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

Joshua Seth Gordon[1], Robert Alan Clements[2], Frederic Paik Schoenberg[3], and Danijel Schorlemmer[4].

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

Keywords. Deviances, Pearson residuals, point processes, seismology, super-thinning, thinned residuals.

[1] joshuagordon@ucla.edu, UCLA Department of Statistics, 8911 Math Sciences Building, Los Angeles, CA 90095-1554

[2] rclements17@gmail.com,

[3] Corresponding author, frederic@stat.ucla.edu, UCLA Department of Statistics, 8125 Math Sciences Building, Los Angeles, CA 90095-1554

[4] danijel.schorlemmer@gfz-potsdam.de, GFZ Helmholtz Centre Potsdam, Helmholtzstrasse 6/7, Building H 7, room 204, 14467 Potsdam, Germany.

1 Introduction

Model evaluation for point processes poses some unique challenges. Unlike time series data, regression style data, or geostatistical data, point process data are not so easily amenable to residual analysis. Indeed, with point process data, the most natural way to define residuals in analogy with other types of processes is by examination of the residual process as suggested in Baddeley et al. (2005). The residual process may be described heuristically as the (possibly standardized) difference between the number of points observed and the number expected according to a given model; the more formal definition of Baddeley et al. (2005) and Zhuang (2006) is presented in Section 2. One may then inspect the integral of this residual process over each cell within some rectangular grid of cells. Such residual plots are problematic, however. As noted in Bray et al. (2014), if the cells are small, then the residuals plots tend for many models to look similar to plots of the data, and thus do not reveal much about whether a model fits well or poorly. If the cells are large, then there is a great loss of information and the residuals say little about the small scale goodness of fit of the model in question. See Daley and Vere-Jones (1988) for a review of point processes and conditional intensities.

Some alternative residual analysis techniques for spatial-temporal point process models have been proposed recently, such as Voronoi residuals, deviances, and super-thinned residuals. These residual methods were reviewed in Clements et al. (2011) and applied to earthquake forecast models from the Regional Earthquake Likelihood Models (RELM) project (Field 2007, Schorlemmer and Gerstenberger 2007) in the framework of the Collaboratory for the Study of Earthquake Predictability (CSEP) described in Jordan (2006). The evaluations in Clements et al. (2011) were made using earthquake occurrence data from January 2006 to September 2009. Here, we review the strengths and weaknesses of these various model evaluation techniques and apply the residual analysis methods to a longer catalog, extending from January 2006 to September 2014.

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

Note that, for the purpose of evaluation and comparison of earthquake forecast models, CSEP currently implements several numerical summary tests, such as the Likelihood-test (L-test) (Schorlemmer et al. 2007) and the Number-test (N-test) (Zechar et al. 2013) to measure the consistency of a forecast with the observation and comparative tests, such as the T- and

W-tests (Rhoades et al. 2013). For instance, the L-test works by first simulating some fixed number s of realizations from the forecast model and comparing the log-likelihood (ℓ) for the observed earthquake catalog (ℓ_{obs}) with that of the simulations (ℓ_j, for $j = 1, 2, \ldots, s$). The quantile score, γ, is defined as the fraction of simulated likelihoods that are less than the observed catalog likelihood:

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

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

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

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

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

This paper is organized as follows: Section 2 presents the observed earthquake occurrences and the forecasted models for comparison, in Section 3 we apply pixel-based Pearson residuals and identify specific locations where each forecast performs well and/or poorly, in Section 4 we compare Models using deviances, Section 5 presents super-thinned residuals, and Section 6 describes Voronoi residuals, demonstrates their utility and introduces a new method of coloring scaling the cells. Section 7 summarizes the results and discusses some of the strengths and weaknesses of these methods.

2 Data and models for comparison

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

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

A: Helmstetter et al. (2007).

B: Kagan et al. (2007).

C: Shen et al. (2007).

All of these models except Model C forecast based exclusively on previous seismicity, whereas Model C incorporates geodetic and geological data as well. Models A, B, and C are five-year models, producing just one forecast of seismicity in each spatial-magnitude grid. Not included are STEP and ETAS Models which are one-day models, producing a forecasted expected number of events in each spatial grid for each day. Note that since not all RELM models in CSEP produce estimates in every pixel within this space-time window, there are different numbers of observed earthquakes occurring in the relevant forecast regions corresponding to different models. As such, Model A had 509 earthquakes observed within its corresponding space-time window, Model B had 343, and Model C had 356. All 5 year models were scaled proportionally to adjust to the temporal window used here.

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

3 Raw and Pearson residuals

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

$$R(B) = \int_B r(t, x, y)\ dtdxdy = \int_B dN - \int_B \hat{\lambda}\ (t, x, y)\ dtdxdy = N(B) - \int_B \hat{\lambda}\ (t, x, y)\ dtdxdy, \quad (1)$$

for any measurable set B. One may then observe these integrated raw residuals over a sequence of pixels B_i. Note that Baddeley et al. (2005) only consider the case of purely spatial point processes characterized by their Papangelou intensities; Zhuang (2006) showed one may nevertheless extend the definition to the spatial-temporal case using the conventional conditional intensity as in (1).

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

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

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

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

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

Model A, shown in Figure 2(a), shares many similarities to Model B in terms of model fit. The largest Pearson residual for Model A (4.57) is located near Mexicali at at the Mexico-

California border (lon ≈ −115.2°W and lat ≈ 32.2°N). In fact, this region contains most of the largest residuals for Model A. Large residuals (such as 3.51) also occur just north of California in the Battle Rock fault zone (lon ≈ −123.8°W and lat ≈ 42.6°N) and (2.59) near the Bare Mountain fault zone (lon ≈ −117.8°W and lat ≈ 36.4°N).

The largest residual for Model C (4.44) occurs in close proximity to the Bare Mountain fault zone (lon ≈ −117.8°W and lat ≈ 36.4°N) as seen in Figure 2(c). There were 10 earthquakes ranging from magnitude 4.2 to 5.19 in this pixel during the temporal window. A cluster of large negative residuals, as seen in dark blue, occurred near the creeping section of the San Andreas fault zone (lon ≈ −120.4°W and lat ≈ 36.1°N). Seismicity was accurately forecasted in the low intensity area in the Eastern most portion of the forecast. Yet as with Models A and B, the majority of the largest Pearson residuals for Model C occurred near the Imperial Fault region.

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

4 Deviances

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

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

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

Figure 3(a) shows the deviances for Model A versus Model B. Model B outperforms Model A in almost all locations where earthquakes were observed. Specifically, Model B appears to perform well along the San Andreas fault (lon ≈ -121.0°W and lat ≈ 36.5°N). Model A only appears to improve upon Model B in the south of the Imperial cluster near the point furthest South in the forecast region (lon ≈ -115.0°W and lat ≈ 32.0°N). Many of the largest deviances in favor or Model B are located in the Imperial cluster. Model B does a much better job in this region while Model A seems to fit better in several selected areas, mostly regions close to known faults. In most locations, however, including the vast majority of locations far from seismicity, Model A offers a better fit, as Model B tends to over-predict events in these locations more than Model A. Overall, the log-likelihood ratio score is -47.45, indicating a significant improvement from Model B compared to Model A.

Note that these results differ substantially from those of Clements et al. (2011), where Model A was observed as generally outperforming Model B.

Deviances for Model A versus Model C are seen in Figure 3(b). Model C appears to forecast the seismicity more accurately in regions where more earthquake clusters were observed. Both Model A and Model C forecast well near the Imperial Fault but Model C appeared to forecast this seismicity more accurately than Model A. In addition, Model C performs much better at the extreme Southern end of the observation window near the Baja, Mexico peninsula (lon $\approx -116.3°$W and lat $\approx 31.8°$N). Overall, Model C offers substantial improvement over Model A with a likelihood ratio score of -63.23.

Deviances for Model B and Model C can be seen in Figure 3(c). Model B forecasts the rate near the Imperial cluster more accurately than Model C, and both models appear to perform well near Mexicali (lon $\approx -115.2°$W and lat $\approx 32.2°$N). White areas indicate regions where both models performed similarly, and we observe many such regions in comparing Models B and C. The area where Model B fits best compared to Model C is near Paso Robles (lon $\approx -120.4°$W and lat $\approx 36.0°$N). Model C fits best compared to Model B near the Bear Mountain Fault (lon $\approx -118.0°$W and lat $\approx 34.0°$N). There are vast regions where Model C slightly outperforms Model B as indicated by the blue regions. Overall Model C fits slightly better than Model B, with a likelihood ratio score of -14.07.

Deviances are useful to compare for forecasts of similar regions and times. An overall score for competing models can be computed using a log-likelihood ratio and allows assessment of the improvement in fit from one model to another. This pixel by pixel comparison enables effortless detection of forecasts in overlapping spatial-temporal windows. While pixel based residual methods such as Pearson residuals and deviances are valuable for gridded forecasts, super-thinned and Voronoi residuals can be especially useful at identifying departures from the data.

5 Super-thinning

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

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

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

Figure 4(a) shows Model A fits well overall despite some clustering in the residuals at very small distances (from $0°$ to $0.1°$) but not much otherwise. Circles indicated observed earthquakes and plus signs indicate simulated points, while lighter points indicate events that occurred earlier in the time and darker points occurred later. There is a significant cluster near the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N) and in the Trinidad fault zone (lon $\approx -124.5°$W and lat $\approx 40.5°$N). However, the centered weighted L-function for the corresponding residuals for Model A, shown in Figure 5(a), reveals that the model performs well is most areas. The L-function is best interpreted along with 95%-confidence bands which are plotted as dash lines. While there is a small amout of inhibition in the residual process but Model A seems to accurately predict the rate of seismicity outside of the interfault zones.

The super-thinned residuals for Model B, shown in Figure 4(b), contain a few significant clusters near the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N), Laguna Salada (lon $\approx -117.1°$W and lat $\approx 32.4°$N), and La Habra (lon $\approx -118.0°$W and lat $\approx 34.0°$N). Indeed, there is significant clustering for Model B up to distances of $0.3°$. As with Model A, the weighted L-function for Model B, shown in Figure 5(b), indicates little inhibition outside of this range. There is also little overprediction as evident by a consistent covering of residual points. This result means enough points were simulated and the residual process is close to what is expected.

As seen in Figure 4(c), there is also significant clustering for Model C up to distances of 0.2°, which occur in similar regions to Models A and B. Clustering for Model C occurs in the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N), Laguna Salada (lon $\approx -117.1°$W and lat $\approx 32.4°$N), and off the coast of Baja California (lon $\approx -117.7°$W and lat $\approx 32.0°$N). Several of the nearly linear patterns in the data also appear in the residuals, such as in the Imperial Cluster, which indicates that the rates in these locations may be misspecified. Indeed, as seen Figure 5(c), at short distances the centered weighted L-function indicates the Model C is under-predicting but the L-functions are within the Poisson bounds for large distances.

5.1 Spatial Temporal Super-thinning

Spatial-temporal super-thinned residuals allow the evaluation of a model's performance over various time domains. The forecast period was split into four equal temporal windows and super-thinned residuals were visualized for Model B to assess the spatial temporal performance of the model. Figure 6 shows these results for Model B. The super-thinned residuals contain clusters near the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N), specifically in temporal window 2 (2 March 2008 to 3 May 2010) and temporal window 3 (3 May 2010 to July 7 2012). Otherwise, there is little over-prediction as evident by a consistent covering of residual points. This result implies enough points were simulated and the residual process in each temporal window is close to what is expected. Figure 7 shows the weighted L-function, corresponding to the four temporal windows. There appears to be significant clustering in temporal window 2 up to distances of 0.3° and in temporal window 3 up to distance of 0.1°. In addition, at distance less than 0.1° in temporal window 4 (2 July 2012 to 2 September 2014), the observed data exhibit greater inhibition than one would expect according to Model B.

To further evaluate each model's temporal performance, the forecast period was split into

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

Spatial-temporal super-thinned residuals are a valuable tool for assessing a model's performance over varying space-time windows. The temporal windows must typically be defined quite arbitrarily, however. For example, the M 7.2 El Mayor earthquake at the end of the second temporal window, as seen in Figure 7(c), and its aftershocks were visible in window 3. Smaller temporal windows can be used, as shown in Figure 8 for instance, but as the size of the temporal window decreases, the number of observed events in each window also decreases resulting in less data available within each window for model evaluation.

6 Voronoi residuals

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

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

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

$$\begin{aligned} \hat{R}_i &:= 1 - \int\limits_{C_i} \hat{\lambda}\, \mathrm{d}\mu \\ &= 1 - |C_i|\bar{\lambda}, \end{aligned} \tag{2}$$

where $\bar{\lambda}$ denotes the mean of the proposed model, $\hat{\lambda}$, over C_i. Following Baddeley et al. (2005), it may be preferable to standardize the residuals, leading e.g. to rescaled Voronoi residuals of the form

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

As with pixel residuals, for each Voronoi residual we can plot the raw residual, scaled residual, or the deviances [Bray et al. 2013].

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

The resulting scaling is trivial computationally, provides useful graphics when applied to earthquake data from Southern California, and the results are very easy to interpret. Figure 9(a) shows the Voronoi cells corresponding to the standardized residuals of the null homogeneous Poisson model, as such the color in each Voronoi cell is defined by the size of the cell. Cells of highly clustered events are colored red and blue cells indicate areas of sparse seismicity. Note that to construct the reference model, all Voronoi cells were truncated by the border of the RELM region, and in computing Voronoi residuals for Models A, B, and

C, the Voronoi cells were constructed spatially and each such cell spans the entire temporal and magnitude window.

Figure 9(b) shows the standardized Voronoi residuals for Model A. Areas towards the edges of the color scale are where Model A performed alarmingly poorly as compared to the null homogeneous Poisson model. Indeed, these are areas where the residual was similar in size to that of the cells of maximum under-prediction or over-prediction of seismicity for the null model. Model A under-predicts seismicity near the Imperial Fault, as indicated by the tight cluster of red cells, and generally tends to over-predict seismicity compared to the reference model. Model A appears to perform well in the Trinidad fault zone (lon $\approx -124.5°$W and lat $\approx 40.5°$N). Model A also substantially outperforms the reference model near Hawthorne, Nevada (lon $\approx -118.70°$W and lat $\approx 38.40°$N).

Figure 9(c) shows Model B similarly under-predicts seismicity in the Imperial Fault region. Model B appears to forecast seismicity accurately near the Campo Indian Reservation (lon $\approx -116.30°$W and lat $\approx 32.40°$N) but generally tends to under-predict seismicity in the Southwestern region of the forecasted area. Model B tends to over-predict seismicity in the Eastern region of the forecast and generally in the Northwestern region. However, within a vast region of over-prediction Model B performs well off the coast of Paso Robles (lon $\approx -121.50°$W and lat $\approx 35.70°$N). The relatively large amount of lighter shades indicates that Model B generally performs well compared to the null model.

Model C (Figure 9(d)) similarly under-predicts in the Imperial Valley, though the vast white shade in this region indicates areas where Model C forecast seismicity accurately compared to the homogeneous Poisson model. For example, Model C appears to perform well near the Channel Islands of California (lon $\approx -119.3°$W and lat $\approx 33.1°$N) and also performs well further South near the Coronado Islands (lon $\approx -117.1°$W and lat $\approx 32.3°$N).

Like Models A and B, Model C appears to under-predict seismicity in the Southwestern area of the forecast region. Alternatively, over-prediction is evident by cluster of large negative residuals located just west of the Sequoia National Forest (lon $\approx -119.5°$W and lat $\approx 36.0°$N) in the Northwestern area of the forecast.

Using the color scaling defined by the null homogeneous Poisson model enables easy comparisons between models. We can see in Figure 9(a) and Figure 9(b) that Model B under-predicts in the eastern portion of the forecast region, whereas Model A tends to over-predict seismicity in this area. In comparing Model A to Model C, one sees that Models A and B both under-predict seismicity near the Channel Islands of California (lon $\approx -119.3°$W and lat $\approx 33.1°$N) whereas Model C appears to forecast seismicity accurately in this region. In addition, Model A tends to over-predict just south (lon $\approx -118.0°$W and lat $\approx 32.0°$N) whereas Models B and C under-predicted seismicity. Compared to Model C, Model B tended to over-predict in areas such as Death Valley (lon $\approx -116.5°$W and lat $\approx 36.3°$N) and near Pomona (lon $\approx -117.8°$W and lat $\approx 34.0°$N). Overall, the Voronoi residuals suggest that Model C forecast seismicity the most accurately of the three models.

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

7 Summary

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

Pixel-based Pearson residuals may be valuable for their ease of interpretation and simplicity of calculation. However, problems can arise due to forecasted conditional intensities of 0 and extreme skew in the standardized residuals rendering them difficult to interpret. Deviances are useful to compare forecasts of similar regions and times. Using a log-likelihood ratio score gives us 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 another.

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

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

It is worth noting that the relative performance of competing models may depend on the diagnostic tool chosen. This is not particular to point processes, and is quite standard in testing and model assessment generally. In the examples shown here, for instance, deviances for Models A and B, shown in Figure 3(a), indicate superior fit of Model A in certain portions of the SouthEastern area of the forecast region. However, Voronoi residuals show that Model A did not perform as well as Model B in certain cells in the same region (Figures 9(b) and 9(c)). A similar result occurs through the comparison of Model B and Model C using deviances and super-thinned residuals. super-thinned residuals for Model B, shown in Figure 4(b), indicate that there is visible clustering near lon $\approx -118.0°$W and lat $\approx 32.0°$N, whereas in Model C, shown in Figure 4(c), that there is less clustering in the same region. Yet when comparing Models B and C using deviances (Figure 3(c)), Model B generally outperforms Model C in this region. These discrepancies can be largely attributable to the different spatial cells used and spatial-temporal scales used in these different residuals.

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

Helmstetter, A. Kagan, Y.Y., and Jackson, D.D. (2007). High-resolution time-independent grid-based forecast M$\geq$5 earthquakes in California. *Seismological Research Letters*, **78**(1), 78-86.

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

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

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

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

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

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

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

Schorlemmer, D., and Gerstenberger, M. C. (2007). RELM testing center. *Seismological Research Letters* **78**, 30-35.

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

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

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

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

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

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. *Bulletin of the Seismological Society of America*, **103**(2A), 787-798.

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.

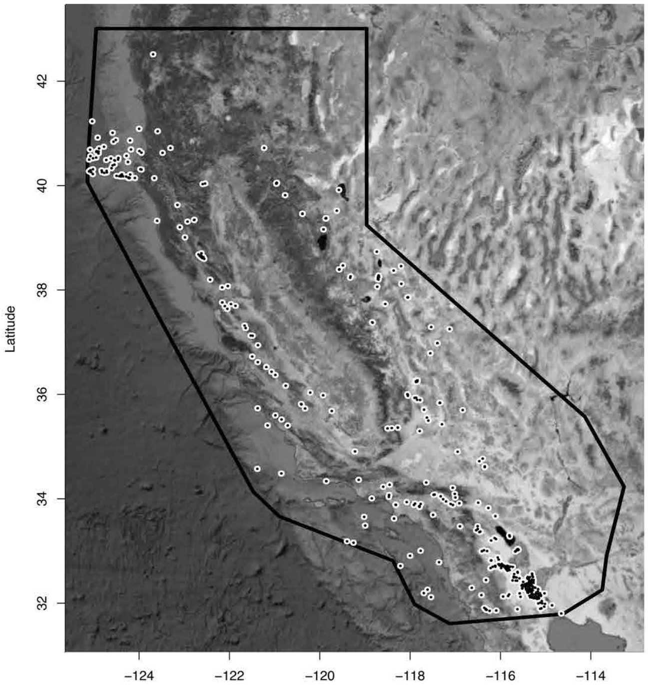
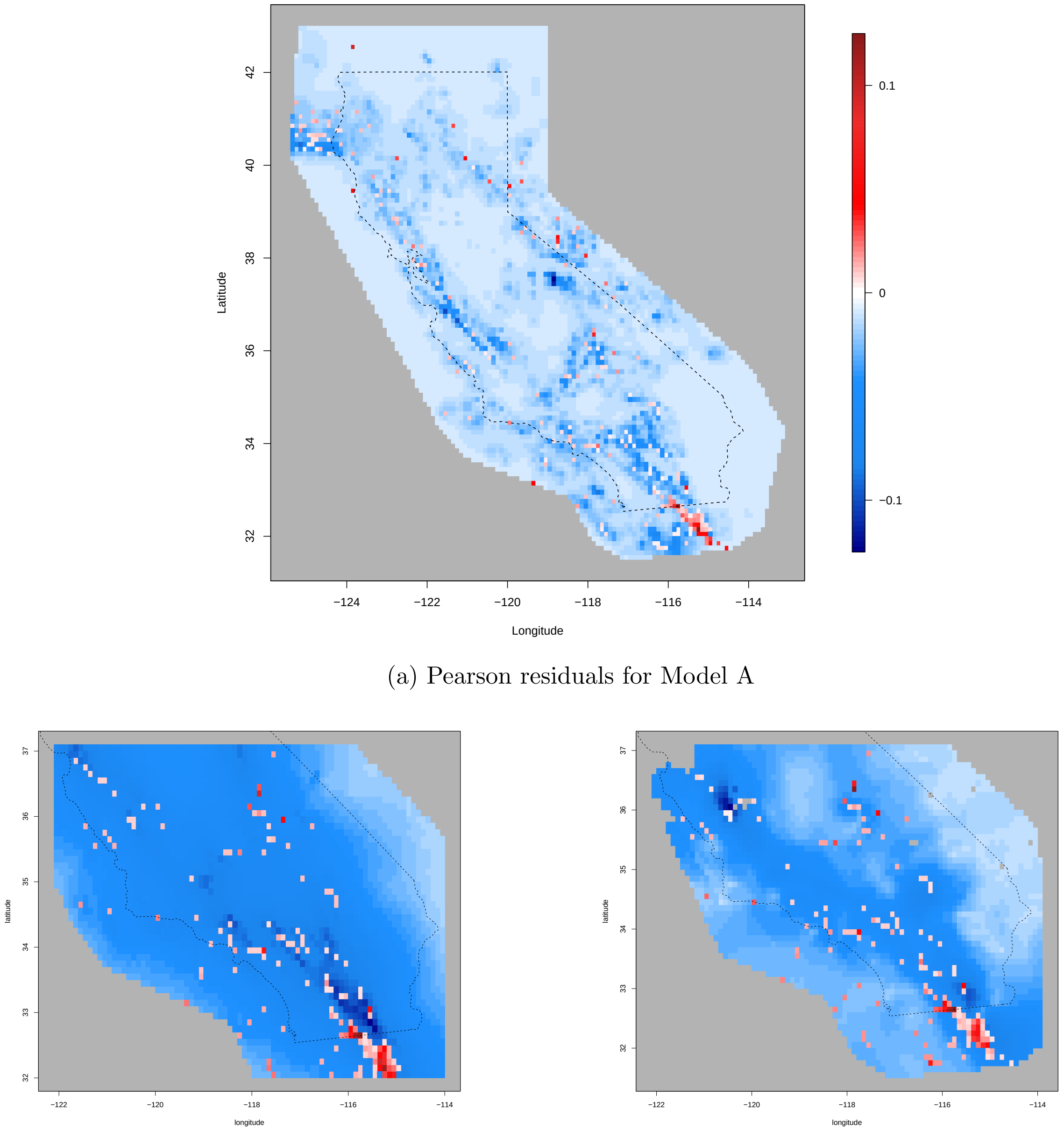


Figure 1: Locations of 510 earthquakes with magnitude $M \geq 3.95$ in the RELM testing region from 1 January 2006 to 2 September 2014.

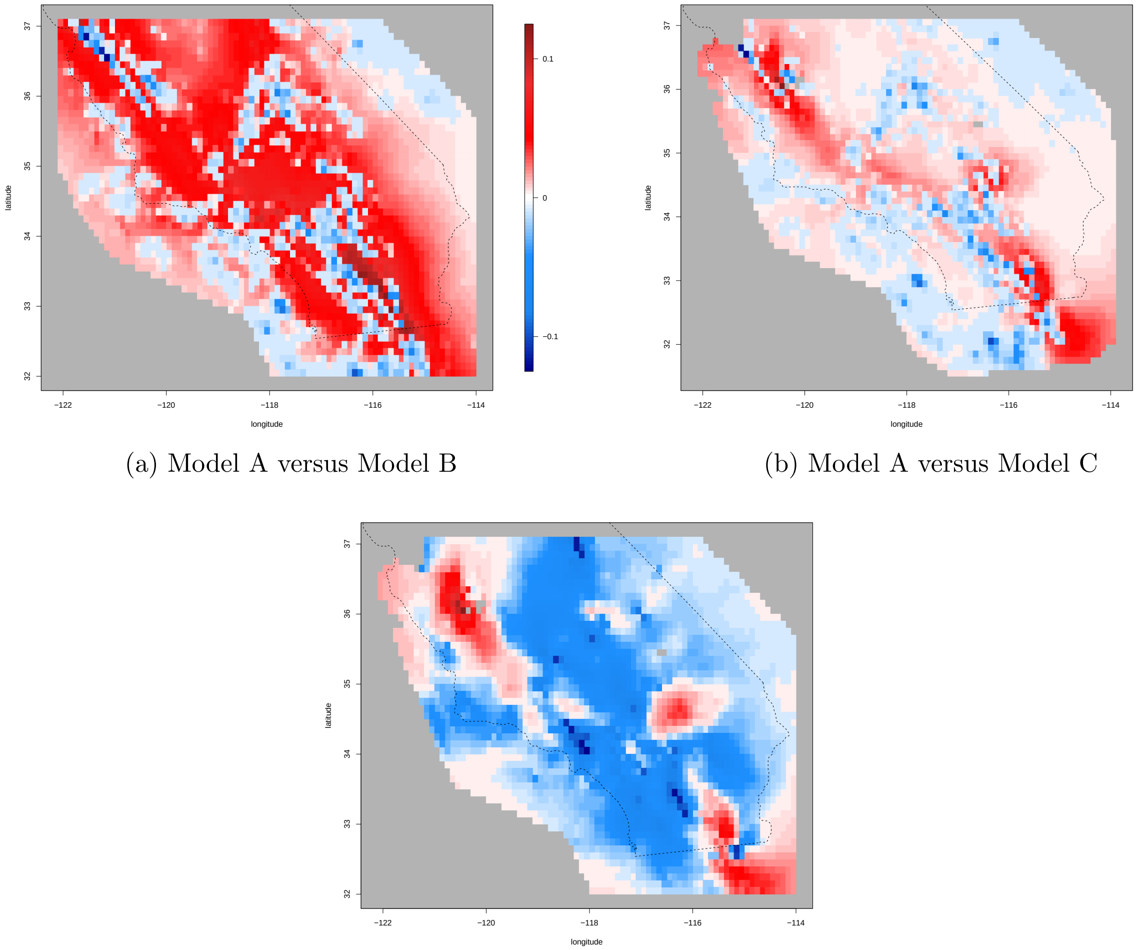


(a) Pearson residuals for Model A

(b) Pearson residuals for Model B

(c) Pearson residuals for Model C

Figure 2: The maximum observed Pearson residual is 4.57 for Model A and is located near Mexicali at at the Mexico-California border (lon ≈ −115.2°W and lat ≈ 32.2°N). The maximum observed Pearson residual is 4.47 for Model B and is located on the pixel that spans the California border with Mexico. The largest residuals for Model C occur in close proximity to the Bare Mountain fault zone (lon ≈ −117.8°W and lat ≈ 36.4°N).

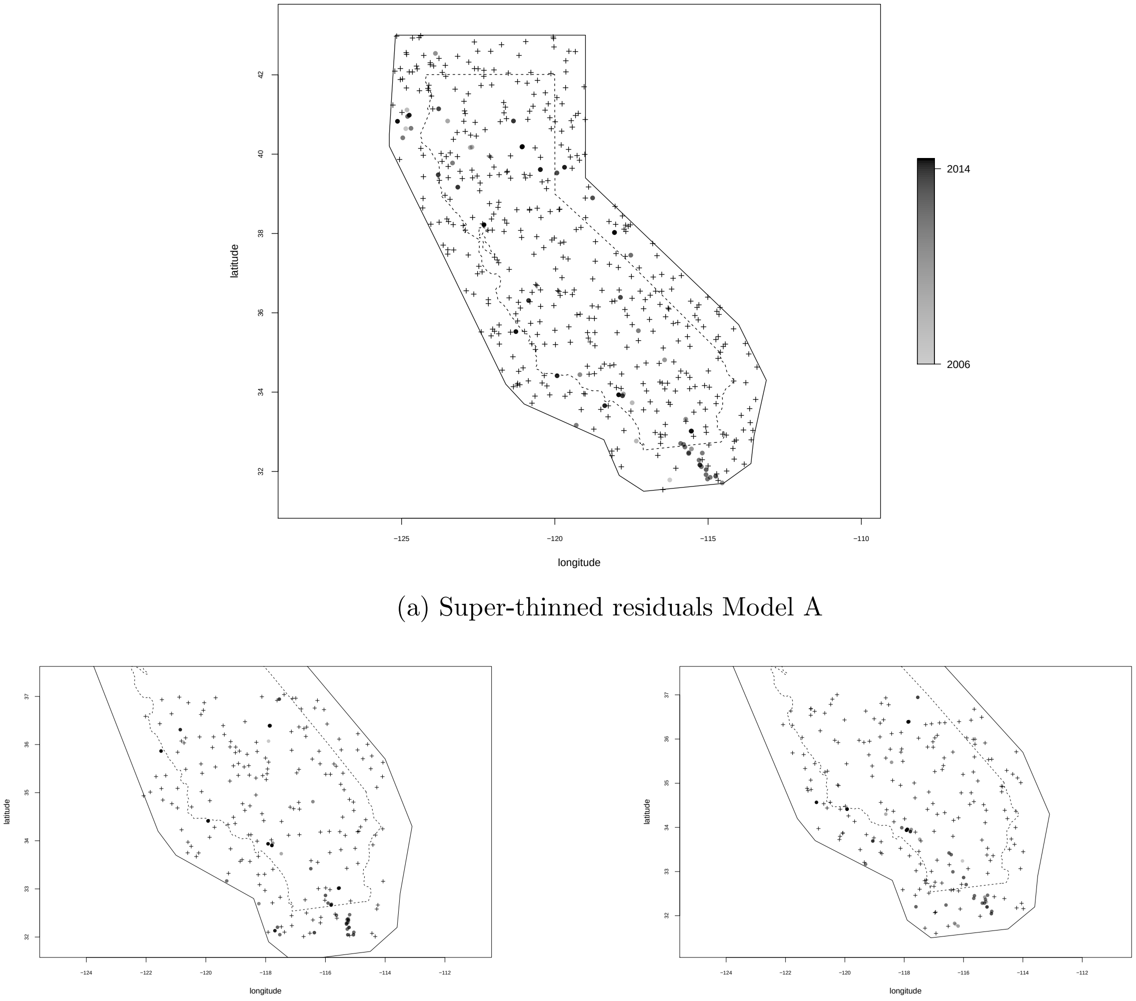


(a) Model A versus Model B

(b) Model A versus Model C

(c) Model B versus Model C

Figure 3: Model B appears to perform well along the San Andreas fault (lon ≈ −121.0°W and lat ≈ 36.5°N) compared to Model A. Both Model A and Model C forecast well near the Imperial Fault but Model C appeared to forecast this seismicity more accurately than Model A. Model B forecasts the rate near the Imperial cluster more accurately than Model C, and both models appear to perform well near Mexicali (lon ≈ −115.2°W and lat ≈ 32.2°N).

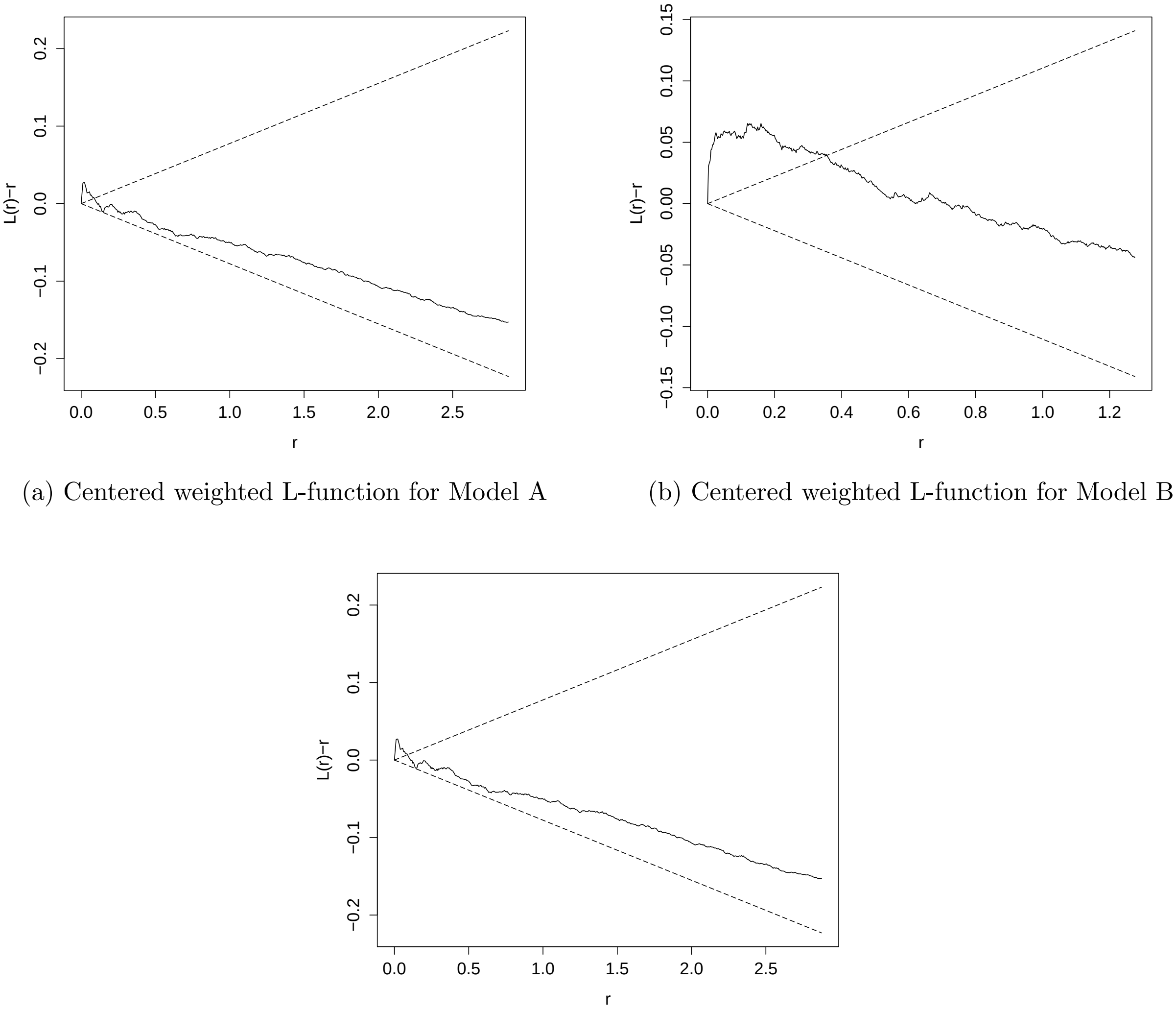


(a) Super-thinned residuals Model A

(b) Super-thinned residuals Model B

(c) Super-thinned residuals Model C

Figure 4: One realization of super-thinned residuals for each Model. Circles indicated observed earthquakes and plus signs indicate simulated points. Lighter points indicate events that occurred earlier in the time, darker points occurred later. There is a significant cluster near the Imperial Fault (lon ≈ −115.2°W and lat ≈ 32.2°N) for Model A. The super-thinned residuals for Model B contain a few significant clusters near the Imperial Fault (lon ≈ −115.2°W and lat ≈ 32.2°N) and Laguna Salada (lon ≈ −117.1°W and lat ≈ 32.4°N). Clustering for Model C occurs in Laguna Salada (lon ≈ −117.1°W and lat ≈ 32.4°N) and off the coast of Baja California (lon ≈ −117.7°W and lat ≈ 32.0°N).

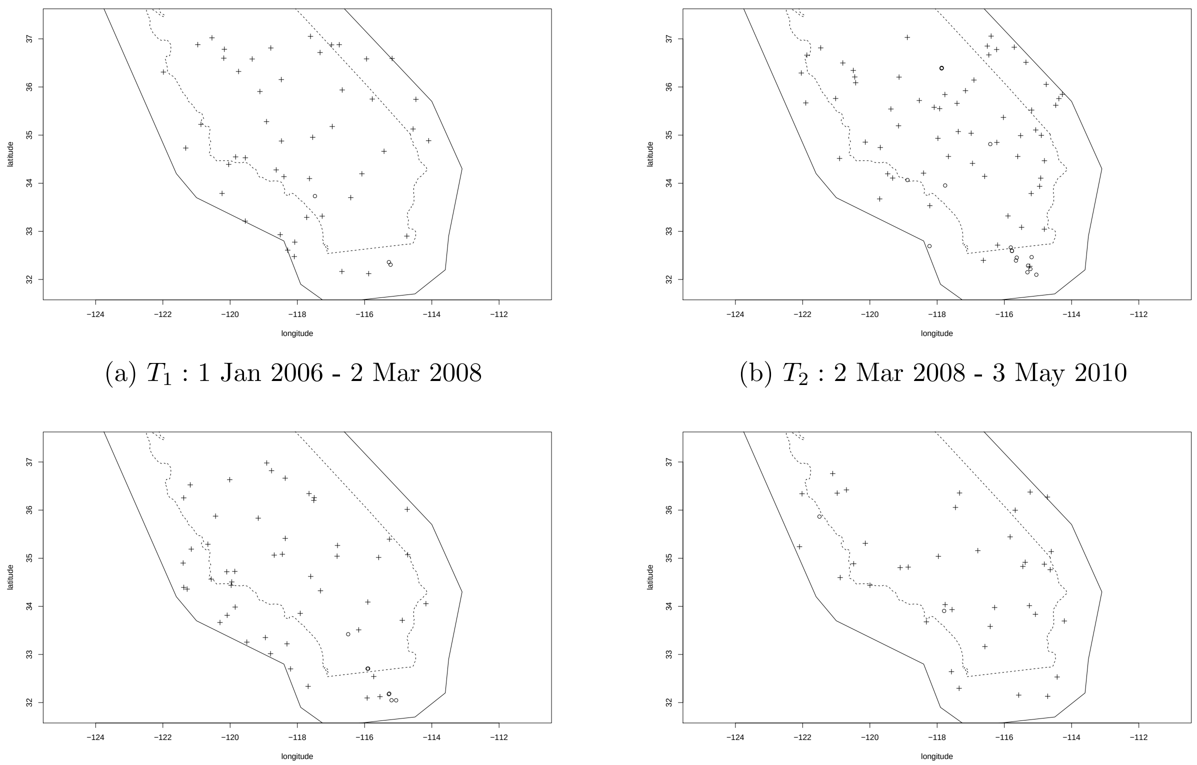


(a) Centered weighted L-function for Model A

(b) Centered weighted L-function for Model B

(c) Centered weighted L-function for Model C

Figure 5: Centered weighted L-function for Model A along with 95%-confidence bands. Model A fits well overall despite some clustering in the residuals at very small distances (from 0° to 0.1°) but not much otherwise. There is significant clustering for Model B up to distances of 0.3°. There is also significant clustering for Model C up to distances of 0.2°, which occurs in similar regions to Models A and B.



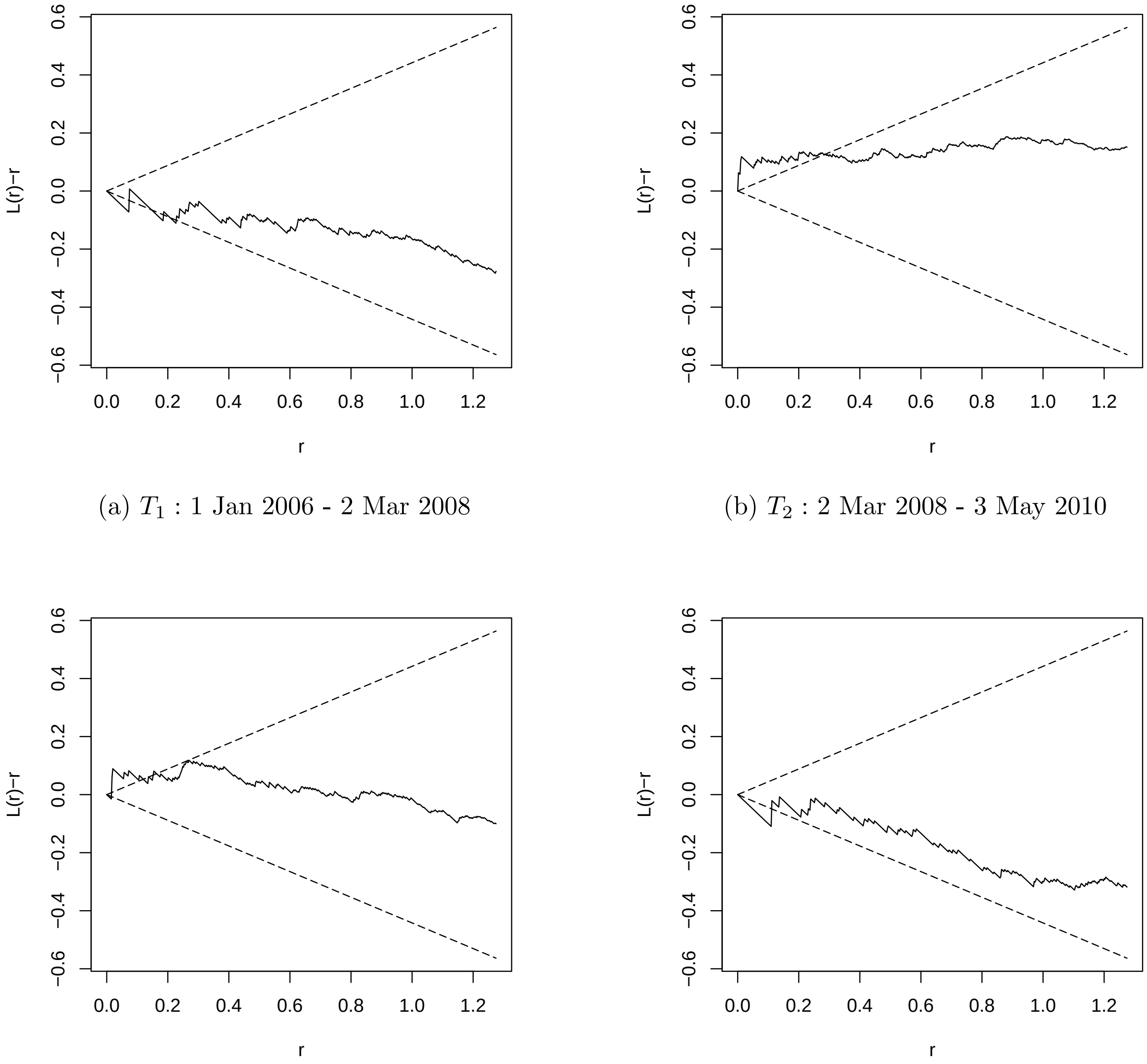
(a) T_1 : 1 Jan 2006 - 2 Mar 2008

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

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

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

Figure 6: One realization of temporal Super-thinned residuals for Model B. Circles indicated observed earthquakes and plus signs indicate simulated points. The super-thinned residuals for Model B contain a few significant clusters near the Imperial Fault (lon $\approx -115.2°$W and lat $\approx 32.2°$N) in temporal window 2 (2 March 2008 to 3 May 2010) and temporal window 3 (3 May 2010 to July 7 2012). Otherwise, the covering of residual points appears to be quite uniform.



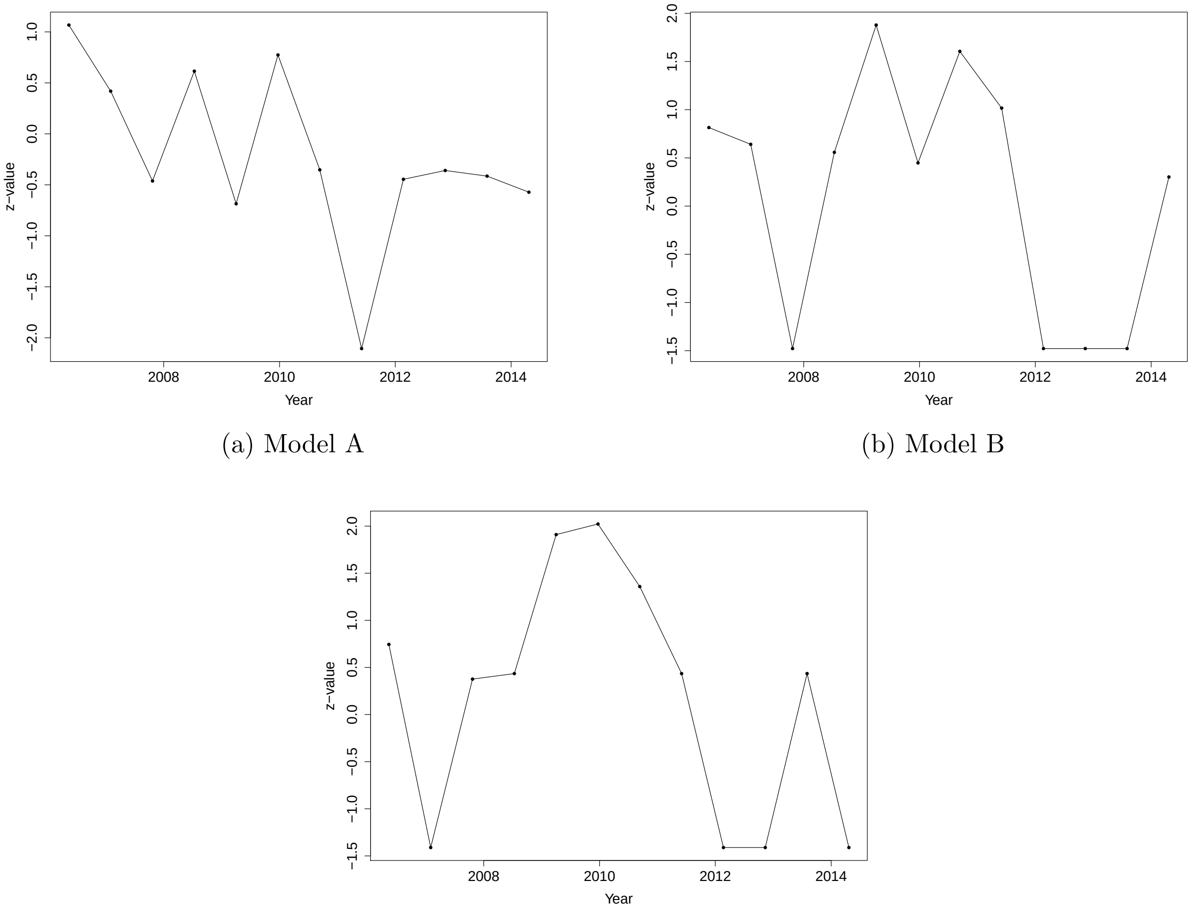
(a) T_1 : 1 Jan 2006 - 2 Mar 2008

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

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

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

Figure 7: Centered weighted L-function for Model B along with 95%-confidence bands. There appears to be significant clustering in temporal window 2 up to distances of 0.3° and in temporal window 3 up to distance of 0.1°. However at distance less than 0.1° in temporal window 4 (2 July 2012 to 2 September 2014), the observed data exhibit greater inhibition than one would expect according to Model B.

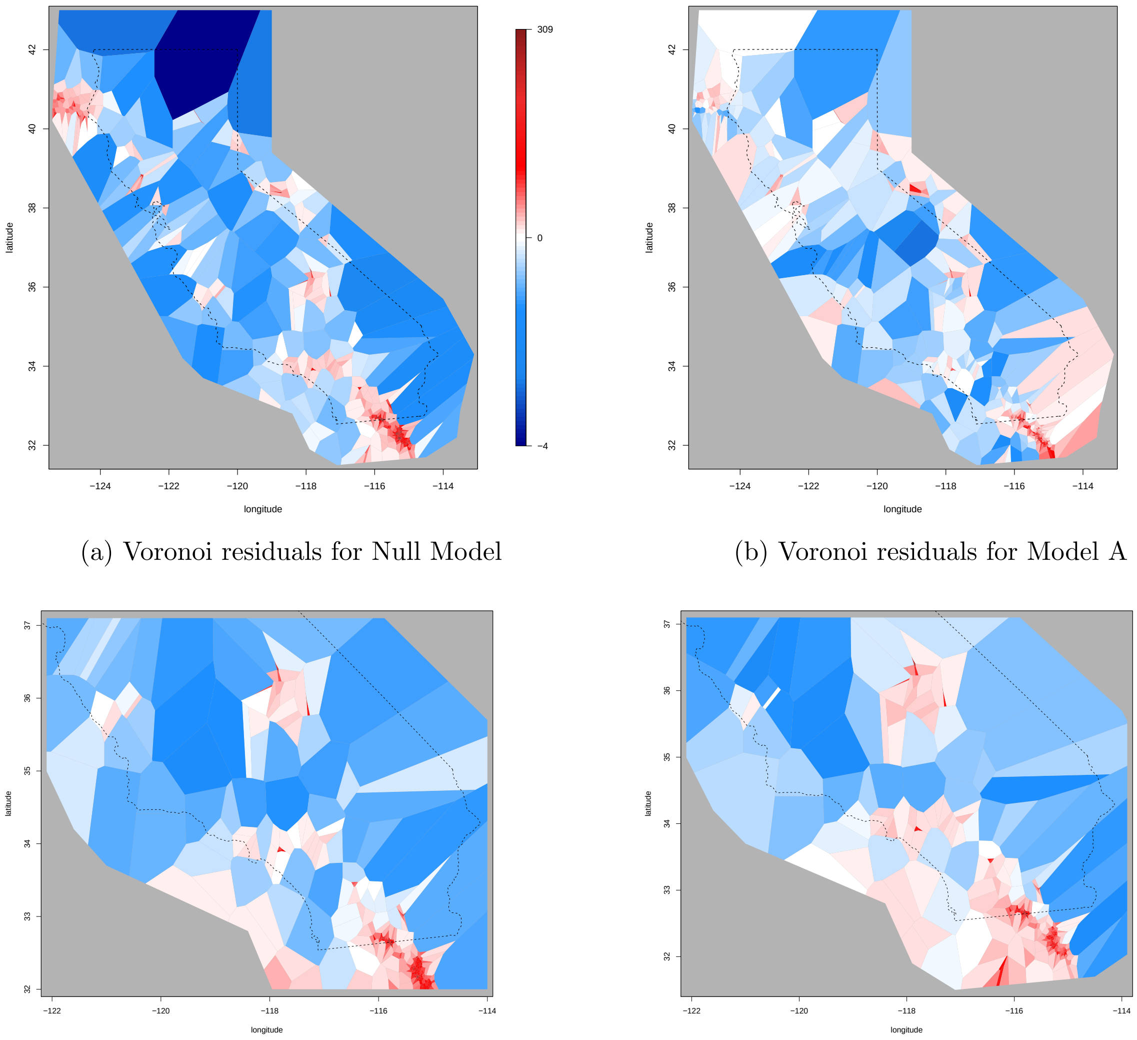


(a) Model A

(b) Model B

(c) Model C

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



(a) Voronoi residuals for Null Model

(b) Voronoi residuals for Model A

(c) Voronoi residuals for Model B

(d) Voronoi residuals for Model C

Figure 9: Voronoi residuals for each Model. When computing the integrated rate over a Voronoi cell, we treat the forecasted rate over each pixel as constant within each pixel. Model A appears to perform well in the Trinidad fault zone (lon ≈ −124.5°W and lat ≈ 40.5°N) while Model B appears to forecast seismicity accurately near the Campo Indian Reservation (lon ≈ −116.30°W and lat ≈ 32.40°N). Model C appears to perform well near the Channel Islands of California (lon ≈ −119.3°W and lat ≈ 33.1°N).