1, 2016 to September 30, 2017, and these methods provide more reliable estimates of the overall fit and more detailed information than the numerical test.

The N-test and the L-test implemented by CSEP provide easy to understand statistics that can be utilized for hypothesis testing, but they cannot identify where a model may be fitting poorly and, hence, have very low power. The N-test does not assess the spatial performance of a model, and the L-test does not differentiate between over-prediction and under-prediction.

Pixel-based residuals have advantages like the simplicity of calculation and ease of interpretation. However, the effects of expected conditional intensities of 0 and high skewness in the standardized residuals increase the difficulty in interpretations. Deviances can be valuable to compare forecasts of similar regions and times. Using a log-likelihood ratio score gives an overall impression of the improvement in fit from the better fitting model and a pixel by pixel comparison enables effortless detection of one model's performance compared

to the other.

One advantage of Voronoi residuals is that the partition of entirely automatic and spatially adaptive, which does not need inputs from the user regarding tuning parameters. Meanwhile, Voronoi residuals are substantially less skewed in comparison with pixel-based methods. Super-thinned residuals may be a promising approach as well, although they may have low power with volatile forecasted intensity.

In addition, the formulation of these models of spatial-temporal 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). The diagnostic tool can determine a model's performance when evaluating point processes. For instance, Voronoi residuals rely on fewer assumptions, so models that do a decent job in areas of low seismicity may outperform models with more accurate forecasts of where earthquakes actually occurred. Moreover, deviances can be misleading if both competing models perform well or poorly in a spatial bin. Hence, deviances are better to be used in combination with Pearson or Voronoi residuals to characterize the overall and relative performance of the two models. Nevertheless, different spatial cells used in these different residuals can lead to discrepancies. For example, deviances for the ETAS and the STEP model in Figure 4.2 suggest a superior fit of the ETAS model in the northwestern area of the forecast region, but Voronoi residuals indicate that the ETAS model did not perform as well as the STEP model in cells in the same region (Figure 5.1(a) and 5.1(b)).

Zechar et al. (2013) suggest using all tests in conjunction since each provides insight into model performance. Based on all evaluation methods applied, the STEP model shows a slightly better performance. In general, both models appear to over-predict seismicity in most locations but under-predict seismicity around Hawthorne, Nevada. Super-thinning indicates that the ETAS model has visible clusters and inhibitions. Meanwhile, Voronoi residuals suggest that the STEP model makes more accurate forecasts in comparison to the ETAS model and this is supported by the results from deviances. The STEP model has less extreme pixel-based residuals, and its spatial distribution of intensity appears to be quite

accurate in areas of low seismicity.

The question of how to improve the residual analysis methods remains open. Different ways of standardizing the residuals were proposed, but it should be noted that the choice of grids is suggested by the experience to be more important. In addition, regarding the Voronoi tessellation, some sampling variability may be induced by the random cell areas and the residuals are dependent, so methods requiring an independent and identically distributed assumption should be used cautiously. A future direction is to consider residuals based on a model-based centroidal Voronoi tessellation providing a partition that creates residuals that are independent of one another if the underlying model is Poisson.

Besides, studies may be conducted in the future to find an optimal choice of k in super-thinning. The parameter k should be chosen in a way that has a high probability to reject a poorly fitted model and a low probability to reject a satisfactory model. Meanwhile, simulations should not heavily determine the assessment of the homogeneity of the super-thinned residuals, so ideally few points may be simulated and many of the observed events are kept. The quality of fit measures under different hypotheses may be compared by future theoretical and simulation research.

REFERENCES

[1] Baddeley, A., Turner, R., Möller, J., Hazelton, M., 2005. Residual analysis for spatial point processes (with discussion). J. R. Stat.Soc. Ser. B 67 (5), 617–666.

[2] Bolt, B., 2006. Earthquakes, 5th ed. W.H. Freeman, New York.

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

[4] Bray, A., Schoenberg, F.P., 2013. Assessment of point process models for earthquake

[5] Brémaud, P., 1981. Point Processes and Queues. Springer, NY. forecasting. Statistical Science, 28(4), 510-520.

[6] Clements, R.A., Schoenberg, F.P., Schorlemmer, D., 2011. Residual analysis methods for space–time point processes with applications to earthquake forecast models in California. Ann. Appl. Stat. 5 (4), 2549–2571.

[7] Clements, R.A., Schoenberg, F.P., Veen, A., 2012. Evaluation of space–time point process models using super-thinning. Environmetrics 23 (7), 606–616.

[8] Daley, D., Vere-Jones, D., 2007. An introduction to the theory of point processes: volume I: Elementary Theory and Methods, Second Edition.

[9] Gerstenberger, M. C., Wiemer, S., Jones, L. M., Reasenberg, P. A., 2005. Real-time forecasts of tomorrow's earthquakes in California. Nature 435 328–331.

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

[11] Jordan, T. H., Jones, L.M., 2010. Operational Earthquake Forecasting: Some Thoughts on Why and How. Seismological Research Letters 81.4, pp. 571–574

[12] Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. J. Am. Stat. Assoc. 83, 9 – 27.

[13] Ogata, Y., 1998. Space-time point-process models for earthquake occurrences, Ann. Inst.Statist. Math., 50(2), 379-402.

[14] Ogata, Y., Zhuang, J., 2006. Space–time ETAS models and an improved extension. Tectonophysics 413 13–23.

[15] Reasenberg, P.A., Jones, L.M., 1989. Earthquake hazard after a mainshock in California, Science,243, 1173-1176.

[16] Reasenberg, P.A., Jones, L.M., 1990. California aftershock hazard forecast, Science, 247, 345-346.

[17] Reasenberg, P.A., Jones, L.M. ,1994. Earthquake aftershocks: Update, Science, 265, 1251-1252.

[18] Schneider, M., Clements, R., Rhoades, D., Schorlemmer, D., 2014. Likelihood- and residual-based evaluation of medium-term earthquake forecast models for California. Geophys. J. Int. 198 (3), 1307–1318.

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

[20] Schorlemmer, D., Gerstenberger, M., Wiemer, S., Jackson, D., Rhoades, D., 2007. Earthquake likelihood model testing. Seismol. Res. Lett. 78, 17–27.

[21] Wong, K., Schoenberg, F.P., 2009. On mainshock focal mechanisms and the spatial distribution of aftershocks. Bull. Seismol. Soc. Amer. 99 (6), 3402–3412.

[22] Zechar, J.D., Schorlemmer, D., Werner, M.J., Gerstenberger, M.C., Rhoades, D.A., Jordan, T.H., 2013. Regional earthquake likelihood models I: First-order results. Bull. Seismol. Soc. Amer. 103 (2A), 787–798.

[23] Zhuang, J., Ogata, Y., Vere-Jones, D., 2004. Analyzing earthquake clustering features by using stochastic reconstruction. J. Geophys. Res. 109 (B5), B05301. doi:10.1029/2003JB002879.

[24] 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.