Title
Partition selection for residuals of spatial point process models

Author
Bray, Andrew

Publication Date
2013

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Partition Selection for Residuals for Spatial Point Process Models

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Andrew Paul Bray

2013

Abstract of the Dissertation

Partition Selection for Residuals for Spatial Point Process Models

by

Andrew Paul Bray

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2013

Professor Frederic Paik Schoenberg, Chair

This dissertation investigates the strengths and weaknesses of the current methods of residual analysis for spatial point process models. The primary focus is on the manner in which the space should be partitioned to form residuals. It proposes a new method whereby the differences between the modeled conditional intensity and the observed number of points are assessed over the Voronoi cells generated by the observations. The resulting residuals are substantially less skewed and can be used to construct diagnostic methods of greater statistical power than residuals based on a regular rectangular grid. These advantages are particularly evident for point processes with conditional intensities close to zero, such as those in seismology. Performance is compared using simulated data and applied to models for Southern California earthquakes.

The dissertation of Andrew Paul Bray is approved.

Hongquan Xu

Mark Handcock

Lixia Zhang

Joakim Ekstrom

Frederic Paik Schoenberg, Committee Chair

University of California, Los Angeles

2013

To my family.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The work presented here is the product of collaboration with many statisticians. I benefitted greatly from reading the work of and having discussions with previous graduate students at UCLA: Robert Clements, David Diez, Ka Wong, and Christopher Barr. I am also very grateful for the patience of graduate students from my own generation who have variously acted as sounding boards, devil's advocates, and companions as we muddle though: Neal Fultz, Josh Embree, Katie McLoughlin, James Molyneux, Patrick McCarthy, Hao Ho, Jeroen Ooms, Jiaschen You, Amy Morrow, and Jung In Kim. From the faculty, I have learned much about the general practice of statistics during conversations with Joakim Ekström, Mark Handcock, and Robert Gould. Finally, I thank my advisor Dr. Schoenberg for his optimism, pragmatism, encouragement, patience, dependability and for being a staunch advocate for all of his students.

Parts of this dissertation will be appearing as Bray and Schoenberg (2014) and are in review as Bray et al. (2014). Specific credit must be given to Ka Wong and Christopher Barr, with whom the idea to use the Voronoi tessellation as the partition for a residual plot originated.

Vita

2005	B.S. in History and Earth & Ocean Sciences, Duke University
2006 – 2007	Research Technician, SALMON Project, University of Alaska
Summer 2009	Statistician Intern, Office of Water, US EPA
Spring 2009 –	Teaching Assistant/Associate, UCLA Department of Statistics
Summer 2010	Statistician Intern, Office of Research and Development, US EPA
Summer 2011 –	Graduate Research Assistant, UtopiaCompression Corporation

Publications

Bray A., Schoenberg F.P. (accepted). *Assessing Point Process Models for Earthquake Forecasting.* Statistical Science.

Bray A., Wong K., Barr C., & Schoenberg F.P. (in review). *Using the Voronoi tessellation to calculate residuals for spatial point process models.* Annals of Applied Statistics.

Nichols K., Schoenberg F.P., Keeley J.E., Bray A. & Diez D. (2011). *The application of prototype point processes for the summary and description of California wildfires.* Journal of Time Series Analysis, 32 420-429.

Bray A.P., Mullin P., Schoenberg F.P., Mac Gibbon K., Romero R., Goodwin T.M. & Fejzo M.S. (2011). *Prenatal Exposure to Hyperemesis Gravidarum Linked to Increased Risk of Psychological/Behavioral Disorders in Adulthood.* Journal of Developmental Origins of Health and Disease, 2:4 200-204.

Khan H.D., Rosati J.A. & Bray A.P. (2011). *Statistical evaluation of data from multi-laboratory testing of a measurement method intended to indicate the presence of dust resulting from the collapse of the world trade center.* Environmental Monitoring and Assessment.

CHAPTER 1

Introduction

1.1 Motivation

The creation of powerful modeling approaches for spatial point processes requires an equally strong effort to develop tools to assess how well the models describe observations from those processes. In the past decade, a framework for spatial point process model assessment has emerged that mirrors that which was developed for linear regression models. A milestone in this development was the publication in 2005 of *Residual analysis for spatial point processes* (Baddeley et al., 2005). In this paper the authors define several residual processes and suggest corresponding diagnostic tools to apply to models of real data. All of the proposed residuals are a function of B: a given bounded closed subset of the bounded region, W, of the plane. Despite its centrality, little guidance is given regarding the selection of B. In the discussion that accompanies the paper, Lawson (2005) points this out:

> The choice of B is a major concern. ... The residuals that are derived will be highly dependent on this choice (i.e. quadrat size, shape, and location) as will the smoothed residual field.

This concern serves as the primary motivation for the research described herein. This dissertation investigates the effect that the choice of partition has on the resulting residual analysis and proposes a new form of partition based on the Voronoi tessellation.

1.2 Background

Residual plots for spatial point processes have two related purposes. The first is to suggest locations or aspects of the model where the fit is poor, so that an incorrectly specified model may be improved. The second is to form the basis of formal testing, i.e. to assess the overall appropriateness of a model to determine if results based on the model may be trusted.

Residual analysis for a spatial point process is typically performed by partitioning the space on which the process is observed into a regular grid, and computing a residual for each pixel. That is, one typically examines aggregated values of a residual process over regular, rectangular grid cells (see, e.g., Clements et al., 2011; Baddeley et al., 2005, 2008). The general form of these aggregated residual measures is a standardized difference between the number of points occurring and the number expected according to the fitted model, where the standardization may be performed in various ways. For instance, for Pearson residuals, one weights the residual by the reciprocal of the square root of the intensity, in analogy with Pearson residuals in the context of linear models.

Baddeley et al. (2005) proposed smoothing the residual field using a kernel function instead of simply aggregating over pixels; in practice this residual field is typically displayed over a rectangular grid, and is essentially equivalent to a kernel smoothing of aggregated pixel residuals. Baddeley et al. (2005) also proposed scaling the residuals based on the contribution of each pixel to the total pseudo-loglikelihood of the model, in analogy with score statistics in generalized linear modeling. Standardization is important for both residual plots and goodness of fit tests, since otherwise plots of the residuals will tend to overemphasize deviations in pixels where the rate is high. Behind the term *Pearson residuals* lies the implication that these standardized residuals should be approximately standard normally distributed, so that the squared residuals, or their sum, are distributed

approximately according to Pearson's χ^2-distribution.

The development of these residual analysis techniques over the past decade, and the fact that such residuals extend so readily to the case of spatial-temporal point processes, may suggest that the problem of residual analysis for such point processes is generally solved. In practice, however, residuals that are examined over a fixed rectangular grid tend to have two characteristics that can limit their effectiveness.

I. When the integrated conditional intensity (i.e. the number of expected points) in a pixel is very small, the distribution of the residual for the pixel becomes heavily skewed.

II. Positive and negative values of the residual process within a particular cell can cancel each other out.

Since Pearson residuals can be standardized to have mean zero and unit (or approximately unit) variance under the null hypothesis that the modeled conditional intensity is correct, one may inquire whether the skew of these residuals is indeed problematic. Consider for instance the case of a planar Poisson process where the estimate of the intensity λ is exactly correct, i.e. $\hat{\lambda}(x, y) = \lambda(x, y)$ at all locations, and where one elects to use Pearson residuals on pixels. Suppose that there are several pixels where the integral of λ over the pixel is roughly 0.01. Given many of these pixels, it is not unlikely that at least one of them will contain a point of the process. In such pixels, the raw residual will be 0.99, and the standard deviation of the number of points in the pixel is $\sqrt{0.01} = 0.1$, so the Pearson residual is 9.90.

This may yield the following effects: 1) such Pearson residuals may overwhelm the others in a visual inspection, rendering a plot of the Pearson residuals largely useless in terms of evaluating the quality of the fit of the model, and 2) conventional tests based on the normal approximation may have grossly incorrect

p-values, and will commonly reject the null model even when it is correct. Even if one adjusts for the non-normality of the residual and instead uses exact p-values based on the Poisson distribution, such a test applied to any such pixel containing a point will still reject the model at the significance level of 0.01.

These situations arise in many applications, unfortunately. For example, in modeling earthquake occurrences, typically the modeled conditional intensity is close to zero far way from known faults or previous seismicity, and in the case of modeling wildfires, one may have a modeled conditional intensity close to zero in areas far from human use or frequent lightning, or with vegetation types that do not readily support much wildfire activity (see eg. Johnson and Miyanishi, 2001; Malamud et al., 2005; Keeley et al., 2009; Xu and Schoenberg, 2011).

These challenges are a result of having a small integrated intensity (characteristic I above), and one straightforward solution would be to enlarge the pixel size such that the expected count in each cell is higher. While this would be effective in a homogeneous setting, in the case of an inhomogeneous process it is likely that this would induce a different problem: cells that are so large that even gross misspecification within a cell may be overlooked, and thus the residuals will have low power. This is the problem of characteristic II. When a regular rectangular grid is used to compute residuals for a highly inhomogenous process, it is generally impossible to avoid either highly skewed residual distributions or residuals with very low power.

These problems have been noted by previous authors, and it is the goal of this manuscript to address them using a new form of residual diagram based on the Voronoi cells generated by tessellating the observed point pattern. The resulting partition obviates I and II above by being adaptive to the inhomogeneity of the process and generating residuals that have an average expected count of 1 under the null hypothesis.

1.3 Notation

The notation used thoughout this dissertation is meant to draws together that used by Stoyan (2005), Baddeley (2007), and Cressie and Wikle (2011).

A *spatial point process* is a stochastic process that generates a countable set of points $\{\mathbf{x}_n\}$ in $\mathbb{R}^2$, satisfying the condition that each bounded subset of $\mathbb{R}^2$ contains a finite number of points. Of interest is often the counting process, $N(B)$, the number of points falling within the Lebesgue-measurable bounded closed set B, which takes values in the non-negative integers. Processes are considered to be *simple* thoughout, i. e. $\mathbf{x}_i \neq \mathbf{x}_j$ if $i \neq j$. A realization from a spatial point process is a *point pattern*, an unordered set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ where each point is a two-element vector. This more compact notation will be used interchagably with the longer form, which considers functions and points of the coordinate pair (x, y).

The mean function of N is represented by the *intensity function*.

$$\lambda(\mathbf{x}) = \lim_{|\mathrm{d}\mathbf{x}| \to 0} \frac{E(N(\mathrm{d}\mathbf{x}))}{|\mathrm{d}\mathbf{x}|}$$

where $\mathrm{d}\mathbf{x}$ is a small region located at $\mathbf{x}$ with volume $|\mathrm{d}\mathbf{x}|$. For the simplest case of a homogeneous Poisson process, the intensity is constant across W, but for earthquake models it is natural to let the intensity vary spatially.

Spatial point processes with more complex spatial interactions (i. e. no Poissonian inter-point independence) have a mean function that is characterized by the Papangelou intensity,

$$\lambda(\mathbf{x}) = \lim_{|\mathrm{d}\mathbf{x}| \to 0} \frac{E(N(\mathrm{d}\mathbf{x} | \sigma\{N(W \setminus \mathbf{x})\}))}{|\mathrm{d}\mathbf{x}|}$$

which specifies the expected infinitesimal rate at point $\mathbf{x}$ conditional on all information in W excluding $\mathbf{x}$. For spatial-temporal processes, the analog is the *conditional intensity*, which conditions on the prior history of events up to time t.

In this dissertation, the distinction between conditional and Papangelou intensities will not be emphasized, as the methods and results here are essentially equivalent for spatial and spatial-temporal point processes.

CHAPTER 2

Scientific Context: Seismology

The need for an alternative method to construct a residual plot arose from the difficulties of using traditional quadrat methods on seismological models, which tend to be strongly inhomogenous. In this section, we introduce the classes of models commonly used in seismology as well as the framework in which model building and testing occurs. In later sections, these models will be used to compare the existing methods of residual analysis with the proposed Voronoi method.

A major goal in seismology is the ability to accurately anticipate future earthquakes before they occur (Bolt, 2003). Anticipating major earthquakes is especially important, not only for short-term response such as preparation of emergency personnel and disaster relief, but also for longer-term preparation in the form of building codes, urban planning, and earthquake insurance (Jordan and Jones, 2010). In seismology, the phrase *earthquake prediction* has a specific definition: it is the identification of a meaningfully small geographic region and time window in which a major earthquake will occur with very high probability. An example of earthquake predictions are those generated by the M8 method (Keilis-Borok and Kossobokow, 1990), which issues an alarm whenever there is a suitably large increase in the background seismicity of a region. Such alarms could potentially be very valuable for short-term disaster preparedness, but unfortunately examples of M8-type alarms, including the notable Reverse Tracing of Precursors (RTP) algorithm, have generally exhibited low reliability when tested prospectively, typically failing to outperform naive methods based simply on smoothed

historical seismicity (Geller et al., 1997; Zechar, 2008).

Earthquake prediction can be contrasted with the related *earthquake forecasting*, which means the assignment of probabilities of earthquakes occurring in broader space-time-magnitude regions. The temporal scale of an earthquake forecast is more on par with climate forecasts and may be over intervals that range from decades to centuries (Hough, 2010).

Many models have been proposed for forecasting earthquakes, and since different models often result in very different forecasts, the question of how to assess which models seem most consistent with observed seismicity becomes increasingly important. Concerns with retrospective analyses, especially regarding data selection, overfitting, and lack of reproducibility, have motivated seismologists recently to focus on prospective assessments of forecasting models. This has led to the development of the Regional Earthquake Likelihood Models (RELM) and Collaborative Study of Earthquake Predictability (CSEP) testing centers, which are designed to evaluate and compare the goodness-of-fit of various earthquake forecasting models.

2.1 A framework for prospective testing

The current paradigm for building and testing earthquake models emerged from the working group for the development of Regional Earthquake Likelihood Models (RELM) in 2001. As described in Field (2007), the participants were encouraged to submit differing models, in the hopes that the competition between models would prove more useful than trying to build a single consensus model. The competition took place within the framework of a prospective test of their seismicity forecasts. Working from a standardized data set of historical seismicity, scientists fit their models and submitted to RELM a forecast of the number of events expected within each of many pre-specified spatial-temporal-magnitude bins. The

first predictive experiment required models to forecast seismicity in California between 2006 to 2011 using only data from before 2006.

This paradigm has many benefits from a statistical perspective. The prospective nature of the experiments effectively eliminates concerns about overfitting. Furthermore, the standardized nature of the data and forecasts facilitates the comparison among different models.

RELM has since expanded into the Collaborative Study of Earthquake Predictability (CSEP), a global-scale project to coordinate model development and conduct prospective testing according to community standards (Jordan, 2006). CSEP serves as an independent entity that provides standardized seismicity data, inventories proposed models, and publishes the standards by which the models will be assessed. This is a particularly compelling framework in which to develop methods of residual analysis because of their essential role in the iterative cycle from model generation, data collection, residual analysis, and subsequent model generation.

2.2 Some examples of models for earthquake occurrences

The first predictive experiment coordinated through RELM considered time-independent spatial point process models, which can be specified by their Papangelou intensity $\lambda(\mathbf{x})$, a function of spatial location $\mathbf{x}$. A representative example is the model specified by Helmstetter et al. (2007) that is based on smoothing previous seismicity. The intensity function is estimated with an isotropic adaptive kernel

$$\lambda(\mathbf{x}) = \sum_{i=1}^{N} K_d\left(\mathbf{x} - \mathbf{x}_i\right)$$

where N is the total number of observed points, and K_d is a power-law kernel

$$K_d(\mathbf{x} - \mathbf{x}_i) = \frac{C(d)}{(|\mathbf{x} - \mathbf{x}_i|^2 + d^2)^{1.5}}$$

where d is the smoothing distance, $C(d)$ is a normalizing factor so that the integral of $K_d()$ over an infinite area equals 1, and $|\cdot|$ is the Euclidean norm. The estimated number of points within the pre-specified grid cells is obtained by integrating $\lambda(\mathbf{x})$ over each cell.

Models of earthquake occurrence that consider it to be a time-dependent process are commonly variants of the epidemic-type aftershock sequence (ETAS) model of Ogata (1988) (see eg. Helmstetter and Sornette, 2003; Ogata et al., 2003; Sornette, 2005; Vere-Jones and Zhuang, 2008; Console et al., 2010; Chu et al., 2011; Wang et al., 2011; Tiampo and Shcherbakov, 2012). According to the ETAS model, earthquakes cause aftershocks, which in turn cause more aftershocks, and so on. ETAS is a point process model specified by its conditional intensity, $\lambda(\mathbf{x}, t)$, which represents the infinitesimal expected rate at which events are expected to occur around time t and location $\mathbf{x}$, given the history H_t of the process up to time t. ETAS is a special case of the linear, self-exciting Hawkes' point process (Hawkes, 1971), where the conditional intensity is of the form

$$\lambda(\mathbf{x}, t|H_t) = \mu(\mathbf{x}, t) + \sum_{t_i<t} g(\mathbf{x} - \mathbf{x}_i, t - t_i; M_i),$$

where $\mu(\mathbf{x}, t)$ is the mean rate of a Poisson-distributed background process that may in general vary with time and space. g is a *triggering function* that indicates how previous occurrences contribute, depending on their spatial and temporal distances and marks, to the conditional intensity λ at the location and time of interest. $(\mathbf{x}_i, t_i, M_i)$ are the origin times, epicentral locations, and moment magnitudes of observed earthquakes.

Ogata (1998) proposed various forms for the triggering function, g, such as the following

$$g(\mathbf{x}, t, M) = K(t + c)^{-p} e^{a(M - M_0)} (|\mathbf{x}|^2 + d)^{-q},$$

where M_0 is the lower magnitude cutoff for the observed catalog.

The parameters in ETAS models and other spatial-temporal point process models may be estimated by maximizing the log-likelihood,

$$\sum_{i=1}^{n} \log\{\lambda(\mathbf{x}_i, t_i)\} - \int_W \int \lambda(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t.$$

The maximum likelihood estimator (MLE) of a point process is, under quite general conditions, asymptotically unbiased, consistent, asymptotically normal, and asymptotically efficient (Ogata, 1978). Finding the parameter vector that maximizes the log-likelihood can be achieved using any of the various standard optimization routines, such as the quasi-Newton methods implemented in the function `optim()` in *R*. The spatial background rate μ in the ETAS model can be estimated in various ways, such as via kernel smoothing seismicity from prior to the observation window or kernel smoothing the largest events in the catalog, as in Ogata (1998) or Schoenberg (2003). Note that the integral term in the loglikelihood function can be cumbersome to estimate, and an approximation method recommended in Schoenberg (2013) can be used to accelerate computation of the MLE.

There are many other earthquake forecasting models quite distinct from the two point process models above. Perhaps most important among these are the Uniform California Earthquake Rupture Forecast (UCERF) models, which are consulted when setting insurance rates and crafting building codes (Field et al., 2009). They are constructed by soliciting *expert opinion* from leading seismologists on which components should enter the model, how they should be weighted, and how they should interact (Marzocchi and Zechar, 2011). Examples of the components include slip rate, geodetic strain rates, and paleoseismic data.

Although the UCERF models draw upon diverse information related to the geophysics of earthquake etiology, commonly used models such as ETAS and its variants rely solely on previous seismicity for forecasting future events. Many attempts have been made to include covariates, but when assessed rigorously, most predictors other than the locations and times of previous earthquakes have been shown not to offer any noticeable improvement in forecasting. Recent examples of such covariates include electromagnetic signals (Jackson, 1996; Kagan, 1997), radon (Hauksson and Goddard, 1981), and water levels (Bakun et al., 2005; Manga and Wang, 2007). A promising exception is moment tensor information, which is now routinely recorded with each earthquake and seems to give potentially useful information regarding the directionality of the release of stress in each earthquake. However, this information appears not to be explicitly used presently in models in the CSEP or RELM forecasts.

CHAPTER 3

Current Methods

This section reviews the suite of model assessment and diagnostic tools that can be used to analyze seismological models, but are useful for spatial point process models more generally. The focus of this dissertation is on residual methods in particular, but each method can play a role in the course of a thorough model assessment.

3.1 Numerical tests

Several numerical tests were initially proposed to serve as the metrics by which RELM models would be evaluated (Schorlemmer and Gerstenberger, 2007). For these numerical tests, each model consists of the estimated number of earthquakes in each of the spatial-temporal-magnitude bins, where the number of events in each bin is assumed to follow a Poisson distribution with an intensity parameter equivalent to the forecasted rate.

The L-test (or Likelihood test) evaluates the probability of the observed data under the proposed model. The numbers of observed earthquakes in each spatial-temporal-magnitude bin are treated as independent random variables, so the joint probability is calculated simply as the product of their corresponding Poisson probabilities. This observed joint probability is then considered with respect to the distribution of joint probabilities generated by simulating many synthetic data sets from the model. If the observed probability is unusually low in the context

of this distribution, the data are considered inconsistent with the model.

The N-test (Number) ignores the spatial and magnitude component and focuses on the total number of earthquakes summed across all bins. If the proposed model provides estimates $\hat{\lambda}_i$ for i corresponding to each of B bins, then according to this model, the total number of observed earthquakes should be Poisson distributed with mean $(\sum_{i=1}^{B} \hat{\lambda}_i)$. If the number of observed earthquakes is unusually large or small relative to this distribution, the data are considered inconsistent with the model.

The L-test is considered more comprehensive in that it evaluates the forecast in terms of magnitude, spatial location, and number of events, while the N-test restricts its attention to the number of events. Two additional data consistency tests were proposed to assess the magnitude and spatial components of the forecasts, respectively: the M-test and the S-test (Zechar et al., 2010). The M-test (Magnitude) isolates the forecasted magnitude distribution by counting the observed number of events in each magnitude bin without regard to their temporal or spatial locations, standardized so that the observed and expected total number of events under the model agree, and computing the joint (Poisson) likelihood of the observed numbers of events in each magnitude bin. As with the L-test, the distribution of this statistic under the forecast is generated via simulation.

The S-test (Spatial) follows the same inferential procedure but isolates the forecasted spatial distribution by summing the numbers of observed events over all times and over all magnitude ranges. These counts within each of the spatial bins are again standardized so that the observed and expected total number of events under the model agree, and then one computes the joint (Poisson) likelihood of the observed numbers of events in the spatial bins.

The above tests measure the degree to which the observations agree with a particular model, in terms of the probability of these observations under the given model. As noted in Zechar et al. (2013), tests such as the L-test and N-test are

really tests of the consistency between the data and a particular model, and are not ideal for comparing two models.

Schorlemmer and Gerstenberger (2007) proposed an additional test to allow for the direct comparison of the performance of two models: the Ratio test (R-test). For a comparison of models A and B, and given the numbers of observed events in each bin, the test statistic R is defined as the log-likelihood of the data according to model A minus the corresponding log-likelihood for model B. Under the null hypothesis that model A is correct, the distribution of the test statistic is constructed by simulating from model A and calculating R for each realization. The resulting test is one-sided and is supplemented with the corresponding test using model B as the null hypothesis.

The T-test and W-test of Rhoades et al. (2011) are very similar to the R-test, except that instead of using simulations to find the null distribution of the difference between log-likelihoods, with the T-test and W-test, the differences between log-likelihoods within each space-time-magnitude bin for models A and B are treated as independent normal or symmetric random variables, respectively, and a t-test or Wilcoxon signed rank test, respectively, is performed.

3.2 Limitations of numerical tests especially for comparing models

Unfortunately, when used to compare various models, such likelihood-based tests suffer from the problem of variable null hypotheses and can lead to highly misleading and even seemingly contradictory results. For instance, suppose model A has a higher likelihood than model B. It is nevertheless quite possible for model A to be rejected according to the L-test and model B not to be rejected using the L-test. Similarly, the R-test with model A as the null might indicate that model A performs statistically significantly better than model B, while the R-test with

model B as the null hypothesis may indicate that the difference in likelihoods is not statistically significant.

The explanation for such results is that the null hypotheses of the two tests are different: when model A is tested using the L-test, the null hypothesis is model A, and when model B is tested, the null hypothesis is model B. The test statistic may have very different distributions under these different hypotheses.

Unfortunately, these types of discrepancies seem to occur frequently, and hence the results of these numerical tests may not only be uninformative for model comparison, but in fact highly misleading. A striking example is given in Figure 4 of Zechar et al. (2013), where the Shen et al. (2007) model produces the highest likelihood of the five models considered in this portion of the analysis, and yet under the L-test has the lowest corresponding p-value of the five models.

3.3 Functional summaries

Functional summaries, i.e. those producing a function of one variable, such as the weighted K-function and error diagrams, can also be useful measures of goodness-of-fit. However, such summaries typically provide little more information than numerical tests in terms of indicating where and when the model and the data fail to agree, or how a model may be improved.

The weighted K-function is a generalized version of the K-function of Ripley (1976), which has been widely used to detect clustering or inhibition for spatial point processes. The ordinary K function, $K(h)$, counts, for each h, the total number of observed pairs of points within distance h of one another, per observed point, standardized by dividing by the estimated overall mean rate of the process, and the result is compared to what would be expected for a homogeneous Poisson process. The weighted version, $K_w(h)$ was introduced for the inhomogeneous spatial point process case by Baddeley and Turner (2000), and is defined similarly

to $K(h)$, except that each pair of points $(\mathbf{x}_i, \mathbf{x}_j)$ is weighted by $1/[\hat{\lambda}(\mathbf{x}_i)\hat{\lambda}(\mathbf{x}_j)]$, the inverse of the product of the modeled unconditional intensities at the points $\mathbf{x}_i$ and $\mathbf{x}_j$. This was extended to spatial-temporal point processes by Veen and Schoenberg (2005) and Adelfio and Schoenberg (1997).

Whereas the null hypothesis for the ordinary K-function is a homogeneous Poisson process, in the case of K_w, the weighting allows one to assess whether the degree of clustering or inhibition in the observations is consistent with what would be expected under the null hypothesis corresponding to the model for $\hat{\lambda}$. While weighted K-functions may be useful for indicating whether the degree of clustering in the model agrees with that in the observations, such summaries unfortunately do not appear to be useful for comparisons between multiple competing models, nor do they accurately indicate in which spatial-temporal-magnitude regions there may be particular inconsistencies between a model and the observations.

Error diagrams, which are also sometimes called receiver operating characteristic (ROC) curves (Swets, 1973) or Molchan diagrams (Molchan, 2010; Kagan, 2009), plot the (normalized) number of alarms versus the (normalized) number of false negatives (failures to predict), for each possible alarm. In the case of earthquake forecasting models an *alarm* is defined as any value of the modeled conditional rate, $\hat{\lambda}$, exceeding some threshold.

Figure 3.1 presents error diagrams for two RELM models, Helmstetter et al. (2007) and Shen et al. (2007). The ease of interpretation of such diagrams is an attractive feature, and plotting error diagrams with multiple models on the same plot can be a useful way to compare the models' overall forecasting efficacy. In figure 3.1 we learn that Shen et al. (2007) slightly outperforms Helmstetter et al. (2007) when the threshold for the alarm is high, but as the threshold is lowered Helmstetter et al. (2007) performs noticeably better. For the purpose of comparing models, one may even consider normalizing the error diagram so that the false negative rates are considered relative to one of the given models

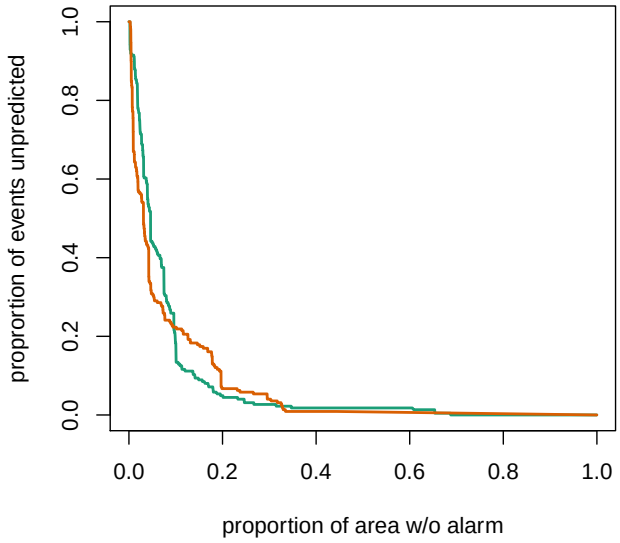

Figure 3.1: Error diagrams for Helmstetter et al. (2007) in blue and Shen et al. (2007) in orange. Model details are in chapter 2 and chapter 6 respectively.

in consideration as in Kagan (2009). This tends to alleviate a common problem with error diagrams as applied to earthquake forecasts, which is that most of the relevant focus is typically very near the axes and thus it can be difficult to inspect differences between the models graphically. A more fundamental problem with error diagrams, however, is that while they can be useful overall summaries of goodness-of-fit, such diagrams unfortunately provide little information as to where models are fitting poorly or how they may be improved.

3.4 Residual methods

Residual analysis methods for spatial-temporal point process models produce graphical displays which may highlight where one model outperforms another or where a particular model does not ideally agree with the data. Some residual methods, such as thinning, rescaling, and superposition, involve transforming the

point process using a model for the conditional intensity λ and then inspecting the uniformity of the result, thus reducing the difficult problem of evaluating the agreement between a possibly complex spatial-temporal point process model and data to the simpler matter of assessing the homogeneity of the residual point process. Often, departures from homogeneity in the residual process can be inspected by eye, and many standard tests are also available. Other residual methods, such as pixel residuals, Voronoi residuals, and deviance residuals, result in graphical displays that can quite directly indicate locations where a model appears to depart from the observations, or where one model appears to outperform another in terms of agreement with the data.

3.4.1 Thinned, superposed, and superthinned residuals

Thinned residuals were first introduced by Lewis and Shelder (1979) for the purpose of simulating spatial-temporal point processes, and were extended for the purpose of model evaluation in Schoenberg (2003). The method involves keeping each observed point independently with probability $b/\hat{\lambda}(\mathbf{x}_i, t_i)$, where $b = \inf_{(\mathbf{x},t)\in\mathcal{W}}\{\hat{\lambda}(\mathbf{x},t)\}$, and $\hat{\lambda}$ is the modeled conditional intensity. If the model is correct, i.e. if the estimate $\hat{\lambda}(\mathbf{x},t) = \lambda(\mathbf{x},t)$ almost everywhere, then the residual process will be homogeneous Poisson with rate b (Schoenberg, 2003). Because the thinning is random, there is no unique thinned pattern to assess. Nonetheless, one may inspect several realizations of thinned residuals and analyze the entire collection to get an overall assessment of goodness-of-fit.

The dual approach was proposed by Brémaud (1981), who suggested superposing a simulated point process onto an observed point process realization so as to yield a homogeneous Poisson process. As indicated in Clements et al. (2012), tests based on thinned or superposed residuals tend to have low power when the model $\hat{\lambda}$ for the conditional intensity is volatile, which is typically the case with earthquake forecasts since earthquakes tend to be clustered in particular spatial-

temporal regions. Thinning a point process will lead to very few points remaining if the infimum of $\hat{\lambda}$ over the observed space is small (Schoenberg, 2003), while in superposition, the simulated points, which are by construction approximately homogeneous, will form the vast majority of residual points if the supremum of $\hat{\lambda}$ is large.

A hybrid approach called super-thinning was introduced in Clements et al. (2012). With super-thinning, a tuning parameter k is chosen, and one thins (deletes) the observed points in locations of space-time where $\hat{\lambda} > k$, keeping each point independently with probability $k/\hat{\lambda}(\mathbf{x}, t)$, and superposes a Poisson process with rate $\hat{\lambda}(\mathbf{x}, t)/k$ where $\hat{\lambda} < k$. When the tuning parameter k is chosen wisely, the method appears to be more powerful than thinning or superposing in isolation.

3.4.2 Rescaled residuals

An alternative method for residual analysis is rescaling. The idea behind rescaled residuals dates back to Meyer (1971) who investigated rescaling temporal point processes according to their conditional intensities, moving each point t_i to a new time $\int_0^{t_i} \hat{\lambda}(t)\,\mathrm{d}t$, creating a transformed space in which the rescaled points are homogeneous Poisson of unit rate under the null hypothesis. Heuristically, the space is essentially compressed when $\hat{\lambda}$ is small and stretched when $\hat{\lambda}$ is large, so that the points are ultimately uniformly distributed in the resulting transformed space. This method was used in Ogata (1988) to assess a temporal ETAS model and extended in Merzbach and Nualart (1986); Nair (1990); Schoenberg (1999); Vere-Jones and Schoenberg (2004) to the spatial and spatial-temporal cases. Rescaling may result in a transformed space that is difficult to inspect if $\hat{\lambda}$ varies widely over the observation region, and in such cases standard tests of homogeneity such as Ripley's K-function may be dominated by boundary effects, as illustrated in Schoenberg (2003).

3.4.3 Pixel residuals

A different type of residual analysis that is more closely analogous to standard residual methods in regression or spatial statistics is to consider the (often standardized) differences between the observed and expected numbers of points in each of various spatial or spatial-temporal pixels or grids, producing what might be called *pixel residuals*. More precisely, the *raw* pixel residual on each pixel grid cell G_i is defined as $N(G_i) - \int \hat{\lambda}(\mathbf{x}, t)\, \mathrm{d}t\, \mathrm{d}\mathbf{x}$, where $N(G_i)$ is simply the number of points (earthquakes) observed in pixel G_i (Baddeley et al., 2005). Baddeley et al. (2005) also proposed various standardizations including Pearson residuals, which are scaled in relation to the standard deviation of the raw residuals: $r_i = \frac{N(G_i) - \int \hat{\lambda}(\mathbf{x},t)\, \mathrm{d}t\, \mathrm{d}\mathbf{x}}{\sqrt{\int \hat{\lambda}(\mathbf{x},t)\, \mathrm{d}t\, \mathrm{d}\mathbf{x}}}$. Although they define these residual processes for a general B, in order to visualize them they are presumably calculated on a regular grid and then smoothed.

The primary difficulty of using a regular pixel parition is that if the pixels are too large, then the method is not powerful to detect local inconsistencies between the model and data, and places in the interior of a pixel where the model overestimates seismicity may cancel out with places where the model underestimates seismicity. On the other hand, if the pixels are small, then the majority of the raw residuals are close zero while those few that correspond to pixels with an earthquake are close to one. These are the problematic characteristics of a fixed grid that are outlined in § 1.2.

In situations where the residuals have a highly skewed distribution, the skew is only intensified by the standardization to Pearson residuals. As a result, plots of the both the raw and the Pearson residuals are not informative, and merely highlight the pixels where earthquakes occur regardless of the fit of the model. The raw or Pearson residuals may be smoothed, as in Baddeley et al. (2005), but such smoothing typically only reveals gross, large-scale inconsistencies between the

model and data and has the rephrased question of how to specify the bandwidth.

If one is primarily interested in comparing competing models, then instead one may plot, in each pixel, the difference between log-likelihoods for the two models, as in Clements et al. (2011). The resulting residuals may be called *deviance residuals*, in analogy with residuals from logistic regression and other generalized linear models. Deviance residuals appear to be useful for comparing models on grid cells and inspecting where one model appears to fit the observed earthquakes better than the other. It remains unclear how these residuals may be used or extended to enable comparisons of more than two competing models, other than by comparing two at a time.

CHAPTER 4

Voronoi Residuals

A partition of W based on the Voronoi tessellation is introduced in order to address the problems caused by characterstics I and II in § 1.2: the skew in the residual distribution caused by pixels with a small integrated intensity and the oversmoothing that occurs in inhomogeneous processes when the pixels are large.

4.1 Voronoi residuals

A Voronoi tessellation 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 spatial or spatial-temporal point pattern $\{\mathbf{x}_n\}$, one may define the cells of its corresponding *Voronoi tessellation*, $\{C_n\}$ as follows.

$$C_i = \{\mathbf{x} \in W : ||\mathbf{x} - \mathbf{x}_i|| \leq ||\mathbf{x} - \mathbf{x}_j||, \forall\, j \neq i\}$$

That is, for each point $\mathbf{x}_i$ of the point process, its corresponding cell C_i is the region consisting of all locations which are closer to $\mathbf{x}_i$ than to any other point of $\{\mathbf{x}_n\}$. The Voronoi tessellation is the collection of such cells. See Okabe et al. (2000) for a thorough treatment of Voronoi tessellations and their properties.

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

are referred to as *Voronoi residuals*.

An immediate advantage of Voronoi residuals compared to conventional pixel-based methods is that the partition is entirely automatic and spatially adaptive. This leads to residuals with a distribution that tends to be far less skewed than pixel-based methods. Indeed, since each Voronoi cell has exactly one point inside it by construction, the raw Voronoi residual for cell i is given by

$$\begin{aligned} \hat{r}_i &:= 1 - \int_{C_i} \hat{\lambda}(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &= 1 - |C_i|\bar{\lambda}, \end{aligned} \tag{4.1}$$

where $\bar{\lambda}$ denotes the mean of the proposed model, $\hat{\lambda}$, over C_i. This raw residual can be scaled in various ways, but it will be left unscaled because the transformation is unnecessary in the method described below.

The distribution of r can approximated using known computation results. For a homogeneous Poisson process, the expected area of a Voronoi cell is equal to the reciprocal of the intensity of the process (Meijering, 1953), and simulation studies have shown that the area of a typical Voronoi cell is approximately gamma distributed (Tanemura, 2003). These properties continue to hold approximately in the inhomogeneous case provided that the conditional intensity is approximately constant near the location in question (Barr and Schoenberg, 2010).

The raw Voronoi residual in (4.1) will therefore tend to be distributed approximately like a modified gamma random variable. More specifically, the second term, $|C_i|\bar{\lambda}$, referred to in the stochastic geometry literature as the reduced area, is well-approximated by a two parameter gamma distribution with a rate of 3.569 and a shape of 3.569 (Tanemura, 2003). The distribution of the raw residuals is therefore approximated by

$$r \sim 1 - X\,; \quad X \sim \text{gamma}(3.569, 3.569). \tag{4.2}$$

By contrast, for pixels over which the integrated conditional intensity is close to zero, the conventional raw residuals are approximately Bernoulli distributed.

The exact distributions of the Voronoi residuals are generally quite intractable due to the fact that the cells themselves are random, but approximations can be made by simulation. Consider the point process defined by the intensity function $\lambda(x, y) = 200x^2|y|$ on the subset $W = [-1, 1] \times [-1, 1]$. Figure 4.1 presents a realization of the process along with the corresponding Voronoi tessellation (top panels) and a regular rectangular pixel grid (bottom panels). Two locations in W were selected for investigation: one with relatively high intensity, the other with relatively low intensity. Residual distributions were simulated by generating 5000 point patterns from the model, identifying the pixel/cell occupied by the location of interest, then calculating the difference between the number of observed points and the number expected in that cell under the same model.

The distribution of Voronoi residuals under the null hypothesis is well approximated by distribution 4.2 at both the high intensity and low intensity locations (figure 4.1). By comparison, the distribution of pixel residuals is that of a Poisson distributed variable with intensity $\int_{G_i} \hat{\lambda}(\mathbf{x})\,\mathrm{d}\mathbf{x}$, for pixel G_i, centered to have mean zero. At the location where the intensity is high, this distribution is moderately skewed, but for the low intensity location the distribution becomes extremely skewed to the point of being effectively a two-valued random variable.

Although distribution 4.2 is a good approximation to the distribution of the Voronoi residuals at a given location across many realizations, this is not the same thing as the distribution of interest: that of the full collection $\{|C_n|\}$ under the null hypothesis. Barr and Diez (2012) showed that this distribution is in fact a mixture of gamma random variables. Nonetheless, experience thus far indicates that 4.2

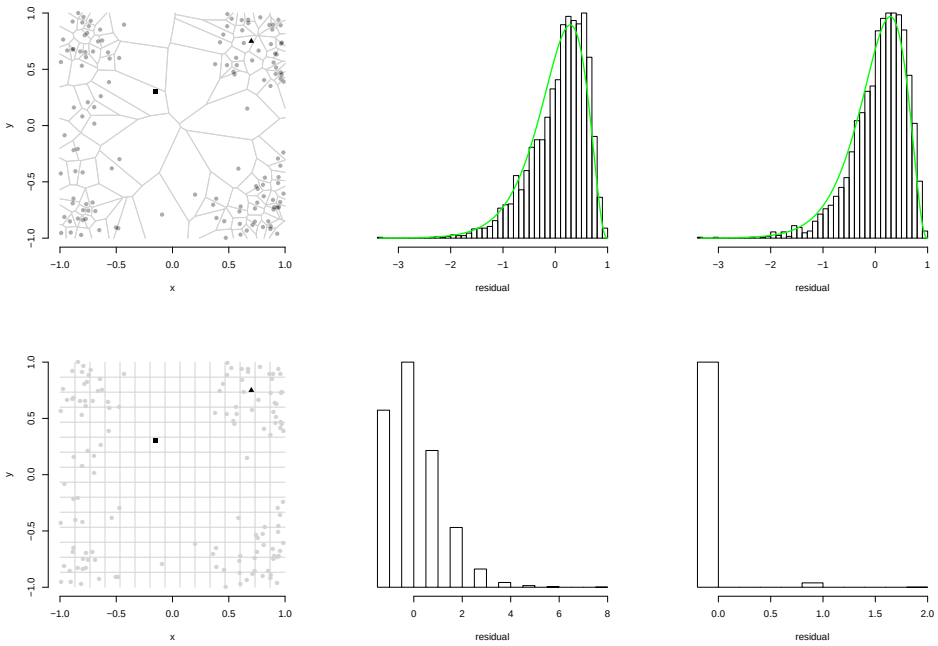

Figure 4.1: Residual distributions under the null hypothesis based on a Voronoi tessellation (top panels) and a pixellated grid (bottom panels). The underlying point process is Poisson with intensity $\lambda(x, y) = 200x^2|y|$. The middle panels show results at location ▲ where $\lambda = 73.5$; the right panels show results at location ■ where $\lambda = 1.35$. The distribution (2) is overlaid in green for the top-middle and top-right plots.

is also a reasonable approximation to this distribution (see e.g. the lower left plot in Figure 3). For a fully rigorous model assessment, one could supplement the analysis by computing the distribution of residuals though repeated simluations and tessellations from the model.

It is also important to note that such residuals are not strictly independent of one another due to the nature of the tessellation. Caution should thus be exercised if methods are used that rely upon asymptotic distributions. Another concern would be a tendency for this method to under- or overstate the degree

of spatial structure in the residuals, as the dependence will be spatially local. Again, however, experience shows that both of these concerns are relatively minor. A second Voronoi residual is introduced in § 7.2 to provide an independent alternative.

4.2 Voronoi Residual Plots

In this section, the construction of the Voronoi residual plot is described and its utility demonstrated using a series of simulations of spatial Poisson processes. The simulations are random samples from a specified *generating model*, $\lambda(\mathbf{x})$. These simulations are then modeled, correctly or incorrectly, by a *proposed model*, $\lambda_0(\mathbf{x})$. Voronoi residuals are then computed and used to assess the degree to which the proposed model agrees with simulations.

4.2.1 Plot construction

The Voronoi residual plot is constructed by first tessellating W based on the observed point pattern $\{\mathbf{x}_n\}$. This was done using the *deldir* package in the R computing environment (Turner, 2011). In order to calculate each r_i, it is necessary to calculate $\bar{\lambda}_0$ (equation 4.1). This was approximated using Monte Carlo integration: randomly generating many points within C_i, evaluating $\lambda(\mathbf{x}_i)$ for each, then taking the mean.

More important than the value of the raw residual is the value that it takes relative to distribution 4.2. Therefore each raw residual is transformed into the value that it takes under the cumulative distribution function of distribution 4.2. A value very close to zero or 1 indicates that the residual is far into the tails of the residual distribution, and therefore would represent marked over- and under-prediction by $\lambda_0(\mathbf{x})$, respectively.

These values and their spatial configuration are visualized using a color map.

Each cell is shown in either red or blue, indicating over- or under-prediction, and then the saturation of the color reflects the magnitude of that residual. Therefore, a region of W where $\lambda_0(\mathbf{x})$ systematically overpredicted will be visualized as a patch of dark red cells, while a residual plot where the true model was fit should appear washed out and spatially unstructured.

Initially, the residual plot used a linear map between residual c.d.f. value and color saturation. These plots had the tendency to overwhelm the viewer because bright colors were assigned to cells with small and moderate residual magnitudes. The goal of the plot is to draw attention to systematic and highly improbable residual values under $\lambda_0(\mathbf{x})$, so this linear map was replaced with an inverse Normal transformation (figure 4.2). Thus only unusually low residuals are colored bright red and unusually high residuals are colored bright blue, while values not in the tails of distribution 4.2 are assigned much more muted colors. Additionally, this transformation can be tuned by the user to reveal more or less constrast between the body and the tails of the distribution by changing the standard deviation of the transformation.

4.2.2 Correctly specified model

We first consider the simplest case, where the proposed model is the same as the generating model, $\lambda_0(\mathbf{x}) = \lambda(\mathbf{x})$. As a result, one expects residuals that are spatially unstructured and relatively small in magnitude; i.e. the only variation should be due to sampling variability.

Figure 4.3 shows a simulation of a spatial Poisson process with intensity $\lambda(x, y) = 200x^2|y|$ on the subset $W = [-1, 1] \times [-1, 1]$, along with its corresponding Voronoi tessellation.

The plot in the upper left panel is a point pattern from the generating model along with the resulting Voronoi tessellation. In the Voronoi residual plot in the

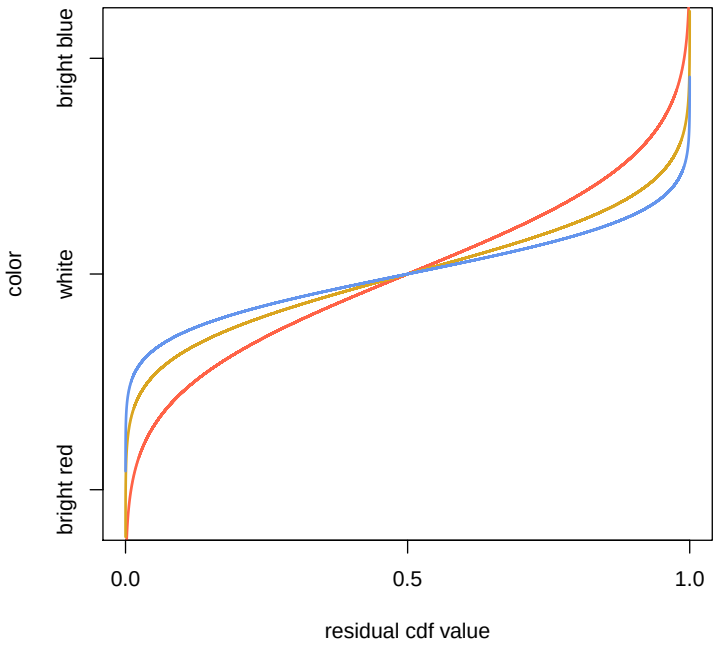

Figure 4.2: The inverse Normal color transformation of residual c.d.f. values at three tuning levels: $\sigma = .75$ (blue), $\sigma = 1$ (gold), $\sigma = 1.5$ (red).

top right panel of figure 4.3, the tiles range from light to moderate hues of red and blue, representing residuals that are within the range expected under the reference distribution. In the lower left panel, a histogram of the raw residuals is shown with a green line representing the reference distribution. The plot in the lower right panel is an alternative way to compare these distribution using a quantile-quantile plot along with 95% point-wise confidence bounds. As would be expected, the histogram and quantile plot of the Voronoi residuals demonstrate that the distribution of the residuals is well approximated by distribution (4.2).

4.2.3 Misspecification

In order to evaluate the ability of Voronoi residuals to detect model misspecification, simulations were obtained using a generating model and then residuals were computed based on a different proposed model. The top left panel of Figure 4.4 displays a realization of a Poisson process with intensity $\lambda(x, y) = 100\mathbb{1}_{\{|X|,|Y|>0.35\}}$.

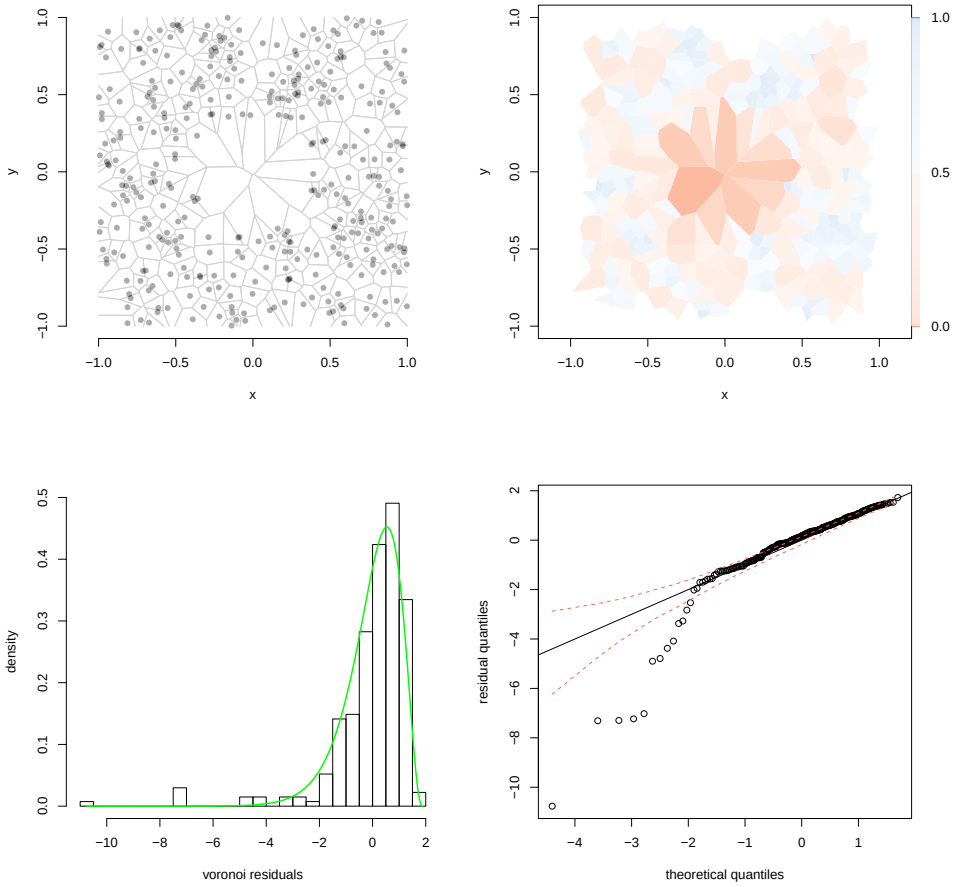

Figure 4.4: Simulated Poisson process with intensity $\lambda(x, y) = 100\mathbb{1}_{\{|x|,|y|>0.35\}}$ with Voronoi tessellation overlaid (top left), Voronoi residual plot for this simulation using a proposed intensity of $\lambda_0(x, y) = 100$ (top right), histogram of the Voronoi residuals, with a green curve tracing the density of the reference distribution (4.2) (bottom left), quantile plot of the Voronoi residuals with respect to distribution (4.2), with pointwise 95% confidence limits obtained via simulation of the proposed model (bottom right). The color scale of the Voronoi residual plot is $\Phi^{-1}\{F(r)\}$, where F is the distribution function of (4.2). Tiles intersecting the boundary of the space are ignored.

The proposed model assumes a constant intensity across the space, $\lambda_0(x, y) = 100$. Because of the lack of points near the origin, the tiles near the origin are larger than expected under the proposed model, and hence for such a cell C near the origin, the integral $\int_C \lambda_0(x, y)\,\mathrm{d}x\,\mathrm{d}y$ exceeds 1, leading to negative residuals of large absolute value. These unusually large negative residuals are evident in the Voronoi residual plot and clearly highlight the region where the proposed model overpredicts the intensity of the process. These residuals are also clear outliers in the left tail of the reference distribution of the residuals, and as a result one sees deviations from the identity line in the quantile-quantile plot in figure 4.4.

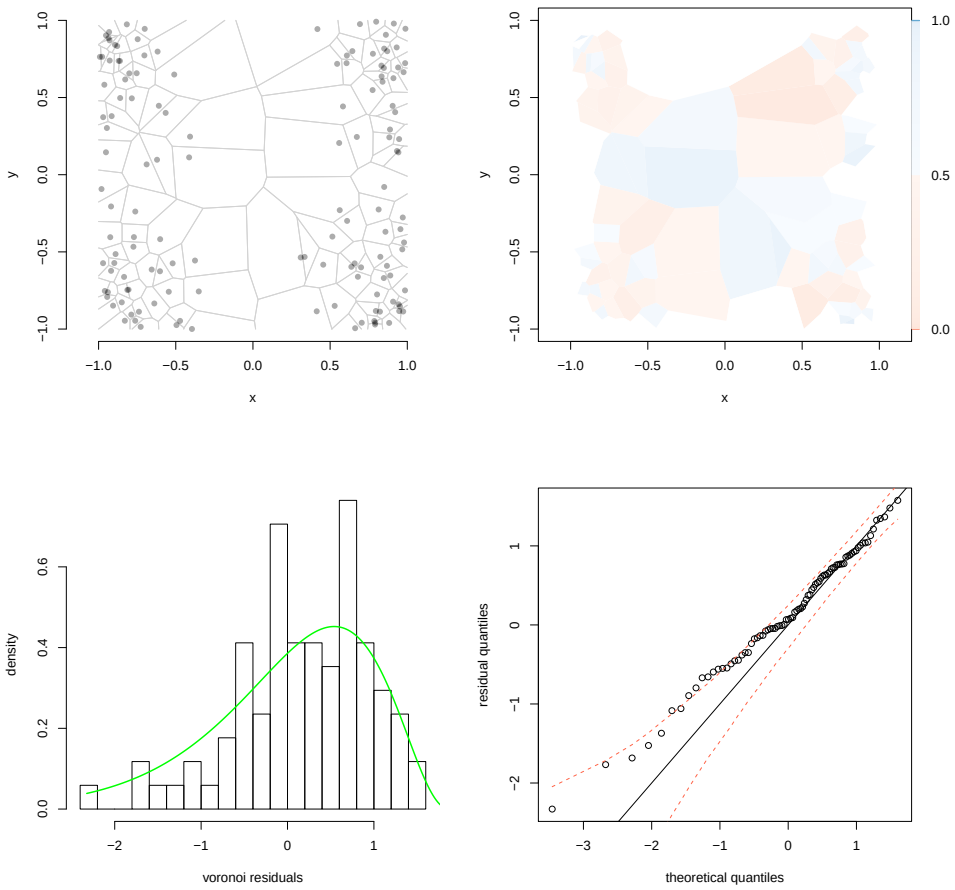

Figure 4.3: Simulated Poisson process with intensity $\lambda(x, y) = 200x^2|y|$ with Voronoi tessellation overlaid (top left), Voronoi residual plot of this simulation (top right), histogram of the Voronoi residuals, with a green curve tracing the density of the reference distribution (4.2) (bottom left), quantile plot of the Voronoi residuals with respect to the distribution (4.2), with pointwise 95% confidence limits obtained via simulation (bottom right). The color scale of the Voronoi residual plot in the top right is $\Phi^{-1}\{F(r)\}$, where F is the distribution function of (4.2). Tiles intersecting the boundary of the space are ignored, as the distribution of these tile areas may differ substantially from the gamma distribution.

CHAPTER 5

Statistical Power

We now consider the manner in which the statistical power of residual analysis using a Voronoi partition differs from that of a pixel partition. In the context of a residual plot, a procedure with low power would generate what appears to be a structureless residual plot even when the model is misspecified. To allow for an unambiguous comparison, here we focus on power in the formal testing setting: the probability that a misspecified model will be rejected at a given confidence level.

5.1 Probability Integral Transform

As was discussed in § 4.1, the distribution of Voronoi residuals under the null hypothesis is well approximated by a modified gamma distribution, while the distribution of pixel residuals is that of a Poisson distributed variable with intensity $\int_{G_i} \hat{\lambda}(\mathbf{x})\,\mathrm{d}\mathbf{x}$, for pixel G_i, centered to have mean zero. To establish a basis to compare the consistency between proposed models and data for these two methods, we utilize the Probability Integral Transform (PIT) (Dawid, 1984). The PIT was proposed to evaluate how well a probabilistic forecast is calibrated by assessing the distribution of the values that the observations take under the cumulative distribution function of the proposed model. If the observations are a random draw from that model, a histogram of the PIT values should appear to be standard uniform.

One condition for the uniformity of the PIT values is that the proposed model be continuous. This holds for Voronoi residuals, which are approximately gamma distributed under the null hypothesis, but not for the Poisson counts from pixel residuals. For such discrete random variables, randomized versions of the PIT have been proposed. Using the formulation in Czado et al. (2009), if F is the distribution function of the proposed discrete model, $X \sim F$ is an observed random count and V is standard uniform and independent of X, then U is standard uniform, where

$$U = F(X-1) + V\left(F(X) - F(X-1)\right), \quad X \geq 1, \tag{5.1}$$

$$U = VF(0), \quad X = 0. \tag{5.2}$$

The method can be thought of as transforming a discrete c.d.f. into a continuous c.d.f. by the addition of uniform random noise.

5.2 Formal Testing

The PIT, both standard and randomized, provides a formal basis for testing for two competing residual methods. For a given proposed model and a given realization of points, the histogram of PIT values, $u_1, u_2, \ldots, u_n$, for each residual method should appear standard uniform if the proposed model is the same as the generating model. The sensitivity of the histogram to misspecifications in the proposed model reflects the statistical power of the procedure.

There are many test statistics that could be used to evaluate the goodness of fit of the standard uniform distribution to the PIT values. Here we choose to use the Kolmogorov-Smirnov (K-S) statistic (Massey, 1951),

$$D_n = \sup_n |F_n(x) - F(x)|,$$

where $F_n(x)$ is the empirical c.d.f. of the sample and $F(x)$ is the c.d.f. of the standard uniform. Since the Voronoi residuals of a given realization are not independent of one another, we use critical values from a simulated reference distribution instead of the limiting distribution of the statistic.

5.3 Simulation Design

Two models were considered for the simulation study. The first was a homogeneous Poisson model on the unit square with intensity λ on $\mathbb{R}^2$. The second was an inhomogeneous Poisson model with intensity

$$\lambda(x, y) = 100 + 200\left(\tilde{x}^{\beta}\tilde{y}^{\beta}c\right), \tag{5.3}$$

on $\mathbb{R}^2$, where $\tilde{x} = \frac{1}{2} - |x - \frac{1}{2}|$ and $\tilde{y} = \frac{1}{2} - |y - \frac{1}{2}|$. The constant c is a scaling constant chosen so that the parenthetical term integrates to one. The result is a function that is symmetric about $x = .5$ and $y = .5$, reaches a maximum at $(.5, .5)$, integrates to 300 regardless of the choice of β, and is reasonably flat along the boundary box. This final characteristic should allow the alternative approach to the boundary problem, described below, to be relatively unbiased. Additionally, it presents inhomogeneity similar to what might be expected in an earthquake setting. Six versions of this model, each with a different choice of β are shown in figure 5.1.

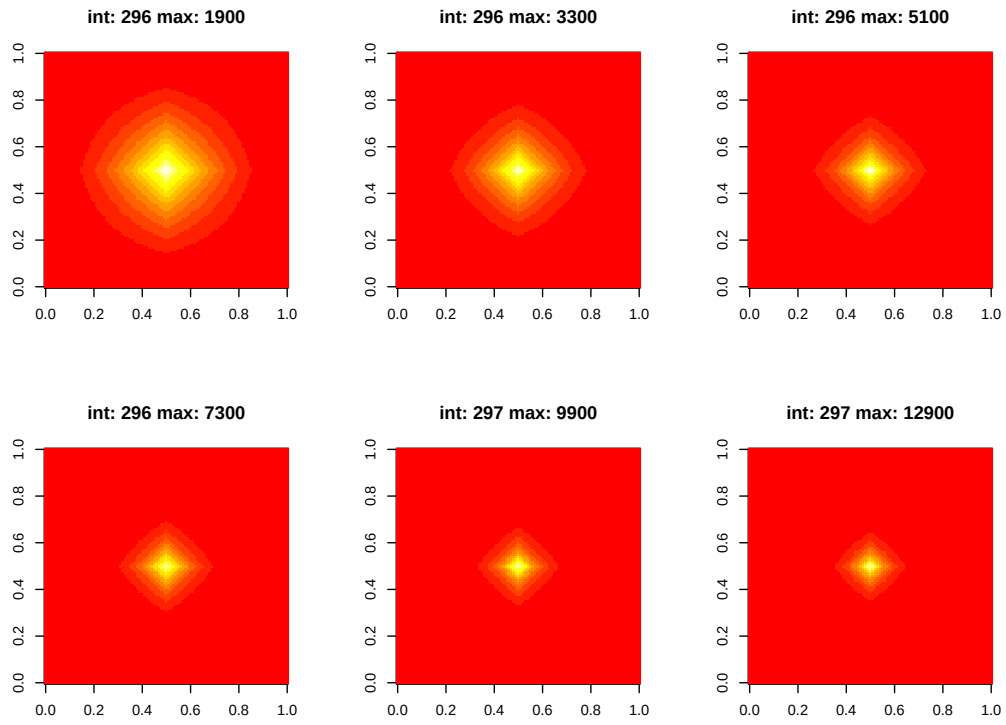

Figure 5.1: Six version of model 5.3. From left to right, top to bottom, $\beta = 2, 3, 4, 5, 6, 7$. Above each model is an approximation of the integral of the function over the unit square along with the maximum intensity at (.5, .5). The color scale is relative to the min and max of each individual plot.

The procedure for the inhomogeneous simulation was as follows. A point pattern was sampled from the true generating model, (5.3) with $\beta = 4$. For a given proposed model, with $\beta = \beta_0$, and a fixed number of pixels n on the unit square $[0, 1]^2$, PIT values were calculated for the counts in each pixel, G_i. The empirical c.d.f. of the PIT transformed residuals was then compared to the c.d.f. of the standard uniform using the K-S test. After many iterations of this procedure, the proportion of iterations with an observed K-S statistic that exceeded the critical value served as the estimate of the power of the method. An analogous procedure was followed for the Voronoi partition, but with the PIT

values calculated by evaluating the Voronoi residuals (4.1) under the modified gamma distribution (4.2).

The homogeneous simulation was conducted in the same manner, but drew samples from a generating model of $\lambda = 500$ and compared them to estimates from a proposed model λ_0.

5.3.1 Boundary effects

It is known that Voronoi cells generated along the boundary of the space do not follow the same distribution as the interior cells. One recourse is to omit them from the analysis, as in § 4.1. Here we consider realizations of the model (5.3) on the entire plane but consider only the distribution of all cells generated by points inside the unit square $[0, 1]^2$. An example of the resulting tessellation is shown in figure 5.2.

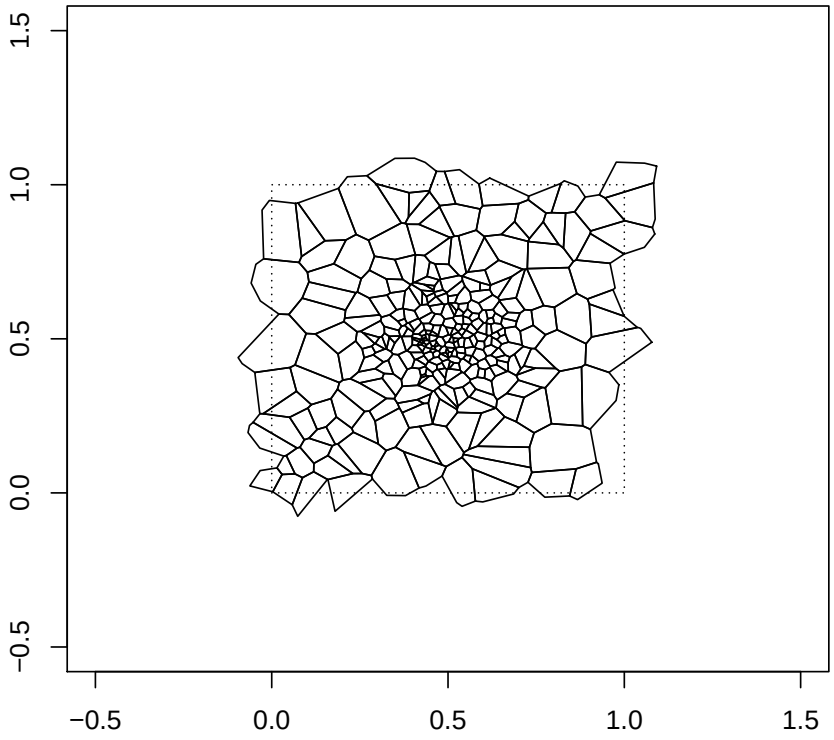

Figure 5.2: The tessellation resulting from generated a point pattern from model 5.3 with $\beta = 4$. The tessellation contains all cells generated by points that lie within the unit square.

The result is a tessellation that both exceeds and is less than the bounding box in places. By using an intensity function that is fairly flat along the boundary, these differences should cancel out and result in a total tessellated area that is roughly the same as that covered by the pixel residuals. This general approach of calculating statistics based on regions outside of W is known as a *boundary plus* method (Baddeley, 2007).

5.4 Results

For the homogeneous model, figure 5.3 shows the resulting estimated power curves for several pixel partitions including, $n \in \{36, 324, 900, 2500\}$. The power of each method was computed for a series of proposed models, $\lambda_0 \in (375, 625)$. The best performance was by the method that used the Voronoi partition, which shows high power throughout the range of misspecification.

For the pixel partitions, $n = 36$ had the highest power but as the number of partitions increased, the K-S test lost its power to detect misspecification. This trend can be attributed to characteristic I: when the space is divided into many small cells, the integrated conditional intensity is very small and the distribution of the residuals is highly skewed. As a consequence, the majority of counts are zeros, so the majority of the PIT values are being generating by $VF(0)$ (equation 5.2), and thus the resulting residuals have little power to detect model misspecification.

As the only misspecification present in the model is in the total number of expected points in the unit square, the most powerful test will be the K-S test performed without any partitioning. This is equivalent to $P(t_{.025} > N(W) > t_{.975})$, where the number of points in the unit square, $N(S) \sim \text{Poi}(\lambda = 500)$, and $t_{.025}$ and $t_{.975}$ are the lower and upper critical values for a Poisson random variable with rate λ_0. This is known in the earthquake forecasting literature as the Number-test (Schorlemmer and Gerstenberger, 2007).

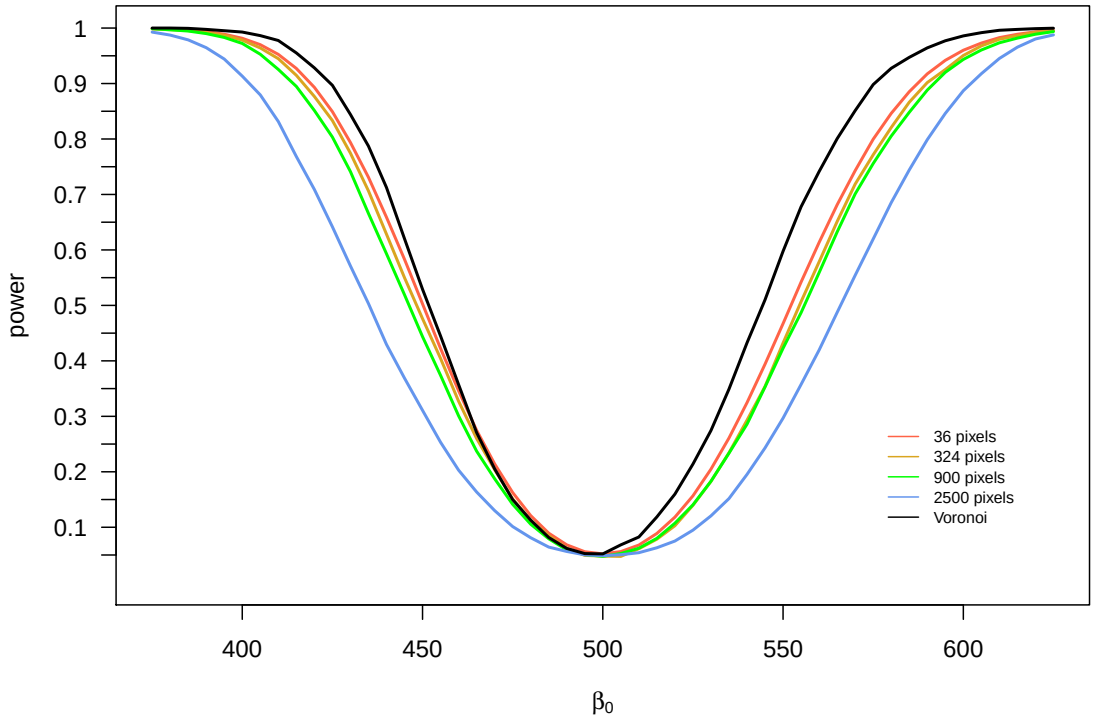

Figure 5.3: Estimated power curves for the K-S test based on five different pixel partitions as well as the Voronoi tessellation. The model under consideration is homogeneous Poisson with a generating intensity of $\lambda = 500$.

For the inhomogeneous case, power curves were computed for a series of proposed models of the form (5), with $\beta_0 \in (.5, 11)$. The results are shown in figure 5.4. The power curve for the Voronoi method presents good overall performance, particularly when the model is substantially misspecified. The Voronoi residuals are not ideally powerful for detecting slight misspecification, however, perhaps because the partition itself is random, thus introducing some variation that is difficult to distinguish from a small change in β.

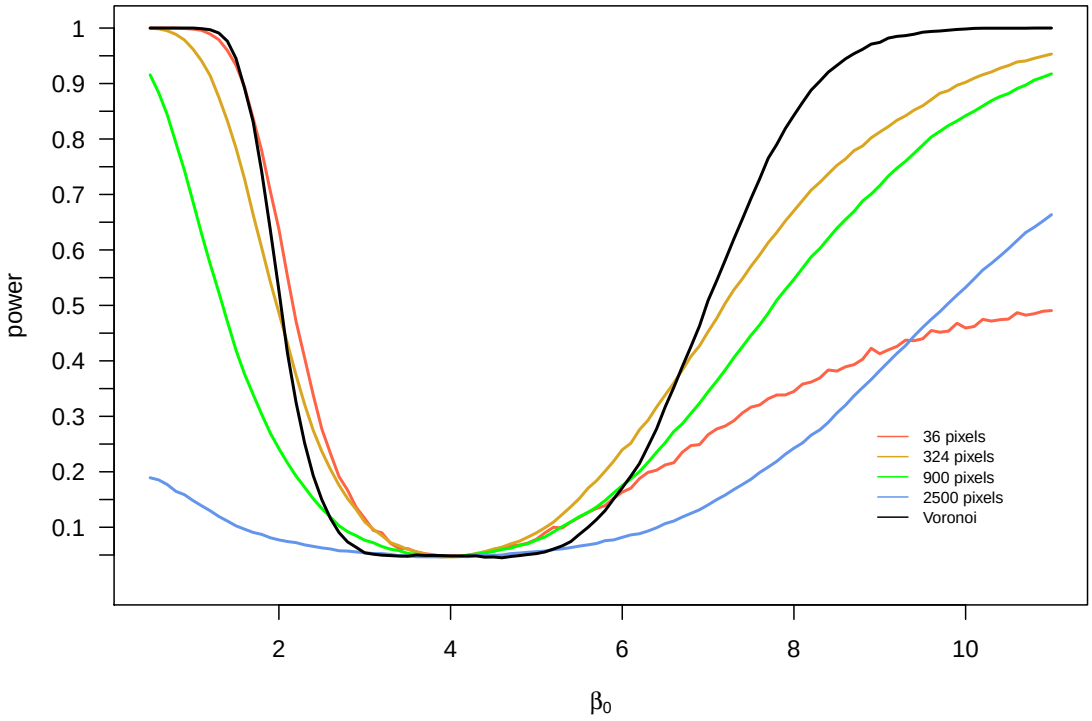

Figure 5.4: Estimated power curves for the K-S test based on five different pixel partitions as well as the Voronoi tessellation. The model under consideration is $\lambda(x, y) = 100 + 200\left(\tilde{x}^{\beta}\tilde{y}^{\beta}c\right)$, where $\tilde{x} = \frac{1}{2} - |x - \frac{1}{2}|$ and $\tilde{y} = \frac{1}{2} - |y - \frac{1}{2}|$. The generating model is $\beta = 4$.

Focusing only on the four pixel methods, the best performance is at $n = 324$ pixels. The poor performance of $n = 36$ in detecting the large positive misspecification is due to the fact that the model becomes more inhomogeneous as β_0 increases, but that inhomogeneity is averaged over cells that are too large (the problem associated with characteristic II in § 1.2). Meanwhile, the poor overall performance of $n = 2500$ is due to the same problem that exists in the homogeneous setting, where the PIT values are dominated by the random uniform noise.

In applications such as earthquake modeling, the use of pixel methods often

result in situations with extremely low intensities in some pixels, similar to the case considered here with $n = 2500$, but perhaps even more extreme. For instance, one of the most successful forecasts of California seismicity (Helmstetter et al., 2007) estimated rates in each of $n = 7682$ pixels in a model that estimated a total of only 35.4 earthquakes above M 4.95 over the course of a prediction experiment that lasted from 2006 to 2011. Estimated integrated rates were as low as 0.000007 in some pixels, and 58% of the pixels had integrated rates that were lower than 0.001. An immediate improvement could be made by aggregating the pixels, but this in turn will average over the strong inhomogeneity along fault lines in the model, which will lower power. For this reason, the Voronoi residual method may be better suited to the evaluation of seismicity models, as well as other processes that are thought to be highly inhomogeneous.

CHAPTER 6

Examples

6.1 RELM

In the present section we apply some of the residual methods discussed above to models and seismicity data from the 5-year RELM prediction experiment that ran from 2006 to 2011. The original experiment called for modelers to estimate the number of earthquakes above magnitude 4.95 that would occur in many pre-specified spatial bins in California. During this time period only 23 earthquakes that fit these criteria were recorded, a fairly small data set from which to assess a model. In order to better demonstrate the methods available in residual analysis, the models that are considered below were recalibrated using their specified magnitude distributions to forecast earthquakes of greater than magnitude 4.0, of which there are 232 on record.

The first model under consideration is one that was submitted to RELM by Helmstetter et al. (2007) and is described in chapter 2. The left panel of figure 6.1 shows the estimated number of earthquakes in every pixel in the greater California region that were part of the prediction experiment. Pixels shaded very light gray have a forecast of near zero earthquakes while pixels shaded black forecast much greater seismicity. The tan circles are the epicenters of the 232 earthquakes in the catalog, many of which are concentrated just South of the Salton Sea, near the border between California and Mexico.

The extent to which the observed seismicity is in agreement with the forecast

can be visualized in the raw pixel residual plot (center panel). The pixels are those established by the RELM experiment. Pixels where the model predicted more events than were observed are shaded in red; pixels where there was underprediction are shown in blue. The degree of color saturation indicates the p-value of the observed residual in the context of the forecasted Poisson distribution. Thus while the Helmstetter et al. (2007) model greatly underpredicted the number of events in the Salton Sea trough (dark blue), it also forecasted a high level of seismicity in several isolated pixels that experienced no earthquakes (dark red). The majority of the pixels are shaded very light red, indicating regions where the model forecast a very low rate of seismicity and no earthquakes were recorded.

The Voronoi residual plot for the Helmstetter et al. (2007) model is shown in the right panel of figure 6.1. It generates far fewer residuals than pixel-based methods (210 versus 7682) because by definition, each tile contains one observation. The spatial adaptivity of this partition is evidenced by the small tiles in regions of high point density and larger tiles in low density regions. The region of consistent underprediction in the Salton Sea trough is easily identified. Unlike the raw pixel residual plot, the Voronoi plot appears to distinguish between areas where the high isolated rates can be considered substantial overprediction (dark red) and areas where, considered in the context of the larger tile, the overprediction is less extreme (light red).

In figure 6.2 we assess how well the Helmstetter et al. (2007) model performs relative to another model in RELM using deviance residuals. The Shen et al. (2007) model is notable for utilizing geodetic strain-rate information from past earthquakes as a proxy for the density (intensity) of the process. $\mu()$ is then an interpolation of this data catalog. The result is a forecast that is generally much smoother than the Helmstetter et al. (2007) forecast, as seen in the left panel of figure 6.2. The center panel displays the deviance residuals for the Helmstetter et al. (2007) model relative to the Shen et al. (2007) model. The color scale is

mapped to a measure of the comparative performance of the two models ranging from 1 (dark blue) indicating better performance of the Helmstetter et al. (2007) model to -1 (dark red) indicating better performance of the Shen et al. (2007) model. This deviance residual plot reveals that the Helmstetter et al. (2007) model's relative advantage is in broad areas off of the main fault lines where the forecast was lower and there were no recorded earthquakes. It appeared to fit worse than the Shen et al. (2007) model, however, just West of the Salton Sea trough region of high seismicity, in a swath off the coast, and in isolated pixels in central California.

The Voronoi deviance plot (right panel) identifies the same relative underperformance of the Helmstetter et al. (2007) model relative to the Shen et al. (2007) model in the central California region and off the coast and is a bit more informative in the areas of higher recorded seismicity. In the Salton Sea trough region, just south of the border of California with Mexico, the Helmstetter et al. (2007) model appears to outperform the Shen et al. (2007) model in a vertical swath on the Western side of the seismicity, while the results on the Eastern side are more mixed. While these regions appear nearly white in the pixel deviance residual plot, suggesting roughly equivalent performance of the models, the aggregation of many of those pixels in the Voronoi plot allows for a stronger comparison of the two models.

The utility of residual methods can be seen by contrasting the residual plots with the error diagram of these same two models (figure 3.1 in chapter 3). While the error diagram and other functional summaries collapse the model and the observations into a new measure (such as the false negative rate), residual methods preserve the spatial referencing, which can help inform subsequent model generation.

6.2 ETAS

In this section we apply Voronoi residual analysis to the spatial-temporal epidemic-type aftershock sequence (ETAS) model of Ogata (1998) (see chapter 3), which has been widely used to describe earthquake catalogs.

There is considerable debate in the seismological community about the best method to estimate the spatial background rate $\rho(x, y)$ (Ogata, 2011; Helmstetter and Werner, 2012; Zhuang et al.). When modeling larger, regional catalogs, ρ is often estimated by smoothing the largest events in the historical catalog (Ogata, 1998), and in such cases a very important open question is how (and how much) to smooth (Schoenberg, 2003; Helmstetter et al., 2007; Helmstetter and Werner, 2012; Zhuang et al.). For a single earthquake-aftershock sequence, however, can one instead simply estimate ρ as constant within a finite, local area, as in Schoenberg (2013)?

A prime catalog to investigate these questions is the catalog of California earthquakes including and just after the 1999 Hector Mine earthquake, which struck east of the city of Barstow. This dataset was analyzed previously in Ogata et al. (2003), and consists of the origin times, epicentral locations, and magnitudes of the 520 earthquakes with magnitude at least 3.0, from latitude 34 to 35, longitude -116 to -117, from October 16, 1999 to December 23, 2000, obtained from the Southern California Seismic Network (SCSN).

The parameters in the model were estimated by maximum likelihood estimation, using the progressive approximation technique described in Schoenberg (2013). For the purpose of this analysis, we focused on the purely spatial aspects of the residuals, and thus integrated over the temporal domain to enable planar visualization of the residuals. The result is a Voronoi tessellation of the spatial domain where for tile C_i, for the integral in equation 4.1, the estimated conditional intensity function $\hat{\lambda}(t, x, y)$ is numerically integrated over the spatial tile C_i

and over the entire time domain from 10/16/1999 to 12/23/2000.

The top left panel of figure 6.3 shows the observation region and epicenter locations, along with their resulting Voronoi tessellation. The high concentration of seismicity along the fault line is obvious. The top right panel shows the residuals, which appear to be generally low to moderate in magnitude with no clear suggestion of systematic over- or under-prediction, indicating general agreement between the ETAS model and the data.

Along the periphery of the fault line, however, there is some evidence of the model overpredicting seismicity, as indicated by several red cells. The histogram and quantile plot of the Voronoi residuals in figure 6.3 reflect this overprediction by highlighting several residuals with values below -3. The lowest of these residuals corresponds to the bright red cell near the top of the residual plot, at a location roughly 35 miles due east of Barstow, California. Looking at an enlarged version of the residual plot (figure 6.4), it is apparent that in the region directly alongside the fault, the residuals are small in magnitude but systematically blue, indicating that the model is underpredicting seismicity. Taken together, these observations suggest that, even for this very local dataset, constraining ρ to be a constant may have resulted in excessive smoothing of the estimated seismicity and hence an underprediction of seismicity close to the fault line.

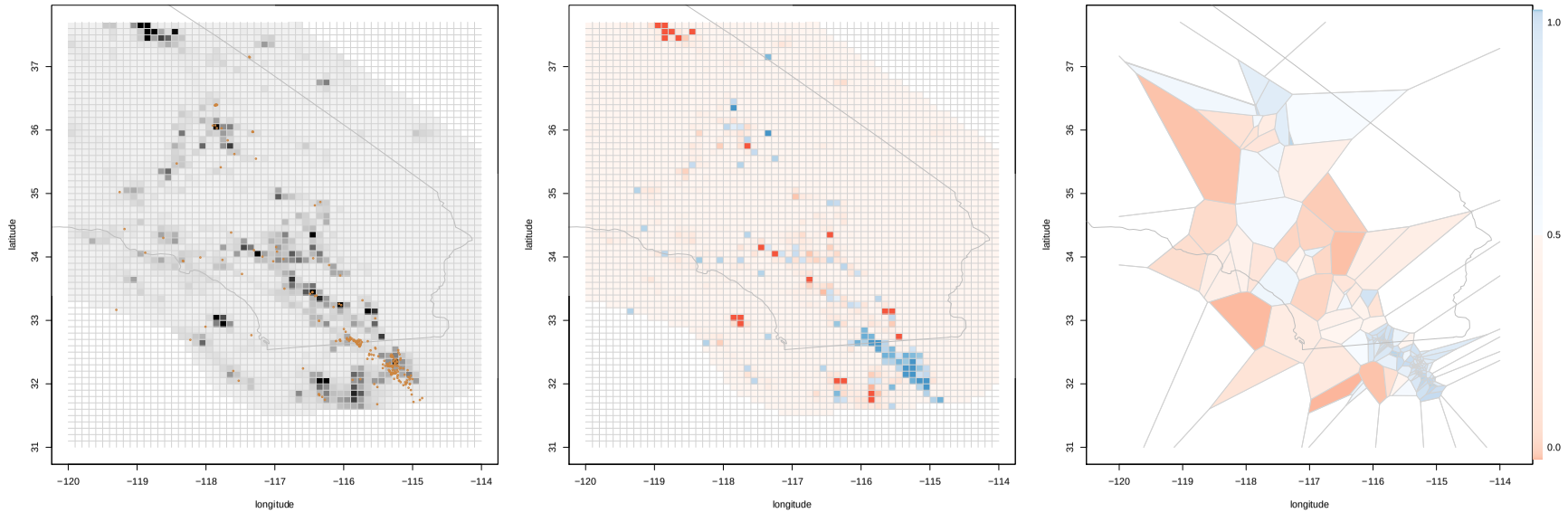

Figure 6.1: (a) Estimated rates under the Helmstetter et al. (2007) model, with epicentral locations of observed earthquakes with $M \geq 4.0$ in Southern California between January 1, 2006 and January 1, 2011 overlaid. (b) Raw pixel residuals for Helmstetter et al. (2007) with pixels colored according to their corresponding p-values. (c) Voronol residuals for Helmstetter et al. (2007) with pixels colored according to their corresponding p-values.

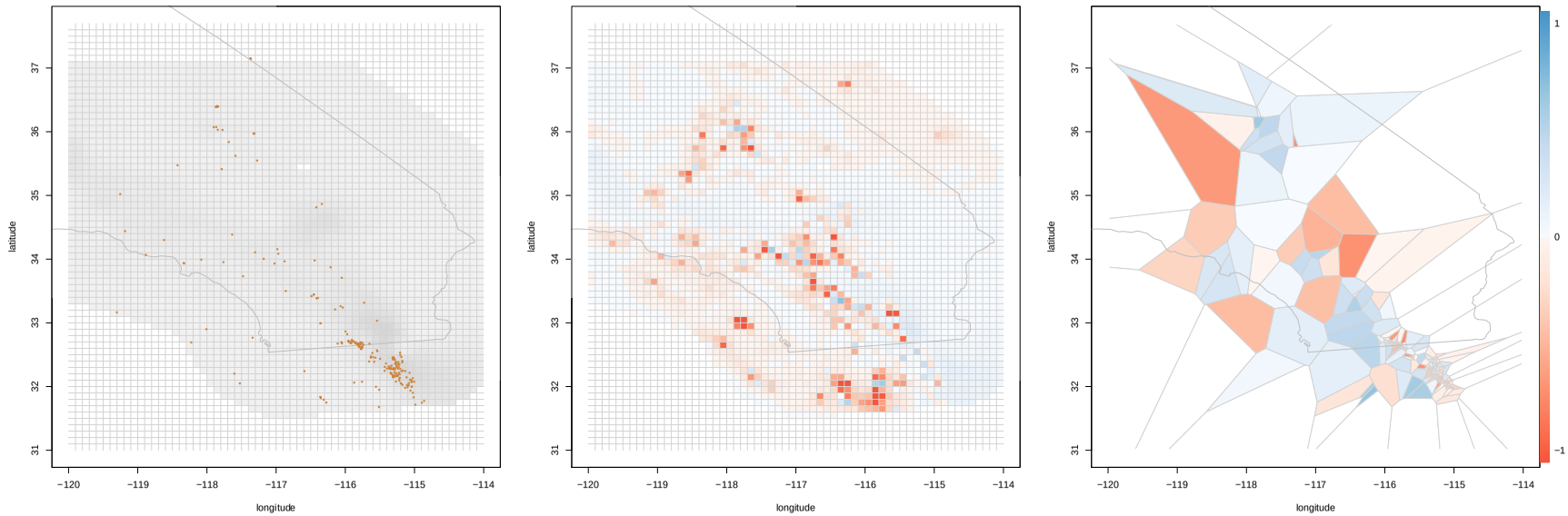

Figure 6.2: (a) Estimated rates under the Shen et al. (2007) model, with epicentral locations of observed earthquakes with $M \geq 4.0$ in Southern California between January 1, 2006 and January 1, 2011 overlaid. (b) Pixel deviance plot with blue favoring model A, Helmstetter et al. (2007), versus model B, Shen et al. (2007). Coloration is on a linear scale. (c) Voronoi deviance plot with blue favoring model A, Helmstetter et al. (2007), versus model B, Shen et al. (2007). Coloration is on a linear scale.

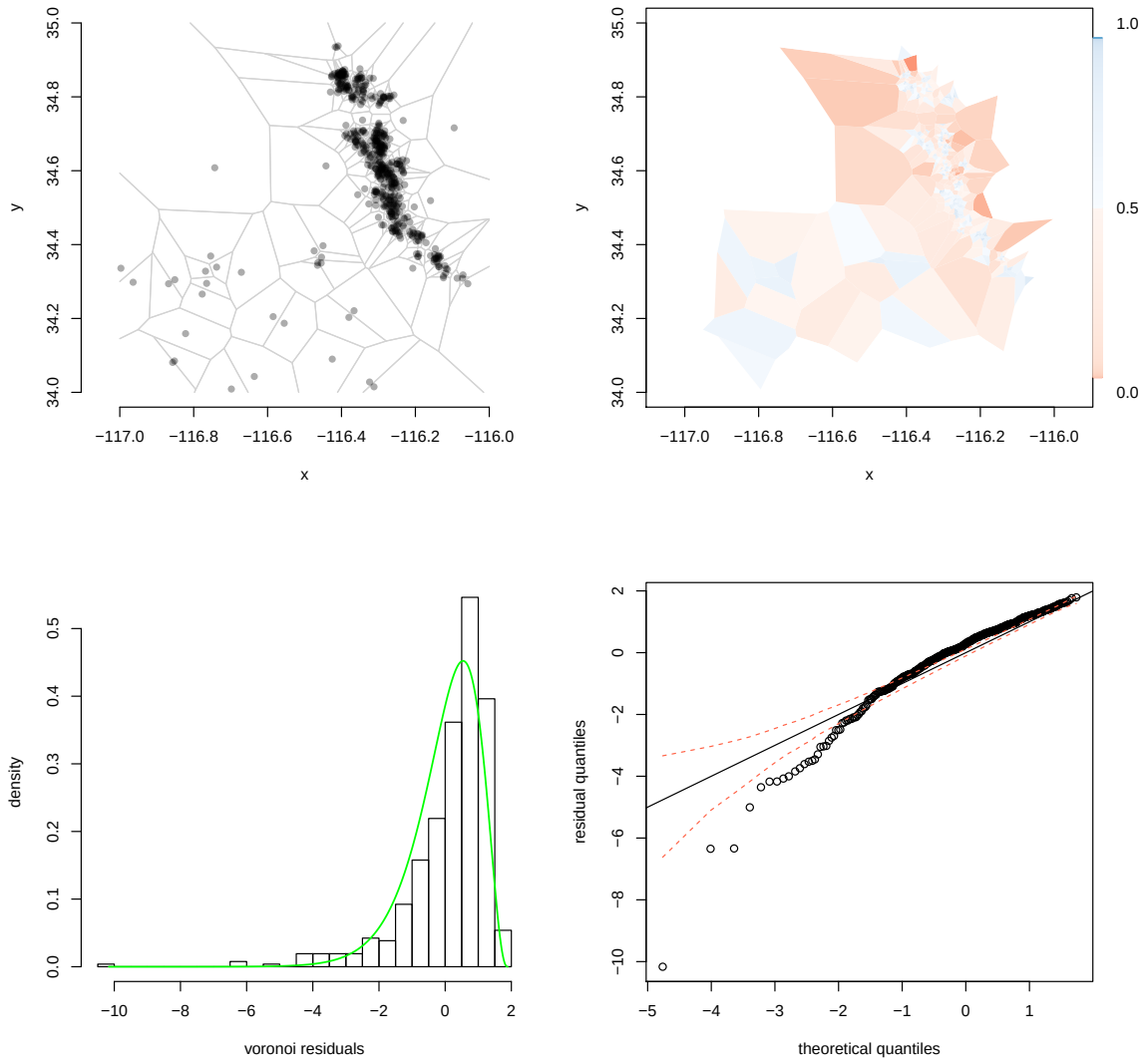

Figure 6.3: Southern California earthquake data (top left), Voronoi residual plot for the ETAS model suggesting over prediction on the in several off-fault sites (top right), histogram of the Voronoi residuals, with a green curve tracing the density of the reference distribution 4.2 (bottom left), quantile plot of the Voronoi residuals with respect to distribution 4.2, with pointwise 95% confidence limits obtained via simulation from the model (bottom right). The scale of the Voronoi residual plot is $\Phi^{-1}\{F(r)\}$, where F is the distribution function of 4.2. Tiles intersecting the boundary of the space are ignored.

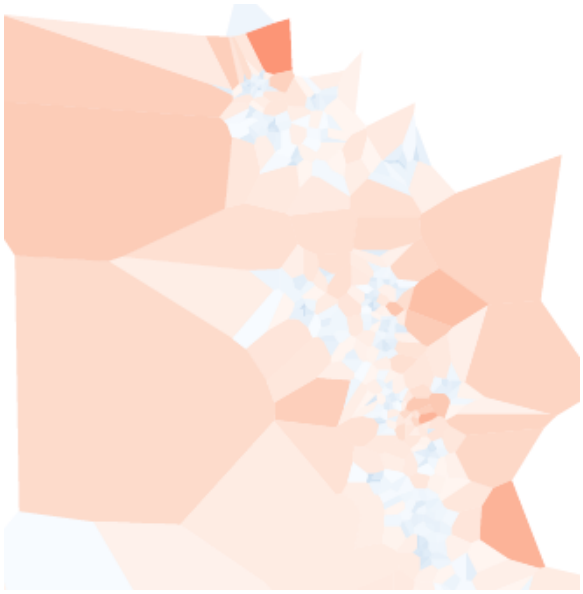

Figure 6.4: An enlarged portion of the residual plot in figure 6.3 showing systematic underprediction (blue) along the fault and overprediction (red) along the periphery.

CHAPTER 7

Model-generated Centroidal Voronoi Partition

While Voronoi residuals have characteristics that make them an effective tool for the assessment of inhomogeneous point process models - they are spatially adaptive and data-driven - other aspects present challenges. Some sampling variability is induced due to the fact that the tessellation originates from a single realization of the point process. Also, a residual will exhibit some depedence with other residuals within its fundamental domain, which may over- or understate the degree of spatial structure. In this section an alternative residual is discussed to address these challenges.

7.1 Centroidal Voronoi Residuals

Barr (2012) presents a Voronoi residual that utilizes successive iterations of regularization to stabilize the tessellation in a bias-variance tradeoff. In a single iteration of the algorithm, each point in the observed pattern $\mathbf{x}_i$ is moved to the centroid of cell C_i weighted by the proposed model, an implementation of the method of Lloyd (1992). That is, find the new point $\mathbf{x}_i^1$ where

$$\mathbf{x}_i^1 = \frac{\int_{C_i} \mathbf{x}\, \lambda_0(\mathbf{x})\, \mathrm{d}\mathbf{x}}{\int_{C_i} \lambda_0(\mathbf{x})\, \mathrm{d}\mathbf{x}}.$$

The new pattern, $\{\mathbf{x}_n^1\}$, is then re-tessellated, and so on, for either 2 or 3 iterations. Barr (2012) shows that the resulting residuals are well-approximated by the Gaussian model, which leads to very accurate tests based on the χ^2 distribution.

7.2 Model-generated partition

In order to achieve independent residuals that preserve some of the properties of the centroidal Voronoi residuals, the following variation is proposed.

7.2.0.1 Generating the partition

Let M be the expected number of points in a cell under H_0 that achieves maximum power under the Kolmogorov-Smirnov test, let $\underline{M} := \int_W \lambda(x)\,\mathrm{d}x$, and let W be a bounded region of interest in $\mathbb{R}^2$.

1. Generate $n = \underline{M}/M$ points in W, forming the pattern $\{\mathbf{x}_n^1\}$.
2. Tesselate W, forming the partition $\{C_n^1\}$.
3. Move each point in $\{\mathbf{x}_n^1\}$ to the model-weighted centroid of its associated cell, defined as the point where.
4. Repeat steps 2 and 3 until $sup\{||\{\mathbf{x}_n^i\} - \{\mathbf{x}_n^{i-1}\}||\} < \epsilon$ for some small distance ϵ, where $||\cdot||$ is the Euclidean distance between successive iterations of the same point.

The final partition that is generated by this algorithm will have cells that all have approximately the same expected count, M, under $\lambda_0(\mathbf{x})$. Several stopping rules could be used; this one was selected for simplicity and relative ease of computation. There is no reason to think that the final partition will be unique for a given $\lambda_0(\mathbf{x})$, M, and W, but it is not clear that this will be problematic. In terms of power, the operative feature of the process is M.

7.2.0.2 Calculating residuals

Once the partition is in place, residuals can be calculated in the manner of Baddeley et al. (2005), where the observed counts in each cell come from the observed

point pattern. Where as Voronoi residuals (4.1) fix the observed count and allow the expected count to vary, these residuals fix the expected count at M and allow the observed count to vary. This should create a set of residuals that are independent of one another and Poisson distributed (if the point process is assumed to be Poissonian).

It is unclear how to best choose M. Is it possible that an analytical solution exists what would maximize the power of the K-S test given a certain effect size. This is an open question that would be interesting to pursue. In any event, this method could then be compared to pixel residuals and regular Voronoi residuals as in figure 5.4.

CHAPTER 8

Conclusion

The proposed residual method described here is naturally spatially adaptive, with cell sizes approximately inversely proportional to the conditional intensity of the underlying point process. These Voronoi residuals may of course be used in tandem with standard, pixel-based residuals, which may in turn be based on a judicious choice of pixel size, or perhaps using a different spatially adaptive grid than the one proposed here.

The importance of selecting the size of the cell on which to compute a residual is not unique to this PIT - K-S statistic testing environment. The discrepancy measure proposed by Guan (2008) is defined on a Borel set of a given shape S. The author emphasizes the importance of choosing an appropriate size for S (page 835), and points out that if the cell is too small or too large, the power will suffer. A related problem arises in the selection of the bandwidth of the kernel used to smooth a residual field (see Baddeley and Turner, 2005, section 13 and discussion).

Although this dissertation has focused on formal testing at the level of the entire collection of residuals, testing could also be performed at the level of individual cells. For the Voronoi partition, this extension is straightforward and is essentially what is being done informally in the shaded residual plots. For any pixel partition, such testing may be problematic, as any pixel with an integrated conditional intensity close to zero would contain zero points with more than 95% probability, so any hypothesis test with $\alpha = 0.05$ using a rejection interval would necessarily have a type I error near 1.

Generating the partition using a tessellation of the observed pattern has advantages and disadvantages. The advantage is that it is adaptive and requires no input from the user regarding tuning parameters. The disadvantages are that some sampling variability is induced by the random cell areas and that the residuals are dependent, so techniques relying upon an i.i.d. assumption must be used cautiously. The centroidal Voronoi variant of Barr (2012) and the model-generated version of § 7.2 may be useful alternatives.

It should also be noted that the standardization methods proposed in Baddeley et al. (2005) may be used with Voronoi residuals, or instead one may elect to plot deviance residuals in each of the Voronoi cells. In general, experience suggests that the standardization chosen for the residuals seems far less critical than the choice of grid.

The paradigm established by RELM and CSEP is a very promising direction for earthquake model development. In addition to requiring the full transparent specification of earthquake forecasts before the beginning of the experiment, the criteria on which these models would be evaluated, namely, the L, N, and R tests, was also established. As the first RELM experiment proceeded, it became apparent that these tests can be useful summaries of the degree to which one model appears to agree with observed seismicity, but that they leave much to be desired. They are not well-suited to the purpose of comparing the goodness-of-fit of competing models or to suggest where models may be improved.

Future prediction experiments will allow for the implementation of more useful assessment tools. Residuals methods, including superthinned, pixel, and Voronoi residuals, seem ideal for comparison and to see where a particular model appears to overpredict or underpredict seismicity. Deviance residuals are useful for comparing two competing models and seeing where one appears to outperforms another in terms of agreement with the observed seismicity. These methods are particularly useful in the CSEP paradigm, as insight gained during one prediction experiment

can inform the building of models for subsequent experiments.

Bibliography

Giada Adelfio and Frederick Paik Schoenberg. Point process diagnostics based on weighted second-order statistics and their asymptotic properties. *Annals of the Institute of Statistical Mathematics*, 61(4):929–948, 1997.

Adrian Baddeley. Spatial point processes and their applications. In *Lecture Notes in Mathematics: Stochastic Geometry*, pages 1–75. Springer Verlag, Berlin Heidelberg, 2007.

Adrian Baddeley and Rolf Turner. Practical maximum pseudolikelihood for spatial point patterns (with discussion). *Australian and New Zealand Journal of Statistics*, 42(3):283–322, 2000.

Adrian Baddeley and Rolf Turner. Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

Adrian Baddeley, Jesper Moller, and Martin Hazelton. Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society, series B*, 67(5):617–666, 2005.

Adrian Baddeley, Jesper Moller, and Anthony G. Pakes. Properties of residuals for spatial point processes. *Annals of the Institute of Statistical Mathematics*, 60:627–649, 2008.

W. H. Bakun, B. Aagaard, B. Dost, and et al. Implications for prediction and hazard assessment from the 2004 Parkfield earthquake. *Nature*, 437:969–974, 2005.

Christopher D. Barr. Voronoi residuals for spatial point process models. *preprint*, 2012.

Christopher D. Barr and David M. Diez. Sizes of Voronoi regions in a spatial network designed by an inhomogeneous Poisson process. *preprint*, 2012.

Christopher D. Barr and Frederic Paik Schoenberg. On the Voronoi estimator for the intensity of an inhomogeneous planar poisson process. *Biometrika*, 94(4): 977–984, 2010.

Bruce Bolt. *Earthquakes*. Freeman, New York, fifth edition, 2003.

Andrew P. Bray and Frederic Paik Schoenberg. Assessment of point process models for earthquake forecasting. *Statistical Science*, 2014. to appear.

Andrew P. Bray, Ka Wong, Christopher D. Barr, and Frederic Paik Schoenberg. Using the Voronoi tessellation to calculate residuals for spatial point process models. *Annals of Applied Statistics*, 2014. in review.

Pierre Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, New York, 1981.

A. Chu, Frederic Paik Schoenberg, P. Bird, David D. Jackson, and Yan Y. Kagan. Comparison of ETAS parameter estimates across different global tectonic zones. *Bulletin of the Seismological Society of America*, 101(5):=2323–2339, 2011.

Robert A. Clements, Frederic Paik Schoenberg, and Danijel Schorlemmer. Residual analysis for space-time point processes with applications to earthquake forecast models in California. *Annals of Applied Statistics*, 5(4):2549–2571, 2011.

Robert A. Clements, Frederic Paik Schoenberg, and Alexander Veen. Evaluation of space-time point process models using super-thinning. *Environmetrics*, 23 (7):606–616, 2012.

R. Console, M. Murru, and G. Falcone. Probability gains of an epidemic-type aftershock sequence model in retrospective forecasting of M$\geq$ 5 earthquakes in Italy. *Journal of Seismology*, 14(1):9–26, 2010.

Noel Cressie and Christopher K. Wikle. *Statistics for spatio-temporal data*. John Wiley and Sons, Hoboken New Jersey, 2011.

Claudia Czado, Tilmann Gneiting, and Leonhard Held. Predictive model assessment for count data. *Biometrics*, 65:1254–1261, 2009.

Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society Series A*, 147(2):278–292, 1984.

Edward H. Field. Overview of the working group for the development of regional earthquake models (RELM). *Seismological Research Letters*, 78:7–16, 2007.

Edward H. Field, T. E. Dawson, K. R. Felzer, and et al. Uniform california earthquake rupture forecast, version 2 (UCERF 2). *Bulletin of the Seismological Society of America*, 99(4):2053–2107, 2009.

R. J. Geller, David D. Jackson, Yan Y. Kagan, and F. Mulargia. Earthquakes cannot be predicted. *Science*, 275(5306):1616–1617, 1997.

Yongtao Guan. A goodness-of-fit test for inhomogeneous spatial Poisson processes. *Biometrika*, 95(4):831–845, 2008.

Egil Hauksson and J. G. Goddard. Radon earthquake precursor studies in Iceland. *Journal of Geophysical Research*, 86:7037–7054, 1981.

A. G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society, Series B*, 33:438–443, 1971.

Agnes Helmstetter and Didier Sornette. Predictability in the epidemic-type aftershock sequence model of interacting triggered seismicity. *Journal of Geophysical Research*, 108:2482–2499, 2003.

Agnes Helmstetter and M. Werner. Adaptive spatio-temporal smoothing of seismicity for long-term earthquake forecasts in California. *Bulletin of the Seismological Society of America*, 102:2518–2529, 2012.

Agnes Helmstetter, Yan Y. Kagan, and David D. Jackson. High-resolution time-independent grid-based forecast M≥5 earthquakes in California. *Seismological Research Letters*, 78(1):78–86, 2007.

Susan Hough. *Predicting the Unpredictable: The Tumultuous Science of Earthquake Prediction*. Princeton University Press, Princeton, NJ, 2010.

David D. Jackson. Earthquake prediction evaluation standards applied to the VAN method. *Geophysical Research Letters*, 23:1363–1366, 1996.

E. A. Johnson and K. Miyanishi. *Forest Fire: Behavior and Ecological Effects*. Academic Press, San Diego, 2001.

Thomas H. Jordan. Earthquake predictability, brick by brick. *Seismological Research Letters*, 77:3–6, 2006.

Thomas H. Jordan and L. M. Jones. Operational earthquake forecasting: Some thoughts on why and how. *Seismological Research Letters*, 81(4):571–574, 2010.

Yan Y. Kagan. Are earthquakes predictable? *Geophysical Journal International*, 131:505–525, 1997.

Yan Y. Kagan. Testing long-term earthquake forecasts: Likelihood methods and error diagrams. *Geophysics Journal International*, 177(2):532–542, 2009.

J. E. Keeley, H. Safford, C. J. Fotheringham, and et al. 2007 Southern California wildfires: lessons in complexity. *Journal of Forestry*, September:287–296, 2009.

Vladimir Keilis-Borok and V. G. Kossobokow. Premonitory activation of earthquake flow: algorithm M8. *Physics of the Earth and Planetary Interiors*, 6(1): 73–83, 1990.

Andrew Lawson. Comment on 'Residual analysis for spatial point processes' by Baddeley, Turner, Moller, and Hazelton. *Journal of the Royal Statistical Society, Series B*, 67:654, 2005.

P. Lewis and G. Shelder. Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly*, 26:403–413, 1979.

S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1992.

B. D. Malamud, J. D. A. Millington, and G. L. W. Perry. Characterizing wildfire regimes in the United States. *Proceedings of the National Academy of Sciences of the USA*, 102(13):4694–4699, 2005.

M. Manga and C. Y. Wang. Earthquake hydrology. In G. Schubert, editor, *Treatise on Geophysics*, volume 4, pages 293–320. Elsevier, 2007.

Werner Marzocchi and Jeremy D. Zechar. Earthquake forecasting and earthquake prediction: different approaches for obtaining the best model. *Seismological Research Letters*, 82(2):442–448, 2011.

F. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 42(253):68–78, 1951.

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

E. Merzbach and D. Nualart. A characterization of the spatial Poisson process and changing time. *Annals of Probability*, 14:1380–1390, 1986.

P. A. Meyer. Démonstration simplifée d'un théorème de Knight. *Springer Lecture Notes in Mathematics*, 191:191–195, 1971.

George Molchan. Space-time earthquake prediction: the error diagrams. *Pure and Applied Geophysics*, 167:907–917, 2010.

M. Nair. Random space change for multiparameter point processes. *Annals of Probability*, 18:1222–1231, 1990.

Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30: 243–261, 1978.

Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83:9–27, 1988.

Y. Ogata. Space-time point process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50:379–402, 1998.

Y. Ogata. Significant improvements of the space-time ETAS model for forecasting of accurate baseline seismicity. *Earth, Planets, and Space*, 63(3):217–229, 2011.

Y. Ogata, L. M. Jones, and S. Toda. When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California. *Journal of Geophysical Research*, 108:2318–2329, 2003.

A. Okabe, B. Boots, K. Sugihara, and S. Chiu. *Spatial Tessellations*. Wiley, Chichester, second edition, 2000.

D. A. Rhoades, Danijel Schorlemmer, M. C. Gerstenberger, and et al. Efficient testing of earthquake forecasting models. *Acta Geophysica*, 59(4):728–747, 2011.

Brian D. Ripley. The second-order analysis of stationary point processes. *Journal of the Royal Statistical Society, Series B*, 39:172–212, 1976.

Frederic Paik Schoenberg. Transforming spatial point processes into Poisson processes. *Stochastic Processes and their Applications*, 81(2):155–164, 1999.

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

Frederic Paik Schoenberg. Facilitated estimation of ETAS. *Bulletin of the Seismological Society of America*, 103(1):1–7, 2013.

Danijel Schorlemmer and M. C. Gerstenberger. RELM testing center. *Seismological Research Letters*, 78(1):30–35, 2007.

Z. K. Shen, David D. Jackson, and Yan Y. Kagan. 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, 2007.

Didier Sornette. Apparent clustering and apparent background earthquakes biased by undetected seismicity. *Journal of Geophysical Research*, 110:B09303, 2005.

Dietrich Stoyan. Fundamentals of point process statistics. In Adrian Baddeley and et al., editors, *Case Studies in Spatial Point Process Modeling*. Springer, New York, 2005.

John A. Swets. The relative operating characteristic in psychology. *Science*, 182 (4116):990–1000, 1973.

Masaharu Tanemura. Statistical distributions of Poisson Voronoi cells in two and three dimensions. *Forma*, 18, 2003.

K. R. Tiampo and R. Shcherbakov. Seismicity-based earthquake forecasting techniques: Ten years of progress. *Tectonophysics*, 522:89–121, 2012.

Rolf Turner. *deldir*. CRAN, 0.0-16 edition, 2011.

A. Veen and Frederic Paik Schoenberg. Assessing spatial point process models for California earthquakes using weighted K-functions: analysis of California earthquakes. In Adrian Baddeley and et al., editors, *Case Studies in Spatial Point Process Models*, pages 293–306. Springer, New York, 2005.

David Vere-Jones and Frederic Paik Schoenberg. Rescaling marked point processes. *Australian and New Zealand Journal of Statistics*, 46(1):133–143, 2004.

David Vere-Jones and J. Zhuang. On the distribution of the largest event in the critical ETAS model. *Physical Review E*, 78:047102, 2008.

Q. Wang, David D. Jackson, and Yan Y. Kagan. California earthquake forecasts based on smoothed seismicity: Model choices. *Bulletin of the Seismological Society of American*, 101(3):1422–1430, 2011.

Haiyong Xu and Frederic Paik Schoenberg. Point process modeling of wildfire hazard in Los Angeles County, California. *Annals of Applied Statistics*, 5(2): 684–704, 2011.

Jeremy D. Zechar. *Methods for evaulating earthquake predictions*. PhD thesis, University of Southern California, 2008.

Jeremy D. Zechar, M. C. Gerstenberger, and D. A. Rhoades. Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts. *Bulletin of the Seismological Society of America*, 100(3):1184–1195, 2010.

Jeremy D. Zechar, Danijel Schorlemmer, M. J. Werner, Matthew C. Gerstenberger, David A. Rhoades, and Thomas H. Jordan. Regional Earthquake Likelihood Models I: First-order results. *Bulletin of the Seismological Society of America*, 103(2):787–798, 2013.

J. Zhuang, David Harte, M. J. Werner, Sebastian Hainzl, and Shiyong Zhou.