Title:
Earthquake prediction: Simple methods for complex phenomena

Author:
Luen, Bradley

Acceptance Date:
01-01-2010

Series:
UC Berkeley Electronic Theses and Dissertations

Degree:
Ph.D., StatisticsUC Berkeley

Advisor:
Stark, Philip B

Committee:
Rice, John, Allen, Richard M

Permalink:
http://escholarship.org/uc/item/22p7f44k

Abstract:

Earthquake prediction: Simple methods for complex phenomena

by

Bradley Luen

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

in

Statistics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Philip B. Stark, Chair
Professor Emeritus John Rice
Associate Professor Richard M. Allen

Fall 2010

Earthquake prediction: Simple methods for complex phenomena

Bradley Luen

Abstract

Earthquake prediction: Simple methods for complex phenomena

by

Bradley Luen
Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Philip B. Stark, Chair

Earthquake predictions are often either based on stochastic models, or tested using stochastic models. Tests of predictions often tacitly assume predictions do not depend on past seismicity, which is false. We construct a naive predictor that, following each large earthquake, predicts another large earthquake will occur nearby soon. Because this "automatic alarm" strategy exploits clustering, it succeeds beyond "chance" according to a test that holds the predictions fixed.

Some researchers try to remove clustering from earthquake catalogs and model the remaining events. There have been claims that the declustered catalogs are Poisson on the basis of statistical tests we show to be weak. Better tests show that declustered catalogs are not Poisson. In fact, there is evidence that events in declustered catalogs do not have exchangeable times given the locations, a necessary condition for the Poisson.

If seismicity followed a stochastic process, an optimal predictor would turn on an alarm when the conditional intensity is high. The Epidemic-Type Aftershock (ETAS) model is a popular point process model that includes clustering. It has many parameters, but is still a simplification of seismicity. Estimating the model is difficult, and estimated parameters often give a non-stationary model. Even if the model is ETAS, temporal predictions based on the ETAS conditional intensity are not much better than those of magnitude-dependent automatic (MDA) alarms, a much simpler strategy with only one parameter instead of five. For a catalog of Southern Californian seismicity, ETAS predictions again offer only slight improvement over MDA alarms.

To Mr. and Mrs. L., and Mr. and Mrs. G.

Contents

List of Figures

List of Tables

Acknowledgments

The chapter "Testing earthquake predictions" was previously published in a different form in *IMS Lecture Notes—Monograph Series. Probability and Statistics: Essays in Honor of David A. Freedman.*

This entire document is a collaboration with Philip Stark. I owe him the greatest of thanks for his insight and patience over the last five years.

Thanks to the other members of my dissertation committee, John Rice and Richard M. Allen, for their careful reading and comments.

Thanks to Steven N. Evans for substantial assistance with the optimal alarm lemma and proof.

Thanks to Alejandro Veen, David Harte, and Stefan Wiemer for code.

Thanks to Shankar Bhamidi, Moorea Brega, the late David A. Freedman, Cliff Frohlich, Robert Geller, Chris Haulk, Mike Higgins, Yan Y. Kagan, Chinghway Lim, Guilherme V. Rocha, Karl Rohe, Frederic P. Schoenberg, and David Shilane for helpful conversations and comments.

Thanks to the staff of the Statistics Department—in particular Angie Fong, who has saved my skin on several occasions.

Finally, thanks to my parents and sister, and to Jennifer Lin and her family, for their personal and occasional financial support.

Chapter 1

Introduction to earthquake predictions and forecasts

1.1 Outline

If the locations and times of large earthquakes could be predicted accurately, lives could be saved, for instance by evacuation. Despite occasional claims of success, no existing method is reliable enough to justify such drastic action. More recently, seismologists have focused on "forecasting" earthquakes. Forecasting is taken to imply less accuracy than prediction, but the distinction is somewhat arbitrary.

The distinction between deterministic and probabilistic forecasts is clearer. What is random? Interpreting probabilistic forecasts is easier if the predictions are considered to be random and the seismicity is not. However, it is more common to consider seismicity to be random. Unfortunately, that makes a frequentist interpretation of probabilistic forecasts impossible. It is possible to interpret the probability of an earthquake as a parameter of a stochastic model. If the model is not testable on a human time scale, the probability assertions are of little value.

To assess whether earthquake predictions or forecasts are successful, we need to define "success." A deterministic prediction in a space-time region is successful if one or more earthquakes occur in the prediction region. It is difficult, however, to determine whether predictive success is "significant." Many statistical tests of forecasts hold predictions fixed and examine their success on randomly generated seismicity. This can be misleading, because predictions usually depend on the observed seismicity: if the observed seismicity were different, the predictions would be different. In chapter 2, we study a naive alarm strategy that simply predicts that an earthquake will be followed by another earthquake within a small space-time window. This predictor may appear to have significant success in a statistical test that employs a null hypothesis of random seismicity—not because the predictor is good, but because under the null hypothesis, clustering is unlikely.

One way of dealing with clustering is to "decluster" earthquake catalogs by removing events that occur close to others. It has often been claimed that declustered catalogs are Poissonian [2, 5]. In chapter 3, we test the hypothesis that times of a declustered catalog are a realisation of a Poisson process, and the much weaker hypothesis that times are exchangeable given the locations. We reject the Poisson hypothesis and, for some catalogs, the exchangeable times hypothesis.

If earthquake sequences are not Poissonian, earthquakes can, to some extent, be successfully forecasted from previous seismicity. The best predictor, in a sense, of a point process is its conditional intensity. If we assume seismicity adheres to a stochastic model, we can make predictions based on the conditional intensity.

A well-known stochastic point process model for earthquake occurrence is the epidemic-type aftershock (ETAS) model. In this model, any earthquake can trigger further shocks; those further shocks may trigger more shocks, and so on. In chapter 4, we examine the simulation and estimation of the model. In chapter 5, we examine prediction for renewal and ETAS models. We address limitations and pathologies of the ETAS model, and compare its predictive success to that of simpler models. Chapter 6 states the implications of this dissertation for statistical seismology. It also describes directions for further work.

The rest of this chapter attempts to clarify the meaning of deterministic and probabilistic earthquake predictions and forecasts. Section 1.2 examines the definitions of earthquake predictions and earthquake forecasts. Section 1.3 explains some terminology in statistical seismology, and gives some notation. Section 1.4 discusses the interpretation of earthquake probabilities and shows, as an example, that it is difficult to interpret the forecasts issued by the Working Group on California Earthquake Probabilities in 2008.

1.2 What are earthquake predictions and forecasts?

For many years, a major criticism of earthquake forecasts was that predictions were vague and hard to test. Allen [6] and Geller [7], among others, have emphasised criteria a prediction must satisfy to be well-defined: it should claim an earthquake with a magnitude in a specified range will occur in a specific time and space window. Predictions may be based on past seismicity alone, extra-seismic variables, or some combination of the two. Some seismologists require a physical basis for predictions so that they can be verified; others find statistical verification more persuasive in the long term. This dissertation focuses on statistical verification.

Many statistical methods of testing prediction and forecasting schemes have been proposed. We discuss the shortcomings of some of them in chapter 2. Tests are generally performed by comparing success of predictors to some baseline. Predicting real seismicity better than a Poisson process can be predicted is a low hurdle, as almost any scheme that makes use of clustering meets this standard. For a scheme

that employs extra-seismic information to be valuable, it should perform better than the best model that uses only past seismicity, and certainly better than a model that uses past seismicity in a very simple way. We give various criteria for measuring performance, in this and subsequent chapters.

Predictions versus forecasts

Geologists often draw a strong distinction between predictions and forecasts. Predictions are taken to be more specific than forecasts, particularly in time. Though they were dealing with volcanic eruptions and not earthquakes, the following definitions by Wright and Pierson [8] give some indication of how the terms are used in the field:

- "A **forecast** is a comparatively imprecise statement of the time, place, and ideally, the nature and size of impending activity."
- "A **prediction** is a relatively precise statement giving the time and place..."

Jackson [9] proposed the following definitions: "Earthquake forecasting means specifying the long-term probability of earthquakes per unit area, magnitude, and time. It may incorporate modest time-dependence. Earthquake prediction, by contrast, means identifying special conditions that make the immediate probability much higher than usual, and high enough to justify unusual action."[1] This and similar definitions leave a substantial grey area in which it is not clear whether something is a prediction or a forecast.

Seismologists in recent years have tended to prefer the term "forecast" where possible, perhaps because the term "prediction" is taken by some to imply an accuracy yet to be achieved. In a debate at nature.com,[2] Robert Geller wrote that "[t]he public, media, and government regard an 'earthquake prediction' as an alarm of an imminent large earthquake, with enough accuracy and reliability to take measures such as the evacuation of cities." He regarded such prediction in the foreseeable future as impossible. Highly publicised failures of schemes trumpeted by their authors as "predictions" may have also contributed to an aversion to the term. The only prediction to have been generally regarded as successful was that of the 1975 Haicheng earthquake, and even that has been brought into question in recent years [9].

Generally speaking, the same statistical tools are used to assess both predictions and forecasts, so to at least some extent, the distinction seems arbitrary. Of course, one-off predictions are not amenable to statistical testing—see section 1.4.1 for an example.

[1] Following Jackson, we consider the calculation of time-dependent hazard as a forecast or prediction; some authors do not.

[2] http://www.nature.com/nature/debates/earthquake/

1.2.1 Types of predictions

Predictions and forecasts may be categorised in several ways. Firstly, predictions and forecasts that are *deterministic* may be distinguished from those that are *probabilistic*. In the former, in each region or subregion of the prediction area, either an earthquake is predicted, or no earthquake is predicted. Probabilistic predictions assign a probability of one or more earthquakes (or, less commonly, a probability distribution for the number of earthquakes) for each region or subregion. There may also be a probability distribution on the magnitude of the earthquakes that occur, or on the maximum magnitude event in each region. It is not clear, however, that seismicity is a stochastic process—the occurrence of earthquakes may be deterministic. Care must thus be taken in interpreting probabilistic predictions. These generally cannot be interpreted in the usual frequentist way: no standard interpretation of probability seems adequate [10], as discussed in the following section.

Earthquake "early warnings" [11], issued after the onset of rupture but up to a minute or so before significant shaking occurs, are sometimes considered predictions; we shall not consider them here.

Deterministic and probabilistic forecasts may be further distinguished by the time (or times) at which they are issued relative to the period that they are forecasting.

- *Periodic forecasts* are issued at regular intervals in time. These will most commonly make a prediction for a length of time equal to the periodicity at which they are issued. For example, a predictor may, at the beginning of each year, give a probability for the occurrence of one or more large earthquakes that year, in a given spatial region. The M8-MSc algorithms [12] declare "times of increased probability" for specified regions every six months, though these times last longer than six months.

- *Moving target forecasts* have regular or irregular updates that supersede previous predictions. We can make an analogy to weather forecasts: a forecast of Wednesday's weather issued on Monday will be superseded by a forecast of Wednesday's weather issued on Tuesday. The California 24-hour aftershock forecast map [13] issues hourly maps giving the "probability of strong shaking" (MMI VI or greater) some time during the next day. The maps show little change from hour to hour unless an earthquake occurs in the region.

- *Alarm strategies* turn an "earthquake alarm" on or off. if the risk of an earthquake in the near future is considered high. If the risk is considered low, the alarm is turned off. These strategies are deterministic. Whether the alarm is on at time t may depend on any data observed up to, but not including, time t. predictions of the VAN group [14] were similar to these, though those predictions were somewhat unspecific as to their space-time extent.

- *Stochastic rate models* give a predicted rate per unit time of earthquakes in a certain magnitude range for the near future. Forecasts from these models may be deterministic or probabilistic. Stochastic point process models such as the ETAS model (see chapter 4) are examples of these.

In meteorology, the following terms are used:

- *Forecast period:* the length of time for which the forecast is valid
- *Lead time:* the length of time between the issue time of the forecast and the beginning of the forecast validity period
- *Forecast range:* lead time plus forecast period; can be thought of as a "horizon."

Thus periodic forecasts have fixed forecast periods; alarms and rate models do not require lead time; while moving target forecasts may make predictions at a range of lead times.

Lead time is important in practice: an accurate forecast issued a month before an earthquake is more useful than the same forecast issued one day in advance. The relative values of these forecasts may be estimated in cost-benefit analyses. Further, moving target predictions may be issued for a (possibly continuous) range of lead times, and a scheme that is successful at one lead time may fail at another. In weather forecasting, the approach seems to be to assess each lead time separately, but the multiple testing complications have to date been insufficiently addressed.

1.3 Terminology and notation

- **Hypocenter (or focus):** The point within the earth where an earthquake rupture starts.
- **Epicenter:** The point on the earth's surface vertically above the hypocenter. The distance between the hypocenter and the epicenter is the *depth* of the earthquake. The locations of hypocenters and epicenters are estimated from measurements taken at multiple seismographic stations—for instance, by measuring the difference between arrival times of longitudinal P-waves and slower transverse S-waves to estimate the distance from each station, and triangulating. All calculated hypocenters and epicenters are estimates.
- **Magnitude:** A numeric characterisation of an earthquake's relative size, according to one of several scales. For instance, the well-known Richter scale (no longer used by seismologists) is based on measurements of the maximum amplitude recorded by a seismograph.

- **Seismic moment** (M_0)**:** A measure of the size of an earthquake. It is the product of the shear modulus (ratio of shear stress to shear strain), the area of fault rupture, and the average displacement during rupture.

- **Moment magnitude** (M_w) **scale:** The magnitude scale preferred by seismologists, as it is applicable to earthquakes of all sizes. It is calculated from seismic moment as

$$M_w = \frac{2}{3} \log_{10} (M_0) - 10.7, \tag{1.1}$$

  where M_0 is in dyne-cm.[3] The constants were chosen to make the scale approximately consistent with older magnitude scales, such as the Richter local magnitude scale.

- **Foreshocks, main shocks, and aftershocks:** In standard usage, the largest (in magnitude) earthquake in a sequence is the *main shock*. Smaller earthquakes following a main shock in a sequence are called *aftershocks*. Smaller earthquakes preceding a main shock are called *foreshocks*. There is no universally accepted definition as to what constitutes an earthquake sequence. Seismologists may thus disagree as to whether a particular earthquake is a foreshock, main shock or aftershock.

  Occasionally, "main shock" is used to refer to the initial shock in a sequence, and not the largest (and then all subsequent events are considered aftershocks). I shall instead refer to these events as "first shocks."

1.3.1 Earthquake catalogs

An earthquake catalog is a list of information, such as estimated hypocenters, times, and magnitudes, about a set of earthquakes in some geographical region. Catalogs may be local, like that of the Southern California Earthquake Data Center, or global, like the Global Centroid Moment Tensor Catalog. All catalogs are incomplete: that is, left-censored with respect to magnitude. Small earthquakes are harder to detect and locate. The level of completeness depends on geographical location. Some regions of the world, such as California, are well-covered by seismographic stations, while other regions, such as the oceans, are covered sparsely. Completeness also depends on time. Because the number of stations has increased and equipment has become more sensitive, magnitude thresholds for completeness are lower for recent data.

[3] One dyne-cm in SI units is 10^{-7} newton metre.

1.3.2 Point process models for seismicity

Suppose we wish to model or predict earthquakes in a *study region* of space-time-magnitude

$$V = A \times (0, T] \times [m_0, \infty). \tag{1.2}$$

Call A the *study area*, $(0, T]$ the *study period*,[4] and m_0 the minimum magnitude. We refer to earthquakes occurring in the study region as *events*.

Assume that the events may be ordered chronologically, without ties. We may model the events in a study region using a temporal point process, or a space-time point process. The point process may be marked or unmarked. In a marked space-time point process, the ith event is characterised by its epicentral latitude X_i and longitude Y_i, its time T_i and its magnitude M_i. We ignore other characteristics of earthquakes, such as depths.

In a marked temporal point process, the ith event is characterised by its time T_i and its magnitude M_i. We know that all events occur in the study area A, but otherwise ignore spatial locations. In an unmarked temporal point process, the ith event is characterised by its time T_i only. An unmarked temporal point process can also be characterised by $N(t)$, its *counting function*:

$$N(t) = \begin{cases} 0 & \text{for } 0 < t < T_1 \\ i & \text{for } T_i \le t < T_{i+1}, i \in \{1, \ldots, n-1\} \\ n & \text{for } T_n \le t \le T, \end{cases} \tag{1.3}$$

where $n = N(T)$ is the number of events in the study region. The counting function is a right-continuous step function with jumps of size 1 at the times of events. We primarily focus on marked and unmarked temporal point process.

Let $\mathcal{F}_{t_0}$ be the σ-algebra [15] generated by the process in the interval $[0, t_0]$ and by $\mathcal{F}_0$ (the information known about the process prior to time 0: for example, observations of events before the start of the study period).

The *intensity* at time t is the expected rate of events at t; this expectation may or may not be conditional on some σ-algebra. The *conditional intensity* $\lambda(t)$ is the expected rate of events conditioned on the history of the process up to (but not including) time t:

$$\lambda(t) = \lim_{\Delta \downarrow 0} \frac{\mathbf{E}[N(t+\Delta) - N(t) | \mathcal{F}_{t-}]}{\Delta}. \tag{1.4}$$

(See Appendix C for a measure-theoretic definition.) We assume $\lambda(t)$ exists and has finite expectation for all $t \in (0, T]$, and that it is absolutely continuous with respect to Lebesgue measure with probability 1 (where the probability distribution is on realisations of the process). The conditional intensity is a measurable function of the

[4] The left side of the interval is open in case time 0 is set to be the time of an event.

random history of the process, up to, but not including, time t—it is *previsible*.

We examine predictions that take the form of *alarm strategies*. These are written as $H(t)$ or $H_t(\omega)$, for temporal point processes ($H(t, x, y, \omega)$ for space-time point processes). The range of H is $[0, 1]$. The value 0 means the alarm is off at time t; the value 1 means the alarm is on. For $0 < H < 1$, the function gives the probability that the alarm is on.[5]

Whether marked or unmarked, temporal or space-time, a point process is just a model. It does not account for all features of real seismicity. These models may, however, be useful for studying patterns of earthquake occurrence, and for forecasting.

1.4 Interpreting earthquake probabilities

Many earthquake forecasts are probabilistic. Usually, it is the seismicity that is assumed to be stochastic, rather than the predictions. This leads to difficulties in interpretation.

For example, working groups of the U.S. Geological Survey have produced estimates of the chance of a magnitude 6.7 or greater earthquake occurring in the San Francisco Bay Area over the following thirty years. The 1999 estimate (for 2000 to 2030) was 0.7 ± 0.1; the 2002 estimate, 0.62 with 95% confidence bounds of 0.37 to 0.87. A 2008 working group estimated the probability of a magnitude 6.7 or greater earthquake occurring in the California area in the time period 2007 to 2036 as "greater than 99%." These probabilities are based on a combination of a wide range of physical and statistical models and simulations, as well as subjective elements. For this reason, and because it is not apparent what concept of probability is being used, interpreting such probabilities is difficult.

Stark and Freedman [10] identified the problems in applying standard definitions of probability to forecasts, in particular that of the 1999 USGS working group. These problems concerned interpretation more than the numerical values. In the frequentist view, the probability of an event is the limit of its relative frequency in repeated trials under the same conditions. For USGS forecasts, no particular thirty-year period can be repeated: it only occurs once.[6] In the (subjective) Bayesian view, probability is a measure of a state of belief on a scale from 0 to 1. However, the USGS forecasts are not Bayesian: they do not start from a prior, nor do they update probabilities using Bayes' rule.[7]

[5] The notion of probability here is different from that in probabilistic forecasts. Here, the alarm is on with some probability; in probabilistic forecasts, probability is intrinsic to the model.

[6] One could postulate an infinite number of worlds, of which one is selected at random, but such a paradigm seems, in the words of Feller, "both uninteresting and meaningless." [16]

[7] Some apparently subjective probabilities are used to weight branches of a decision tree to select models in a Monte Carlo simulation of seismicity. However, these probabilities cannot represent degrees of belief. For instance, it is implausible that a Poisson model for seismicity is correct, but

Working group	Forecast period	Prob. of M $\geq$ 6.7 event	95% confidence range
WG99	2000-2030	70%	50% to 90%
WG02	2002-2031	62%	37% to 87%
WGCEP 2007	2007-2036	63%	41% to 84%

Table 1.1: USGS-sanctioned 30-year Bay Area earthquake forecasts. Probability forecasts of WG88 (50%) and WG90 (67%) were for events of magnitude "about 7" and may not be directly comparable.

A promising avenue for interpretation of earthquake forecasts is that a probability may be viewed as a property of a mathematical model designed to describe some real-world system. George Box said that "all models are wrong, but some are useful." [17] The usefulness is dependent on the degree of agreement with the real-world system.

Such a model-based interpretation of probability may be the most natural paradigm for earthquake probabilities. A stochastic model is proposed. That model produces as output a number, called a "probability." That number—which exists only in the model—is interpreted as having something to do with the real world, namely, it is taken to be the chance of an earthquake in some region of space, time, and magnitude. However, predictions for earthquakes are difficult to test, because of the time scales involved; most of the USGS working group forecasts, for example, cannot be subjected to a meaningful statistical test, as no repetition is possible. This makes the predictions of little value unless we have good reason to be confident in the model.

1.4.1 Example: The 2008 USGS Working Group forecast

The USGS working groups that created long-term forecasts for regions of California were succeeded by a group commissioned to develop a statewide forecast. The Working Group on California Earthquake Probabilities (referred to as WGCEP 2007, although the forecast was not published until 2008) consisted of scientists and engineers from a variety of disciplines. It was sponsored by the USGS, the California Geological Survey, and the Southern California Earthquake Center. The group in turn sought the opinions of the broader seismological community. As well as a forecast for the entire state, forecasts for subregions—Northern and Southern California and the San Francisco and Los Angeles regions—were to be generated, for minimum magnitude thresholds from 6.7 to 8.0. A key purpose was to create a model that could be used by the California Earthquake Authority to set earthquake insurance rates. Estimation of shaking or ground motion, however, was outside the scope of the project.

---

that model is given positive weight in the decision tree. Bayes' rule is mentioned as an alternative approach that is not pursued.

Date	Location	Magnitude
April 18th, 1906	San Francisco	7.8
April 21st, 1918	San Jacinto	6.8
Jan 22nd, 1923	Humboldt County	7.2
June 29th, 1925	Santa Barbara	6.8
Nov 4th, 1927	Lompoc	7.1
Dec 31st, 1934	Cerro Prieto	7.0
May 19th, 1940	Imperial Valley	6.9
July 21st, 1952	Kern County	7.5
Nov 8th, 1980	Humboldt County	7.3
Oct 18th, 1989	Loma Prieta	6.89
April 25th, 1992	Cape Mendocino	7.15
June 28th, 1992	Landers	7.29
Oct 16th, 1999	Hector Mine	7.12

Table 1.2: M $\geq$ 6.7 earthquakes in the California earthquake catalog used by WGCEP 2007.

California is on the boundary of the Pacific and North American plates. The San Andreas fault, a strike-slip transform fault, runs through the state and forms the boundary between the plates. The physical reality is far more complex than this single fault: hundreds of other known faults exist in the state.

The forecast by WGCEP 2007 incorporated ideas from geodesy, geology, seismology, and paleoseismology. Data concerning relative plate movements and fault locations and offsets, the history of observed earthquakes, and reconstructions of the history of unobserved earthquakes were inputs. All this very different information was then combined.

WGCEP 2007's primary finding was that "the chance of having one or more magnitude 6.7 or greater earthquakes in the California area over the next 30 years is greater than 99%," and that "the likelihood of at least one even more powerful quake of at least magnitude 7.5 or greater in the next 30 years is 46%." A magnitude 7.5 or greater event was considered more likely in the southern half of the state than the northern half (37% chance against 15%). In the twentieth century, 13 events of magnitude 6.7 or greater are thought to have occurred in the California area (see Table 1.2). Note that there are uncertainties in all the magnitudes, with generally larger uncertainties for earlier events. The 28-year stretch from 1952 to 1980 without a $M \geq 6.7$ event suggests a 30-year stretch without such an event is plausible, so the 99% probability is a strong statement. The San Andreas and Hayward faults were considered to be at elevated risk: large earthquakes were thought to be more likely in the period of study than the historical rates of earthquakes on this fault would

suggest.

Let us set aside the issue of whether the model used makes sense, and focus on interpreting the output probabilities. The working group drew an analogy between their probabilities and the annual chance of being killed by lightning ("about 0.0003%"). A problem with interpreting the lightning figure is that it questionably equates a frequency with a probability. Each "trial" is an individual person-year; thus not all trials are the same. The chance of being killed by lightning should vary from person to person, depending on their location and their propensity to be outdoors during thunderstorms.

The working group also drew an analogy between their forecast and weather forecasts. However, weather forecasts are testable. They can be compared to observed weather on a daily basis. If, over a span of hundreds or thousands of days, weather forecasts are found to be consistent with observations, we can have confidence in the forecasting method. In contrast, large earthquakes in a particular region are uncommon, occurring on a time scale of decades or longer. Earthquake forecasts on the scale of days involve extremely small probabilities. This makes earthquake forecasts more difficult to test and to interpret than weather forecasts. Some attempts to evaluate earthquake forecasts have circumvented this issue by dividing a large study area into many smaller cells of space and time. A difficulty is that the dependence structure of seismicity between such cells is unknown and difficult to estimate.

The WGCEP 2007 forecasts are for one thirty-year period only. Consider the forecast that the chance of one or more magnitude 6.7 or greater earthquakes in the California area over the next 30 years is greater than 99%. If no such earthquake occurs, there is strong evidence that the forecast was wrong. If such an earthquake does occur, however, it is not strong evidence that the "99%" figure was accurate—such an observation would also be consistent with a forecast probability of 80%, or 30%. The "99%" is thus impossible to validate by itself. If the same method were used to create a sequence of consecutive thirty-year forecasts of large California earthquakes, validation might be possible. Even a small number of thirty-year periods, however, exceeds a human lifespan.

As the study says, "Californians know that their State is subject to frequent—and sometimes very destructive—earthquakes." It is not clear that the study helps the public to understand or to prepare for such events.

A further source of confusion is that it not clear whether the randomness in the WGCEP forecast is assumed to be in the Earth or in the model. Seismicity is complex, but there is no evidence that it is inherently random. Even if it were, there is no correct stochastic model for seismicity. This is a major issue in statistical tests of earthquake predictions, as many tests assume an inappropriate null hypothesis of random seismicity. We examine this issue in the following chapter.

Chapter 2

Testing earthquake predictions[1]

2.1 Introduction

Earthquake prediction has roots in antiquity [19]. Predictions have been based on a variety of seismic and non-seismic phenomena, including animal behavior [20, 21, 22, 23, 19]; water level, temperature and composition in wells and springs [24, 25]; electric and magnetic fields and radio waves on the ground and in the air [26, 27]; electrical resistivity of the ground and the air [28, 29, 27]; cloud formations or other atmospheric phenomena [30, 19]; infrared radiation [27]; pattern recognition [31, 32]; temporal clustering [33, 34]; and variations in the rate or pattern of seismicity [35].

There are large research efforts directed towards predicting earthquakes, such as the Collaboratory for the Study of Earthquake Predictability.[2] "Even a stopped clock is right twice a day," and almost any method for predicting earthquakes will succeed occasionally—whether the method has merit or not. Indeed, prominent geophysicists disagree about whether earthquake prediction is possible in principle.[3] How, then, ought we decide whether a method for predicting earthquakes works?

Earthquake predictions have been assessed using ideas from statistical hypothesis testing: a test statistic is compared to its distribution under a null hypothesis [37, 33, 38, 39, 40, 41, 42, 43]. The null hypothesis is rejected at significance level α if the test statistic exceeds the $1 - \alpha$ quantile of its distribution under the null hypothesis. If the null hypothesis is rejected, researchers tend to conclude—erroneously—that the predictions must have merit.

The null hypothesis can be rejected for many reasons. A Type I error might

[1] Previously published in a different form in *IMS Lecture Notes—Monograph Series. Probability and Statistics: Essays in Honor of David A. Freedman* [18]. Some notation in this chapter differs from the rest of the dissertation.

[2] http://www.cseptesting.org/

[3] See, e.g., http://www.nature.com/nature/debates/earthquake/equake_frameset.html and [36].

occur. Or the null hypothesis could be false, but in a way that does not imply that the predictions work. For example, the null hypothesis might use a poor model for seismicity. Or the null hypothesis might not account for how the predictions depend on seismicity. We explore that possibility below.

Conclusions ultimately depend on details of the null hypothesis that can be difficult to justify, or that are known to contradict the data. For example, Stark and Freedman [10] argue that standard interpretations of probability do not make sense for earthquakes—especially for large events, the most important to predict. For rare events, such as large earthquakes, there are not enough data to test or discriminate among competing stochastic models. Models are often calibrated using empirical scaling laws that tie the rates of occurrence of large earthquakes to the rates for smaller earthquakes. Generally, these rules of thumb are themselves fitted to data from other parts of the world: applying them to a region as small as the San Francisco Bay area, for example, is questionable. Thus, stochastic models for earthquake occurrence do not seem like a good foundation for evaluating earthquake predictions, especially predictions of large earthquakes.

Moreover, Stark [44, 45] argues that testing predictions using a stochastic model for seismicity and conditioning on the predictions tends to be misleading, and that it is preferable to treat seismicity as fixed and compare the success of the predictions with the success of a simple rule. Consider rain forecasts as an analogy. The rule "if it rains today, predict that it will rain tomorrow; otherwise, predict that it will not rain tomorrow" works pretty well. If a meteorologist cannot do better, the meteorologist's predictions have little value.

The seismic analogue is "if there is an earthquake with magnitude greater than the threshold M_τ, predict that there will be an earthquake with magnitude M or above within time t and distance d of the first." Here, M, t and d might depend on the location or magnitude of the first earthquake. Kagan [46] calls this the "automatic alarm" strategy, and uses it to evaluate earthquake predictions for Greece (the VAN predictions). The approach can also include a stochastic element to make a "semi-automatic alarm" strategy: Stark [44, 45] compares the VAN predictions to the rule: "If there is an earthquake with magnitude $\geq M_\tau$, toss a (biased) coin. If the coin lands heads, predict that there will be another earthquake with magnitude $\geq M$ within time t and distance d of the first. If the coin lands tails, do not make a prediction."

2.2 Phenomenology of earthquakes

See Bolt [47] for a lay review. The *epicenter* of an earthquake is the point on Earth's surface directly above the earthquake's *focus*, the place that the motion nucleates. Epicenters and foci are not known exactly: they are estimated from ground motion at seismographic observing stations around the globe. Sizes of earthquakes

are also estimated from ground motion measured at seismographic stations. There are many measures of earthquake size, including several definitions of "magnitude."

An earthquake *catalog* is a list of the estimated locations, times, and magnitudes of earthquakes found by a given authority, such as the U.S. Geological Survey. Earthquake catalogs are incomplete below some magnitude (left-censored in magnitude) because smaller events are harder to identify and locate. Moreover, unless some minimum number of stations detect ground motion, the algorithms used to locate earthquakes do not even conclude that there was an earthquake. (The incompleteness as a function of magnitude tends to decrease with time, as equipment becomes more sensitive and networks more extensive.)

Earthquakes occur to depths of 700 km or so in Earth's mantle [48]; however, most earthquakes and almost all large earthquakes occur within a few tens of kilometres of the surface. Earthquakes cluster in space. Most earthquakes occur on pre-existing faults. With very few known exceptions, epicenters of large earthquakes are close to the margins of tectonic plates, because it takes large strains—the relative motions of plates—to produce large earthquakes. Indeed, most large earthquakes occur in a relatively narrow band around the Pacific Ocean, the "ring of fire."

Earthquakes also cluster in time: large earthquakes invariably have aftershocks; some have foreshocks; and there are "swarms" of moderate-to-large earthquakes. Defining foreshocks and aftershocks is difficult. The terms "foreshock" and "aftershock" are often taken to imply a causal connection to a main shock. Unfortunately, earthquake physics is largely a mystery. Proximity in space and time can be coincidental rather than causal. One cannot tell whether an earthquake is the main shock or a foreshock of a larger event except—at best—in retrospect.[4] And stochastic models for earthquakes can produce spatio-temporal clustering without physical foreshocks or aftershocks per se (for example, gamma renewal models [50], discussed in chapter 5.3.5).

The most common stochastic model for seismicity takes the epicenters and times of shocks[5] above some threshold magnitude to be a realisation of a spatially inhomogeneous but temporally homogeneous Poisson process. The spatial heterogeneity reflects tectonics: some regions are more active seismically than others. The temporal homogeneity is justified by appeal to the lengthy time scale of plate tectonics (tens of thousands of years) relative to the time scale of observation, which is on the order of centuries. The seminal reference on stochastic models for seismicity is Vere-Jones [51], which considers temporal and marked temporal processes, but not spatial processes. Some more recent models use branching processes [52, 53].

[4] Identifying an event as a foreshock or aftershock is a causal inference based on association in time and space. Causal conclusions from associations in non-experimental data are highly suspect. See, e.g., Freedman [49].

[5] Sometimes this is restricted to main shocks, which are difficult to separate from foreshocks and aftershocks, as noted above and in chapter 3.

2.3 Tests of earthquake predictions

There are two categories of earthquake predictions: *deterministic* or *binary predictions*, which are of the form "there will be an earthquake of magnitude $M \geq 6$ within 100 km of San Francisco, CA, within the next 30 days;" and *probabilistic predictions*, which are probability distributions, or statements of the form "there is a 90% chance of an earthquake of magnitude $M \geq 6$ within 100 km of San Francisco, CA, in the next 30 days." Stark and Freedman [10] point out some difficulties in interpreting probabilistic predictions, using the USGS prediction for the San Francisco Bay Area as an example; here we concentrate on deterministic predictions.

To keep the exposition simple, we take the goal to be to predict all earthquakes that exceed some threshold magnitude M, that have epicenters in some region R of Earth's surface, and that occur during some time period T. We examine several statistical approaches to testing whether predictions have merit.[6]

Let Q denote the total number of earthquakes of magnitude $\geq M$ with epicenters in R during T. Let A denote the number of alarms (predictions). The jth alarm is characterised by V_j, a connected region of space, time and magnitude, and a value p_j, the probability the prediction assigns to the occurrence of an earthquake in V_j. Deterministic predictions take $p_j = 1$. They assert that an earthquake will occur in V_j. Probabilistic forecasts assign a probability $p_j \in (0, 1)$ to the occurrence of an earthquake in V_j. Let δ_j be the duration of V_j. When δ_j is on the order of weeks, V_j is generally considered a "prediction." When δ_j is on the order of a year or more, V_j is generally considered a "forecast." However, some authors use "prediction" to mean deterministic prediction, and "forecast" to mean probabilistic prediction, regardless of the time horizon.

Let λ_j denote the historical rate of earthquakes in the spatial and magnitude range—but not the temporal range—covered by V_j.[7] The historical rates $\{\lambda_j\}_{j=1}^{A}$ enter into some tests, as we shall see. Let S_j indicate whether the jth alarm is successful:

$$S_a \equiv \begin{cases} 1, & \text{if there is an earthquake in } V_j, \\ 0, & \text{otherwise;} \end{cases} \tag{2.1}$$

and for $k = 1, \ldots, Q$ let P_k denote whether the kth earthquake is predicted:

$$P_k = \begin{cases} 1, & \text{if the } k\text{th event is in some } V_j,\ j = 1, \ldots, A, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

[6] Statistical terminology is used in some unfamiliar ways in the geophysical literature. For example, "significance" and "confidence" sometimes denote 100% minus the P-value, rather than the chance of a type I error for a fixed-size test (e.g., [42, p. 193] and [54, pp. 723, 731], which also confuses the P-value with the chance that the null hypothesis is true). "Random probabilities" are sometimes fixed parameters [38, p. 3773], and "parameters" sometimes means statistics [55, p. 263].

[7] Typically, λ_j is the empirical rate over a span of a decade or more over a spatial region that includes V_j.

Let $S \equiv \sum_{j=1}^{A} S_j$ denote the number of successful alarms, and let $P \equiv \sum_{k=1}^{Q} P_k$ denote the number of earthquakes that are predicted successfully. The number of false alarms is $F = A - S$ and the number of earthquakes that are missed—not predicted—is $M = Q - P$. Let V be the volume of space-time studied,

$$V \equiv \int_{R,T} dr\, dt, \tag{2.3}$$

and let V_A denote the total space-time volume of alarms,

$$V_A \equiv \int_{\cup_{j=1}^{A} V_j} dr\, dt. \tag{2.4}$$

The fraction of the study volume covered by alarms is $v = V_A/V$. Generally, the smaller v is, the more informative the alarms are, but this can be distorted by spatial heterogeneity of the distribution of earthquakes in R.[8] The success rate of the predictions is $s = S/A$; the fraction of earthquakes successfully predicted is $p = P/Q$; the false alarm rate is $f = F/A$; and the rate of missed events (failures to predict) is $m = M/Q$. If we raise an alarm for the entire study volume and time V, we can ensure that $s = p = 1$, but then $v = 1$, so the alarms are not informative.

Predictions are generally evaluated using a combination of s, p, f, m, and v. Prediction methods can be ranked by adjusting their tuning parameters so that their values of v are equal, then comparing their values of p, or vice versa. For a given alarm volume v, the method with largest p is best. For a given value of p, the method with the smaller v is best. Some evaluation strategies fix p and compare values of f, or vice versa.

2.3.1 Testing strategies

A common strategy for evaluating earthquake predictions statistically is to compare the success of the predictions on the observed seismicity with the success of the same predictions on random seismicity (e.g., [39, 56, 57, 58]). This strategy does not make sense because predictions usually depend on past seismicity: if the seismicity had been different, the predictions would have been different.[9]

[8] To account for the spatial heterogeneity of events, some authors use normalised counting measure in space—based on the historical occurrence of events in a given volume—rather than Lebesgue measure. See, e.g., Kossobokov et al. [42].

[9] This is a bit like the Monte Hall or *Let's Make a Deal* problem [59, Ch. 10]. A prize is hidden at random behind one of three doors. The contestant picks a door. The host then reveals that the prize is not behind one of the two doors the contestant did not pick. The contestant is now allowed to switch his guess to the third door. Should he? Some erroneous arguments assume that the door the host opens is independent of which door conceals the prize. That is not a good model for the game, because the host never opens the door that hides the prize: which door the host opens depends on

Several stochastic models for seismicity are common for testing predictions.

1. Some studies model seismicity by a homogeneous Poisson process with intensity equal to the mean historical rate in the study region (e.g., [41]). Some studies condition on the number of events and model seismicity as uniform over the study region or subsets of the study region [60, 61, 41].

2. Some studies use a spatially heterogeneous but temporally homogeneous Poisson process model, with rate in the spatial region R_j equal to the historical rate λ_j [33, 42].

3. Some studies condition on the observed locations of past events, but model the times of the events as Poisson or uniform [42, 62].

4. Some studies condition on the observed locations and the observed times, but model the times as exchangeable [46, 63]. That is, if the observed time of the jth event in the catalog is $t_j, j = 1, \ldots, Q$, then, according to the model, it is equally likely that the times would have been $t_{\pi(j)}, j = 1, \ldots, Q$, where π is any permutation of $\{1, \ldots, Q\}$.

In the last approach (the permutation model, sometimes called "randomising a catalog"), times of events in the study region are exchangeable, conditional on the observed locations.[10]

There are variations on these approaches. For example, some researchers try to remove putative aftershocks from the catalogs (e.g., [64, 46, 40]). This is called "declustering." The most common method for declustering is to make spatio-temporal holes in a catalog: after each event, all smaller events that occur within a given time interval and epicentral distance are deleted. The time interval and distance can depend on the magnitude of the event [65, 66, 67]. (For more on declustering, see chapter 3.) It is common to assume that a declustered catalog is a realisation of a temporally homogeneous Poisson process.[11] Assessments of earthquake predictions are known to be sensitive to details of declustering and to spatial variability of the rate of seismicity [68, 69, 70, 56, 44, 45, 71, 72].

Another approach to testing is to compare the success rate of predictions with the (theoretical or empirical) success rate of random predictions that do not use any seismic information [73]. This seems to be a straw-man comparison because such random predictions ignore the empirical clustering of seismicity.

---

the contestant's guess and on which door hides the prize. Similarly, holding the prediction fixed regardless of the seismicity is not a good model for earthquake prediction. Whether a prediction is issued for tomorrow typically depends on whether there is an earthquake today.

[10] See chapter 3.4.1 for an explanation of conditional exchangeability.

[11] This kind of declustering produces a process that has less clustering than a Poisson process because it imposes a minimum distance between events—see chapter 3.4.1.

2.3.2 Jackson, 1996

Jackson [38] reviews methods for testing deterministic and probabilistic predictions. The approach to testing deterministic predictions is based on a probability distribution for the number of successful predictions, in turn derived from a null hypothesis that specifies $P(S_j = 1)$, $j = 1, \ldots, A$. Jackson does not say how to find these probabilities, although he does say that usually the null hypothesis is that seismicity follows a Poisson process with rates equal to the historical rates. He assumes that $\{S_j\}_{j=1}^{A}$ are independent, so S is the sum of A independent Bernoulli random variables. Jackson advocates estimating the P-value, $P(S \geq S_{\text{observed}})$, by simulating the distribution of the sum of independent Bernoulli variables, and mentions the Poisson approximation as an alternative. See Kagan and Jackson [39] for more discussion of the same approaches. Both articles advocate a likelihood-ratio test for evaluating probabilistic forecasts. They also propose a variant of the Neyman-Pearson testing paradigm in which it is possible that both the null hypothesis and the alternative hypothesis are rejected, in effect combining a goodness-of-fit test of the null with a likelihood ratio test against the alternative.

2.3.3 Console, 2001

Console [55] addresses deterministic predictions and probabilistic forecasts. His discussion of deterministic predictions includes several statistics for comparing alternative sets of predictions. His discussion of probabilistic forecasts is based on the likelihood approach in Kagan and Jackson [39], described above. The likelihood function assumes that predictions succeed independently, with known probabilities. For Console, the null hypothesis is that seismicity has a Poisson distribution [55, p. 266]. He gives one numerical example of testing a set of four predictions on the basis of "probability gain," but no hint as to how to determine the significance level or power of such tests. His test rejects the null hypothesis if more events occur during alarms than are expected on the assumption that seismicity has a homogeneous Poisson distribution with true rate equal to the observed rate. Console also mentions selecting prediction methods on the basis of a risk function, and Bayesian methods. The loss function Console contemplates is linear in the number of predicted events, the number of unpredicted events, and the total length of alarms, all of which are treated as random. He does not address estimating the risk from data, but it seems that any estimate must involve stochastic assumptions about Q, S, F and M.

2.3.4 Shi, Liu & Zhang, 2001

Shi, Liu and Zhang [40] evaluate official Chinese earthquake predictions of earthquakes with magnitude 5 and above for 1990–1998. They divide the study region into 3,743 small cells in space, and years of time. In a given cell in a given year, either

an earthquake is predicted to occur, or—if not—that is considered to be a prediction that there will be no event in that cell during that year. They define the R-score as

$$R = \frac{\#\text{ cells in which earthquakes are successfully predicted}}{\#\text{ cells in which earthquakes occur}} - \frac{\#\text{ cells with false alarms}}{\#\text{ aseismic cells}}, \tag{2.5}$$

which measures the concordance of the binned data with predictions of occurrence and of non-occurrence. In computing the R-score, they first decluster the catalog using the method of Keilis-Borok et al. [65]. Their hypothesis tests use the R-score as the test statistic. They compare the R-score of the actual predictions on the declustered catalog with the R-score of several sets of random predictions, generated as follows:

1. Condition on the number of cells in which earthquakes are predicted to occur. Choose that many cells at random without replacement from the 3,743 cells, with the same chance of selecting each cell; predict that earthquakes of magnitude 5 or above will occur in those randomly-selected cells.

2. To take spatial heterogeneity into account, for the jth cell, toss a p_j-coin, where p_j is proportional to the historical rate of seismicity in that cell. If the jth coin lands heads, predict that an earthquake of magnitude 5 or above will occur in the jth cell. Toss coins independently for all cells, $j = 1, \ldots, 3743$. The constant of proportionality is the ratio of the number of cells for which the actual predictions anticipate events, divided by the historical annual average number of cells in which events occur. This produces a random number of predictions, with predictions more likely in cells where more events occurred in the past.

3. Condition on the number of cells in which earthquakes are predicted to occur. Choose that many cells at random without replacement from the 3,743 cells. Instead of selecting cells with equal probability, select the jth cell with probability p_j, with p_j set as in (2.2). Predict that earthquakes of magnitude 5 or above will occur in those randomly-selected cells.

The third approach is a blend of the first two approaches: the number of simulated predictions each year is forced to equal the actual number of predictions, but the chance of raising a prediction in the jth cell depends on the historical rate of seismicity in the jth cell. None of these three comparison methods depends on the observed seismicity during the study period, 1990–1998. In particular, none exploits clustering, which is presumed to have been eliminated from the catalog.

2.4 Some claims of successful predictions

2.4.1 Wyss and Burford, 1987

Wyss and Burford [41] claim to have predicted the magnitude $M_L = 4.6$ earthquake that occurred on 31 May 1986 near Stone Canyon, California, about a year before it occurred, using "seismic quiescence," an anomalous paucity of earthquakes over some period of time. They examine the rates of earthquakes on different sections of the San Andreas fault and identify two fault sections in which the rate dropped compared with the rates in neighboring sections. They say that "the probability [of the prediction] to have come true by chance is $< 5\%$." The probability they calculate is the chance that an earthquake would occur in the alarm region, if earthquakes occurred at random, independently, uniformly in space and time, with rate equal to the historic rate in the study area over the previous decade. That is, their null hypothesis is that seismicity follows a homogeneous Poisson process with rate equal to the historical rate; clustering is not taken into account.

2.4.2 VAN predictions based on Seismic Electrical Signals

There has been a lively debate in the literature about whether predictions made by Varotsos, Alexopoulos and Nomicos (VAN) [26] of earthquakes in Greece succeeded beyond chance. See volume 23 of *Geophysical Research Letters* (1996). The participants did not even agree about the number of earthquakes that were predicted successfully, much less whether the number of successes was surprising. Participants disagreed about whether the predictions were too vague to be considered predictions, whether some aspects of the predictions were adjusted post hoc, what the null hypothesis should be, and what tests were appropriate.

2.4.3 Kossobokov et al., 1999

Kossobokov, Romashkova, Keilis-Borok and Healy [42] claim to have predicted four of the five magnitude 8 and larger earthquakes that occurred in the circum-Pacific region between 1992 and 1997. They say "[t]he statistical significance of the achieved results is beyond 99%." (From context, it is clear that they mean that the P-value is $< 1\%$.) Their predictions are based on two algorithms, M8 and MSc, which track the running mean of the number of main shocks; the difference between the cumulative number of main shocks and a windowed trend in the number of main shocks; a measure of spatial clustering of main shocks derived from the distance between shocks and the diameters of the sources in a temporal window; and the largest number of aftershocks of any event in a temporal window. These are used to identify "times of increased probability," which are predictions that last five years.

The declustering method described above was used to classify events as main shocks or aftershocks.

Kossobokov et al. [42] calculate statistical significance by assuming that earthquakes follow a Poisson process that is homogeneous in time but heterogeneous in space, with an intensity estimated from the historical rates of seismicity, $\{\lambda_j\}$. Kossobokov et al. [42] condition on the number of events that occur in the study area, which leads to a calculation in which locations and times are iid across events, the epicenters and times are independent of each other, the temporal density of earthquake times is uniform, and the spatial distribution of epicenters is given by the historical distribution between 1992 and 1997. Their calculation does not take temporal clustering into account, and it conditions on the predictions. They calculate the chance that S or more of the Q events would occur during alarms to be $\sum_{x=S}^{Q} {}_Q C_x \pi^x (1-\pi)^{S-x}$, where π is the normalised measure of the union of the alarms. The measure is the product of the uniform measure on time and counting measure on space, using the historical distribution of epicenters in the study volume to define the counting measure.

2.5 A naive predictor

In this section we exploit the empirical clustering of earthquakes in time to construct a predictor that succeeds far beyond chance according to tests that hold predictions fixed and treat seismicity as random. The chance model for seismicity uses the observed times and locations of earthquakes, but shuffles the times: according to the null hypothesis, the times of events are exchangeable given their locations and magnitudes. However, the predictions are the same for all elements of the null. We can simulate from the null model by randomly permuting the list of observed times relative to the list of observed locations and magnitudes [46, 63]. That is, if the locations, magnitudes, and times of the events in the catalog are $\{(r_j, M_j, t_j)\}_{j=1}^{Q}$, we take the $Q!$ outcomes $\{(r_j, M_j, t_{\pi(j)})\}_{j=1}^{Q}$ (as π ranges over all $Q!$ permutations of $\{1, \ldots, Q\}$) to be equally likely under the null hypothesis, given the predictions. We do not claim that this predictor or this null hypothesis is good. Rather, we claim that this approach to testing is misleading.

We apply the approach to the Global Centroid Moment Tensor (CMT) catalog[12] for 2004 and for 2000–2004. We make two sets of predictions:

(i) After each earthquake of (body-wave) magnitude M_τ or greater, predict that there will be another earthquake of magnitude M_τ or greater within 21 days, and within 50 km epicentral distance.

[12] http://www.globalcmt.org/CMTsearch.html

(ii) After each earthquake of magnitude M_τ or greater, predict that there will be an earthquake within 21 days and within 50 km that is at least as large as any within 50 km within 21 days prior to its occurrence.

Predictor (ii) is equivalent to predictor (i) if an event is deemed eligible for prediction only if there is no larger event within 50 km in the 21 days leading up to the event.

Let M_j be the magnitude of the jth event that triggers an alarm; let t_j be the time of the jth event that triggers an alarm; and let R_j be the set of points on Earth's surface that are within 50 km of the epicenter of the jth event that triggers an alarm. Recall that an alarm is a connected region V_j of space, time, and magnitude. For predictor (i),

$$V_j = R_j \times [t_j, t_j + 21 \text{ days}] \times [M_\tau, \infty), \tag{2.6}$$

while for predictor (ii),

$$V_j = \{R_j \times [t_j, t_j + 21 \text{ days}] \times [M_j, \infty)\} \setminus \bigcup_{k:M_k > M_j} \{R_k \times [t_k, t_k + 21 \text{ days}] \times [M_k, \infty)\}. \tag{2.7}$$

An event at time t and epicenter r with magnitude M is predicted by the second set of predictions if and only if

$$M \in \bigcap_j \{[M_j, \infty) : (t, r) \in [t_j, t_j + 21 \text{ days}] \times R_j\}. \tag{2.8}$$

Predictor (i) tends to predict more aftershocks: large events trigger alarms that contain some of their aftershocks. Predictor (ii) prevents aftershocks from being predicted by main shocks; however, it does not prevent aftershocks with magnitude M_τ or larger from predicting still larger aftershocks of the same main event, provided the predicted aftershock is the largest event in the preceding 21 days, within 50 km. Predictors (i) and (ii) generate the same number of alarms and have the same total duration, but not the same extent of space and magnitude. Note that these alarms need not be disjoint.

We consider two values of the threshold magnitude M_τ: 5.5 and 5.8. We compare the number of events successfully predicted by these two predictors with the distribution of the number that would be predicted successfully if seismicity were "random." Using the CMT catalog, we generate a set of alarms. Holding those alarms fixed, we see how successful the alarms would be in predicting random seismicity—generated by randomly permuting the times in the CMT catalog.

Table 2.1 summarises the results.[13] Under the null hypothesis, both prediction methods (i and ii) succeed well beyond chance for CMT data from the year 2004

[13]These simulations differ from the simulations in the previously published version of the paper. In the latter, the naive predictions were not fixed; instead, they depended on the simulated seismicity.

and the years 2000–2004, for both values of the threshold magnitude. That is not because the prediction method is good; rather, it is because the stochastic model in the null hypothesis fails to take into account the empirical clustering of earthquakes and the dependence of the predictions on the seismicity. Holding predictions fixed as seismicity varies randomly does not make sense.

2.6 Discussion

Interpreting earthquake predictions is difficult. So is evaluating whether predictions work. To use a statistical hypothesis test, something must be assumed to be random, and its probability distribution under the null hypothesis must be known. Many studies model seismicity as random under the null hypothesis. That approach has serious drawbacks, and details of the stochastic model, such as spatial heterogeneity, independence or exchangeability, matter for testing. Most null hypotheses used in tests ignore the empirical clustering of earthquakes. Some try to remove clustering with ad hoc adjustments as a prelude to probability calculations. It is often assumed that the resulting data represent a realisation of a Poisson process. In the next chapter, we show this is implausible. The standard approach to testing—hold the predictions fixed while seismicity varies randomly according to some stochastic model—does not take into account that in practice, the predictions would be different if the seismicity were different. The result is that simple-minded schemes, such as the "automatic alarm strategy," succeed well beyond chance in hypothesis tests. This is not because the predictions are good: it is because the tests are bad.

year	M_τ	events	succ	succ w/o	max sim	P-value (est)	τ
2004	5.5	445	95	30	20	$< 10^{-4}$	4.0×10^{-4}
2004	5.8	207	24	7	7	0.002	1.8×10^{-4}
2000–2004	5.5	2012	320	85	40	$< 10^{-4}$	3.6×10^{-4}
2000–2004	5.8	995	114	29	18	$< 10^{-4}$	1.8×10^{-4}

Table 2.1: Simulation results using the Global Centroid Moment Tensor (CMT) catalog. We seek to predict events with body-wave magnitude M_τ and above. "Events" is the total number of events in the time period with magnitude at least M_τ. Each event with body-wave magnitude M_τ or greater triggers an alarm. In each row, the number of alarms is equal to the number of events in column 3. The spatial extent of the alarm is a spherical cap of radius 50 km centred at the epicenter of the event that triggers the alarm. The temporal extent of the alarm is 21 days, starting at the time of the event that triggers the alarm. We set the magnitude extent of alarms in two ways. Column 4, 'succ,' is the number of successful predictions using predictor (i): it is the number of events with magnitude at least M_τ that are within 21 days following and within 50 km of the epicenter of an event with magnitude M_τ or greater. Column 5, 'succ w/o,' is the number of successful predictions using predictor (ii): it is the number of events that are within 21 days following and within 50 km of the epicenter of an event whose magnitude is at least M_τ but no greater than that of the event in question. Events that follow within 21 days of a larger event are not counted; this is intended to reduce the number of predictions satisfied by aftershocks. Column 6, 'max sim,' is the largest number of successful predictions in 10,000 random permutations of the times of the events in the Global CMT catalog, holding the alarms and the locations and magnitudes of events in the catalog fixed. The alarms are those corresponding to column 5—predictor (ii) in the text—that is, an event is eligible for prediction only if its magnitude exceeds that of every event within 50 km within the 21 days preceding it. Column 7, 'P-value (est),' is the estimated P-value for predictor (ii): the fraction of permutations in which the number of successful predictions was greater than or equal to the observed number of successful predictions for the CMT catalog. Column 8, 'τ,' is an upper bound on the fraction of the study region (in space and time) covered by alarms; it is not adjusted for overlap of alarms.

Chapter 3

Are declustered earthquake catalogs Poisson?

3.1 Overview

Earthquake catalogs are highly clustered in space and time. Some seismologists have claimed they are able to process catalogs by thinning events from clusters so that the declustered catalog is Poisson. However, the tests of the Poisson null hypothesis that they use are weak. In this chapter, we perform tests that reject the hypothesis that the declustered catalogs are Poisson. In fact, declustered catalogs may not even have exchangeable times.

3.1.1 Earthquakes cluster in space and time

Earthquakes do not occur uniformly in space or in time. Even along a single fault, some subregions are more active seismically than others. In a spatially and temporally homogeneous Poisson process, the expected rate is constant in space and time. While realisations of a homogeneous Poisson process will show some heterogeneity, observed data are far more heterogeneous in space and time than is likely in such a process.

To model spatial heterogeneity, some seismologists [37] have fitted spatially heterogeneous, temporally homogeneous Poisson processes to seismicity. The rate of events in any region is estimated from the historical rate of events, smoothed geographically. Such models can account for the spatial clustering of earthquakes, but not the temporal clustering. Temporally heterogeneous Poisson models are unhelpful, since we do not know in advance which times will have higher rates of events.

Many models for seismicity include clustering [74], either explicitly, like the ETAS model (discussed in detail in chapter 4), or implicitly, like some renewal process models.

3.1.2 Declustering to fit simple models

Instead of modelling the clustered seismicity, a different approach is to delete some events from the catalog and fit a temporally homogeneous stochastic model to the remaining events. The thinning of clusters is known as *declustering*. We refer to catalogs yet to be declustered as *raw catalogs*, and to catalogs from which clustered events have been removed as *declustered catalogs*. Declustered catalogs are often used to estimate "background" rates of seismicity—usually meaning the rate of "main shocks," vaguely defined [75, 76].

Declustering methods may be divided into several classes [77, 76]: main shock window methods, linked-window methods, stochastic declustering methods, and others. Main shock window methods remove the earthquakes in a space-time window around every main shock, where the definition of "main shock" varies from method to method. These methods can be thought of as "punching holes" adjacent to main shocks in the catalog. Gardner and Knopoff [75, 2] created a widely-used set of rules to determine window sizes, with larger main shocks having larger windows in space and time. We examine declustering using the Gardner-Knopoff windows in section 3.2.1.

Linked-window methods calculate a space-time window for every event in the catalog—not just for main shocks. An event is included in a cluster if and only if it falls within the window of at least one other event in that cluster. For example, suppose that earthquake B occurred in the window of earthquake A and earthquake C occurred in the window of earthquake B. Then earthquakes A, B, and C are all in the same cluster, regardless of whether earthquake C was within the window of earthquake A. After clusters are determined, a declustered catalog is created by reducing every cluster to a single event—for instance, by removing all events in a cluster except the first, by removing all events except the largest, or by replacing the whole cluster with a single "equivalent event" with magnitude representing the summed moment of all events in the cluster. The most widely used linked-window method is that of Reasenberg [67], discussed in section 3.2.2.

Stochastic declustering methods employ a random element to decide whether to remove a particular shock. Applying such a method twice to the same raw catalog may not produce the same declustered catalog both times. The best-known method of this type was introduced in 2002 by Zhuang, Ogata, and Vere-Jones [76], though the underlying ideas date back several decades [74]. We discuss this method in section 3.2.4.

Other methods, such as the "waveform similarity approach" of Barani, Ferretti, Massa, and Spallarossa [78], do not fit into any of the three classes described above; we do not consider them here.

3.1.3 Are declustered catalogs Poisson?

It has often been claimed that declustered catalogs are "Poissonian" [2, 5]. The basis for the claim is that a test of the hypothesis that the times follow a Poisson process does not reject that hypothesis. The test that has been used ignores the spatial locations of events and largely ignores the temporal order of events: it partitions the catalog period into (arbitrary) time intervals, counts the number of intervals with k events (up to some arbitrary maximum), then performs a chi-square test comparing these counts with those expected if seismicity were a realisation of a temporally homogeneous Poisson process (with rate estimated from the data) [2, 78]. This test is approximate and ad hoc. It has limited power against many alternatives, and ignores space.

If seismicity follows a Poisson process, then, conditional on the number of events that occur, the times of those events are independent, identically distributed (iid) uniform random variables. This condition can be tested directly, without estimating a rate or making arbitrary choices of intervals and bins, using a Kolmogorov-Smirnov test. This exact test does not require parameter estimation or arbitrary choices of intervals or bins—it retains all the temporal information about the events. It is more powerful than the chi-square test against many, but not all, plausible alternatives. However, like the chi-square test, it ignores the spatial locations of events. Section 3.3 tests the hypothesis that event times in catalogs of Southern Californian seismicity declustered using Gardner-Knopoff windows are Poisson. A portmanteau chi-square and Kolmogorov-Smirnov test rejects at level 0.05: we would conclude that the declustered catalogs are not Poisson in time.

What about spatio-temporal processes? We know a priori that catalogs declustered using window methods cannot be Poisson in space-time. If two events are very close in space-time in a raw catalog, at least one will be deleted by the declustering. In a Poisson process, two events may occur arbitrarily close to one another in space and time.

However, the declustered catalogs may have some of the simple properties of Poisson processes. In a temporally homogeneous Poisson process, the times are iid uniform. This implies a weaker condition: that conditional on the locations of events, the process has *exchangeable* times. This means that, given a set of n catalog locations and n catalog times, all $n!$ assignments of the times to the locations are equally likely.

In sections 3.4 and 3.5, we use methods based on the work of Romano [79, 80] to test whether, conditional on the locations and times of the events in declustered catalogs, all permutations of the times are equally likely. This is a much weaker hypothesis than the hypothesis that events follow a spatially heterogeneous, temporally homogeneous Poisson process. We performed this test on three declustered catalogs of Southern Californian earthquakes. One catalog was declustered using Reasenberg's method. The second used Gardner-Knopoff windows in a main shock window method; the last used Gardner-Knopoff windows in a linked-window method. The estimated

M	$L(M)$ (km)	$T(M)$ (days)
2.5	19.5	6
3.0	22.5	11.5
3.5	26	22
4.0	30	42
4.5	35	83
5.0	40	155
5.5	47	290
6.0	54	510
6.5	61	790
7.0	70	915
7.5	81	960
8.0	94	985

Table 3.1: Window radius and duration as functions of magnitude, as given by Gardner and Knopoff [2]. For an event of magnitude M, the radius is $L(M)$ and the duration is $T(M)$. For values of M falling between values given in the table, the sizes of the window are linearly interpolated.

P-value for the test of the catalog declustered using Reasenberg's method was 0.003. The estimated P-values for the tests of the catalogs declustered using the Gardner-Knopoff windows were 0.025 and 0.069 for the main shock window method and the linked-window method respectively. Detailed test results are given in section 3.5.

3.2 Declustering methods

3.2.1 Gardner-Knopoff windows

Gardner and Knopoff [75, 2] gave radii and durations for windows as increasing functions of magnitude. These windows can be used in a main shock window declustering algorithm [5, 76, 81] or in a linked-window algorithm [77].

Table 3.1 gives window sizes from their 1974 paper, which superseded those given in their 1972 paper. For magnitudes not given in the table, the sizes are linearly interpolated. Gardner and Knopoff stated they did not have "any strong affection for these particular windows," which they found by visually scanning for clusters in a catalog of Southern Californian earthquakes, then fitting functions relating duration and radius to main shock magnitude.

There are a number of ways to use a set of windows to decluster catalogs. The following list is not exhaustive.

- **Method 1:** Remove every event that occurred in the window of some other event [2].
- **Method 2:** Divide the catalog into clusters as follows: include an event in a cluster if and only if it occurred within the window of at least one other event in the cluster. In every cluster, remove all events except the largest [2].
- **Method 3:** Consider the events in chronological order. If the ith event falls within the window of a preceding larger shock that has not already been deleted, delete it. If a larger shock falls within the window of the ith event, delete the ith event. Otherwise, retain the ith event [75].

Methods 1 and 2 are linked-window methods. Gardner and Knopoff found that using Methods 1 and 2 with their windows produced "remarkably similar" declustered catalogs when applied to a raw Southern Californian catalog [2]. Method 3 is a main shock window method.

3.2.2 Reasenberg declustering

The most widely used linked-window declustering method is that of Reasenberg [67]. Reasenberg claimed that Gardner-Knopoff windows are excessively large and that declustering using these windows removes too many events. He gave his own windows, called "interaction zones." The name emphasises that they are based on physics.

We give the formulae for Reasenberg's windows, and the physical assumptions behind them, momentarily. For now, suppose we have calculated a window for every earthquake in a raw catalog. If one earthquake is followed by a second earthquake within the window of the first, the two events are considered to be related—that is, in the same cluster. If a third event is in the window of either the first event or the second event, it belongs to the same cluster. And so on. An event in the window of any prior event joins the cluster to which that prior event belongs.

Like the Gardner-Knopoff windows, the Reasenberg windows are spatially larger for events with greater magnitude. Unlike the Gardner-Knopoff windows, the Reasenberg windows are temporally *smaller* for events with greater magnitude. However, the space-time size of a cluster with many events may be much larger than the space-time size of the window of any event in the cluster.

The following subsections derive formulae for the spatial and temporal extent of windows.

Spatial extent of Reasenberg windows

Reasenberg estimated the *source dimension* of an event as the radius of a circular fault that would generate the same seismic moment as that event, assuming a con-

stant stress drop of 30 bars. In the public version of Reasenberg's code,[1] the source dimension in kilometres for an event of magnitude M is modelled as

$$r(M) = 0.011 \times 10^{0.4M}. \tag{3.1}$$

This radius is capped at 30 km.[2]

The windows are calculated in chronological order, event by event. Before we find the window of the ith event, we must find which, if any, existing cluster the event belongs to by determining whether it falls within the window of any previous event. Suppose the ith event is in cluster J. Let M_i be the magnitude of the ith event, and $M^*_{J,i}$ be the magnitude of the largest event in the Jth cluster up to and including the time of the ith event. (If the ith event does not fall in the window of any previous event, it is in a new cluster, and $M^*_{J,i} = M_i$.) Reasenberg's method uses hypocentral distance (the Gardner-Knopoff method uses epicentral distance). The window of the ith event is spatially spherical with radius

$$Qr(M_i) + r(M^*_{J,i}), \tag{3.2}$$

where Q is a constant. Q is typically taken to be 10. This is an estimate of the maximum distance over which "stress-relieving processes" such as afterslip can act. For $Q = 10$, Reasenberg's windows are generally smaller in epicentral extent than the Gardner-Knopoff windows, especially for low magnitude earthquakes not in the same cluster as a large main shock.

Reasenberg acknowledged the physics are oversimplified in this window calculation. The geometry of faults is neither spherically nor circularly symmetric—it varies from fault to fault. It is also not clear what value the proportionality constant Q should take, though Reasenberg contended that both the number of clusters identified and the number of earthquakes in each cluster were "remarkably insensitive" to choices of Q. Nor is it clear that the form of equation (3.2) is a good model for the maximum distance over which stress-relieving processes can act.

Temporal extent of Reasenberg windows

Let t be the time after a particular main shock. Suppose the raw catalog consists of events above magnitude $M_{\min}$, and is complete after time t_0 has elapsed since the last main shock. Assume that the rate of aftershocks with magnitude at least $M_{\min}$ at time t after the main shock, occurring close in space to the main shock, is

[1] ftp://ehzftp.wr.usgs.gov/cluster2000/cluster2000x.f

[2] For the physics connecting the radius of a circular fault and the magnitude of an earthquake, see Kanamori and Anderson [82]. Alternative expressions for the fault radius are sometimes used. For example, Helmstetter, Kagan, and Jackson [83], following the empirical work of Wells and Coppersmith [84], used an uncapped radius of $r(M) = 0.01 \times 10^{0.5M}$.

approximately

$$\frac{\delta n}{\delta t} \approx \frac{C}{t}, \text{ for } t > t_0, \tag{3.3}$$

where C is a function of $M_{\max}$, the magnitude of the main shock, and of $M_{\min}$. Equation (3.3) is Omori's law, a well-established empirical relationship between the rate of aftershocks and the time after the main shock. We discuss Omori's law further in chapter 4. Reasenberg stated that C was approximately empirically related to $M_{\max}$ and $M_{\min}$ as

$$C = 10^{2(\Delta M - 1)/3}, \tag{3.4}$$

where $\Delta M \equiv M_{\max} - M_{\min}$. The expected number of events between times t_i and T is then

$$\begin{aligned} n(t_i, T) &\equiv 10^{2(\Delta M-1)/3} \int_{t_i}^{T} \frac{1}{t} dt \\ &= 10^{2(\Delta M-1)/3} \log\left(\frac{T}{t_i}\right). \end{aligned}$$

Reasenberg modelled each aftershock sequence as a time-dependent Poisson process with rate as given in equation (3.3) and C as given in equation (3.4). Suppose we have observed an aftershock at time t_i after a main shock. Under Reasenberg's assumptions, the probability of observing at least one event in the interval $(t_i, t_i + \tau)$ is

$$1 - \exp\left[-n(t_i, t_i + \tau)\right] = 1 - \left(\frac{t_i + \tau}{t_i}\right) \exp\left[-10^{2(\Delta M-1)/3}\right]. \tag{3.5}$$

Reasenberg claimed this probability is

$$P = 1 - \exp\left[-10^{2(\Delta M-1)/3}\left(\frac{\tau}{t_i}\right)\right]. \tag{3.6}$$

The right-hand sides of (3.5) and (3.6) are approximately equal when τ is small compared to t_i. Rearranging equation (3.6) to make τ the subject,

$$\tau = \frac{-\ln(1-P)}{10^{2(\Delta M-1)/3}} t_i. \tag{3.7}$$

Reasenberg concludes that if an earthquake occurs at a time t_i after the main shock, the chance that the next event in the sequence will occur in the interval $(t_i, t_i + \tau]$ is P.

The temporal duration of the window of the event at time t_i is

$$\tau' \equiv \begin{cases} \tau_{\min}, & \text{if } \tau < \tau_{\min} \\ \tau, & \text{if } \tau_{\min} \leq \tau \leq \tau_{\max} \\ \tau_{\max}, & \text{if } \tau > \tau_{\max}. \end{cases} \tag{3.8}$$

Reasenberg used the values $\tau_{\min} = 1$ day and $\tau_{\max} = 10$ days. In contrast, Gardner-Knopoff windows may be hundreds of days long.

To summarise, in Reasenberg's declustering scheme, the window of an event is spatially spherical with radius given by equation (3.1), and lasts a duration τ' after the event. If one earthquake falls in the window of another, both are in the same cluster.

After the clusters are identified, each is replaced with a single *equivalent event*. This event has the following properties:

- Its time is the time of the largest event in the cluster.
- Its seismic moment is the sum of the moments of all the events in the cluster. That is, moment magnitudes are converted to seismic moments and summed; then the sum is converted to a moment magnitude, as in equation (1.1).
- Its hypocenter is the unweighted centroid of the hypocenters of the events in the cluster.

The magnitude threshold below which a catalog is incomplete is often higher following large main shocks than it is during periods with no large events. The current version of the Reasenberg algorithm allows the minimum magnitude cut-off inside a cluster to depend on the size of the main shock [83]. The cut-off is governed by two parameters: x_{meff}, the magnitude cut-off outside of clusters, and x_k. Inside a cluster, the cut-off is

$$x_{\text{meff}} + x_k M_{\max}. \tag{3.9}$$

Code for Reasenberg declustering is freely available online (see footnote 1). The implementation in the MATLAB package ZMAP allows eight parameters to be set:

- $\tau_{\min}$ (default 1 day)
- $\tau_{\max}$ (default 10 days)
- P: the probability in equation (3.6) (default 0.95)
- x_{meff}: minimum magnitude cut-off outside of clusters (default 1.5)
- x_k: coefficient for main shock magnitude in equation (3.9) (default 0.5)

- r_{fact}: same as Q in equation (3.2) (default 10)
- epicentral error, assumed to be the same for all events (default 1.5 km)
- depth error, assumed to be the same for all events (default 2 km)

If epicentral or depth errors are specified, they are converted to a hypocentral error and the spatial radius of every window is reduced by twice that distance. Changing any of these parameters may result in a different declustered catalog.

Reasenberg's declustering algorithm is complex, requiring many parameter choices, and, if it is to be justified physically, many assumptions. The spatial radius of the window of an event approximates the maximum hypocentral distance at which an aftershock caused by that event can occur. This radius depends heavily on an arbitrary parameter Q. The temporal duration of a window is chosen so that if a host of stylised assumptions hold, the next event in that cluster will occur within the window 95% of the time. In fact, there is empirical evidence the assumptions do not hold. For example, times of aftershocks are not a realisation of a Poisson process with rate given by Omori's law (see chapter 4).

3.2.3 Comparison of windows

Davis and Frohlich [77] evaluated five window declustering schemes. Three of these—their own, that of Reasenberg [67] and that of Shlien-Toksöz [85]—were linked-window methods. The two others — that of Gardner-Knopoff [2] and that of Knopoff-Kagan-Knopoff [86]—could be applied either as main shock window methods or as linked-window methods. In the Knopoff-Kagan-Knopoff method, only events with magnitude greater than some minimum threshold are eligible to be classified as main shocks. The window sizes do not depend on magnitude in the Shlien-Toksöz and Davis-Frohlich methods, but do in the other three methods.

Davis and Frohlich created synthetic raw catalogs by simulating main shocks and aftershocks using a method they devised. They examined the performance of the declustering schemes when applied to these catalogs. The methods were scored for correct and incorrect identification of the aftershocks. The Gardner-Knopoff and Knopoff-Kagan-Knopoff methods scored higher as linked-window methods than as main shock window methods. While the scores of all the linked-window methods were similar, the Shlien-Toksöz and Davis-Frohlich methods scored marginally higher than the others. This success, it should be noted, was for simulations, and may not reflect performance in reality.

3.2.4 Stochastic declustering

Main shock window and linked-window methods make arbitrary choices for the sizes of windows. Different choices give different declustered catalogs and different

estimates of the background seismicity.

An alternative is *stochastic declustering*. "Stochastic" indicates a random element is used to determine which events to delete. Many stochastic declustering methods fit a model that gives for each event in the catalog a likelihood of being a background event. The following algorithm is common:

1. Set $j = 1$.
2. Using the fitted model, find the likelihood ϕ_j that the jth event in the raw catalog is a background event.
3. Generate a uniform random number U_j in $[0, 1]$. These are independent for all j.
4. If $U_j < \phi_j$, label the jth event as a background event and retain it. Otherwise consider it an offspring event and delete it.
5. If $j = N$, the set of retained events is the declustered catalog; stop. Otherwise, set $j = j + 1$ and go to step 2.

Zhuang, Ogata, and Vere-Jones [76] proposed a declustering method that fits a branching process model called the epidemic-type aftershock (ETAS) model [87] to raw catalogs. The ETAS model distinguishes between two types of event. A *background event* is not directly caused by previous earthquakes. An *offspring event* is "triggered" by exactly one preceding event. Both background and offspring events may trigger offspring. The goal of declustering is to delete offspring events, but not background events.

The ETAS model is the subject of chapter 4; equation (4.4) gives its standard functional form. For now, suppose we have fitted the ETAS model to a raw catalog with n events.

Marsan and Lengliné [88] proposed a similar method that fits a nonparametric Hawkes branching process to a raw catalog. (The ETAS model is a special case of a Hawkes process. See chapter 4.2.1 for further discussion of Hawkes processes.)

The RELM group [89] also used a declustering scheme that included stochastic elements. They independently sampled parameters for Reasenberg declustering from prior distributions. For example, $\tau_{\min}$ was sampled uniformly from the interval $[0.5, 2.5]$; $\tau_{\max}$ was sampled uniformly from the interval $[3, 15]$. They then declustered a raw catalog using the sampled parameters. For each raw catalog, they independently repeated the sampling and declustering 10,000 times. The 10,000 declustered catalogs were then combined into a single probabilistic catalog. For each event, the probability of being a main shock was equal to the proportion of the 10,000 declustered catalogs in which it was retained. This method avoids the problem of using one arbitrary parameter combination but creates a larger problem by using an arbitrary joint prior for the parameters.

3.3 Tests for homogeneous Poisson times

The title of Gardner and Knopoff's 1974 paper was "Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?" The abstract, in its entirety, was "Yes." To show this, they used a chi-square test, which we describe in section 3.3.1. In this problem, the assumptions of a standard chi-square test are not satisfied. Moreover, the test has little power against plausible alternatives. The Kolmogorov-Smirnov test, described in section 3.3.2, can be also used to test the hypothesis that seismicity follows a temporally homogeneous Poisson process. The assumptions of the Kolmogorov-Smirnov test are satisfied. While the test does not have good power against all alternatives, it has better power than the chi-square test against many plausible alternatives.

Both the chi-square test and Kolmogorov-Smirnov test use only the sequence of declustered catalog times $\{t_1, \ldots, t_n\}$, ignoring the spatial locations of events. We give a spatio-temporal test in section 3.4.2.

3.3.1 Chi-square test

Both Gardner and Knopoff [2] and Barani, Ferretti, Massa, and Spallarossa [78] performed chi-square tests of the hypothesis that catalogs they declustered were "Poissonian." Details were not stated completely, but we believe the tests proceed as follows.

1. Partition the study period into K time intervals of length T/K. The choice of K (or, equivalently, the choice of the length of time intervals) is ad hoc.

2. For $k \in \{1, \ldots, K\}$, count the number of events in the Kth interval:

$$N_k \equiv \sum_{i=1}^{n} \mathbf{1}((k-1)T/K < t_i \leq kT/K). \tag{3.10}$$

3. Pick $B > 2$, the number of "bins." The choice of B is ad hoc. For $b \in \{0, \ldots, B-2\}$, count the number of intervals containing b events:

$$O_b \equiv \sum_{k=1}^{K} \mathbf{1}(N_k = b). \tag{3.11}$$

Also count the number of intervals with $B-1$ or more events:

$$O_{B-1} \equiv \sum_{k=1}^{K} \mathbf{1}(N_k \geq B-1). \tag{3.12}$$

4. Estimate the rate of events per interval for the Poisson process. One simple estimate is
$$\hat{\lambda} = n/K. \tag{3.13}$$
As we show below, this is not necessarily the best estimate for the chi-square test.

5. Assuming $\bar{\lambda}$ is the true rate, calculate the expected number of intervals with b events for $b \in \{0, \ldots, B-2\}$:
$$E_b \equiv K e^{-\hat{\lambda}} \frac{\hat{\lambda}^b}{b!}. \tag{3.14}$$
Find the expected number of intervals with $B-1$ or more events:
$$E_{B-1} \equiv K - \sum_{b=0}^{B-2} E_b. \tag{3.15}$$

6. Calculate the chi-square statistic:
$$\chi^2 \equiv \sum_{b=0}^{B-1} \frac{(O_b - E_b)^2}{E_b}. \tag{3.16}$$
Calculate the (approximate) P-value by finding the corresponding quantile of a chi-square distribution with d degrees of freedom:
$$P \equiv 1 - \frac{\gamma(d/2, \chi^2/2)}{\Gamma(d/2)}. \tag{3.17}$$
The choice of d is discussed below. Note that if E_b is too small for some value of b, the approximation in equation (3.17) will be poor.

The results of the test can depend on choices K and B and on the method of estimating λ.

Gardner and Knopoff performed the chi-square test on a number of declustered catalogs. For instance, they tested a declustered catalog of earthquakes with magnitude at least 3.8 occurring in the Southern Californian Local Area from 1932 to 1971. The raw catalog had 1751 events; the declustered catalog had 503 events. They divided the forty-year period into ten-day intervals, and counted the number of events in each interval. They found the number of intervals with b events for all b. They found the chi-square statistic, and compared this to a chi-square distribution with 2 degrees of freedom.

They did not explicitly state the number of bins they used; we are uncertain whether they used $d = B - 1$ or $d = B - 2$. Heuristically, one degree of freedom is lost in estimating λ. From our simulations below, we believe that $d = B - 2$ gives the better approximation.

Does the chi-square approximation hold?

Under the null hypothesis of the simple chi-square test [90] for goodness of fit, the probability of an interval falling in bin b is known and identical for all K intervals, and the numbers of events in the intervals are independent. The distribution of counts in the B bins is multinomial. The asymptotic distribution of the test statistic is chi-square with $B - 1$ degrees of freedom.

The test here departs from the assumptions of the simple chi-square test. The bin probabilities are not known. Instead, the rate parameter of the Poisson process is estimated, and the bin probabilities calculated from this. However, conditional on the estimate of the rate, the numbers of events in the intervals are no longer independent, and the distribution of bin counts is not multinomial.

In the case where the bin probabilities are given by a p-parameter distribution, and the bin probabilities are calculated from an efficient estimator based on the likelihood of the bin counts, the test statistic has an asymptotic chi-square distribution with $B - p - 1$ degrees of freedom [90]. The estimate $\hat{\lambda} = n/K$ cannot be derived from the bin counts alone, so that if this is the estimate Gardner and Knopoff use in their test, it is not apparent that the chi-square statistic has an asymptotic chi-square distribution. The MLE based on the bin counts solves

$$\frac{\sum_{b=1}^{B-2} bO_b}{\lambda} + \frac{\sum_{i=B-2}^{\infty} \lambda^i/i!}{\sum_{j=B-1}^{\infty} \lambda^j/j!} O_{b-1} = K. \tag{3.18}$$

However, this must be solved numerically.

We simulated 10^6 independent realisations of a rate 1 Poisson process over a period of 500 time intervals, then counted the number of intervals with 0, 1, 2, or 3 or more events. For every realisation, we performed the above chi-square test with four bins, estimating the rate as $\hat{\lambda} = n/K$. The null hypothesis is true, so a level 0.05 test should reject 5% of the tests. In fact, a chi-square test with 2 degrees of freedom rejected 5.01% of the time. In contrast, a chi-square test with 3 degrees of freedom rejected 2.02% of the time. Similar tests with different rates and different numbers of bins confirm that setting the degrees of freedom to $B-2$ is the better approximation. However, we only simulated a limited variety of processes.

Power of the chi-square test

Even if the chi-square test has the correct asymptotic level, it has low power against many plausible alternatives. To see this, suppose all the intervals with events occurred at the beginning of the study period. Under the null, this would be very unlikely. However, because the chi-square test does not account for the order of the intervals, it will not reject if the observed counts are close to the expected counts for each bin. The Kolmogorov-Smirnov test, described in the following subsection, has more power against this and similar inconsistencies with the Poisson times hypothesis.

Furthermore, the chi-square test does not incorporate the spatial locations of the events. Even if the sequence of times is consistent with a temporally homogeneous Poisson process, the set of locations and times is not necessarily consistent with a spatially heterogenous, temporally homogeneous Poisson process: while the time distribution of events over the entire region may appear uniform, events in small subregions may be more clustered or more dispersed in time than is likely under a temporally homogeneous Poisson model. See section 3.5.1 for a trivial example of a process that is homogeneous in time but heterogeneous in space-time.

3.3.2 Kolmogorov-Smirnov tests

The Kolmogorov-Smirnov test compares the empirical distribution of the times of a random variable to a specified reference distribution function $F(x)$. The test rejects when a statistic quantifying the "distance" between $F(x)$ and the empirical distribution function $F_n(x)$ is large. The Kolmogorov-Smirnov (K-S) statistic is

$$D_n \equiv \sup_x |F_n(x) - F(x)| . \tag{3.19}$$

The Kolmogorov-Smirnov test has been used by Reasenberg and Matthews [91, 92] as part of a suite of tests (also including the Cramer-von Mises and Anderson-Darling tests and a test based on the difference in rate between intervals and their complements) of uniformity of declustered earthquake sequences preceding main shocks. If the tests are generated by a homogeneous Poisson process, then conditional on n, they are drawn independently from a uniform distribution on $(0, T]$. So $F(x) = t/T$, and the K-S statistic is

$$D_n = \sup_x \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(t_i \le t) - t \right| . \tag{3.20}$$

This statistic requires neither parameter estimation nor the partition of time into intervals. The test has asymptotic power 1 against any fixed alternative that the data are iid $F' \neq F$.

The Dvoretzky-Kiefer-Wolfowitz inequality, as tightened by Massart [93], states

Process	Chi-square test power	K-S power
Heterogeneous Poisson	0.1658	1
Gamma renewal	1	0.0009

Table 3.2: Estimated power of level 0.05 tests of homogeneous Poisson null hypothesis for two temporal point processes, estimated from 10,000 simulations of each process. The chi-square test is described in section 3.3.1. It uses ten-day intervals and four bins. The Kolmogorov-Smirnov test is described in section 3.3.2. In the "Heterogeneous Poisson" process, events occur at rate 0.25 per ten days for twenty years, then at rate 0.5 per ten days for a further twenty years. The Kolmogorov-Smirnov test rejects in all simulations, while the chi-square test usually does not reject. In the "Gamma renewal" process, the times between events are independent and follow a gamma distribution with shape 2 and rate 1. The chi-square test rejects in all simulations, while the Kolmogorov-Smirnov test rarely rejects. The two tests are powerful against different alternatives.

that

$$P(D_n > x) \leq 2\exp\left(-2nx^2\right). \tag{3.21}$$

We can use this inequality to calculate conservative P-values for the hypothesis that times are homogeneous Poisson.

3.3.3 Tests on simulated data

In this subsection we estimate the power of the chi-square and Kolmogorov-Smirnov tests for uniform times against contrasting alternatives. All tests in this section are at level 0.05. Results are summarised in Table 3.2.

Consider realisations of the following temporally heterogeneous Poisson process over a forty-year study period:

- For the first twenty years of the study period, events occur at rate 0.25 per ten days.
- For the next twenty years of the study period, events occur at rate 0.5 per ten days.

We applied the following tests at level 0.05 to 10,000 simulations of the process:

- A chi-square test with ten-day intervals and four bins.
- The Kolmogorov-Smirnov test.

The chi-square test rejected in 1658 of the 10,000 simulations. The power of the test is low because the chi-square test does not take the order of the intervals into account. The Kolmogorov-Smirnov test rejected in all 10,000 simulations. The increased rate in the second half of the study period is an obvious departure from uniformity.

Now consider a process in which the inter-event times are independent with a Gamma(2,1) distribution (where the time unit is ten days). We simulated 10,000 forty-year realisations of this process. The chi-square test rejected in all 10,000 simulations. The test notes that more intervals have one event and fewer have 2 or more events than would be expected under the Poisson hypothesis. The Kolmogorov-Smirnov test rejected in only 9 of the 10,000 simulations. The power is low because the "non-uniformity" is spread out through the study period.

The Kolmogorov-Smirnov test is sensitive to the distribution across the whole study period, but not to local variations. The chi-square test, in contrast, is sensitive to local variations but not to the overall shape of the distribution of times. Neither test uses spatial information: they test whether a catalog is Poisson in time, not in space-time.

3.3.4 Tests on declustered catalogs

Gardner and Knopoff applied Method 1 of section 3.2.1 using their windows to thin raw catalogs of seismicity in Southern California [2]. They carried out a number of tests on different declustered catalogs using a variety of bin widths. None of the tests gave a significant value for the χ^2-statistic.

We did not have the catalog used by Gardner and Knopoff. We instead used the SCEC catalog from 1932 to 1971, covering the same time period and approximately the same spatial region as the Gardner-Knopoff study. We declustered a catalog of Southern Californian seismicity using the Gardner-Knopoff windows. The Gardner-Knopoff raw catalog contains 1,751 events with magnitude at least 3.8, whereas our raw catalog contains 1,556 such events.

We created three declustered catalogs, using the three methods from section 3.2.1. Our declustered catalogs contained 437, 424, and 544 events after applying Methods 1, 2, and 3 respectively. The Gardner-Knopoff declustered catalog of events with magnitude at least 3.8 contained 503 events.

We tested the Poisson hypothesis using the chi-square test with $B = 4$ and $d = 2$ and the Kolmogorov-Smirnov test. We accounted for multiple testing using Bonferroni's inequality, which is conservative. We rejected the null hypothesis if either the chi-square test or the Kolmogorov-Smirnov test gives a P-value of less than 0.025. This ensures that if the null hypothesis is true, a Type I error has probability no greater than 0.05. We also compare chi-square tests using $\hat{\lambda} = n/k$ to the tests using the solution of (3.18) as the estimate of λ. In the cases we studied, the differences between these two estimates are negligible.

Method	Chi-square P-value	MLE chi-square P-value	K-S P-value	Reject?
1	0.087	0.087	0.012	Yes
2	0.297	0.295	0.0064	Yes
3	6×10^{-6}	4×10^{-6}	0.022	Yes

Table 3.3: P-values for tests of Poisson null hypothesis for Southern Californian seismicity declustered using Methods 1, 2, and 3 from section 3.2.1. "Chi-square P-value" is for the test using n/K as the estimate for λ. "MLE chi-square P-value" is for the test using the maximum likelihood estimate of λ from the bin counts (solving (3.18)). The hypothesis is rejected at level 0.05 if the P-value from either the chi-square test or the Kolmogorov-Smirnov test is less than 0.025.

Results are given in Table 3.3. We rejected the null hypothesis for all three declustering methods: this contradicts Gardner and Knopoff's conclusion. In all three methods, the difference between the empirical distribution of times and a uniform distribution was large enough to cause the Kolmogorov-Smirnov test to reject at level 0.025. The chi-square test would reject at level 0.025 in one of the three cases.

3.4 Space-time distribution

3.4.1 Weakening the Poisson hypothesis

In a spatially heterogeneous, temporally homogeneous Poisson process, the marginal distribution of times is Poisson. The tests in the previous section reject the hypothesis that the marginal distribution of times is Poisson. Hence they also reject the hypothesis that the data come from a spatially heterogeneous, temporally homogeneous Poisson process.

In fact, we know a priori that a catalog declustered using a window method cannot be a realisation of such a process. In a spatially heterogeneous, temporally homogeneous Poisson process, two events may occur arbitrarily close to one another with strictly positive probability. A window declustering method, on the other hand, will not leave any pairs of events closer than some minimum distance in space-time (this distance depends on the method and the catalog magnitude threshold). If a raw catalog contains two events very close in space and time, the later event will fall within the window of the former, and one or both of them will be deleted.

Consider a spatially heterogeneous, temporally homogenous Poisson process on spatial domain A on time $(0, T]$. The space-time rate is the product of the marginal spatial rate and the constant temporal rate. In any Poisson process, conditional on the number of events, the events are iid, with probability density proportional to the space-time rate. Conditional on the locations of events, the marginal distribution of

times is just the normalised temporal rate: that is, conditional on the locations, the times are iid uniform.

We shall test a weaker condition: conditional on the locations, times are *exchangeable*. Suppose a catalog, not necessarily in chronological order, contains n earthquakes. Let the location of the ith event be (x_i, y_i), where x_i is longitude and y_i is latitude. We do not consider depths. Let T_i be a random variable representing the time of the event at (x_i, y_i). Let Π be the set of all $n!$ permutations of $\{1, \ldots, n\}$. We say the process has *exchangeable times* if, conditional on the locations,

$$\{T_1, \ldots, T_n\} \stackrel{d}{=} \{T_{\pi(1)}, \ldots, T_{\pi(n)})\} \tag{3.22}$$

for all permutations $\pi \in \Pi$.

Under a null hypothesis of exchangeability, given the locations $\{(x_i, y_i)\}$ and the times $\{t_i\}$, no assignment of times to locations is more or less likely than any other. (It follows that conditional on both the observed locations and observed times, the locations and times are independent.) For example, suppose we condition on the occurrence of earthquakes at locations A, B, and C and at times $1, 2$, and 3. If the times are exchangeable, the following pairings of locations and times are equally likely: $\{(A, 1), (B, 2), (C, 3)\}$, $\{(A, 1), (B, 3), (C, 2)\}$, $\{(A, 2), (B, 1), (C, 3)\}$, and so on for all six permutations. In a process that is not exchangeable, assignments in which events close in space are also close in time may be more likely (space-time clustering, which does not occur in a Poisson process) or less likely (which also does not occur in a Poisson process).

The remainder of this section outlines a nonparametric test of the hypothesis of exchangeable times. In section 3.5.3, we perform the test on several declustered catalogs.

3.4.2 Testing earthquake catalogs for exchangeable times

Romano [79, 80] used empirical process theory to develop methodology for bootstrap and randomisation tests of "nonparametric" hypotheses such as independence, symmetry, and exchangeability. The tests evaluate the absolute differences in measures of individual sets under the empirical distribution and under a transformation of the empirical distribution. The sets for which the probabilities are evaluated are required to be a Vapnik-Chervonenkis (VC) class.[3] The transformation maps distributions that do not satisfy the null onto distributions that do. For distributions satisfying the null, the transformation is the identity. The test statistic is proportional to the largest absolute difference in measure over all sets in the VC class. We give general details of Romano's methodology in Appendix A. Here, we focus on testing the specific hypothesis of exchangeable times.

[3] Or, more generally, a Glivenko-Cantelli class.

Let $\hat{P}_n$ be the empirical measure and $\tau(\hat{P}_n)$ be the transformation of the empirical measure into the set of exchangeable measures. The measure $\hat{P}_n$ has mass $1/n$ at every observed point of longitude, latitude, and time (x_i, y_i, t_i). The measure $\tau(\hat{P}_n)$ is the average of the (hypothetical) empirical measures for all $n!$ permutations of the data. We informally call $\tau(\hat{P}_n)$ the *empirical null measure*. Since after a permutation operation any spatial location may be paired with any observed time, the empirical null measure has support at every one of the n^2 points (x_i, y_i, t_j) for $1 \le i, j \le n$. All permutations are equally likely, so the empirical null measure places equal weight $1/n^2$ on every one of these points.

Let $\mathbf{V}$ be the set of lower-left quadrants in $\mathbf{R}^3$. Then $\mathbf{V}$ is a VC class. Identify each quadrant $\{(-\infty, x] \times (-\infty, y] \times (-\infty, t]\}$ by its corner (x, y, t). Let the test statistic be the supremum of the distance between $\hat{P}_n$ and $\tau(\hat{P}_n)$ evaluated over all lower-left quadrants:

$$\sup_{V \in \mathbf{V}} |\hat{P}_n(V) - \tau(\hat{P}_n)(V)|. \tag{3.23}$$

This is a generalisation of the Kolmogorov-Smirnov statistic (3.19) from one dimension to three dimensions.

The empirical measure of a lower-left quadrant with corner (x, y, t) is

$$\hat{P}_n(V) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \le x, y_i \le y, t_i \le t). \tag{3.24}$$

By conditional independence, the empirical null measure of a lower-left quadrant with corner (x, y, t) is

$$\tau(\hat{P}_n)(V) = \frac{1}{n!} \sum_{\pi_j \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \le x, y_i \le y, t_{\pi_j(i)} \le t) \tag{3.25}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i \le x, y_i \le y) \cdot \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}(t_k \le t). \tag{3.26}$$

To find the maximum distance over all lower-left quadrants, it is sufficient to find the maximum distance over a set of n^3 quadrants: those with corners (x_i, y_j, t_k), for $0 \le i, j, k \le n$. To see this, we classify every quadrant as one of two types according to its corner (x', y', t'). Quadrants of the first type have x' less than the minimum longitude in the catalog, or y' less than the minimum latitude, or t' less than the minimum time. Both the empirical measure and the empirical null measure of these quadrants will be zero. This means we do not have to consider quadrants of this type when finding the greatest distance between the two measures.

Quadrants of the second type have $x' \ge \min\{x_i\}$, $y' \ge \min\{y_j\}$, and $t' \ge \min\{t_k\}$. Let $I(x') = \sum \mathbf{1}(x_i \le x')$; that is, the number of points with x-coordinate less than

or equal to x'. Similarly, let $J(y') = \sum \mathbf{1}(y_i \le y')$ and $K(t') = \sum \mathbf{1}(t_i \le t')$. Also record the order statistics of the points in each of the three dimensions: let $x_{(L)}$ be the Lth largest value of x observed in the catalog, $y_{(L)}$ be the Lth largest value of y in the catalog, and $t_{(L)}$ be the Lth largest value of t in the catalog.

Following equation (3.24), the empirical measure of a lower-left quadrant V' with corner (x', y', t') is

$$\hat{P}_n(V') = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \le x', y_i \le y', t_i \le t') \tag{3.27}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \le x_{(I(x'))}, y_i \le y_{(J(y'))}, t_i \le t_{(K(t'))}). \tag{3.28}$$

This holds because no event in the catalog has an x-value between $x_{(I(x'))}$ and x', or a y-value between $y_{(J(y'))}$ and y', or a t-value between $t_{(K(t'))}$ and t'. (If a point between $x_{(I(x'))}$ and x' existed, there would be more than $I(x')$ points with an x-coordinate no greater than x', yielding a contradiction.)

Similarly, following (3.26), the empirical null measure of a quadrant V' with corner (x', y', t') is

$$\tau(\hat{P}_n)(V') = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \le x_{(I(x'))}, y_i \le y_{(J(y'))}) \cdot \frac{1}{n}\sum_{k=i}^{n} \mathbf{1}(t_i \le t_{(K(t'))}). \tag{3.29}$$

That is, for every lower-left quadrant with non-zero empirical measure or empirical null measure, there exists a lower-left quadrant with a corner of the form (x_i, y_j, t_k) for $0 \le i, j, k \le n$ that has the same empirical measure and empirical null measure. So to find the supremum of the absolute difference between these two measures over all lower-left quadrants, we need only find the maximum over the set of n^3 lower-left quadrants with corners of that form. Moreover, it is sufficient to consider the order (or ranks) of the events in each of the three dimensions.

The exact conditional null distribution of the test statistic is found by computing the statistic for all permuted catalogs. Since there are $n!$ permutations, this is computationally unfeasible for catalogs with more than a dozen or so events. We instead look at $N = 1000$ randomly permuted catalogs for each test. Under the null hypothesis, the $N+1$ values obtained from the catalog test statistic and the statistics for the permuted catalogs are equally likely. So a test that rejects the null when the catalog statistic is larger than the 50th-largest value of the 1000 statistics for the permuted catalogs has estimated level 0.05.[4] The estimated P-value is the proportion

[4]To be more accurate, the estimated level is 50/1001. This is the probability under the null that the catalog test statistic is one of the 500 largest elements of the concatenated vector of length 1001, ignoring ties. Also note that this assumes sampling without replacement. In practice, we sample

of the elements of the concatenated vector that are greater than or equal to the test statistics for the unpermuted catalog.

The P-value is an estimate because of sampling variability. A sample of 1000 permutations only allows the sampling distribution to be determined with limited accuracy. If the P-value estimated from all $n!$ permutations were 0.05, the standard error of the P-value estimated from 1000 permutations would be $\sqrt{.05 \times .95/1000} \approx 0.007$. We would not like the result of a test based on a sample of permutations to differ from the result of a test based on all permutations. Therefore, if the P-value estimated after 1000 permutations was between 0.01 and 0.1, we continued sampling permutations up to a total of 10,000. (We do not sample 10,000 permutations for all tests because of the computational expense.) This two-step process introduces some bias for P-values near 0.01 or 0.1, but negligible bias for P-values near 0.05.

3.4.3 Test algorithm

We used the statistical computing software R to perform tests for exchangeable times on raw and declustered catalogs. R code for the test is given in Appendix B. These are the steps of the algorithm.

1. Sort the catalog of longitudes, latitudes, and times in time order: this facilitates step 3. Label the sorted points $\{x_i, y_i, t_i\}$ for $i \in \{1, \ldots, n\}$. Find the longitude and latitude ranks of every event. The longitude rank of the ith event is

$$\sum_{j=1}^{n} \mathbf{1}(x_j \le x_i).$$

2. Find the empirical *spatial* measure for all lower-left quadrants in $\mathbf{R}^2$ with corners

$$(y_i, x_j), \quad 1 \le i, j \le n. \tag{3.30}$$

In the R implementation in Appendix B, this spatial distribution is stored in the matrix `xy.upper`. In that matrix, the entry indexed by (i, j) is

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i \le y, x_j \le x). \tag{3.31}$$

This is the number of events in the catalog with latitude less than the latitude of the ith event in the catalog and with longitude less than the longitude of the jth event in the catalog. (Note that here we store points as (y_i, x_j) to be consistent with the indexing of R.)

with replacement. Since the sampling fraction is miniscule, the difference between sampling with and without replacement is negligible.

3. Find the absolute differences between the empirical measure (3.24) and the empirical null measure (3.26) for all n^3 lower-left quadrants with corners

$$(x_j, y_i, t_k), \quad i, j, k \in \{1, \ldots, n\}.$$

Find the maximum value of all these differences; this is the test statistic S. Because an object with n^3 elements may cause memory problems in R, the code finds the distances for quadrants with corners (x_j, y_i, t_k) for every value of k successively; that is, we find

$$S = \max_k \left[\max_{j,i} \left| \hat{P}(V(j,i,k)) - \tau(\hat{P})(V(j,i,k)) \right| \right],$$

where $V(j,i,k)$ is the lower-left quadrant with corner (x_j, y_i, t_k).

4. Set $N_1 = 1000$ and $N_2 = 10{,}000$. Set the permutation counter $H = 1$.

5. Create a random permutation $\{1, \ldots, n\}$. Apply this permutation to the locations while keeping times fixed. (We could apply the permutation to the times while keeping locations fixed, but this would require the additional step of re-sorting the catalog in time order.) The spatial measure has not changed, but its indexing has, so also apply the permutation to both the rows and the columns of `xy.upper`.

6. As in step 3, find the absolute differences between the empirical measure and the empirical null measure for the n^3 lower-left quadrants. Let S_H be the maximum value of all these distances.

7. If $H = N_1$, go to the next step. If $H = N_2$, go to step 9. Otherwise set $H = H + 1$ and go to step 5.

8. Estimate a preliminary P-value

$$\hat{P} = \sum_{h=1}^{N_1} \frac{1}{N_1} \mathbf{1}(S_h \geq S).$$

If $\hat{P} < 0.01$, reject the hypothesis and stop. If $\hat{P} > 0.1$, do not reject the hypothesis and stop. If $0.01 \leq \hat{P} \leq 0.1$, set $H = H + 1$ and go to step 5.

9. Compare S to the distribution of S_H. The estimated P-value is

$$\hat{P} = \sum_{h=1}^{N_2} \frac{1}{N_2} \mathbf{1}(S_h \geq S).$$

Reject the hypothesis if this value is less than 0.05. Stop.

3.5 Test cases and results

3.5.1 Tests for exchangeable times on simulated catalogs

We first trialled the test on data we knew to be a realisation of a spatial point process in which times are exchangeable, given the locations. We applied the permutation test to a simulation of 500 points of a uniform process on $[0,1] \times [0,1] \times [0,1]$. This process satisfies the null hypothesis: conditional on the number of points, the locations and times are independent and identically distributed.

Figure 3.1 plots the two spatial co-ordinates against each other, as well as each of the spatial co-ordinates against time. No surprising clustering is evident in any case, as expected from the independence of the co-ordinates. Figure 3.2 compares the simulation test statistic to its sampling distribution under the null, estimated from 1000 random permutations of the catalog. The simulation test statistic is near the centre of the histogram. The test gave an estimated one-tailed P-value of 0.546. The hypothesis that the process is exchangeable is not rejected at the 0.05 level. No Type I error occurred; the test result was as expected.

We then performed a test on data we knew to be a realisation of a spatial point process in which times are *not* exchangeable, given the locations. We generated such a set of points using the following algorithm:

1. Generate event times as a homogeneous Poisson process of rate 500 on $(0, 1]$.

2. For events with time in $(0, 0.5]$, generate spatial co-ordinate x uniformly in $(0, 0.5]$. For events with time in $(0.5, 1]$, generate spatial co-ordinate x uniformly in $(0.5, 1]$.

3. For events with time in $(0, 0.25] \cup (0.5, 0.75]$, generate spatial co-ordinate y uniformly in $(0, 0.5]$. For events with time in $(0.25, 0.5] \cup (0.75, 1]$, generate spatial co-ordinate y uniformly in $(0.5, 1]$.

The times of these simulated events are a realisation of a homogeneous Poisson process, and the expected number of events in any spatial region during time $(0, 1]$ is proportional to the area of the region. The process is not, however, homogeneous in space-time. Some space-time regions cannot contain points. For example, at spatial co-ordinates $x = 0.1, y = 0.1$, events can only occur for times in $(0, 0.25]$, and not in $(0.25, 1]$. Thus some combinations (x_i, y_i, t_k) are impossible in this process, and times are not exchangeable. Figure 3.3 plots the two spatial co-ordinates against each other, as well as each of the spatial co-ordinates against time. The plots of t against x and y show there are space-time regions in which points do not occur.

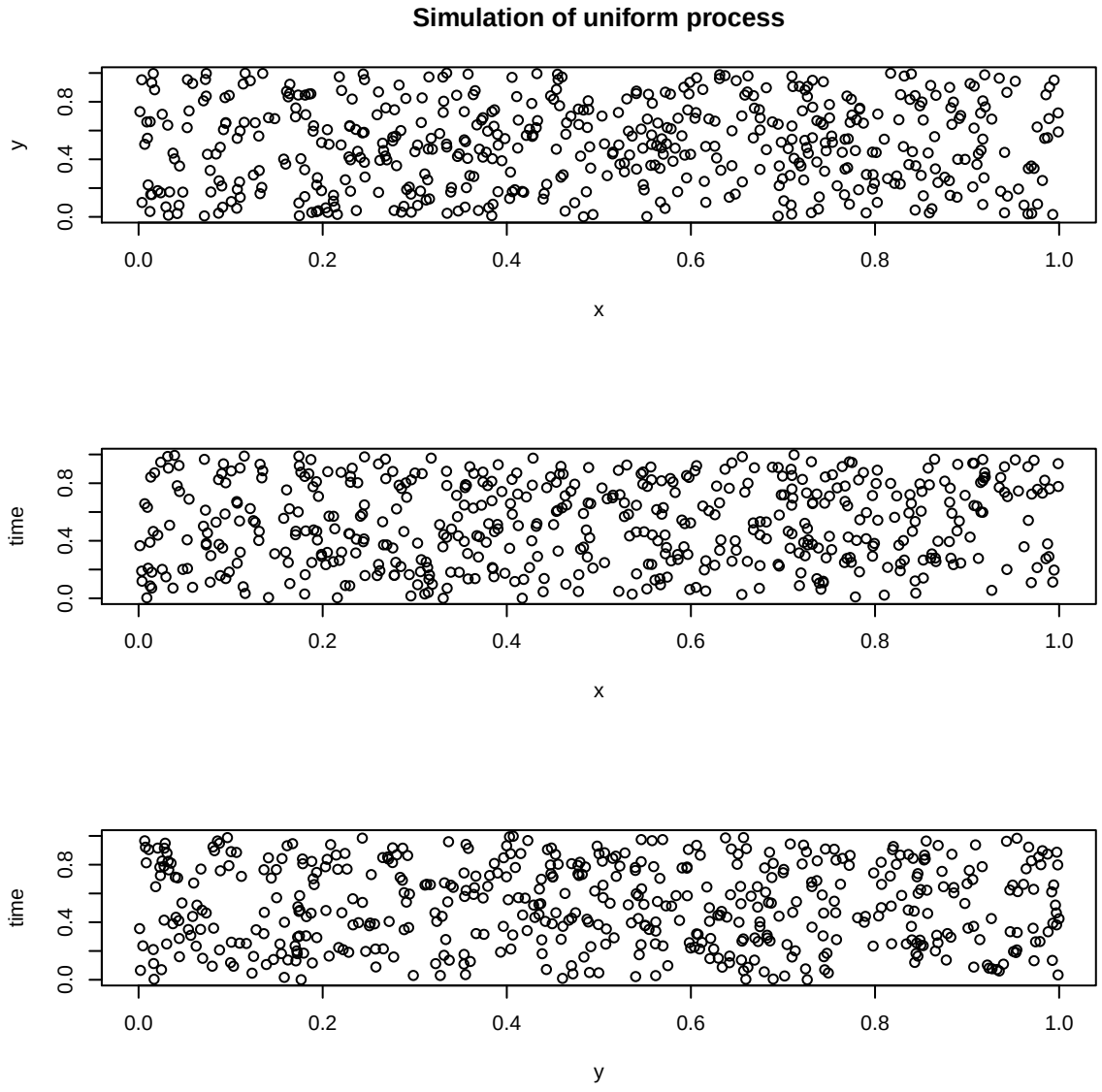

Figure 3.1: Realisation of a uniform process of 500 events on $[0,1] \times [0,1] \times [0,1]$. Conditional on the number of events, x, y and t are all independent.

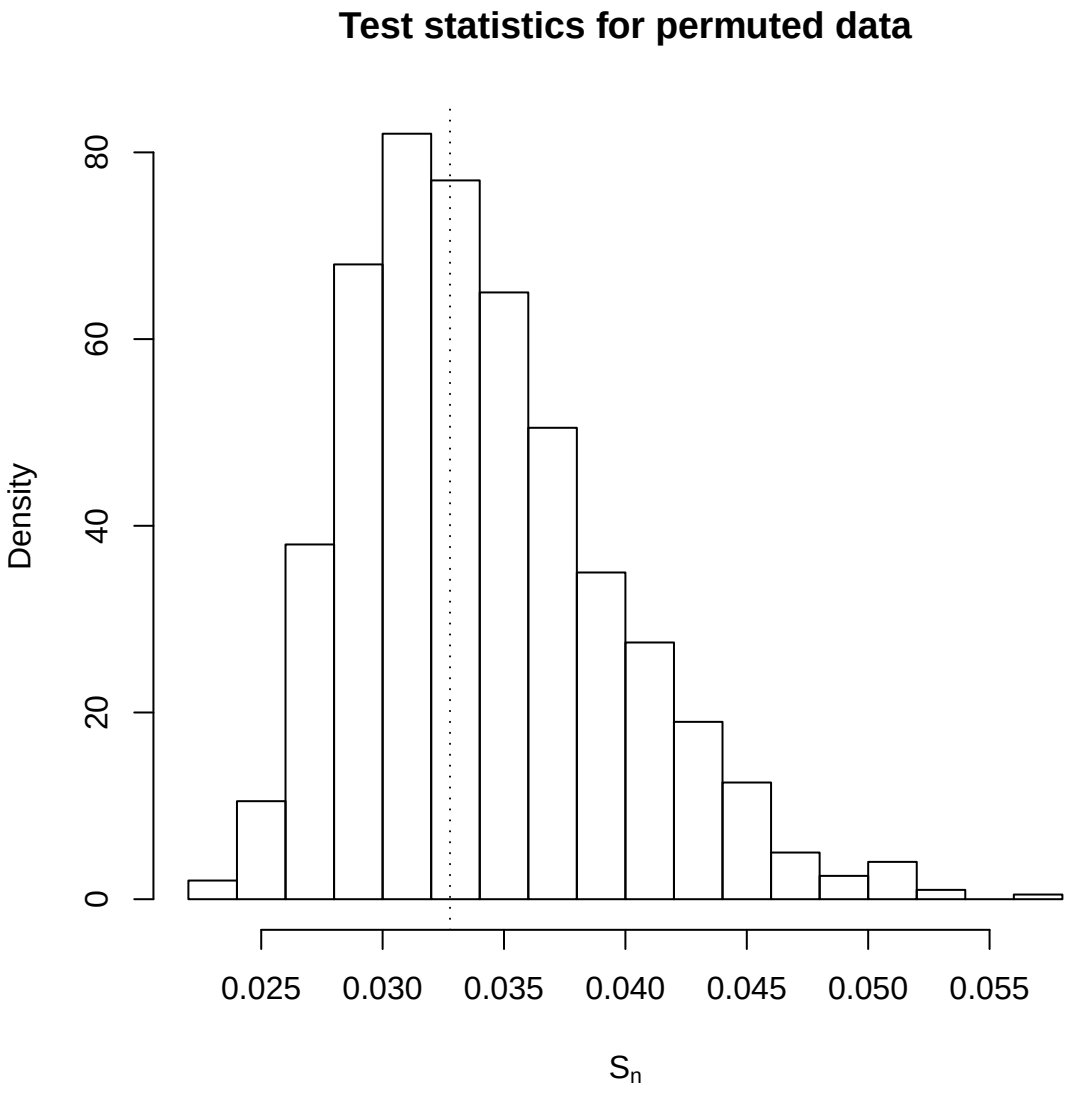

Figure 3.2: Estimated sampling distribution of the test statistic (3.23) for the uniform catalog depicted in Figure 3.1. The distribution is estimated from 1000 permutations of the catalog. The test statistic for the original catalog is represented by the dashed line. The estimated one-tailed P-value is 0.546; the hypothesis of exchangeable times is not rejected.

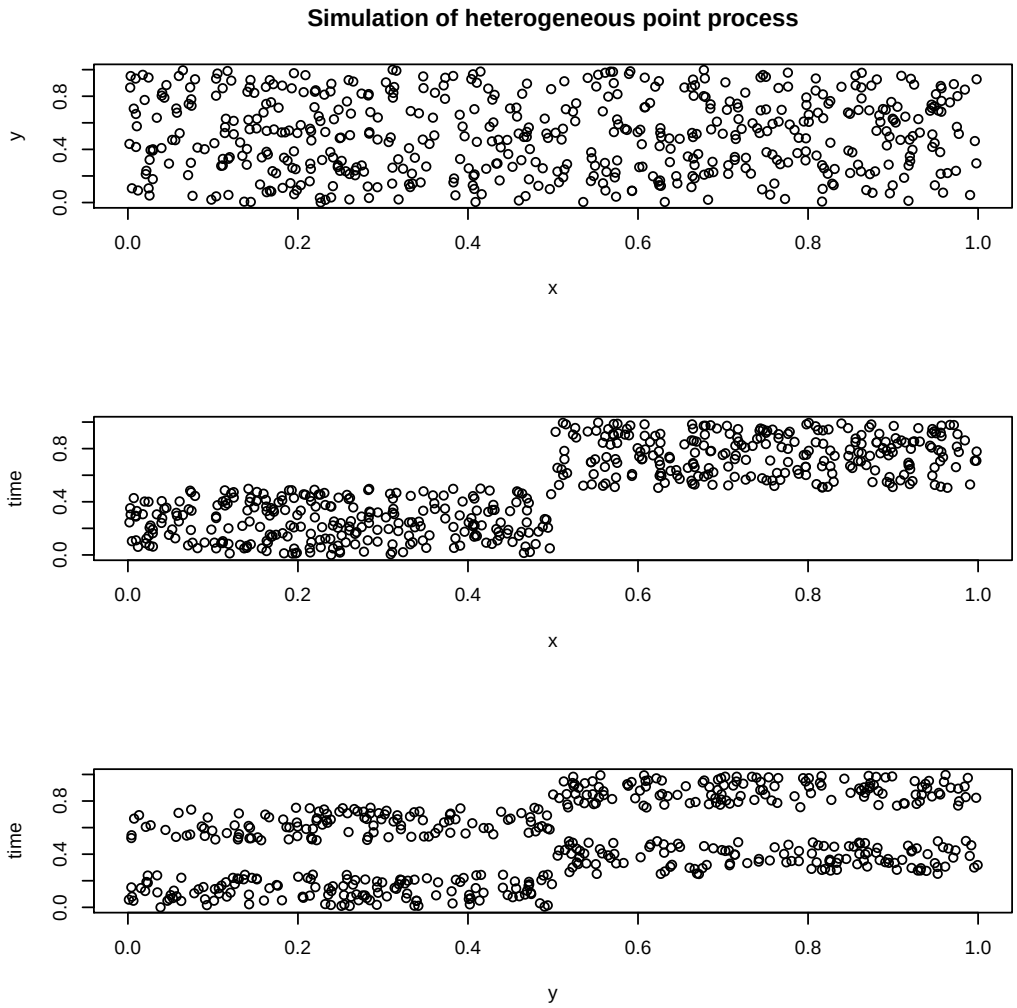

Figure 3.3: Realisation of the heterogeneous point process described in section 3.5.1, with 491 events. The event times $\{t_i\}$ were generated as a homogeneous Poisson process of rate 500 on $(0, 1]$. For each event with $t_i \in (0, 0.5]$, x_i was generated independently from a uniform distribution on $(0, 0.5]$. For each event with $t_i \in (0.5, 1]$, x_i was generated independently from a uniform distribution on $(0.5, 1]$. For each event with $t_i \in (0, 0.25] \cup (0.5, 0.75]$, y_i was generated independently from a uniform distribution on $(0, 0.5]$. For an event with $t_i \in (0.25, 0.5] \cup (0.75, 1]$, x_i was generated independently from a uniform distribution on $(0.5, 1]$. The process of times is homogeneous Poisson; however, at any given location, events may only occur at certain times. For example, an event at location $x, y < 0.5$ can occur for $t \in (0, 0.25]$ but not for $t \in (0.25, 1]$. Event times are therefore not exchangeable.

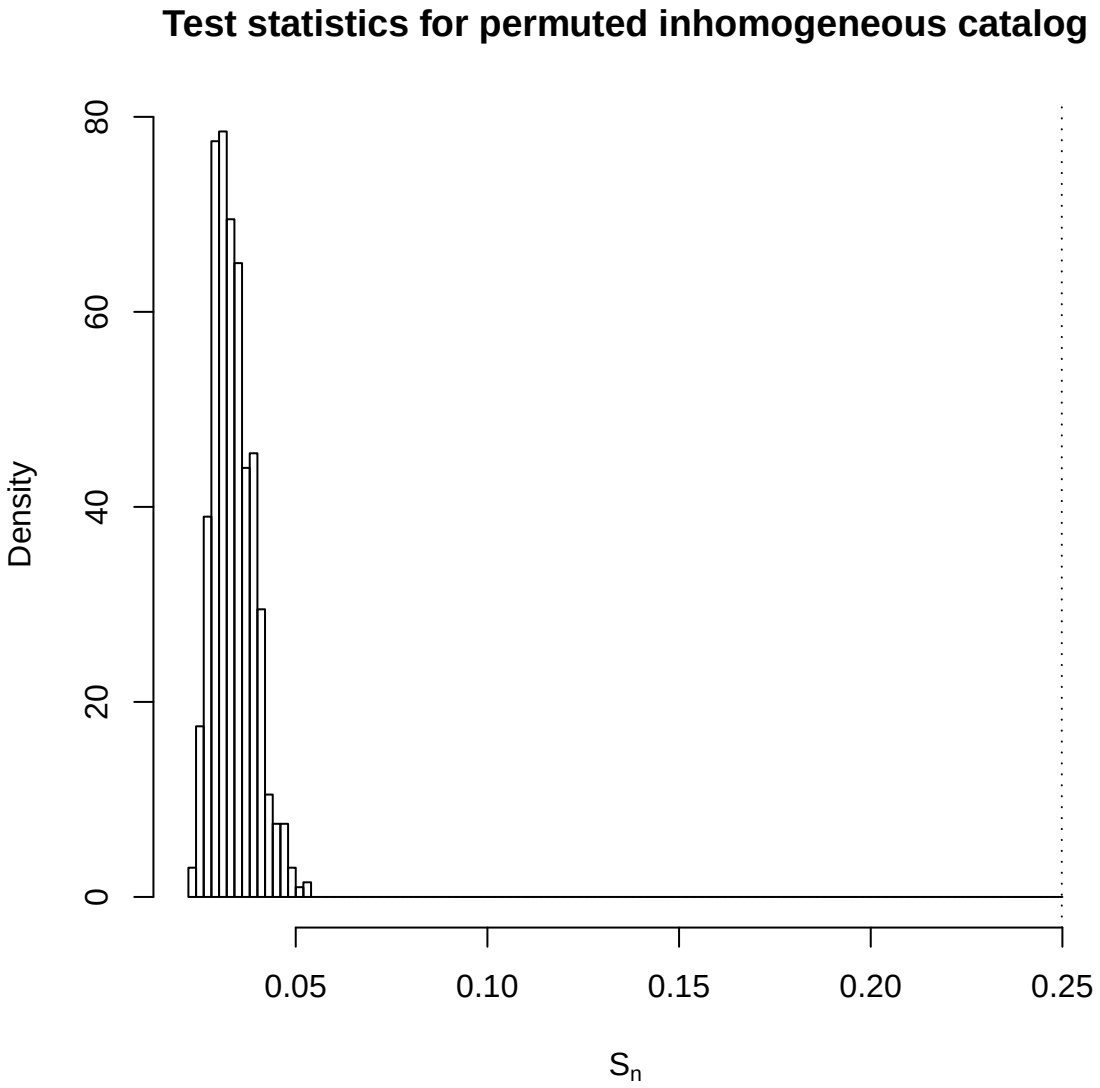

Figure 3.4: Estimated sampling distribution of the test statistic (3.23) for the heterogeneous catalog depicted in Figure 3.3. The distribution is estimated from 1000 permutations of the catalog. The test statistic for the original catalog, represented by the dashed line, exceeds by far any of the statistics for the permuted catalogs. The estimated one-tailed P-value is less than 0.001; the hypothesis of exchangeable times is rejected.

Figure 3.4 compares the simulation test statistic to its null sampling distribution, as estimated from 1000 random permutations of the catalog. The simulation test statistic is larger than any of the statistics for the permuted catalogs. The test gave an estimated one-tailed P-value of less than 0.001. The hypothesis that the process has exchangeable times is rejected at the 0.05 level. The test detects the obvious: times are not exchangeable. This example shows that catalogs with times that come from a temporally homogeneous Poisson process may still exhibit clustering in space-time. A test for a Poisson process that only includes times will not detect this.

3.5.2 Tests of exchangeable times on recent SCEC catalogs

Our first test on a real catalog was on a Southern California Earthquake Center (SCEC) catalog[5] of 753 events of magnitude 2.5 or greater in Southern California during the year 2009. Figure 3.5 plots the locations of the events over a map of Southern California. The events are not uniformly distributed over the spatial region; there is a greater density of events near the San Andreas and related faults. The null hypothesis of our test permits spatial heterogeneity. The test conditions on locations, so spatial heterogeneity in itself will not cause a rejection.

Figure 3.6 compares the test statistic for the raw catalog to its null sampling distribution, as estimated from 1000 random permutations of the catalog. The raw catalog test statistic is larger than any of the statistics for the permuted catalogs. The test gave an estimated one-tailed P-value of less than 0.001. The hypothesis that the raw catalog is exchangeable is rejected at the 0.05 level.

What if we decluster this catalog? We applied the implementation of Reasenberg's algorithm in Stefan Wiemer's ZMAP package for MATLAB[6] to decluster the 2009 SCEC catalog of events of magnitude 2.5 or greater, using default parameters, as given in section 3.2.2. After declustering, 475 events remained in the catalog. (Small changes in the parameters do not result in the reclassification of more than a few events.) Figure 3.7 plots the locations of the events. Again, there is clustering around known faults. Figure 3.8 compares the test statistic for the declustered catalog to its null sampling distribution, as estimated from 1000 random permutations of the catalog. The declustered catalog test statistic is out in the right tail. The test gave an estimated one-tailed P-value of 0.003. The hypothesis that the declustered catalog is exchangeable is rejected at the 0.05 level.

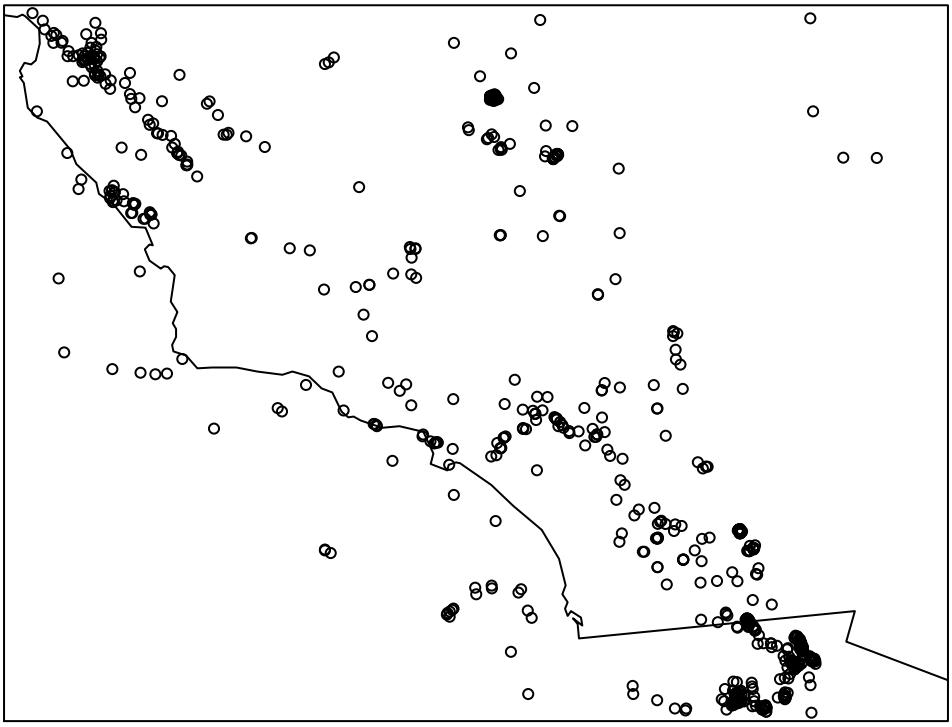

Figure 3.5: Raw SCEC catalog of events of magnitude 2.5 or greater in Southern California during year 2009. The catalog contains 753 events. The events are not spatially homogeneous.

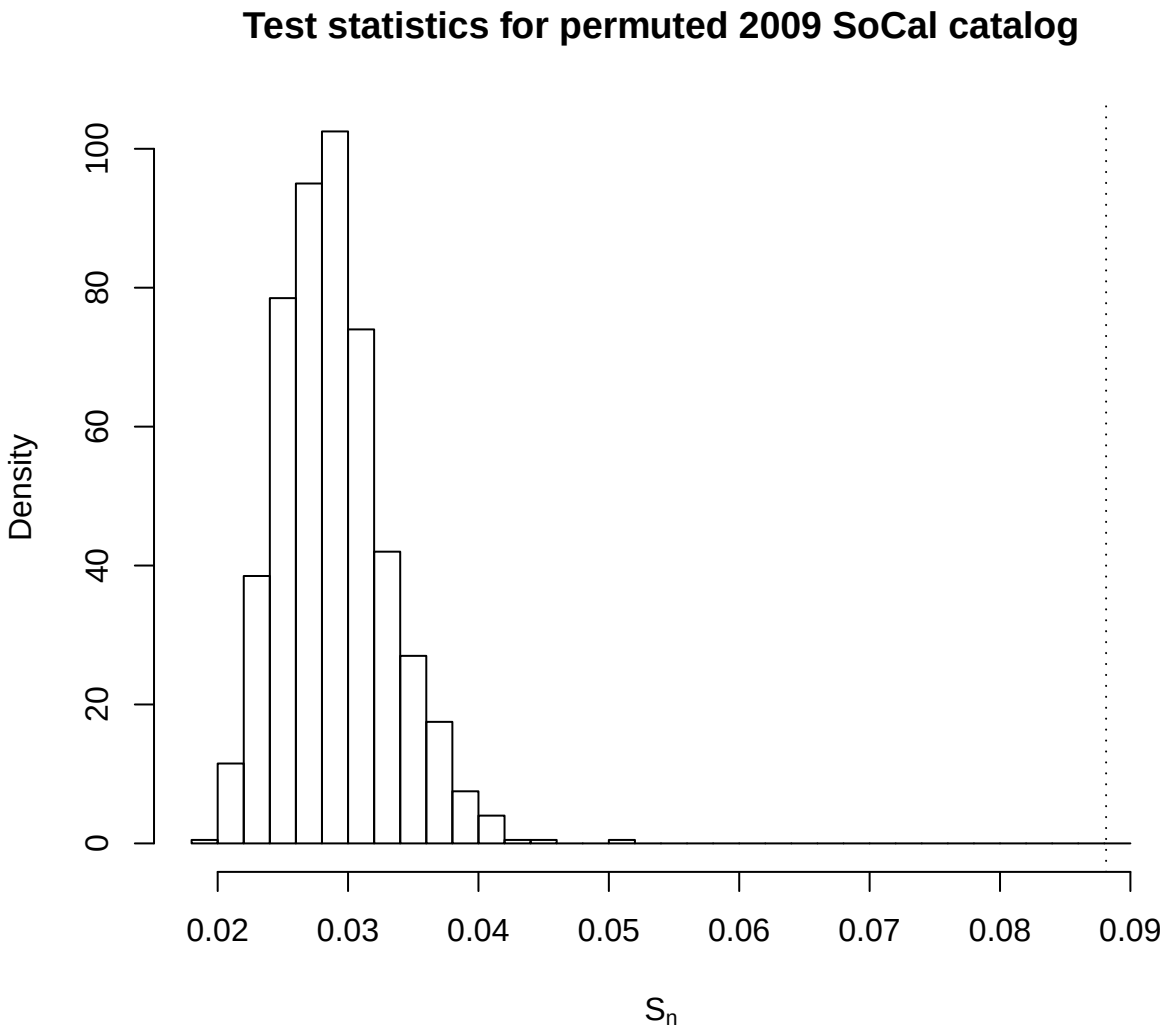

Figure 3.6: Estimated sampling distribution of the test statistic (3.23) for the raw SCEC catalog of events of magnitude 2.5 or greater in Southern California during year 2009. The distribution is estimated from 1000 permutations of the catalog. The test statistic for the original catalog, represented by the dashed line, exceeds by far any of the statistics for the permuted catalogs. The estimated one-tailed P-value is less than 0.001; the hypothesis of exchangeable times is rejected.

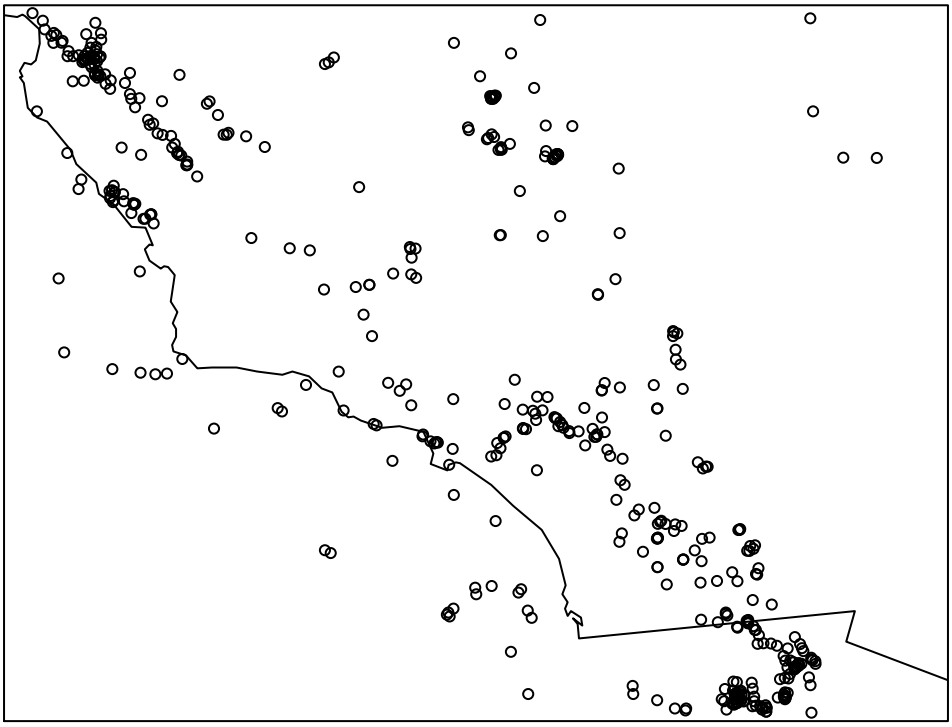

Figure 3.7: SCEC catalog of events of magnitude 2.5 or greater in Southern California during year 2009, declustered using Reasenberg's method. The declustered catalog contains 475 events.

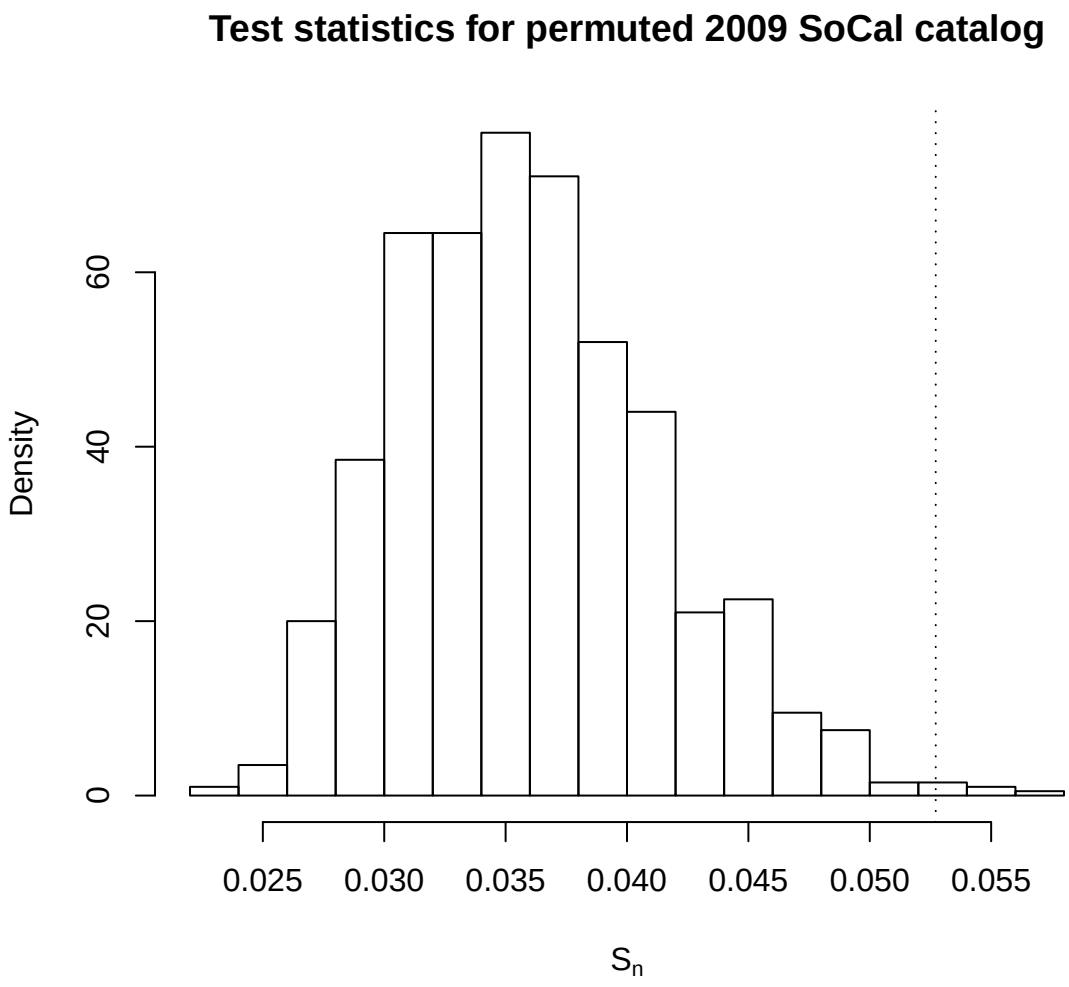

Figure 3.8: Estimated sampling distribution of the test statistic (3.23) for the SCEC catalog of events of magnitude 2.5 or greater in Southern California during year 2009, declustered using Reasenberg's method. The distribution is estimated from 1000 permutations of the catalog. The test statistic for the declustered catalog, represented by the dashed line, is greater than almost all of the statistics for the permuted catalogs. The estimated one-tailed P-value is 0.003; the hypothesis of exchangeable times is rejected at level 0.05.

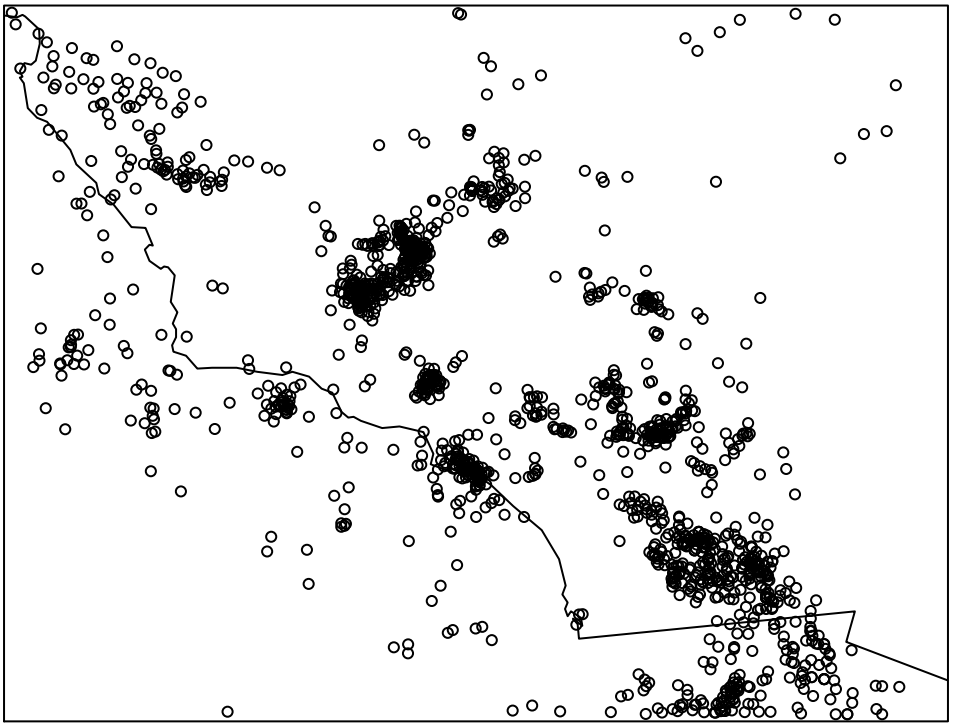

Figure 3.9: Raw SCEC catalog of events of magnitude 3.8 or greater in Southern California from 1932 to 1971. The catalog contains 1,556 events.

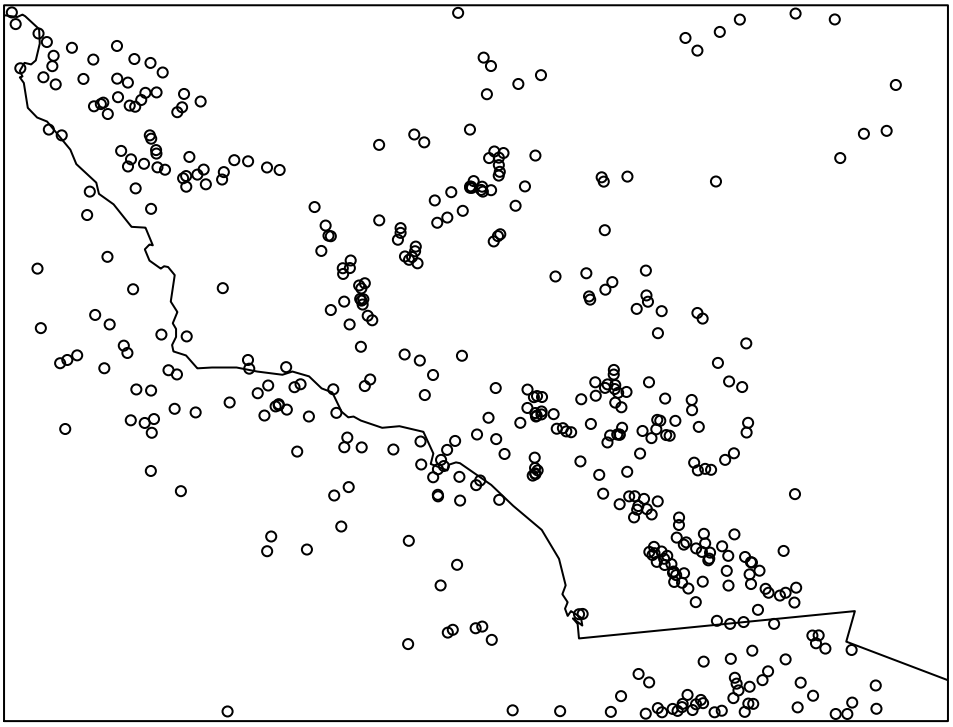

Figure 3.10: SCEC catalog of events of magnitude 3.8 or greater in Southern California from 1932 to 1971, declustered using Gardner-Knopoff windows in a linked-window method. The declustered catalog contains 424 events.

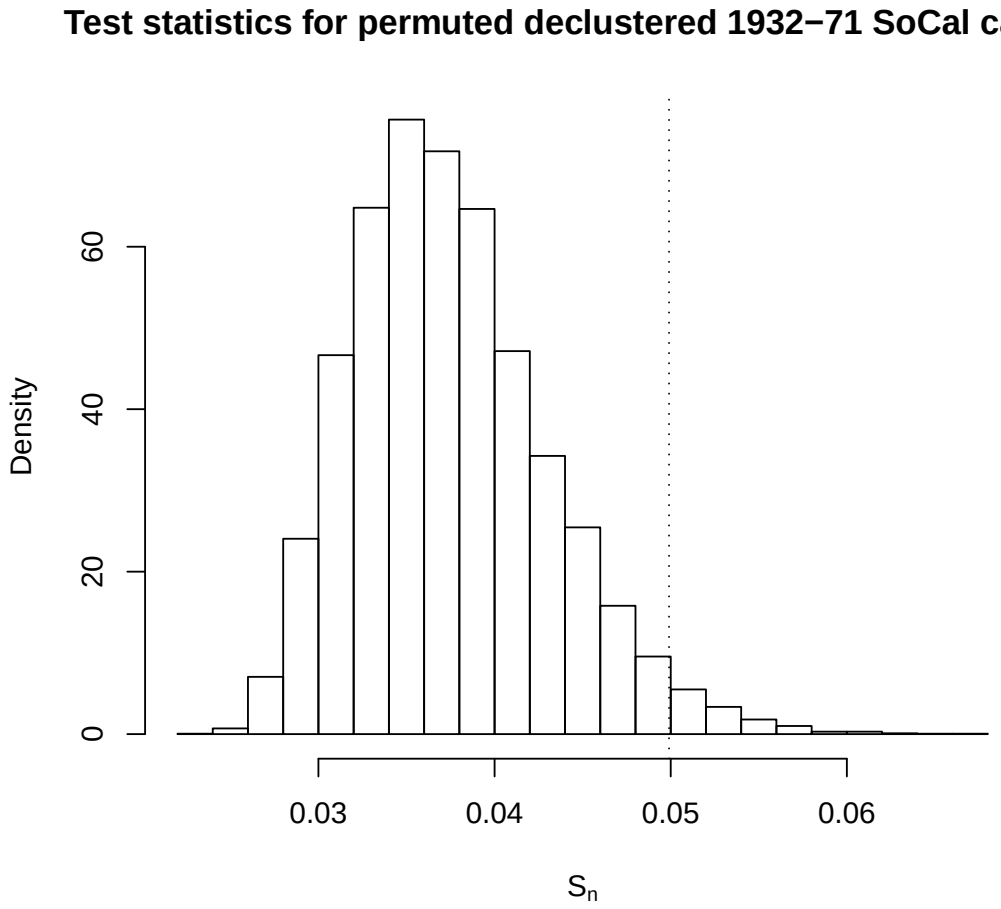

Figure 3.11: Estimated sampling distribution of the test statistic (3.23) for the 1932-1971 Southern Californian catalog of events of magnitude 3.8 or greater, declustered using Gardner-Knopoff windows in a linked-window method. The distribution is estimated from 10,000 permutations of the catalog. The test statistic for the declustered catalog, represented by the dashed line, exceeds most of the statistics for the permuted catalogs. The estimated one-tailed P-value is 0.005; the hypothesis of exchangeable times is rejected at level $\alpha = 0.05$.

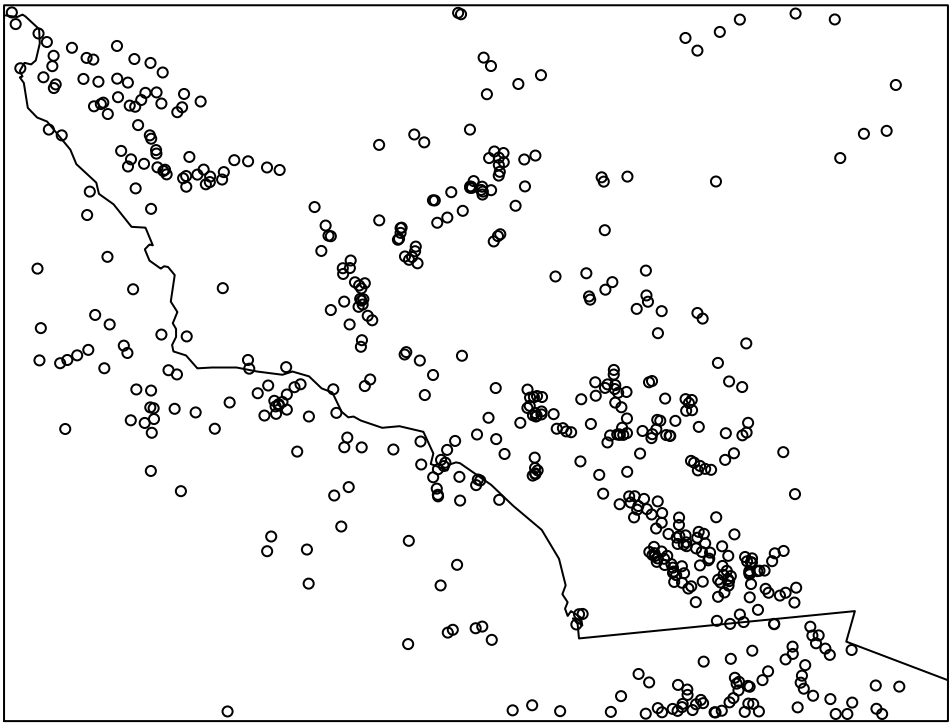

Figure 3.12: SCEC catalog of events of magnitude 3.8 or greater in Southern California from 1932 to 1971, declustered using Gardner-Knopoff windows in a main shock window method. The declustered catalog contains 544 events.

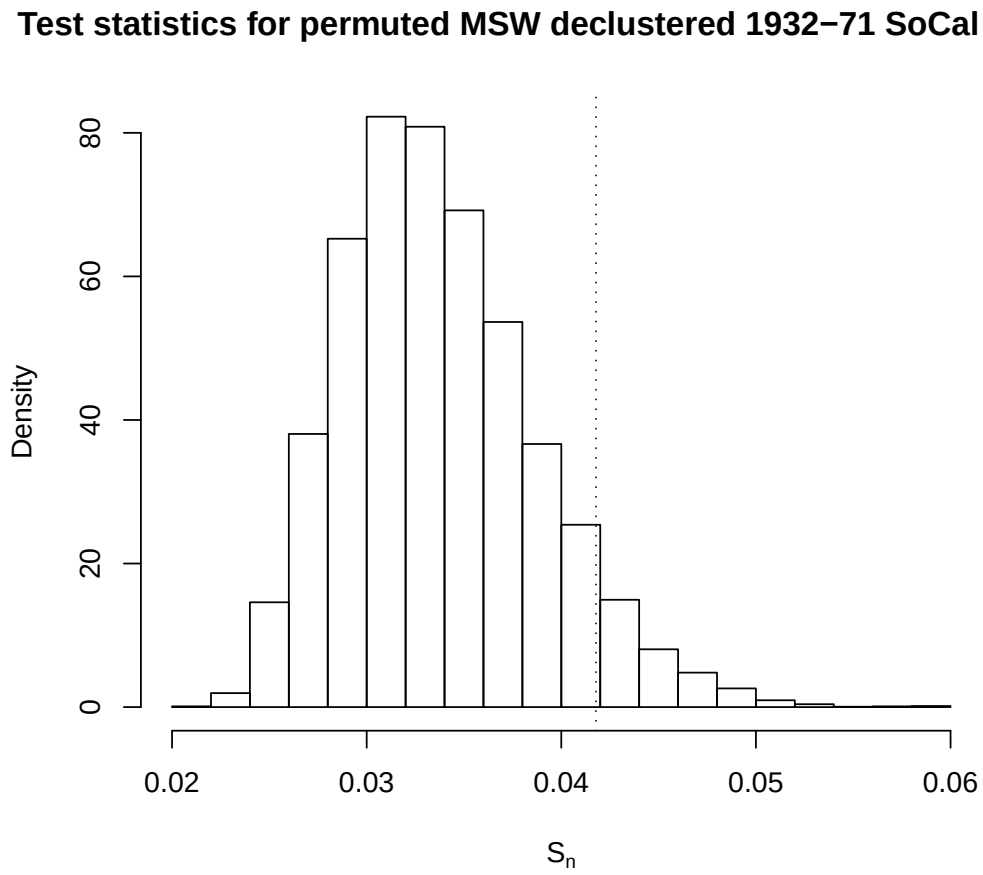

Figure 3.13: Estimated sampling distribution of the test statistic (3.23) for the 1932-1971 Southern Californian catalog of events of magnitude 3.8 or greater, declustered using Gardner-Knopoff windows in a main shock window method. The distribution is estimated from 10,000 permutations of the catalog. The test statistic for the declustered catalog, represented by the dashed line, exceeds many but not all of the statistics for the permuted catalogs. The estimated one-tailed P-value is 0.069; the hypothesis of exchangeable times is not rejected at level $\alpha = 0.05$.

3.5.3 Tests of exchangeable times on catalogs declustered using Gardner-Knopoff windows

As in section 3.3, we used the raw SCEC catalog of events of magnitude 3.8 or greater in Southern California from 1932 to 1971. Figure 3.9 plots the locations of the events. We applied the Gardner-Knopoff windows to the raw catalog in two ways. The first was the linked-window method Gardner and Knopoff used in their 1974 paper [2]. This was Method 1 from section 3.2.1. Secondly, we used the Gardner-Knopoff windows in a main shock window method. This was Method 3 from section 3.2.1.

The first method left a declustered catalog of 424 shocks. Figure 3.10 plots the locations of the events. Figure 3.11 compares the test statistic for the declustered catalog to its null sampling distribution, as estimated from 10,000 random permutations of the catalog. The declustered catalog test statistic is far out in the right tail. The test gives an estimated one-tailed P-value of 0.005. The hypothesis that the declustered is exchangeable is rejected at the 0.05 level.

The second method left 544 main shocks. Figure 3.12 plots the locations of the events. Figure 3.13 compares the test statistic for the declustered catalog to its null sampling distribution, as estimated from 10,000 random permutations of the declustered catalog times. The declustered catalog test statistic is out in the right tail. The test gives an estimated one-tailed P-value of 0.069. This is a small P-value, but not small enough to reject the hypothesis that the declustered is exchangeable at the 0.05 level.

3.6 Discussion

"Ok, so why do you decluster the catalog?"

This is a question posed in the online FAQ for the Earthquake Probability Mapping Application of the USGS.[7] The reasons given are "to get the best possible estimate for the rate of mainshocks," and that "the methodology [of the Earthquake Probability Mapping Application] requires a catalog of independent events (Poisson model), and declustering helps to achieve independence." It is not clear, however, that modelling only main shocks is better than modelling all shocks or that declustered catalogs consist of independent events.

Accurate estimation of the rate of main shocks requires an unambiguous definition of "main shocks." Often, main shocks are taken to be the set of events that remain after a catalog is declustered—a circular definition. A deterministic declustering method will produce the same declustered catalog every time it is applied to

[5] http://www.data.scec.org/catalog_search/data_mag_loc.php
[6] http://www.earthquake.ethz.ch/software/zmap
[7] http://earthquake.usgs.gov/learn/faq/?faqID=280

a particular raw catalog. Different methods, however, will produce different declustered catalogs. Furthermore, the complexity of some declustering methods makes it difficult to interpret what it means for an event to be classified as a main shock or an aftershock. Instead of using a declustering method to identify main shocks, it seems preferable to state a simple and clear definition of "main shock," then identify main shocks in this way—or to model all large events, main shocks or otherwise, because all large shocks can do damage.

We tested whether Southern Californian earthquake catalogs, declustered using the Gardner-Knopoff windows, followed a temporally homogeneous Poisson hypothesis. The chi-square test used by Gardner and Knopoff did not always reject, but the Kolmogorov-Smirnov test did. Neither test takes spatial locations into account.

We knew a priori that declustered catalogs are not truly Poisson in space and time. But declustered catalogs may still have properties in common with Poisson processes. In a spatially heterogeneous, temporally homogeneous Poisson model, two events may occur arbitrarily close to one another with strictly positive probability. As we stated in section 3.4.1, very close events cannot occur in a catalog declustered using a window method, so the catalogs are not exactly Poisson. We instead tested the weaker hypothesis that declustered catalogs had exchangeable times, given the locations. For catalogs declustered using the Reasenberg method and the Gardner-Knopoff windows in a linked-window method, one-tailed tests at level $\alpha = 0.05$ rejected the hypothesis of exchangeable times. This suggests that linked-window declustered catalogs are not even exchangeable. This may be because too many events are being removed, meaning that events close in space are unlikely to be close in time. However, the hypothesis of exchangeable times was not rejected at the $\alpha = 0.05$ level for a one-tailed test of a catalog declustered using Gardner-Knopoff windows applied as a main shock window method. This is not evidence that main shock window declustering is superior to linked-window declustering—the hypothesis of exchangeable times could be rejected for a larger catalog declustered using main shock windows.

Future work will include testing other declustering methods. Particularly interesting candidates for testing are the stochastic declustering methods of Zhuang et al., and of Marsan and Lengliné. The Hawkes process models in those methods assume that the background seismicity is a realisation of a Poisson process—in particular, they might have exchangeable times. Thus it seems plausible that declustered catalogs will resemble realisations of a heterogeneous Poisson process. More generally, one could create a declustered catalog that appeared exchangeable by removing events at random from the raw catalog until the remaining catalog passes a test for exchangeable times. This is trivially guaranteed to work: a catalog with one event cannot fail our test. This Procrustean declustering is unlikely to be useful.

Furthermore, simulations suggest that some of the events deleted are main shocks. Sornette and Utkin [94] simulated catalogs from a stationary marked ETAS process (see chapter 4) in which background events triggered offspring events. They then

employed the declustering method of Zhuang et al., which uses the parametric form of the ETAS model, to classify events as background or offspring. They found that the declustering was "rather unreliable" for distinguishing between the two types of event. If a declustering method has limited success in correctly classifying background and offspring events when the model assumed by the method is true, the value of declustering real catalogs is questionable. There is, on the other hand, inherent value in modelling clustered events. Large aftershocks may do just as much damage as main shocks. Removing them from the catalog after they occur will not repair the damage they cause.

Chapter 4

ETAS simulation and estimation

4.1 Introduction

Earthquake times in the ETAS model

As we saw in chapter 3, earthquakes cluster temporally. The modified Omori law and the Gutenberg-Richter law are empirical relationships widely accepted by seismologists as governing aftershock occurrence. The *modified Omori law*, or Omori-Utsu law [95], proposes that the rate of aftershocks $\Delta n/\Delta t$ at time t after a main shock approximately follows

$$\frac{\Delta n}{\Delta t} = \frac{K}{(t+c)^p}. \tag{4.1}$$

Here, K, c and p are assumed constant for a particular sequence, but vary between sequences. Sequences following larger main shocks will tend to have larger estimated values of K [96]. When estimated for real sequences, c is almost always positive. Estimated values of p have been found to fall between 0.3 and 2 for observed sequences.

If the true rate of aftershocks followed the modified Omori law, the total number of aftershocks would be the integral of the right-hand side of (4.1) from zero to infinity. If $p > 1$, the integral is finite. If $p \leq 1$, the integral is infinite. This seems unphysical, yet the value of p estimated for many finite aftershock sequences is less than 1.

The epidemic-type aftershock, or ETAS, model, devised by Ogata [52, 3], is a model for earthquakes with magnitude greater than or equal to some cut-off m_0. All earthquakes with magnitude $\geq m_0$ may trigger further shocks with magnitude $\geq m_0$. (Earthquakes with magnitude less than m_0 are not in the model.) The model classifies shocks as *background events*, which are not triggered by previous seismicity, and *offspring events*, which are directly triggered by one preceding earthquake. Offspring events may have their own offspring ("aftershocks of aftershocks").

We refer to a triggering event as a *parent*. Events directly triggered by a parent are *children*. Events triggered directly or indirectly by some event are *descendants* of that

event. An event can have 0 or 1 parents and any number of children and descendants. The ETAS model thus fits a *branching structure* to earthquake catalogs. In fact, the ETAS model is a version of a marked branching point process model called the Hawkes process.

In the ETAS model, background events are generated by a Poisson process with rate μ. For every event, the rate of offspring shocks decreases in time according to the modified Omori law with the same value of p. The rate of aftershocks of an earthquake with magnitude m_i is proportional to $10^{\alpha m_i}$, where α is a non-negative parameter. That is, the expected rate of offspring of an earthquake increases exponentially with magnitude. In the temporal ETAS model, an earthquake at time t_i with magnitude m_i, where m_i is greater than some minimum magnitude m_0, triggers further events with magnitude greater than or equal to m_0 at rate

$$\phi(t, t_i, m_i) = \frac{K \cdot 10^{\alpha(m_i - m_0)}}{(t - t_i + c)^p} \tag{4.2}$$

where K, α, c and p are parameters. The total intensity λ of earthquakes with magnitude at least m_0 at some time t is the background rate μ of earthquakes with magnitudes $\geq m_0$ not triggered by previous events, modelled as a constant rate Poisson process; plus the rates of earthquakes triggered by previous events.

We can characterise the process by its *conditional intensity* function (see equation (1.4)). The ETAS conditional intensity is

$$\lambda(t|\mathcal{F}_t) = \mu(m_0) + \sum_{i=1}^{N(t)} \phi(t, t_i, m_i) \tag{4.3}$$

$$= \mu(m_0) + \sum_{i=1}^{N(t)} \frac{K \cdot 10^{\alpha(m_i - m_0)}}{(t - t_i + c)^p} \tag{4.4}$$

where t_i and m_i are the time and magnitude of the ith earthquake and $\mathcal{F}_t$ is the σ-field generated by all previous earthquakes with magnitude $\geq m_0$.

The ETAS parameters are usually taken to be constant over the study region. The model can be extended to become a space-time model. The background rate μ may be allowed to vary spatially [4]; as may other parameters. An ETAS model for times, magnitudes, and locations may incorporate a distribution for offspring locations relative to the location of the parent. We briefly discuss space-time ETAS models in section 4.2.2.

Magnitudes in the ETAS model

Empirically, the relative frequencies of earthquakes at magnitudes up to about 8 follow the *Gutenberg-Richter (GR) law*. The relationship is

$$N \propto 10^{-bm}. \tag{4.5}$$

Here N is the number of events at a magnitude m or larger in a given region of space and time, while b is a constant, often assumed to equal 1. If $b = 1$, there are ten times as many magnitude $M-1$ earthquakes as magnitude M earthquakes. The GR law has been found to fit well both main shock sequences and aftershock sequences.

The magnitudes of all events in the ETAS model are drawn independently at random from a probability distribution with density $p(M)$. The rate at which an earthquake at time t_i with magnitude m_i triggers earthquakes of a *particular* magnitude M is

$$\phi(t, t_i, m_i, M) = p(M)\phi(t, t_i, m_i). \tag{4.6}$$

An offspring event may be larger than its parent.

The magnitude distribution in the ETAS model is almost always a form of the GR distribution[1] on $[m_0, \infty)$. The standard GR probability density is

$$p(M|m_0) = b \ln 10 \cdot 10^{-b(M-m_0)}. \tag{4.7}$$

Equation (4.7) implies a non-trivial chance of an earthquake of $M > 10$—when no such earthquakes have ever been observed—including the possibility of an $M = 15$ event, equivalent to tearing the earth in half. Thus the GR distribution is often truncated, so that the maximum possible magnitude is $m_1 < \infty$. The probability density is

$$p(M|m_0, m_1) = \frac{b \ln 10 \cdot 10^{-b(M-m_0)}}{1 - 10^{-b(m_1-m_0)}}. \tag{4.8}$$

Whether the magnitude distribution is truncated does not affect ETAS parameter estimates (aside from b) when standard methods are used. However, truncation has a huge effect on model properties, such as the branching ratio and the expected rate of events. We prefer to use the truncated distribution, as the GR law is empirical, and there is no empirical evidence for arbitrarily large earthquakes.

The ETAS model has been studied both analytically [97] and through simulation [4], and has been fitted to real seismicity [3, 1]. When parameters are estimated from data, those estimates often imply a non-stationary "explosive" process in which an earthquake is expected to generate an infinite number of offspring. Some properties

[1] In this chapter, we use the terms "Gutenberg-Richter distribution" and "GR distribution" to refer to both the empirical law and the probability distribution that results from sampling from events obeying the empirical law.

of various forms of the model are outlined in section 4.2.

The ETAS model is more complex than renewal process models. In a renewal process, inter-event times are independent; in the ETAS model, they are not. In a renewal process, conditional intensity depends only on the time since the last event; in the ETAS model, conditional intensity depends on the entire history. Despite its complexity, the ETAS model is a simplification of earthquake occurrence. There is no reason to believe seismicity exactly follows the ETAS parametric form, and the assumption that every earthquake is directly triggered by at most one previous event is not considered correct by seismologists. Real earthquakes are thought instead to be caused by complex interactions of fault systems.

Section 4.3 examines the simulation of ETAS processes. The fastest current method uses the branching structure of the model and simulates the process as a sum of heterogeneous Poisson processes. The rate of earthquake generation will be too low if the simulation does not account for background events occurring before the start of the simulation that produce offspring events after the simulation begins—*edge effects*. When the Omori parameter p is greater than 2, there are algorithms that give "perfect simulation" with no edge effects. Unfortunately, when Omori's law is fitted to data, the estimate of p is almost always less than 2. However, by using long "burn-in" times, we can ensure that edge effects are small.

ETAS models have often been estimated by maximising the likelihood numerically. The log-likelihood is often very flat, so conventional numerical ascent methods often fail to converge to the global maximum. Estimation can instead be viewed as an incomplete data problem in which the branching structure of the process is unobserved. Veen and Schoenberg [4] noticed this, and applied an "EM-type" algorithm to estimate the model. We outline their method in section 4.4. We find that in simulations, estimates produced by this method can be inaccurate, even when the number of events is large. Furthermore, in simulations where the identification of events as background or offspring events is known, fitting the ETAS model does a poor job of identifying which events are background and which are offspring [94].

In this chapter, we examine the estimation of the temporal ETAS model, applying the Veen-Schoenberg method to both real and simulated catalogs. The bias and variance of parameter estimates on simulated ETAS models varies considerably, depending on the simulation parameters. Furthermore, some properties of real catalogs, such as the distribution of inter-event times, are not reproduced by ETAS models with estimated parameters. In chapter 5, we examine prediction when seismicity follows an ETAS model.

4.2 ETAS model variations and their properties

The ETAS model is a form of Hawkes process. It was devised by Ogata [52, 3] based on work by Kagan and Knopoff [98, 99]. In this section, we give some

background on Hawkes processes, space-time generalisations of ETAS, and properties of the ETAS model.

4.2.1 Hawkes processes

The ETAS model is perhaps the mostly widely used *linear marked Hawkes process* model [100, 101, 102]. "Linear" indicates that the contributions to conditional intensity of past earthquakes stack linearly. Hawkes processes have been used frequently as models in forestry and epidemiology [103], and recently as a model of YouTube video viewing [104]. They were used to model earthquake occurrence before the development of the ETAS model [105].

If $N(t)$ is a Hawkes process, the conditional intensity $\lambda(t)$ satisfies

$$P(N(t) - N(s) = 1|N(s)) = \lambda(s)(t-s) + o(t-s) \quad (4.9)$$

$$P(N(t) - N(s) > 1|N(s)) = o(t-s) \quad (4.10)$$

for $s \leq t$. The process is *self-exciting*: heuristically, every event may trigger later events. The intensity $\lambda(t)$ of events at time t is determined by the process up to time t:

$$\lambda(t) = \mu + \int_{-\infty}^{t} g(t-u, m(u))dN(u). \quad (4.11)$$

Here, μ is a non-negative background rate and $m(u)$ is the mark at time u. The mark only exists at times at which an event occurs. Conditional on the times of the events, the marks are generated independently at random. The *kernel function* g satisfies $g(s,m) \geq 0$ for $s \geq 0$ and $g(s,m) = 0$ for $s < 0$.

The *branching ratio*, n, is

$$n = \mathbf{E}\int_{0}^{\infty} g(s,m)ds \quad (4.12)$$

when this exists. The expectation is over the mark distribution. The branching ratio is the expected number of children for every event. If $n < 1$, the *expected rate* is finite and identical for all t:

$$\mathbf{E}\lambda(t) = \frac{\mu}{1-n}. \quad (4.13)$$

The kernel for the ETAS model is

$$g(s,m) \equiv \frac{K \cdot 10^{\alpha(m-m_0)}}{(s+c)^p}. \quad (4.14)$$

4.2.2 Generalisations to space-time

There are several proposed extensions of ETAS to space-time [98, 52, 97]. The spatial aspect may be independent of the magnitude of the triggering earthquakes, or it may scale with magnitude [106]. Kagan and Jackson [107] used a modified ETAS model to generate "short-term seismic hazard estimates." One of several extensions suggested by Ogata [52] uses circularly symmetric aftershock densities, in which the squared distance between an offspring event and its parent follows a Pareto distribution. This leads to the following conditional intensity function:

$$\lambda(t,x,y|\mathcal{F}_t) = \mu(m_0,x,y) + \sum_{i=1}^{N(t)} \frac{K_0 \cdot 10^{\alpha(m_i-m_0)}}{(t-t_i+c)^p((x-x_i)^2+(y-y_i)^2+d)^{1+\rho}} \tag{4.15}$$

where (x_i, y_i) are the co-ordinates of the ith epicenter, while $d > 0$ and $\rho > 0$ are additional parameters governing the spatial distribution. The σ-field $\mathcal{F}_t$ is determined by $\{t_i, x_i, y_i, m_i : i \leq N(t)\}$. This chapter focuses on the temporal, non-spatial model.

4.2.3 Properties of the ETAS model

Branching ratios

The expected number of children of a *particular* earthquake with magnitude m is

$$n_m = \frac{K \cdot 10^{\alpha(m-m_0)}}{(p-1)c^{p-1}} \tag{4.16}$$

if $p > 1$. This varies immensely with m: for $\alpha = 0.5$ (a typical value), an earthquake with magnitude $m+2$ will on average generate ten times as many aftershocks as one with magnitude m.

The ETAS branching ratio is the expected number of children, averaged over all initial magnitudes. If we assume that magnitudes are independent and every magnitude is drawn from an untruncated GR distribution, then from (4.12) this is

$$n(m_0) = \frac{K}{(p-1)c^{p-1}} \frac{b}{b-\alpha}, \tag{4.17}$$

provided $p > 1$ and $\alpha < b$. If $p \leq 1$ or $\alpha \geq b$, an initial shock is expected to have an infinite number of children. Then n is infinite.

If magnitudes follow a truncated GR distribution, the branching ratio is

$$n(m_0,m_1) = \frac{K}{(p-1)c^{p-1}} \frac{b}{b-\alpha} \frac{1-10^{(\alpha-b)(m_1-m_0)}}{1-10^{-b(m_1-m_0)}}, \tag{4.18}$$

provided $\alpha \neq b$. For this to be finite, p must be greater than 1, but α need not be

less than b. For the special case where $\alpha = b$,

$$n(m_0, m_1) = \frac{Kb \log 10 \cdot (m_1 - m_0)}{(p-1)c^{p-1}\left(1 - 10^{-b(m_1 - m_0)}\right)}. \tag{4.19}$$

If n is finite, every primary aftershock will, on average, generate n secondary aftershocks; every secondary aftershock will, on average, generate n third-order aftershocks; and so on. So if $n < 1$, the total number of aftershocks that an initial shock is expected to generate is a geometric series. If the magnitude of the shock is known to be m, and subsequent magnitudes follow an untruncated GR distribution, the expected total number of aftershocks is

$$\frac{n_m}{1-n} \tag{4.20}$$

If the magnitude of the initial shock is drawn at random from the untruncated GR distribution, the expected total number of aftershocks is

$$\frac{n}{1-n} \tag{4.21}$$

If the background rate of shocks is μ, the average rate of shocks, including aftershocks, is

$$\mu\left(1 + \frac{n}{1-n}\right) \tag{4.22}$$

provided $n < 1$, which in turn requires $p > 1$ and $\alpha < b$. Similar expressions can be found for truncated GR.

When n is infinite, the expected number of children of a random-magnitude earthquake is infinite. When n is finite but greater than 1, a random-magnitude earthquake has a finite expected number of children, but an infinite expected number of total descendants: the process is unstable. When n is less than 1, the process is stable.

In 24 catalogs studied by Helmstetter and Sornette [97], the branching ratio calculated from parameters estimated for ETAS models with untruncated GR is less than 1 six times, greater than 1 but finite eight times, and infinite ten times. Under truncated GR, the branching ratio is less than 1 thirteen times, greater than 1 but finite five times, and infinite six times.

4.3 Simulation

As with many Hawkes process models, analytical results for ETAS are limited. Many aspects of the model must be studied through simulation. A standard point process simulation method is *thinning*. Events are generated at a higher rate than the model specifies and accepted or deleted ("thinned") with some probability. Ogata [108]

gave a thinning algorithm for simulating Hawkes processes in which the conditional intensity (4.11) is non-increasing between events; the intensity may increase at times of events, as it does in the ETAS model. The probability depends on the history of the process, as well as the rate at which events are generated. As the probability of rejection may be high, such an approach may take too much computation to be practical.

As alternatives to thinning, faster and more accurate algorithms that take advantage of the branching structure have been recently been developed [109]. We discuss these in the following subsection. We shall see that simulations of ETAS models using parameter values in the ranges that arise in seismology require very long burn-in times to approach stationarity.

4.3.1 Simulation using the branching structure

If Q is a stationary marked Hawkes process, it may be decomposed into a marked background process $\bar{Q}$ and a set of marked offspring processes $Q_0^{(n)}$. The background times form a homogeneous Poisson process on $\mathbf{R}$ with intensity equal to the background rate μ. The nth offspring process $Q_0^{(n)}$ consists of times $\{T_l^{(n)}\}$ and magnitudes $\{M_l^{(n)}\}$ of aftershocks of all orders of the nth event in $\bar{Q}$. The "background" events of $Q_0^{(n)}$ are the first-order aftershocks of the nth event in $\bar{Q}$. The kth order aftershocks in $Q_0^{(n)}$ are the $(k+1)$th order aftershocks in Q. Background events in $Q_0^{(n)}$ occur at a non-constant rate governed by $g(s = t - t_n, m_n)$, the kernel of the n event of $\bar{Q}$. Thus the offspring processes are also Hawkes processes, with intensity

$$\lambda_0^n(s) = g(s, m_n) + \sum_l g\left(s - T_l^{(n)}, M_l^{(n)}\right). \tag{4.23}$$

The event times of the offspring processes are on $\mathbf{R}^+$, with time measured since the event that triggered the process. As before, the marks are drawn from the GR distribution, and are independent conditional on their times.

Let N be the process of times of Q, $\bar{N}$ be the process of times of $\bar{Q}$ and $N_0^{(n)}$ be the process of times of $Q_0^{(n)}$. Then if S_n are the event times of $\bar{N}$,

$$N(t) = \bar{N}(t) + \sum_n N_0^{(n)}(t - S_n), \tag{4.24}$$

where $N_0^{(n)}(t)$ is defined as zero for $t < S_n$. If the branching ratio $\mathbf{E}\int_0^\infty g(t, m)dt$ is less than 1, the component branching processes $N_0^{(n)}$ almost surely have a finite number of points and N is a stationary process.

We can use this decomposition to efficiently simulate ETAS processes. The event times S_n can be interpreted as background earthquakes (*immigrants* in the branching

process literature). The event times $\{T_l^{(n)}\}$ are aftershocks (*offspring* or *descendants*). An offspring event can be larger than its parent.

In ETAS, the process of children of an event (t_i, m_i), is a heterogeneous Poisson process. The number of children, if it has finite expectation, will have a Poisson distribution. The number of children before time T always has finite expectation

$$\int_{t_i}^{T} g(t - t_i, m_i)dt = \frac{K \cdot 10^{\alpha(m_i - m_0)}}{1 - p} \left[(T - t_i + c)^{1-p} - c^{1-p}\right] \tag{4.25}$$

and a Poisson distribution. Given the number of children of a parent at time t_i, the times between the parent and its children are independent and follow a probability distribution on $(0, T - t_i]$ proportional to the modified Omori law. The distribution function is

$$F(t) = \frac{(t - t_i + c)^{1-p} - c^{1-p}}{(T - t_i + c)^{1-p} - c^{1-p}} \tag{4.26}$$

for $t > t_i$. This suggests the following algorithm for simulating a temporal ETAS process on the interval $(0, T]$ generation by generation.

1. Generate the number of background events N_0 from a Poisson distribution with mean μT.

2. Generate the times of the background events $T_{0,i}, i \in \{1, \ldots, N_0\}$ by choosing $N_0(t)$ times uniformly at random on $(0, T]$. Generate their magnitudes $M_{(0,i)}$ with density (4.7) for untruncated GR or density (4.8) for truncated GR.

3. Let g be the generation of aftershock; set $g = 1$.

4. For every event in the previous generation $(T_{g-1,i}, M_{g-1,i})$, generate the number of children $N_{g,i}$ from a Poisson distribution with mean as in (4.25) and $t_i = T_{g-1,i}$. Thus, in the gth generation of aftershocks, there will be $N_g = \sum N_{g,i}$ events. Generate the times of these events by generating uniform random variables $U_{g,i,j}$ on $(0, 1]$ and performing an inverse probability transform (inverting (4.26)) onto $(T_{g-1,i}, T]$:

$$T_{g,i,j} = \left[U_{g,i,j}(T - T_{g-1,i} + c)^{1-p} + (1 - U_{g,i,j})c^{1-p}\right]^{\frac{1}{1-p}} + T_{g-1,i} - c. \tag{4.27}$$

5. To reduce the number of indices, relabel the generation g event times $(T_{g,i}), i \in \{1, \ldots, N_g\}$. Generate magnitudes $(M_{g,i})$ with density (4.7) for untruncated GR or density (4.8) for truncated GR. If $N_g = 0$, stop. Otherwise set $g = g + 1$ and go to step 4.

4.3.2 Avoiding edge effects in simulation

Bravaccino et al. [110] derived bounds for the tails of the distribution of cluster length (from background event to last descendant) for Hawkes processes satisfying certain conditions. When the bounds hold, *perfect simulation*, with no edge effects, is possible. For the ETAS model, however, these bounds only hold for $p > 2$. Unfortunately, the value of p estimated for real data is almost always less than 2.

We instead use a *burn-in* period $T_b < \infty$. We simulate the background process on $(-T_b, T]$. We then simulate the aftershock sequence of all events of the background process up to time T. We consider the simulation to be the process on $(0, T]$.

Suppose $p \leq 2$. Then there are almost surely events that occur before time $-T_b$ that have offspring after time 0, no matter the size of T_b. However, if T_b is large, the rate at time 0 of aftershocks of background events before $-T_b$ will be small compared to the background rate. Simulation using the branching structure is inexpensive, so we can use very large values of T_b.

4.4 Estimation

Maximum likelihood (ML) parameter estimates for ETAS are intractable analytically. Ogata [3, 1] used the Davidon-Fletcher-Powell optimisation algorithm to find numerical ML parameter estimates. This method requires a careful choice of starting parameter values. Ogata recommended starting with unexceptional values like $K = 0.01, p = 1.3$, and μ chosen to give a rate of background events equal to a quarter of the total rate of events.

Ogata fitted the untruncated temporal ETAS model to 24 earthquake catalogs, mostly Japanese. We summarise the estimated parameter values in Table 4.1. Typically five parameters—μ, K, α, c, p—are estimated. In many of the 24 models, b was set to be 1. Likewise, m_0 is not estimated, but is selected to be some magnitude level above which the catalog is thought to be complete. Only five of the 24 fitted sets of parameters give a point process that is stationary.

Numerical maximum likelihood estimation may be unreliable because of the flatness of the log-likelihood function, or because of multimodality. Veen and Schoenberg [4] gave an alternative EM-type estimator for the space-time ETAS model with conditional intensity (4.15). In a personal communication, Veen provided us with code for similar estimation of the temporal model (4.4). We consider the temporal case here.

4.4.1 Maximising the complete log-likelihood

Dempster et al. [111] named and explained the expectation-maximisation (EM) algorithm for parameter estimation in incomplete data problems (though earlier au-

Parameter	Minimum	LQ	Median	UQ	Maximum
m_0	2.5	3.975	5	5.4	7
b	0.9	1	1	1	1.2
μ (per day)	0	0.00735	0.0215	0.04775	0.59
K	0.0002	0.00875	0.0155	0.047	5.2
c (days)	0.003	0.01075	0.025	0.135	11.6
p	0.85	1	1.115	1.327	3.5
α	0.155	0.5975	0.725	0.95	1.37

Table 4.1: Summary of parameter estimates for temporal ETAS models fitted using maximum likelihood by Ogata [3, 1] for 24 catalogs. The columns give minimum, lower quartile, median, upper quartile and maximum estimates for every parameter. In many of the 24 models, b was set to be 1. m_0 is also not estimated, but is selected to be some magnitude level above which the catalog is thought to be complete.

thors had proposed versions of the algorithm [112]). The algorithm essentially finds maximum likelihood parameter estimates for models with unobserved variables. In practice, EM has been found to provide more stable estimates than conventional numerical MLE in a variety of cases. However, EM estimates may still sometimes converge to local maxima.

Fitting the ETAS model may be considered an incomplete data problem, with the branching structure described by a set of unobservable variables. This approach was proposed by Veen and Schoenberg [4] based on the "stochastic reconstruction" declustering method of Zhuang et al. [76, 113], described in chapter 3.2.4. Veen and Schoenberg gave a method for the estimation of spatio-temporal ETAS models; below we give a method for temporal-only models.

Consider an ETAS model with a rate μ of background events on some area. Let θ be the parameter vector $\{\mu, K, \alpha, c, p\}$. Let $u_i = j$ if the ith earthquake was the child of the jth earthquakes ($j < i$); let $u_i = 0$ if the ith earthquake was a background event. Suppose u_i is known for all i. Let $N_b = \sum_i 1(u_i = 0)$ be the number of background events on $(0, T]$; this has a Poisson distribution with mean μT. Let l_i be the number of children of the ith event. The expected number of children by time T is

$$G_i(\theta) = \int_{t_i}^{T} g(t - t_i; m_i)dt, \tag{4.28}$$

as calculated in equation (4.25). The complete log-likelihood depends on $G_i(\theta)$. Like Veen and Schoenberg, when we calculate the likelihood, for computational reasons

we approximate (4.28) by

$$G_i(\theta) = \int_{t_i}^{\infty} g(t - t_i; m_i)dt = n_m. \tag{4.29}$$

The likelihood of N_b, the observed number of background events, is

$$P(N_b) = \frac{\exp[-\mu T](\mu T)^{N_b}}{N_b!}. \tag{4.30}$$

The likelihood that the ith event has l_i children is

$$P(l_i) = \frac{\exp[-G_i(\theta)](G_i(\theta))^{(l_i)}}{l_i!} \tag{4.31}$$

The product $P(N_b)\prod_i P(l_i)$ gives the likelihood of the observed "tree" structure. The *complete* likelihood also includes terms for the times of the events, specified in terms of times between the triggering and triggered events. (Times of background events do not enter the likelihood, since the background rate is the same at all times.) It is

$$L_c(\theta) = P(N_b)\prod_i P(l_i) \prod_{i:u_i \neq 0} \frac{g(t_i - t_{u_i}; m_{u_i})}{G_{u_i}(\theta)}. \tag{4.32}$$

Taking logs and simplifying, we get the complete log-likelihood:

$$\begin{aligned} l_c(\theta) = &- \log(N_b!) - \mu T + N_b \log(\mu T) + \sum_i [-\log l_i! - G_i(\theta) + l_i \log(G_i(\theta))] \\ &+ \sum_{i:u_i \neq 0} [\log(p-1) + (p-1)\log c - p\log(t_i - t_{u_i} + c)]. \end{aligned} \tag{4.33}$$

Still assuming that the $\{u_i\}$ were known, we could maximise by taking partial

derivatives and setting to zero. The partial derivatives are

$$\frac{\partial l_c}{\partial \mu} = \frac{N_b}{\mu} - T \tag{4.34}$$

$$\frac{\partial l_c}{\partial K} = -\frac{1}{K}\sum_i (G_i(\theta) - l_i) \tag{4.35}$$

$$\frac{\partial l_c}{\partial \alpha} = -(\log 10)\sum_i [(m_i - m_0)(G_i(\theta) - l_i)] \tag{4.36}$$

$$\frac{\partial l_c}{\partial c} = \sum_{i:u_i \neq 0}\left(\frac{p-1}{c} - \frac{p}{t_i - t_{u_i} + c}\right) + \frac{p-1}{c}\sum_i (G_i(\theta) - l_i) \tag{4.37}$$

$$\frac{\partial l_c}{\partial p} = \sum_{i:u_i \neq 0}\left(\frac{1}{p-1} + \log c - \log(t_i - t_{u_i} + c)\right) + \left(\frac{1}{p-1} + \log c\right)\sum_i (G_i(\theta) - l_i)\,. \tag{4.38}$$

Setting the right-hand side of (4.35) to zero requires $\sum_i (G_i(\theta) - l_i) = 0$, so setting the right-hand sides of (4.37) and (4.38) to zero implies

$$\sum_{i:u_i \neq 0}\left(\frac{p-1}{c} - \frac{p}{t_i - t_{u_i} + c}\right) = 0, \tag{4.39}$$

$$\sum_{i:u_i \neq 0}\left(\frac{1}{p-1} + \log c - \log(t_i - t_{u_i} + c)\right) = 0. \tag{4.40}$$

The algorithm roughly set out below estimates parameters that maximise the complete log-likelihood.

1. Set initial values.
2. Estimate μ.
3. Keeping other parameter estimates at their current values, set (4.35) and (4.36) to zero and solve numerically for K and α.
4. Keeping other parameter estimates at their current values, solve (4.39) and (4.40) numerically for c and p.
5. If estimates have changed by less than some stopping criterion, finish; otherwise return to step 2.

4.4.2 EM-type estimation

Since the parents $\{u_i\}$ are in practice unobservable, the complete log-likelihood cannot be directly calculated in practice. Instead, an EM algorithm computes the probability that the ith earthquake was triggered by the jth earthquake, conditional on the current parameter estimates and on the observed events up to t_i, for all $j < i$. The algorithm calculates an *expected* complete log-likelihood, weighted by those probabilities. Conditional on $\mathcal{F}_t$ (the σ-field generated by times and magnitudes, but not the branching structure, up to time t), $P(u_i = j)$ is zero for $j \geq i$, and

$$P(u_i = j) = \frac{g(t_i - t_j, m_j)}{\mu + \sum_{r=1}^{i-1} g(t_i - t_r, m_r)} \tag{4.41}$$

for $1 \leq j < i$. The model probability that an event was a background shock is

$$P(u_i = 0) = \frac{\mu}{\mu + \sum_{r=1}^{i-1} g(t_i - t_r, m_r)}. \tag{4.42}$$

Then the EM-type algorithm iterates between computing the expected complete log-likelihood as above (the E-step), and maximising it (the M-step). One version of this algorithm is:

1. Set the counter $j = 1$. Set the parameter vector θ to sensible initial values.
2. Use (4.41) and (4.42) and the current estimate of θ to calculate the triggering probabilities.
3. Substitute $\sum_i P(u_i = 0)$ for N_b in (4.34) and set to zero; solve for μ.
4. Substitute $\sum_{s \geq i+1} P(u_s = i)$ for l_i in (4.35) and (4.36) and set to zero; solve for K and α.
5. Substitute $\sum_{s \geq i+1} P(u_s = i)$ for l_i in (4.39) and (4.40) and solve for c and p.
6. If the changes in parameter estimates are small, stop; otherwise, return to step 2.

Variations to increase the speed of the algorithm are possible.

4.4.3 Example: Southern California seismicity

Veen and Schoenberg fitted a space-time ETAS to the SCEC catalog of events of magnitude 3 and greater occurring in Southern California from January 1st, 1984 to

Parameter	VS spatial estimate	Temporal estimate
μ	0.165 per day	0.330 per day
K	0.0423	0.0225
α	0.449	0.688
c	0.0192 days	0.0377 days
p	1.22	1.39
$\mathbf{E}\lambda$	0.97 per day	0.92 per day

Table 4.2: Comparison of parameter estimates for space-time and temporal ETAS models fitted to Southern Californian seismicity. The models are fitted to the SCEC catalog of magnitude 3 or greater events, from January 1st, 1984, to June 17th, 2004. The column "VS spatial estimate" gives temporal parameter estimates derived from the Veen and Schoenberg [4] spatio-temporal ETAS parameter estimates.. In this column, μ is the integral of Veen and Schoenberg's spatial background rate estimate over the area of study; $K = \pi K_0/(\rho d^\rho)$ converts Veen and Schoenberg's space-time parameters to a temporal parameter by integration; and α is relative to base 10 (instead of base e). The column "Temporal estimate" gives parameter estimates using a temporal ETAS model. The estimates are quite different. The space-time model uses locations to help determine branching structure, which may be responsible for the larger clusters in that model. The average rate of events $\mathbf{E}\lambda$ differs by 5% between the two models; this may be due in part to small differences in the catalogs, and in part to rounding error.

June 17th, 2004. We fitted a temporal ETAS model to the same catalog using code provided by Veen [114]. Table 4.2 shows the parameter estimates are very different.

4.4.4 Variability of estimates

Veen and Schoenberg studied the variability of the parameter estimates of their EM-type algorithm using simulations of a space-time ETAS model with parameters as given in Table 4.3. The parameters were intended to model earthquakes of magnitude at least 2 in Southern California between longitudes $-122°$ and $-114°$ and latitudes $32°$ and $-37°$, and were chosen based on the work of Ogata and discussions with seismologists. In their model, the squared distance between an offspring event and its parent follows a Pareto distribution, as in equation (4.15). Converting their space-time ETAS parameters to temporal ETAS parameters, we obtain a value $K = 0.00345$. Using a GR distribution truncated at 8 for magnitude, the parameters give a branching ratio of 0.953, implying stationarity. A background shock, on average, generates 21 descendants. Small changes in almost any of these parameters can lead to large changes in the expected total number of aftershocks per back-

Parameter	Value
m_0	2
m_1	8
μ	0.032 per day in the region
α	1
b	1
c	0.01 days
p	1.5
d	0.015
ρ	0.8
K_0	3.05×10^{-5}

Table 4.3: "Typical" space-time ETAS parameter values used by Veen and Schoenberg [4] to simulate Southern Californian seismicity.

Param.	True value	200,000 day burn-in		10,000 day burn-in		No burn-in	
		Mean	RMSPE	Mean	RMSPE	Mean	RMSPE
μ	0.0008	0.00077	11.9	0.00078	13.1	0.00078	13.1
K	0.00345	0.00338	36.7	0.00339	43.0	0.00327	38.1
α	1	1.02	13.6	1.000	13.9	1.008	11.8
c	0.01	0.013	98.5	0.013	87.5	0.014	223
p	1.5	1.546	12.4	1.573	15.0	1.582	30.6

Table 4.4: Results of fitting ETAS model to three sets of 100 simulated catalogs, each of length 100,000 days: one set with 200,000 days burn-in per catalog, one set with 10,000 days burn-in per catalog, and one with no burn-in. The simulation parameters were as in Table 4.3. The magnitude distribution is truncated GR with $2 \leq M \leq 8$. "Mean" is the mean of estimates. "RMSPE" is root mean square percentage error. Note that the estimator failed to converge for one of the catalogs with no burn-in; this is not reflected in the table.

Param.	True value	Mean	RMSPE
μ	0.1687	0.3365	99.9
K	0.04225	0.03407	19.9
α	0.4491	0.1956	56.5
c	0.01922	0.04125	115.8
p	1.222	1.507	23.4

Table 4.5: Results of fitting ETAS model to 100 simulated catalogs, each of length 20 years (7305 days). The burn-in time was 1000 years. The simulation parameters, in column "True value," were those fitted by Veen and Schoenberg for Southern California seismicity [4]. The magnitude distribution was truncated GR with $3 \leq M \leq 8$. "Mean" is the mean of estimates. "RMSPE" is root mean square percentage error. Note that the estimator failed to converge for one of the catalogs; this is not reflected in the table.

Parameter	Catalog length (days)		
	50,000	100,000	200,000
μ	17.9	11.9	9.1
K	74.7	36.7	30.1
α	21.6	13.6	7.0
c	358	98.5	34.0
p	32.0	12.4	5.4

Table 4.6: Root mean square percentage error for EM-type algorithm parameter estimates for simulated ETAS catalogs with 200,000 days burn-in. One hundred catalogs of each of the lengths 50,000 days, 100,000 days and 200,000 days were simulated and fitted using Veen and Schoenberg's EM-type algorithm. The simulation parameters were the temporal ETAS parameters in Table 4.3. The magnitude distribution was truncated GR with $2 \leq M \leq 8$. Note that the estimator failed to converge for three of the length 50,000 catalogs; this is not reflected in the table.

ground shock. Veen and Schoenberg found substantially reduced bias in fitting using the EM-type algorithm compared to conventional maximum likelihood—in the latter case, convergence was poor if tolerance levels and stopping criteria were not optimal, whereas the EM estimates were more stable. The variability of EM estimates was still high, however, particularly in K_0, α, and d.

Table 4.4 summarises estimation results for simulations using the parameters in Table 4.3. Estimates of μ, α, and p were fairly accurate—usually within 20% of true values. However, estimates of K and c were less accurate. Overall, estimates are less accurate than Veen and Schoenberg found for their space-time ETAS estimates, primarily due to bias. The space-time model is perhaps better at picking out the correct branching structure. Simulations made without any burn-in time gave very poor estimates for c and p when the temporal ETAS model was fitted. However, estimates of α were, on average, slightly better for simulations with no burn-in. While the 200,000-day burn-in simulations gave the most accurate estimates for μ, K, and p, the 10,000-day burn-in simulations gave more accurate estimates of c.

Table 4.5 summarises estimation results for a different set of parameters, equivalent to those estimated by Veen and Schoenberg for Southern California seismicity for a space-time ETAS model. The simulations were for 20 years (7,305 days). The results here are notably poorer. Table 4.6 indicates that estimation is more accurate for larger catalogs. Error decreases as roughly $\sqrt{n}$ for estimates of μ and K, and more quickly than this for α, c, and p.

4.4.5 Goodness-of-fit

How well do estimated ETAS models fit seismicity? That is, can we test the hypothesis that observed seismicity is a realisation of an ETAS process with parameters equal to those fitted to the observed seismicity? Veen and Schoenberg [115] have proposed a test of spatial goodness-of-fit. We assess a temporal ETAS model by applying a test that treats times and magnitudes separately.

The temporal component of the test considers only the times between successive events. The distribution of inter-event times in the estimated ETAS model may be approximated through simulation. Figure 4.1 plots the empirical distribution of times between magnitude 3 or greater events in the SCEC catalog from January 1st, 1984, to June 17th, 2004 (7,474 days). It compares this to the inter-event time distribution for the estimated ETAS model, and for a gamma renewal process. The catalog has a greater proportion of inter-event times of less than two hours than occur in either model.

We use the Kolmogorov-Smirnov statistic—the supremum of the differences between the ETAS cumulative distribution function for inter-event times and the empirical cumulative distribution function for a sample of inter-event times—in a test of goodness-of-fit. When the empirical data is the Southern Californian inter-event

times, the Kolmogorov-Smirnov statistic is 0.151. Since the inter-event times in ETAS are dependent, we cannot use the Dvoretzky-Kiefer-Wolfowitz inequality 3.21 or similar bounds to find the P-value; instead, we use simulation to sample from the null distribution for the statistic. We use the fitted ETAS model to simulate 100,000 catalogs, each of length 7,474 days (with 200,000-day burn-in). The Kolmogorov-Smirnov statistics for the 100,000 simulated catalogs ranged from 0.004 to 0.082. The P-value for the null hypothesis that the data are a realisation of an ETAS process is less than 10^{-5}. Note that this test is conservative, as the ETAS parameters are estimated from the data.

The magnitude distribution was not significantly different from truncated GR (Kolmogorov-Smirnov P-value 0.6768).

4.4.6 Classification

As mentioned in chapter 3, the ETAS model can be used to classify events as background or offspring—essentially by solving equations (4.41) and (4.42) in the E-step of the EM-type algorithm above. Sornette and Utkin [94] simulated ETAS catalogs. They used stochastic declustering to classify events as background or offspring. This essentially uses a similar method to the EM-type algorithm above to estimate the branching variables $\{u_i\}$. They found that classification was "rather unreliable."

4.5 Summary

The ETAS model is an intuitively appealing clustering model for seismicity. It is based on the well-established modified Omori and Gutenberg-Richter empirical laws. It allows aftershocks, and aftershocks of aftershocks, but is still an oversimplification of earthquake physics. Simulation from the model requires care—if the Omori parameter p is less than 2, very long burn-in times are needed. When the model is fitted to real data using numerical maximum likelihood, the parameter estimates often imply the process is explosive: every earthquake is expected to trigger an infinite number of aftershocks.

Veen and Schoenberg showed that an EM-type algorithm performs better than conventional numerical maximum likelihood for estimating a space-time ETAS model. However, the EM-type algorithm for temporal ETAS does not estimate parameters accurately from "realistic" simulated catalogs where the model is true—even for catalogs over 500 years long. Furthermore, for real data, the fitted ETAS model does not provide a statistically adequate fit to observed earthquake inter-event times. Short inter-event times are significantly more common than the model predicts. In addition, classification of events as background or offspring using ETAS models is inaccurate—even when the ETAS model is true.

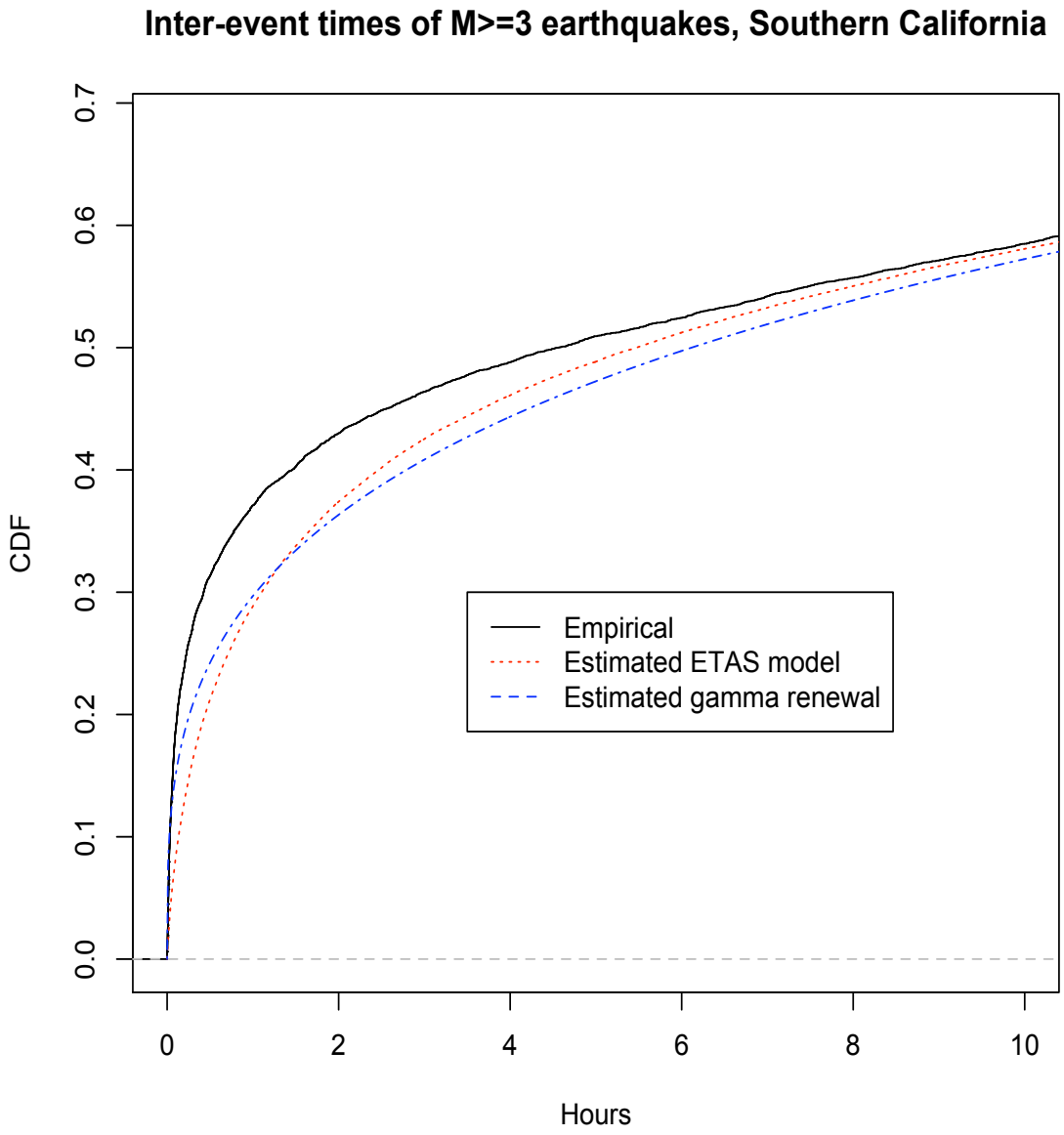

Figure 4.1: Cumulative distribution functions of inter-events times attached. The empirical inter-event distribution (SCEC catalog of Southern Californian $M \geq 3$ earthquakes, 1984-2004, $n = 6958$) is significantly different from both the fitted ETAS and gamma renewal models (in both cases, the P-value is less than 0.00001 for a test using the Kolmogorov-Smirnov test statistic). Empirically, there are more inter-event times under 2 hours than either fitted model would suggest. Beyond 12 hours, the difference in empirical distributions is small (not pictured).

If ETAS models did a good job of predicting real seismicity, they would still be of value. We examine prediction in the following chapter.

Chapter 5

Prediction of renewal processes and the ETAS model

Earthquake predictions have been based on fitting ad hoc stochastic models of earthquake occurrence—such as the ETAS model discussed in the previous chapter—to observed seismicity. The predictions are then constructed to be optimal when the stochastic model is true and the estimated parameter values are correct. However, the statistical and physical justification of such models is quite weak. For many stochastic models of earthquake occurrence, simple automatic alarms, which turn on for a fixed interval after each sufficiently large event, are optimal. That is not the case for ETAS. However, when the ETAS model is correct, the optimal predictions perform only slightly better than a far simpler prediction scheme, magnitude-dependent automatic alarms, which turn on after each sufficiently large event for an interval that depends exponentially on the magnitude of that event.

5.1 Introduction

Earthquakes are incredibly complicated phenomena. *Ab initio* physical models of earthquake occurrence remain relatively crude; stochastic phenomenological models are common [74, 116, 117, 87], including Poisson processes; more general renewal processes, including gamma, lognormal, Weibull, and others; and branching processes like the ETAS model of the previous chapter. The connection between stochastic descriptive models and the underlying physics is tenuous, and the utility of such stochastic models for predicting real, destructive, earthquakes remains unproved. It is implausible that any of these models is correct; in fact, there is statistical evidence that none is.

If we pretend that seismicity really does adhere to a stochastic process model, we could answer some interesting questions. This chapter addresses a few. How predictable are earthquakes? What methods attain or approach the theoretical limits

on predictive accuracy? Are there simple predictors that perform comparably to theoretically optimal predictors on real seismicity? Answering these questions is interesting because predictions that use extra-seismic data should improve on simple methods that do not. If a method uses electrical signals [26], groundwater level and temperature [118], or animal behavior [22, 23] to predict earthquakes, but does not perform better than a simple method that bases predictions on past seismicity, it is of little value.

In this chapter we study the limits of earthquake predictability on the assumption that earthquakes follow a stochastic process—a convenient fiction. We study *alarms* that are on when an earthquake is predicted, and off when no earthquake is predicted. We characterise optimal alarms in terms of the conditional intensity of the stochastic process. We study the accuracy of simple earthquake predictions, *automatic alarms*, which turn on after each sufficiently large earthquake and remain on for a window of time w unless there is another sufficiently large earthquake while the alarm is on. In that case, the alarm is refreshed so that it remains on for a period w after that new earthquake. The window w may be the same for all earthquakes—*simple automatic alarms*—or may depend on the magnitude of the earthquake—*magnitude-dependent* automatic alarms.

Automatic alarms with constant w are in fact optimal for a large class of stochastic process models, including Poisson and gamma renewal models for $\gamma \geq 1$. Indeed, they are nearly optimal for almost every[1] stochastic model of seismicity we have seen in the literature. Moreover, automatic alarms are simple and have an intuitive justification: seismicity tends to cluster in time and space, so after each earthquake, it is reasonable to expect more earthquakes near and soon. In general, larger earthquakes are more frequently followed by further shocks. We find these observations to be a strong argument for using automatic alarms as a touchstone method for more complicated predictions and predictions that rely on extra-seismic data: if those predictions do not substantially outperform automatic alarms, their complexity might not be worthwhile.

One measure of the success of an alarm strategy is the proportion of earthquakes that occur when the alarm is on. However, we can trivially ensure this proportion is always one by turning the alarm on for the entire region. When studying the success of an alarm strategy, we therefore also consider the *alarm fraction*—that is, the fraction of space-time that the alarm is on, either in expectation or for observed seismicity. We wish to find the alarm strategy that maximises the expected number of earthquakes that occur when the alarm is on, subject to fixing the alarm fraction.

Section 5.3 examines prediction of renewal processes. In these processes, the times between earthquakes are independent and identically distributed. In some cases, the

[1] They are not optimal for processes in which the conditional intensity increases with time after earthquakes. Models based on the *seismic gap hypothesis* [119] are of this form. For such models, *delayed automatic alarms*, which turn on after waiting a period after sufficiently large earthquakes, are optimal.

automatic alarm strategy is the optimal strategy; in others, it is the worst possible strategy. The performance of the best-fitting renewal model is a widely applicable measure of predictability.

Section 5.4 examines prediction of ETAS processes. The difference in predictive performance between the optimal strategy and a magnitude-dependent automatic alarm strategy is quite small, even when the true model is ETAS. Section 5.5 examines optimal ETAS and automatic alarm prediction for Southern Californian seismicity. A magnitude-dependent automatic alarm strategy, with two parameters and one degree of freedom, does almost as well as the hard-to-fit five-parameter temporal ETAS model.

Section 5.6 makes recommendations for the future use of the ETAS model. The ETAS model has some predictive power. However, since a magnitude-dependent automatic alarm strategy made predictions that were almost as good for the data we examined, the value of the more complex model is limited.

5.2 Alarms, conditional intensity, and the error diagram

We wish to model the sequence of earthquakes with magnitude at least m_0 in a geographical study area A over a study period of time $(0, T]$ as a stochastic point process. We ignore earthquakes with magnitude less than m_0. An *event* is an earthquake with magnitude at least m_0 occurring in the study area during the study time. In a marked temporal point process, a set of events is characterised by the event times $\{T_i\}$ and magnitudes $\{M_i\}$; the spatial locations are all in A but are otherwise ignored. In a marked space-time point process, events are characterised by their times, locations, latitudes, and longitudes.

Consider predictors that take the form of "alarms" in time or in space-time. At any point in the study region $A \times (0, T]$, the alarm is either on or off, depending on previous seismicity. Roughly speaking, we want the alarm to be on when and where events occur, and to be off over large regions where there are no events.

An *alarm strategy* $H(t) \in \{0, 1\}$ is a rule to determine when and where the alarm is on, with 1 representing "on" and 0 representing "off." (Later we will consider alarms that may be on with some probability in $[0, 1]$.) The rule must be specified before the beginning of the study period, but the alarms themselves may not be. An example is the *simple automatic alarm strategy*. For predictions in time, this strategy turns on the alarm for a fixed duration of time w after each observed event. For predictions in space-time, the strategy turns on the alarm for duration w after each observed event in an area of radius r around that event. We discuss automatic alarm strategies further in section 5.2.4.

5.2.1 Error diagrams

A simple way to evaluate alarm strategies is to use a scoring system that rewards correct forecasts and punishes incorrect ones. Molchan [120] characterised the performance of an alarm $H(t)$ using two statistics: $\hat{\tau}$, the alarm fraction (fraction of space-time taken up by alarms), and $\hat{\nu}$, the fraction of events occurring outside of alarms (fraction of failures to predict). For a temporal point process on $(0, T]$:

$$\hat{\tau} \equiv \frac{\int_0^T H(t)dt}{T}, \tag{5.1}$$

$$\hat{\nu} \equiv \frac{\int_0^T H(t)dN(t)}{N(T)}. \tag{5.2}$$

(We may instead consider $\tau \equiv \mathbf{E}\hat{\tau}$ and $\nu \equiv \mathbf{E}\hat{\nu}$ when these expectations exist; see section 5.2.2.) Two other statistics that may be of interest are the number of times the alarm is turned on and an event occurs before the alarm is turned off; and the number of times the alarms is turned on and no event occurs before the alarm is turned off.

Not all forecasts are expressed as alarms. For example, some probabilistic forecasts partition the study region into *cells*. For each cell, they give a probability that at least one event will occur in that cell. We can convert such a probabilistic forecast to an alarm by comparing the forecast probabilities to some threshold probability p_0. If the forecast probability for a cell exceeds p_0, then turn on the alarm over that cell; otherwise, do not turn on the alarm. We can find $\hat{\nu}$ and $\hat{\tau}$ for the resulting alarm. If the set of forecast probabilities is determined prior to the start of the study period, we could choose the threshold p_0 in advance to give a particular value of $\hat{\tau}$. However, the full set of forecast probabilities is not always known at the start of the study period—for instance, if the forecast probability for a cell may depend on the seismicity between the start of the study period and the start of the cell. In this case, we can either choose p_0 before the study and find $\hat{\tau}$ after the study, or else choose $\hat{\tau}$ before the study and find the p_0 that gives that value of $\hat{\tau}$ after the study.[2] This methodology allows different forecasting schemes to be compared if they have the same value of $\hat{\tau}$. The probabilistic nature of the forecasts is not utilised in such a comparison.

The *error diagram*, or *Molchan diagram* [120], is a plot of a set of points $(\hat{\tau}, \hat{\nu})$ obtained by considering a continuum of alarm thresholds. For the probabilistic forecasts outlined above, it calculates $(\hat{\tau}, \hat{\nu})$ for all thresholds p_0. The error diagram is a variation of the receiver operating characteristic (ROC) curves first used to measure the performance of radar image analysis in the Second World War, and now widely used

[2] The test is still prospective if the forecasting method and the method for finding p_0 are determined before the start of the study period and are not changed thereafter.

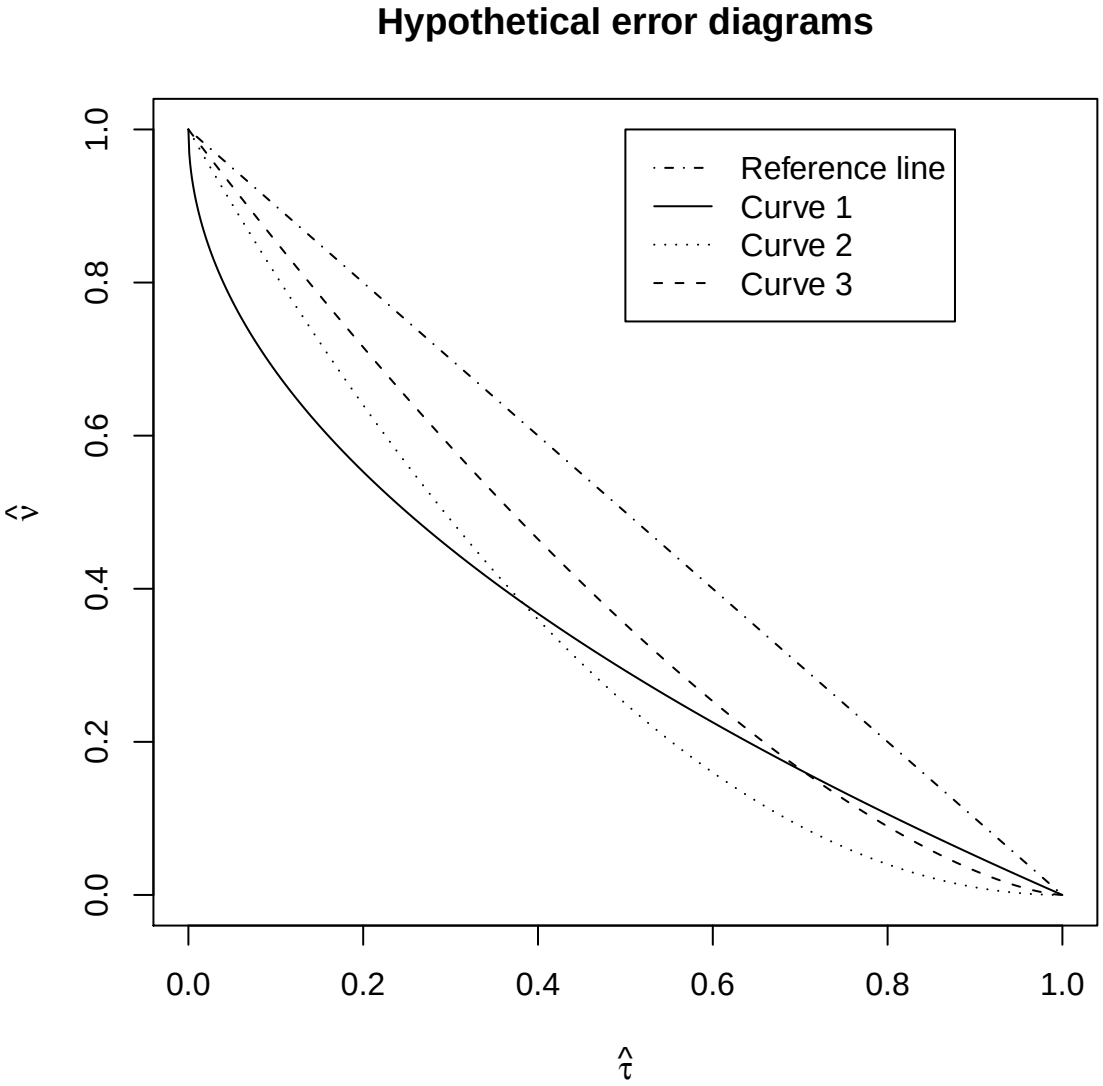

Figure 5.1: Three hypothetical error diagram curves, compared to a reference line. On the x-axis, $\hat{\tau}$ is the proportion of time covered by alarms. On the y-axis, $\hat{\nu}$ is the proportion of events that do not fall within alarms, a measure of error. Curve 1 gives lowest error for $\hat{\tau} < 0.38$. Curve 2 gives lowest error for $\hat{\tau} > 0.38$. Curve 3 is dominated by curve 2, and never gives lowest error. The reference line shows an error diagram for "random guessing": $\hat{\nu} = 1 - \hat{\tau}$.

in weather forecasting [121]. Unlike ROC curves, by convention the error diagram y-axis gives failures, not successes.

The error diagram is widely used in seismology [117, 122]. It may be constructed for any family of predictors that can produce alarms for multiple values of $\hat{\tau}$. Figure 5.1 displays several error diagrams. If one error diagram gives a lower value of $\hat{\nu}$ than another for all values of $\hat{\tau}$, then the former *dominates* the latter. Often, however, one forecast performs better for some thresholds, while another performs better for others (see Figure 5.1). To choose between forecasts, we can determine summary statistics based on their error diagrams. We could use the area under the curve for this purpose—the smaller the area, the better the forecast. The smallest possible area is 0; the largest possible area is 1. Using a random predictor (see the following subsection), we would expect an area of 0.5. Note that this is a fairly arbitrary way of choosing between forecasts. When using forecasts to make decisions, a cost-benefit analysis may be more informative.

We assume that alarm regions are nested. That is, if $\hat{\tau}_1 < \hat{\tau}_2$, the alarm region for $\hat{\tau}_1$ is a subset of the alarm region for $\hat{\tau}_2$. This results in error diagrams that are non-increasing in $\hat{\tau}$.

Drawing an error diagram for a particular forecast requires a rule to determine the regions in which alarms are declared. For example, suppose a model gives a predicted conditional intensity of events $\lambda(t)$ for a temporal point process on $(0, T]$. One rule is to declare alarms whenever the conditional intensity exceeds some threshold Λ. We may calculate values of $\hat{\tau}$ and $\hat{\nu}$ for the resulting alarms. By considering all possible values for the threshold Λ, we obtain a set of $(\hat{\tau}, \hat{\nu})$ pairs—that is, an error diagram. Level sets nest, satisfying the requirement that alarm regions nest. A higher value of $\hat{\tau}$ means the alarm is on more often, requiring a lower conditional intensity threshold.

If there is a Λ such that the level set $\{\lambda = \Lambda\}$ has positive measure, there will be a gap in the error diagram between points corresponding to strategies that just include and just exclude this level set in alarms. We may leave these gaps blank, or interpolate, or use a tiebreaking strategy to determine which parts of a region with constant λ are added to the alarm region first.

5.2.2 Expected error diagrams

An empirical error diagram for some alarm strategy plots the observed fraction of unpredicted events $\hat{\nu}$ for each value of the alarm fraction $\hat{\tau}$. If we are predicting a point process model for seismicity, and not seismicity itself, we may calculate *expected error diagrams*. These plot ν, the expected fraction of unpredicted events, against τ, the expected alarm fraction, for a family of alarm strategies. The expectation is over realisations ω of the point process.

An alarm strategy $H_t(\omega)$ may, in general, depend on the history of the process up to, but not including, time t. This history is denoted by the σ-field $\mathcal{F}_t$, a function of

t and ω. For a temporal point process on $(0, T]$,

$$\tau \equiv \mathbf{E}\left[\frac{\int_0^T H_t(\omega)dt}{T}\right], \tag{5.3}$$

$$\nu \equiv \mathbf{E}\left[\frac{\int_0^T H_t(\omega)dN(t)}{N(T)}\right]. \tag{5.4}$$

Example: random predictions

In a homogeneous temporal Poisson process, the conditional intensity of events is constant: it does not depend on the history of the process. If, before the study period begins, we choose an alarm set that covers a proportion τ of the study region, the expected proportion of events that occur while the alarm is on is τ. The expected error diagram is thus the line

$$\nu = 1 - \tau. \tag{5.5}$$

Now suppose we have a non-Poisson temporal point process. Consider the following random alarm strategy:

1. Divide the time interval into K subintervals of equal length T/K, where the ith subinterval is
$$((i-1)T/K, iT/K].$$
2. Randomly select j of the subintervals. Set $H = 1$ (that is, turn on the alarm) in these subintervals; set $H = 0$ otherwise. The probability an alarm is declared for a particular interval is $\tau = j/K$.

Let the number of events in the ith subinterval be N_i. Then

$$N(T) = \sum_{i=1}^{K} N_i.$$

The number of successfully predicted events in the ith subinterval is N_i if $H = 1$ for that subinterval, and 0 otherwise. The expected number of successfully predicted events in the interval, conditional on N_i, is τN_i. The expected total number of successfully predicted events is

$$\sum_i \tau \mathbf{E} N_i = \tau N(T).$$

For all $N(T) > 0$, if a fraction τ of the study period is covered with alarms by this strategy, we expect a fraction τ of the events that occur on $(0, T]$ to be captured by alarms, and a fraction $1 - \tau$ to be missed. The expected error diagram is thus given

by equation (5.5), and is the same as for predictions of a Poisson process. Note that this strategy is one of many random alarm strategies that give the same result in expectation.

If the process is not Poisson, a prediction scheme that exploits knowledge of the history of the process should do better than this: for instance, by exploiting clustering. Such a predictor should have $1 - \nu > \tau$, and, in a sufficiently long test, $1 - \hat{\nu} > \hat{\tau}$.

5.2.3 The optimal alarm strategy

Out of all alarm strategies that have the alarm on an expected proportion $\tau \in [0, 1]$ of the time, what strategy maximises the expected number of events that occur when the alarm is on? Below, we show turning on the alarm if and only if the conditional intensity $\lambda(t)$ is greater than some threshold Λ is optimal.[3]

Optimal alarm lemma: Let $N(t)$ be an orderly[4] temporal point process. Let $\mathcal{F}_t \equiv \sigma\{N_s : 0 \le s \le t\}$ be the σ-field generated by the process up to time t.

Let μ be the product measure of Lebesgue measure on $[0, T]$ and the probability measure on the space Ω of realisations ω of the point process. A stochastic process $H = H_t(\omega)$ is *adapted* if H_t is $\mathcal{F}_t$-measurable for each $t \in [0, T]$. The *previsible* or *predictable* σ-field $\mathcal{P}$ is the σ-field on $[0, T] \times \Omega$ generated by the left-continuous, adapted processes. A stochastic process H is said to be *previsible* if $(t, \omega) \mapsto H_t(\omega)$ is $\mathcal{P}$-measurable. If H is previsible, then for any $0 \le t \le T$ the restriction of H to $[0, t] \times \Omega$ is $\mathcal{F}_{t-}$-measurable, where $\mathcal{F}_{t-} \equiv \sigma\{N_s : 0 \le s < t\}$. In particular, if H is previsible, then H_t is $\mathcal{F}_{t-}$-measurable for all $t \in [0, T]$.

Similarly, for $t \in [0, T]$, let μ_t be the product measure of Lebesgue measure on $[0, t]$ and the probability measure on $\mathcal{F}_t$. We consider previsible functions of the form $H_t(\omega)$ that take values in $[0, 1]$ and are progressively measurable with respect to μ_t for $t \in [0, T]$. Then H_t is the probability that an alarm is on at time t, conditional on the history of the process up to but not including time t.

We wish to find H maximising

$$\mathbf{E}\left[\int_0^T H_t \, dN(t)\right] \tag{5.6}$$

subject to

$$\mathbf{E}\left[\frac{1}{T}\int_0^T H_t \, dt\right] = \tau \tag{5.7}$$

for a fixed $\tau \in [0, 1]$. Equation (5.6) gives the expected number of events that occur during alarms.

[3] We received a huge amount of assistance from Steven N. Evans for this section.

[4] In an *orderly* point process, events cannot occur simultaneously.

The *dual previsible projection* is the unique non-negative, non-decreasing, right-continuous $\mathcal{F}_t$-previsible process L such that

$$\int_0^t X_s N(ds) - \int_0^t X_s L(ds) \tag{5.8}$$

is an $\mathcal{F}_t$-martingale for all bounded previsible processes X. Assume L is absolutely continuous with respect to Lebesgue measure on $[0,T]$, almost surely with respect to the probability measure on Ω. Then the *conditional intensity process* λ defined by $L_t = \int_0^t \lambda(s)\,ds$ is previsible. Also assume that $\mathbf{E}\lambda(t)$ exists and is finite for all $t \in [0,T]$. Then:

(i) **Existence:** There exists an H satisfying (5.7) with

$$H_t = \begin{cases} 1 & \text{when } \lambda(t) > \Lambda \\ 0 & \text{when } \lambda(t) < \Lambda \end{cases} \tag{5.9}$$

almost everywhere with respect to μ, for some constant $\Lambda \in [0,\infty]$.

(ii) **Sufficiency:** If a function H satisfies (5.7) and (5.9) for some Λ, then it maximises (5.6) subject to (5.7).

(iii) **Necessity:** If a function H maximises (5.6) subject to (5.7), then for some Λ it satisfies (5.9) almost everywhere with respect to μ.

Thus the *optimal alarm strategy* is to turn on the alarm when the conditional intensity is greater than (or possibly equal to) some threshold Λ. "Optimal" means that out of all strategies with alarm fraction τ, this strategy maximises the expected rate per unit time of events that occur when the alarm is on. This is assumed by Helmstetter and Sornette [97], and proved by Molchan [123] for stationary processes by analogy to the the Neyman-Pearson lemma [124, 90]. We give a proof for both stationary and non-stationary processes in Appendix C.

Let the conditional intensity $\lambda(t)$ have distribution function

$$F_t(x) = P(\lambda(t) \le x). \tag{5.10}$$

Suppose this distribution function is continuous. Then for any x, the measure of the set $\{\lambda(t) = x\}$ is zero. An optimal alarm strategy turns on the alarm when $\lambda(t) > \Lambda$ and turns off the alarm when $\lambda(t) < \Lambda$. The threshold Λ is chosen to satisfy

$$\tau = \frac{1}{T}\int_0^T [1 - F_t(\Lambda)]dt. \tag{5.11}$$

The expected fraction of events occurring outside alarms is

$$\nu = \frac{\int_0^T \int_0^\Lambda x dF_t(x)\, dt}{\int_0^T \int_0^\infty x dF_t(x)\, dt}. \tag{5.12}$$

Plotting ν against τ for all thresholds Λ gives an *optimal error diagram*.

If $F_t(\Lambda)$ is discontinuous, randomisation (over and above sample-path randomness) may be necessary to obtain some values of τ. Suppose there exists $\Lambda \geq 0$ such that $P(\lambda(t) = \Lambda) > 0$. Consider the following prediction strategy:

- Before the prediction period starts, flip a coin that comes up heads with probability $q \in [0, 1]$. If the coin comes up heads, turn on the alarm when $\lambda(t) = \Lambda$. If the coin comes up tails, do not turn on the alarm when $\lambda(t) = \Lambda$.
- When $\lambda(t) > \Lambda$, turn on the alarm.
- When $\lambda(t) < \Lambda$, do not turn on the alarm.

For a stationary point process, the proportion of time covered by alarms is then

$$\tau = 1 - F_t(\Lambda) + qP(\lambda = \Lambda).$$

We can choose q to give any value of τ between $P(\lambda(t) > \Lambda)$ and $P(\lambda(t) \geq \Lambda)$. The expected fraction of events successfully predicted is

$$1 - \nu = \lim_{\epsilon \downarrow 0} \int_{l+\epsilon}^\infty x dF_t + qlF(l). \tag{5.13}$$

This is equivalent to linear interpolation between two points on the error diagram—the point given by declaring an alarm when $\lambda > \Lambda$ and the point given by declaring an alarm when $\lambda \geq \Lambda$. Note there are non-randomised strategies that give the same ν for each τ.

5.2.4 Automatic alarms

Automatic alarms for temporal processes

Consider a naive predictor of a temporal point process that, after an event with magnitude M_i at time T_i, turns on an alarm for duration $w(M_i)$ immediately following the event—that is, on the interval $(T_i, T_i + w(M_i)]$. If no event occurs while the alarm is on, the alarm is switched off until another event occurs. If the Jth event occurs during an alarm, the alarm is extended to last until the time

$$\max_{j \leq J} (T_j + w(M_j)). \tag{5.14}$$

This is an *automatic alarm strategy* [125, 44] for a temporal point process. We call the case where the window duration w is the same for all magnitudes the *simple automatic alarm strategy*, and the case where the duration depends on magnitude the *magnitude-dependent automatic alarm strategy*, or *MDA alarm strategy*.

Since earthquakes cluster in time, the simple automatic alarm strategy will, in the long run, predict earthquakes more successfully than a strategy that assigns alarms uniformly at random within the study region. This success is not surprising, but it provides a reference level of success that more complex predictors should surpass to be useful.

Empirically, the rate of seismicity is elevated for longer after large earthquakes than after small earthquakes. In an MDA alarm strategy, the window duration should increase with magnitude. An MDA alarm strategy that scales sensibly with magnitude should outperform the simple automatic alarm strategy. We choose to use

$$w(M) = ku^M. \tag{5.15}$$

The parameter u can be fitted to data to optimise some goodness-of-fit or predictive criterion. For example, in section 5.4.4, we choose u to minimise the area under the error diagram when MDA alarms are applied to a training data set. Of course, other window functions could be used. However, this exponential form seems intuitively reasonable for most magnitudes, as seismic moment scales exponentially with the moment magnitude scale.[5]

We can construct error diagrams for automatic alarm strategies. For the simple automatic alarm strategy, we obtain a set of values of $\hat{\tau}$ and $\hat{\nu}$ by considering all values of w in $[0, \infty)$. For the MDA alarm strategy, we assume the form (5.15), fix u, and consider all values of k in $[0, \infty)$. We can compare the error diagram given by some prediction strategy in some study region to that given by simple and MDA alarm strategies. A useful prediction strategy should give lower values of $\hat{\nu}$ than both simple and MDA alarm strategies for some or all values of $\hat{\tau}$. A prediction strategy is of little interest if it is dominated by simple automatic or MDA alarms.

Automatic alarms for space-time processes

An automatic alarm strategy for a marked space-time point process turns on an alarm of duration $w(M_i)$ in a spatial area $r(x_i, y_i, M_i)$ following each event. That is,

$$H(x, y, t) = \begin{cases} 1, & \text{if } (x, y, t) \in \cup\{R(x_i, y_i, M_i) \times (T_i, T_i + w(M_i)]\} \\ 0, & \text{otherwise,} \end{cases} \tag{5.16}$$

[5] A cap on alarm size to prevent alarms lasting decades may be required if very large earthquakes—say $M > 8$—are observed.

where $R(x_i, y_i, M_i)$ is a spatial region containing the epicenter of the ith event. This region can be chosen in a number of ways—it may be circular with radius scaling with magnitude, or it may reflect the geometry of faults. The "naive predictor" (i) of chapter 2.5 is an example of a simple automatic alarm strategy in space-time. In that strategy, the regions $R(x_i, y_i, M_i)$ are circular and centred at the epicenter, and the durations and radii of alarms are the same for all events.

5.3 Prediction of renewal processes

Renewal processes are temporal point processes in which the times between occurrences of events are independent and identically distributed. They may be marked or unmarked. The distribution of the renewal times may be defined by a distribution function, or by a *hazard function* giving the conditional intensity as a function of the time since the last event.

5.3.1 Alarms for a renewal process

Suppose that events occur according to a renewal process on $\mathbf{R}$, and that the process is in its stationary state at time 0. Let $D(x)$ be the distribution function of the inter-event times and $\delta(x) \equiv D'(x)$ be the probability density; assume $D(x)$ is continuous. The hazard function at time t is the limit of the probability that x falls between t and $t + \Delta t$ divided by Δt, conditional on $x \geq t$:

$$h(t) \equiv \lim_{\Delta t \downarrow 0} \frac{D(t + \Delta t) - D(t)}{\Delta t 1 - D(t)} \tag{5.17}$$

$$= \frac{\delta(t)}{1 - D(t)}. \tag{5.18}$$

Denote by $\lambda(t)$ the conditional intensity at time t. We have

$$\lambda(t) = h(B_t), \tag{5.19}$$

where B_t is the length of time back to the last event before time t. It is well-known [126] that B_t has density $\bar{\lambda}(1 - D)$, where $\bar{\lambda}$ is the reciprocal of the expected inter-event time. As before, the optimal alarm strategy turns on the alarm when the conditional intensity exceeds some threshold—that is, when $h(B_t)$ is large.

Suppose that h is monotone decreasing with inverse function η. Then

$$P(\lambda(t) > x) = P(B_t < \eta(x)) \tag{5.20}$$

$$= \bar{\lambda} \int_0^{\eta(x)} (1 - D(u))\, du. \tag{5.21}$$

Therefore, if we want the alarm to be on an expected fraction τ of the time, we optimally have the alarm on when the time back to the most recent event is less than $w(\tau)$, where $w(\tau)$ solves

$$\tau = \bar{\lambda} \int_0^{w(\tau)} (1 - D(u))\, du. \tag{5.22}$$

This is the simple automatic alarm strategy for a renewal process. The simple automatic alarm strategy is optimal when the hazard is decreasing. A necessary but not sufficient condition for this is that $\delta(x)$ be decreasing.

What proportion of events will occur during alarms under this strategy? If an event occurs at time t, then it will occur during an alarm if $B_t < w(\tau)$. Given that an event occurs at time t, the conditional distribution of B_t is just the inter-event time distribution. The expected proportion of events caught by automatic alarms of length w is therefore $D(w)$. If the measure of every level set of λ is zero, any value of τ can be obtained without randomisation.

If the hazard is decreasing, then so is the density function, because $\delta = h(1 - D)$ is the product of two decreasing functions. Consequently, if the hazard is decreasing, a strategy that declares an alarm when the density is large is equivalent to one that declares an alarm when the hazard is large. The density function can, however, decrease without the hazard decreasing. The hazard is

$$h(x) = -\frac{d}{dx} \log(1 - D), \tag{5.23}$$

so

$$\frac{d}{dx} h(x) = -\frac{d^2}{dx^2} \log(1 - D). \tag{5.24}$$

So the hazard is decreasing if and only if the log of $(1 - D)$ is convex.

When an unmarked renewal process model is fitted to real seismicity, the hazard function is generally decreasing (Figure 5.3). If the hazard is not always decreasing, then the optimal strategy is still to turn on the alarm when the hazard is large, but this will no longer give an automatic alarm strategy. If, for example, the hazard has a single maximum away from 0, the optimal strategy may be to wait a time after the observation of an event; if no event has occurred, then turn on the alarm for a fixed length of time.

5.3.2 Marked renewal processes

In a marked renewal process, inter-event times are still iid, but the events have marks. Most simply, the marks may be drawn independently from some distribution. For instance, a marked renewal process model of earthquakes may have magnitudes drawn independently from a truncated Gutenberg-Richter distribution. If the

inter-event times and the magnitudes are independent, then the conditional intensity depends only on the time since the last event, and the analysis of section 5.3.1 still holds. More generally, the hazard function may depend on the magnitude of the last event.

5.3.3 Success of automatic alarms

This subsection deals with the success of the simple automatic alarm strategy applied to unmarked renewal processes. The work in this subsection is based on the work of Kagan [117] and Molchan [123].

Error diagram

Consider a simple automatic alarm strategy that declares an alarm for a duration w after each event. Suppose the renewal times have differentiable distribution function $D(x)$ and density function $\delta(x)$. The average rate of events is $\bar{\lambda}$. From the previous subsection, we have

$$\nu(w) = 1 - D(w) \tag{5.25}$$

and $\tau(w)$ as given in equation (5.22).

We can find an alternative expression for $\tau(w)$ as follows. Consider $(T_i, T_{i+1}]$, the time interval between the ith and $(i+1)$th events. If $T_{i+1} - T_i \leq w$, the entire interval is covered by an alarm. If $T_{i+1} - T_i > w$, only the first w of the interval is covered by an alarm. So the expected time between $(T_i, T_{i+1}]$ that is covered by an alarm is $w(1 - D(w)) + \int_0^w [x\delta(x)]dx$. The expected fraction of time covered by alarms is

$$\tau(w) = \bar{\lambda} \left[\int_0^w x\delta(x)\, dx + w(1 - D(w)) \right]. \tag{5.26}$$

This is equal to the right-hand side of equation (5.22).

5.3.4 Success of general and optimal alarms

Suppose that after the ith event occurs at time T_i we declare an alarm for the period

$$A_i = (\max\{T_i + w_1, T_{i+1}\}, \min\{T_i + w_2, T_{i+1}\}] \tag{5.27}$$

for some $w_1 < w_2, w_1 \in [0, \infty), w_2 \in (0, \infty]$. That is, if no event occurs in $(T_i, T_i + w_1]$, the alarm is switched on at time $T_i + w_1$. (If an event occurs before the alarm is switched on, the event is missed and the time until an alarm begins is reset to $T_{i+1} + w_1$.) The alarm remains on until time $T_i + w_2$, or until another event occurs, whichever is sooner.

For a renewal process, the expected proportion of time covered by the alarm under this strategy is

$$\tau(w_1, w_2) \equiv \tau(w_2) - \tau(w_1); \tag{5.28}$$

that is, $\tau(w_1, w_2)$ is the alarm fraction for automatic alarms of length w_2 minus the alarm fraction for automatic alarms of length w_1. Similarly, the expected proportion of events that fall within the alarm is

$$\begin{aligned} 1 - \nu(w_1, w_2) &\equiv (1 - \nu(w_2)) - (1 - \nu(w_1)) && (5.29) \\ &= \nu(w_1) - \nu(w_2). && (5.30) \end{aligned}$$

As with any other point process, the optimal alarm for a renewal process is on when the conditional intensity is above some threshold. As stated earlier, the conditional intensity for a renewal process is the hazard function of the time since the last event:

$$\lambda(t) = \frac{\delta(B_t)}{1 - D(B_t)}. \tag{5.31}$$

If the hazard is unimodal in x, and the mode is neither 0 nor ∞, an optimal predictor is of the form (5.27) and satisfies

$$\lambda(w_1) = \lambda(w_2). \tag{5.32}$$

5.3.5 Example: Gamma renewal processes

Daley and Vere-Jones [127] and Kagan [117] previously studied the predictability of gamma renewal processes. We confirm their results here.

Suppose the inter-event times are independent with gamma distributions with shape parameter κ and rate parameter β. The probability density function of the inter-event times is

$$\delta(x) = \frac{\beta^\kappa \exp(-\beta x)}{\Gamma(\kappa)} x^{\kappa - 1}. \tag{5.33}$$

The distribution function is

$$D(x) = \frac{\gamma(\kappa, \beta x)}{\Gamma(\kappa)}, \tag{5.34}$$

where $\gamma(\kappa, \beta x)$ is the lower incomplete gamma function of κ and βx [128]. The mean rate of events is $\bar{\lambda} = \beta / \kappa$.

Kagan [117] derived expressions for ν and τ for the automatic alarm strategy for a gamma renewal process. The proportion of events not predicted is

$$\nu(w) = 1 - F(w) = 1 - \frac{\gamma(\kappa, w\beta)}{\Gamma(\kappa)}. \tag{5.35}$$

The fraction of time taken up by alarms is

$$\tau(w) = \frac{\beta}{\kappa}\left(\frac{\gamma(1+\kappa, \beta w)}{\Gamma(\kappa)} + w\nu(w)\right). \tag{5.36}$$

The hazard is

$$h(x) = \frac{x^{\kappa-1}\beta^{\kappa}\exp(-\beta x)}{\Gamma(\kappa) - \gamma(\kappa, \beta x)}. \tag{5.37}$$

For $\kappa < 1$, this is strictly decreasing, so automatic alarms are optimal. For $\kappa > 1$, this is strictly *increasing*, so automatic alarms are *anti*-optimal: they are worse than random guessing. In this case, the optimal predictor is of the form (5.27) with $w_2 = \infty$. For $\kappa = 1$, hazard is constant: the process is homogeneous Poisson.

5.3.6 Applications of renewal processes

Renewal processes are widely used to model earthquake sequences [117, 127, 129, 130, 131, 132, 50, 133]. Parametric distributions, such as the gamma and lognormal, are commonly used for inter-event times. Alternatively, one could nonparametrically model the hazard function (or inter-event time distribution). Figure 5.3 shows an estimated hazard function for Southern Californian earthquakes with magnitude ≥ 3. We use the R function "muhaz" to calculate hazard from inter-event times using kernel-based smoothing. The hazard gives an estimate of the expected rate of events under a renewal model as a function of the time since the last event. The data are the times of events in the SCEC catalog from 1984 to 2009. The estimated hazard generally decreases as the time since the last event increases. We believe the rise in estimated hazard for times of more than two weeks after the last event occurs because the data are sparse—periods of more than two weeks without a magnitude 3 earthquake in Southern California are rare. The geographic pooling will obscure effects on long time scales. To see periodicity or characteristic earthquakes, we would have to isolate smaller regions.

In reality, inter-event times are not independent. Nevertheless, by allowing for clustering, a renewal model should be more successful than a Poisson model at predicting real seismicity. Clustering corresponds to a hazard that decreases, at least initially—so there might not be much difference between the optimal strategy and a simple automatic alarm strategy.

5.4 Automatic alarms and ETAS predictability

If seismicity really were an ETAS process, how well could it be predicted based on past seismicity alone? This depends on whether the true parameters of the model are known, or whether they must be estimated. If the ETAS parameters are known,

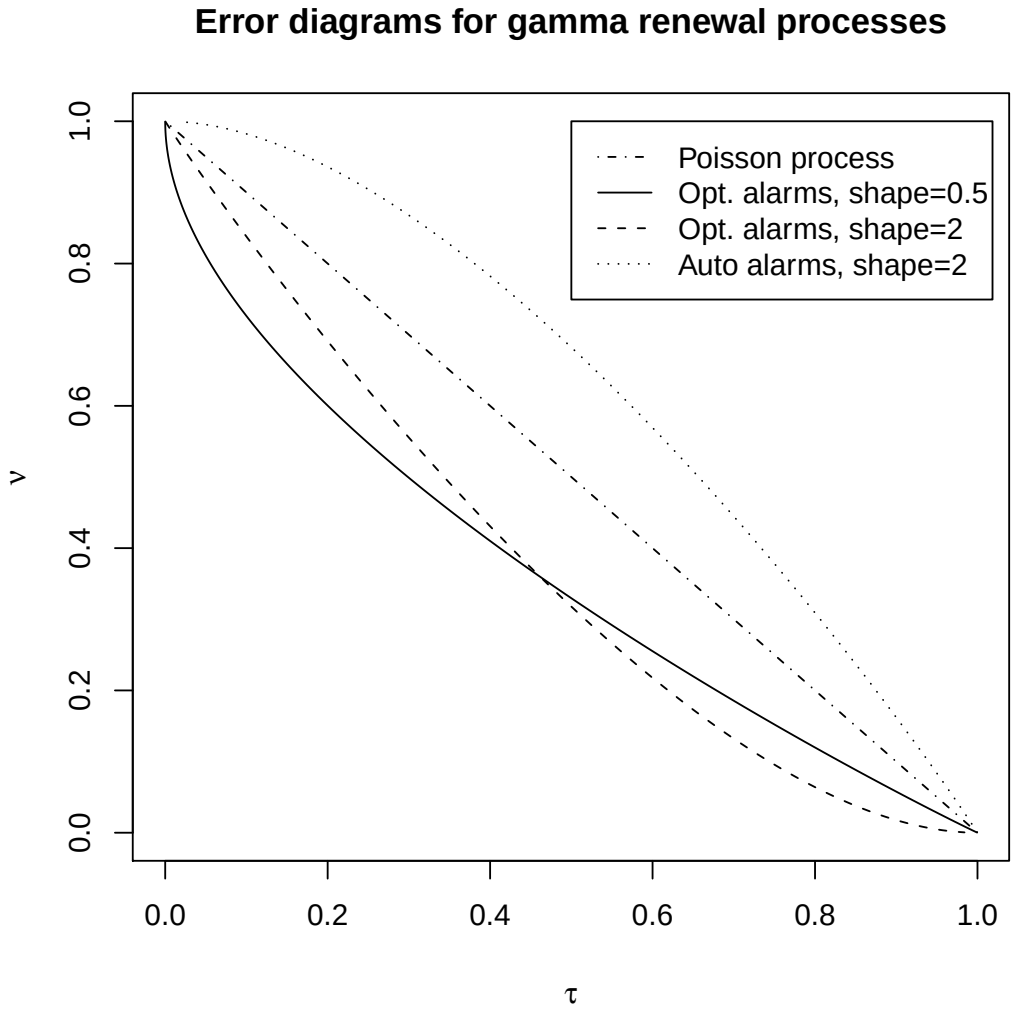

Figure 5.2: Error diagrams for gamma renewal processes. The dashed-dotted straight line is the expected error diagram for a gamma renewal process with shape $\kappa = 1$, i.e., a Poisson process. The solid line is the expected error diagram for automatic alarms for a gamma renewal process with $\kappa = 0.5$. For this process, automatic alarms are optimal. It is below the line for the Poisson, showing some predictive success. The dotted line is the expected error diagram for automatic alarms for a gamma renewal process with $\kappa = 2$. It is above the line for the Poisson, showing the strategy does worse than random guessing. The dashed line is the expected error diagram for optimal alarms for a gamma renewal process with $\kappa = 2$. It is the automatic alarm error diagram for that process rotated 180 degrees about $(0.5, 0.5)$. It is below the line for the Poisson.

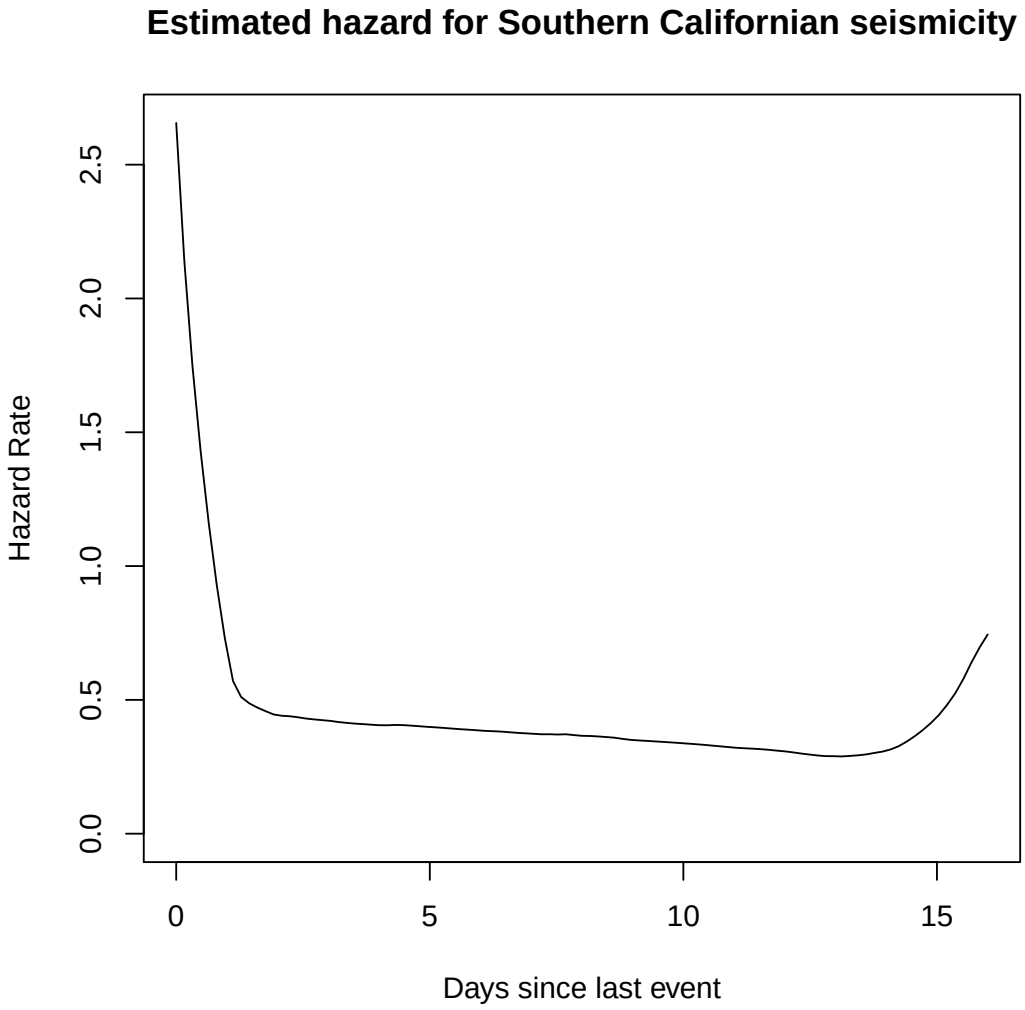

Figure 5.3: Estimated hazard function for inter-event times between earthquakes with magnitude 3 or greater in Southern California. The hazard is estimated from the SCEC catalog from 1984 to 2009, using the "muhaz" smoothing function in R. See section 5.3.6 for notes on the estimation of hazard. Note that only 60 out of 8093 inter-event times are longer than 10 days, so estimation for that region is poor. For shorter times, the hazard function is decreasing.

the conditional intensity predictor is optimal. If the amount of data available is limited, the accuracy of conditional intensity predictions using estimated parameters is weakened.

Suppose the true parameter values are known. In this case, we may calculate the conditional intensity (4.4) at any point in time. An optimal predictor will turn on an alarm if and only if the conditional intensity is above some threshold. We may simulate future seismicity from the process, and determine τ, the proportion of time the alarm is on, and ν, the proportion of events missed by the alarms. Finding τ and ν for a continuum of thresholds gives an error diagram, commonly used to examine predictive success. In this section, we apply this method to simulations of ETAS models. In section 5.5, we shall do the same for real seismicity.

5.4.1 Previous work

Helmstetter and Sornette [134] examined the "intrinsic limits" of predictability in the ETAS model by studying simulated temporal ETAS catalogs with parameters $m_0 = 3, \mu = 1, \alpha = 0.8, c = 0.001, b = 1$ and branching ratio $n = 0.8$ (they did not explicitly state K and p). They attempted to examine how well ETAS seismicity could be predicted if the parameter values were known, setting aside any errors in model fitting. They claimed that the conditional intensity function is the best predictor of the process; Molchan [123] established conditions under which this is true (see also section 5.2.3). If the times and magnitudes of events up to time u have been observed, and the parameters are known, the process from time u onwards can be simulated with the exact probability law.

Molchan and Keilis-Borok [135] used the maximum distance between the error diagram and the diagonal line $\nu = 1 - \tau$ as a measure of predictive success or predictability. They found that the predictability for temporal ETAS simulations claimed by Helmstetter and Sornette was broadly comparable to the predictability of simpler renewal process models. Recall that, in terms of the error diagram, fitting a renewal process model with decreasing magnitude-independent hazard is equivalent to using a simple automatic alarm strategy.

5.4.2 ETAS conditional intensity and the error diagram

Recall that an expected error diagram for a point process alarm strategy plots the expected proportion of unpredicted earthquakes ν as a function of the expected alarm fraction τ. An error diagram where ν is low for most values of τ indicates a highly predictable process, while a curve close to the line $\nu = 1-\tau$ indicates unpredictability.

In optimal error diagrams, predictions are based on perfect knowledge of the conditional intensity. These give the largest improvements over making random predictions independently from the process history. More generally, conditional intensity is un-

known: we instead construct a model and estimate conditional intensity from data. Then the estimate is used to construct the predictor, which may be tested on further data.

In section 5.2.3, we showed that the alarm strategy that minimises expected ν for a given expected τ turns on the alarm if and only if the conditional intensity is greater than some cut-off value $\Lambda(\tau)$. The optimal error diagram quantities τ and ν can be calculated from the distribution of the conditional intensity, if this is known. Let the distribution function of λ be $F(\Lambda) = P(\lambda \leq \Lambda)$. The optimal strategy turns on the alarm when the conditional intensity λ is greater than Λ, and

$$\tau(\Lambda) = 1 - F(\Lambda) \tag{5.38}$$

$$\nu(\Lambda) = \frac{\int_0^\Lambda \lambda dF(\lambda)}{\int_0^\infty \lambda dF(\lambda)}. \tag{5.39}$$

The alarm fraction τ is an increasing function of Λ, while ν is a decreasing function of Λ. For ETAS, the distribution of $\lambda(t)$ is continuous for all t.

The simple automatic alarm strategy and the optimal alarm strategy do not give identical prediction regions for the ETAS model. The conditional intensity may remain high for quite some time after a large number of earthquakes have occurred, whereas the automatic alarm strategy may turn off the alarm if no earthquake has occurred in some time, even if there have many earthquakes in the medium-term past. Similarly, conditional intensity may be high in the medium-term following a large earthquake, even if immediate aftershocks appear to have stopped. These two situations—large numbers of earthquakes and a high magnitude earthquake—often occur simultaneously.

Figure 5.4 displays how the ETAS conditional intensity differs from that of a homogeneous Poisson process and a renewal process. Figure 5.5 shows the empirically-determined distribution of the conditional intensity of an ETAS model with parameters $\mu = 0.01, K = 0.00345, \alpha = b = 1, = 0.01, p = 1.5, m_0 = 5, m_1 = 8$.

5.4.3 Predicting ETAS simulations

We simulated ETAS seismicity for several sets of parameter values, and compared the success of the optimal predictor to that of simple automatic alarms. The simulation parameters correspond to values fitted to several Japanese catalogs by Ogata [3, 1] via maximum likelihood; they appear in Table 5.1. Five sets of parameters correspond to stationary ETAS processes; the sixth ("East of Izu") does not. Most simulations are of length 100,000 days, following a burn-in period of 10,000 days. The burn-in period is intended to reflect the length of historical catalogs, and may not be long enough to reach stationarity (when a stationary state exists). In addition, the simulation length may not be long enough to observe distribution tails

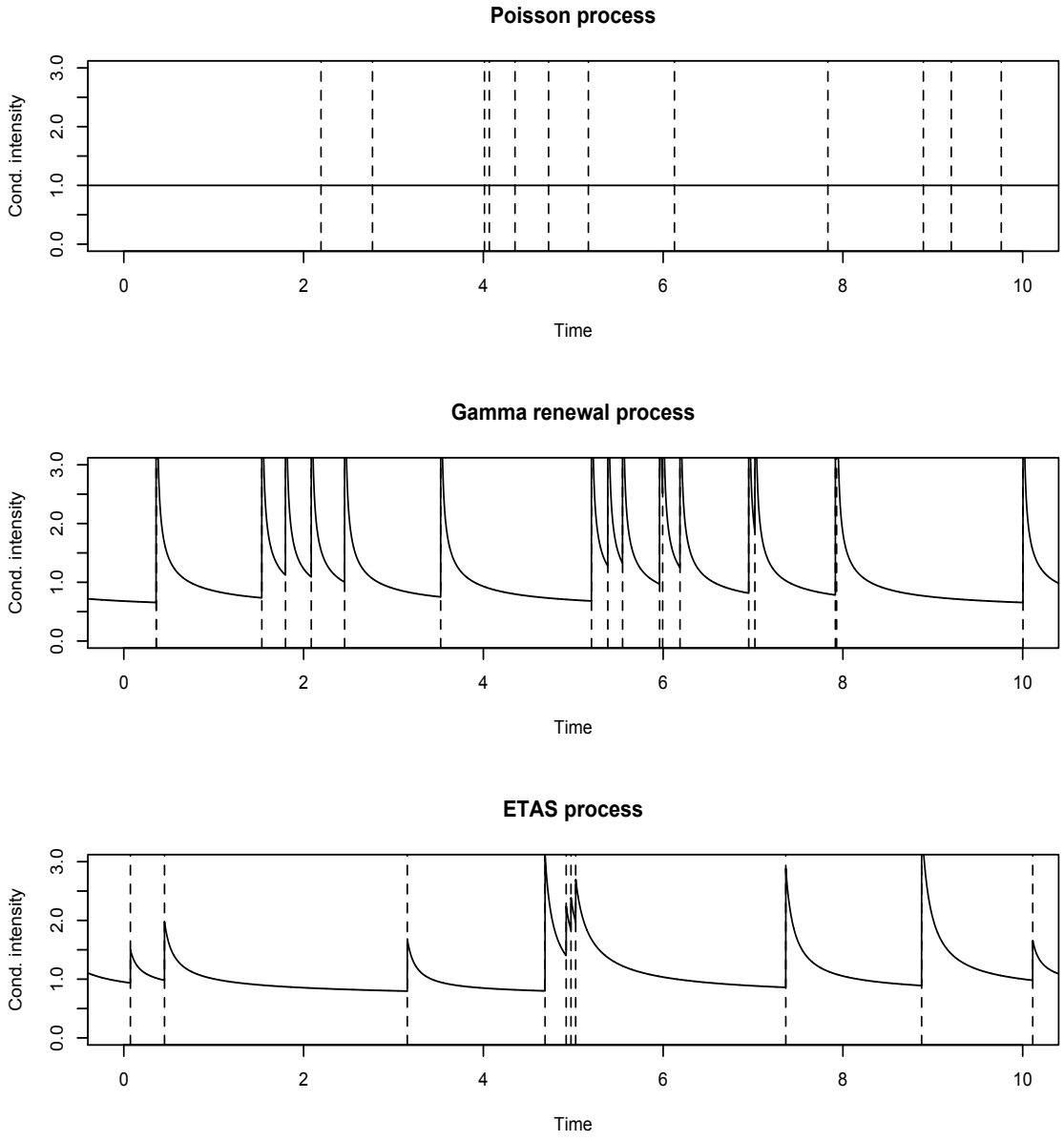

Figure 5.4: Conditional intensities for simulated point processes. The top graph is for a Poisson process. The middle graph is for a gamma renewal process with shape parameter 0.5 and rate parameter 0.5. The bottom graph is for an ETAS process with $b = 1, \mu = 0.5, K = 0.04, c = 0.1, \alpha = 0.5, p = 1.1$. Each process has an expected rate of one event per unit time. The vertical dotted lines show times of events.

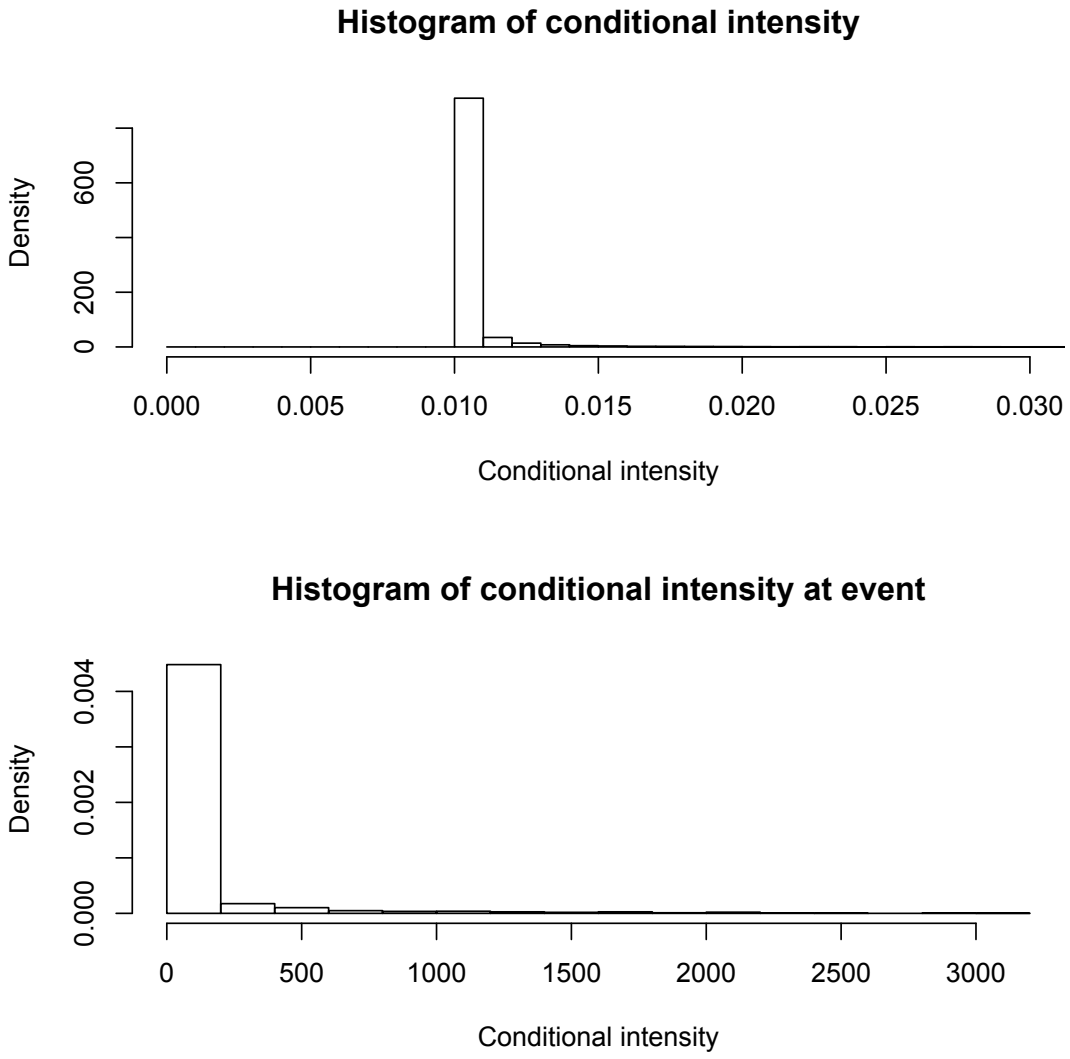

Figure 5.5: Top: empirical conditional intensity distribution of the ETAS model with parameters $\mu = 0.01, K = 0.00345, \alpha = b = 1, = 0.01, p = 1.5, m_0 = 5, m_1 = 8$. The graph excludes the 1.2% of the time the conditional intensity exceeded 0.03. The minimum value of the conditional intensity is 0.1, the background rate. The conditional intensity is rarely much larger than the background rate. Bottom: empirical distribution of conditional intensity just before the occurrence of an event, for ETAS model with parameters as above. For 46% of events, conditional intensity was less than 0.03. However, some events occur when conditional intensity is in the hundreds or thousands. (Note that the x-axis scale is different from the top graph.)

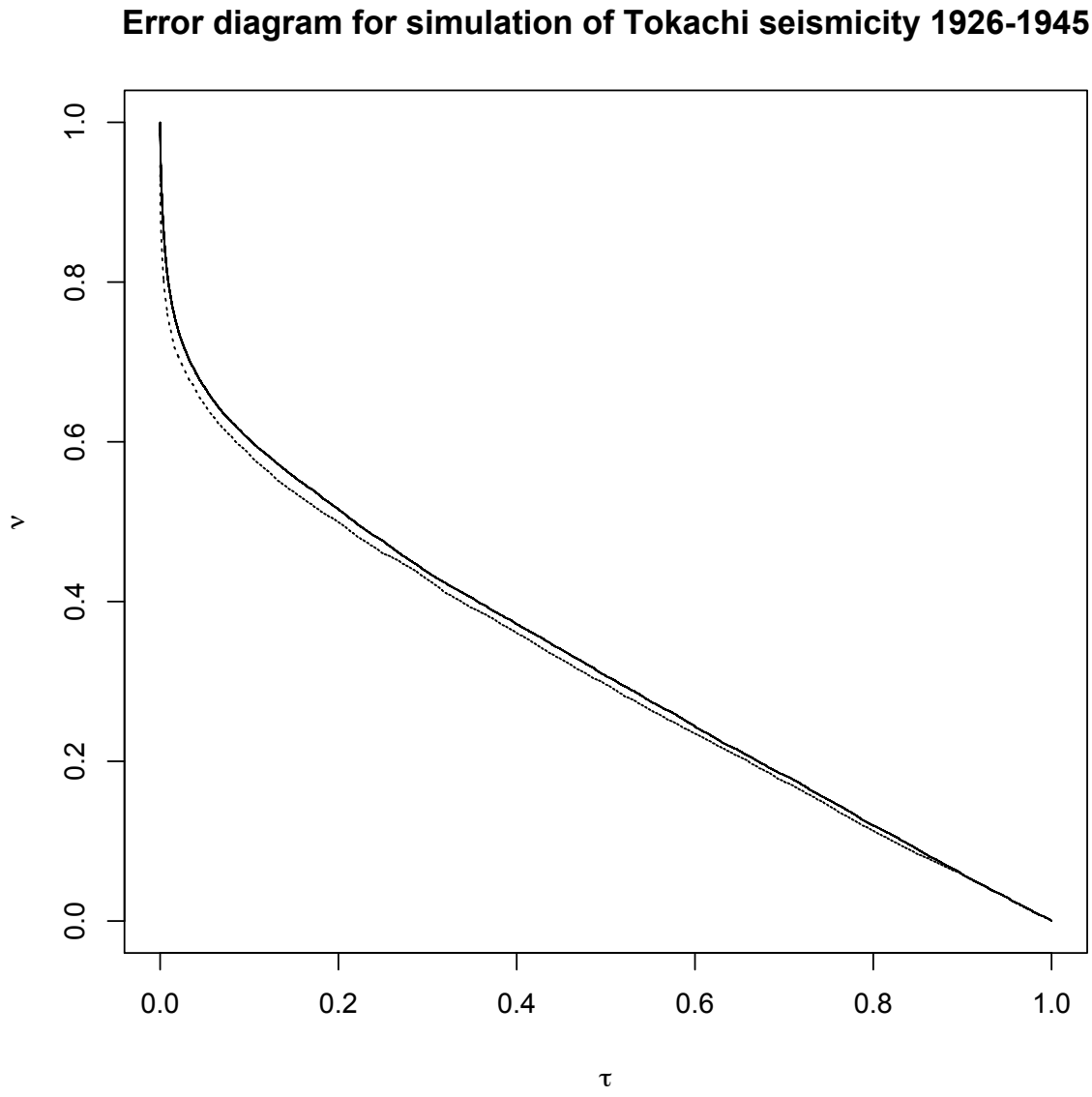

Figure 5.6: Error diagram for a simple automatic alarm strategy (solid line) and conditional intensity predictor (dotted line) for a 200,000 (with 10,000 day burn-in) day simulation of Tokachi seismicity based on parameters estimated by Ogata [1] from the catalog from 1926-1945. The simulation parameters were $m_0 = 5, m_1 = 9, b = 1, \mu = 0.047, K = 0.013, c = 0.065, \alpha = 0.83, p = 1.32$. On the x-axis, τ gives the fraction of time covered by alarms; on the y-axis, ν gives the fraction of earthquakes of magnitude 5 or greater not predicted. The 10th percentile of interarrival times is 40 minutes, the median is 4.3 days, and the 90th percentile is 34 days.

caused by extreme events.

Figure 5.6 gives an error diagram for one simulation. The optimal predictor outperforms the simple automatic alarm strategy for all values of τ. However, for any particular value of τ, the difference between the two predictors is small. Now, a small difference may be important. For example, if even one in a thousand large earthquakes could be predicted with certainty, it would be a huge achievement. In this case, it is not clear how much value is represented by the improvement in prediction of the optimal ETAS predictor over simple automatic alarms—particularly as this is a best-case scenario for ETAS, since the simulated catalog follows the ETAS model.

Table 5.2 summarises results for all six sets of simulation parameters. In every case, a large proportion of shocks occur within a few hours of a previous shock, and most shocks occur within a few days of their parent. The improvement of the optimal predictor over the automatic alarm strategy never exceeds 5.4%, and is typically below 2%. For ETAS seismicity, the optimal strategy is consistently better than the simple automatic alarm strategy, but the improvement is modest.

Catalog	m_0	μ	K	α	c	p
Tokachi 1926-45	5.0	0.047	0.013	0.83	0.065	1.32
Izu Peninsula	2.5	0.022	0.035	0.17	0.003	1.35
Tokachi 1952-61	5	0.032	0.021	0.72	0.059	1.10
Matsuhiro swarm	3.9	0.0006	0.092	0.27	0.13	1.14
East of Izu	2.9	0.59	0.016	0.31	0.009	1.73

Table 5.1: Parameters estimated by Ogata [3, 1] for Japanese earthquake catalogs. The Gutenberg-Richter parameter b was assumed to be 1 in every case. The estimates for "East of Izu" imply an explosive process; the other sets imply a process with a stationary state. We use these parameter estimates for simulations; the results are given in Table 5.2.

5.4.4 Magnitude-dependent automatic alarms for ETAS

We can also assess the success of magnitude-dependent automatic alarms at predicting ETAS processes. Figure 5.7 plots error diagrams for predictors of a simulation of a ten-year catalog of Southern Californian seismicity, with parameters $m_0 = 3, \mu = 0.1687, K = 0.04225, \alpha = 0.4491, c = 0.1922, p = 1.222$. The MDA alarms were of the form $k \times 3.7^M$, with the base 3.7 chosen to minimise the area under the error diagram for a training set. Area under the error diagram for some other predictors are given in Table 5.3. In the test set, the area under the MDA alarm curve is 0.242. This is slightly better than simple automatic alarms (0.252) and slightly worse than the optimal conditional intensity predictor using the true

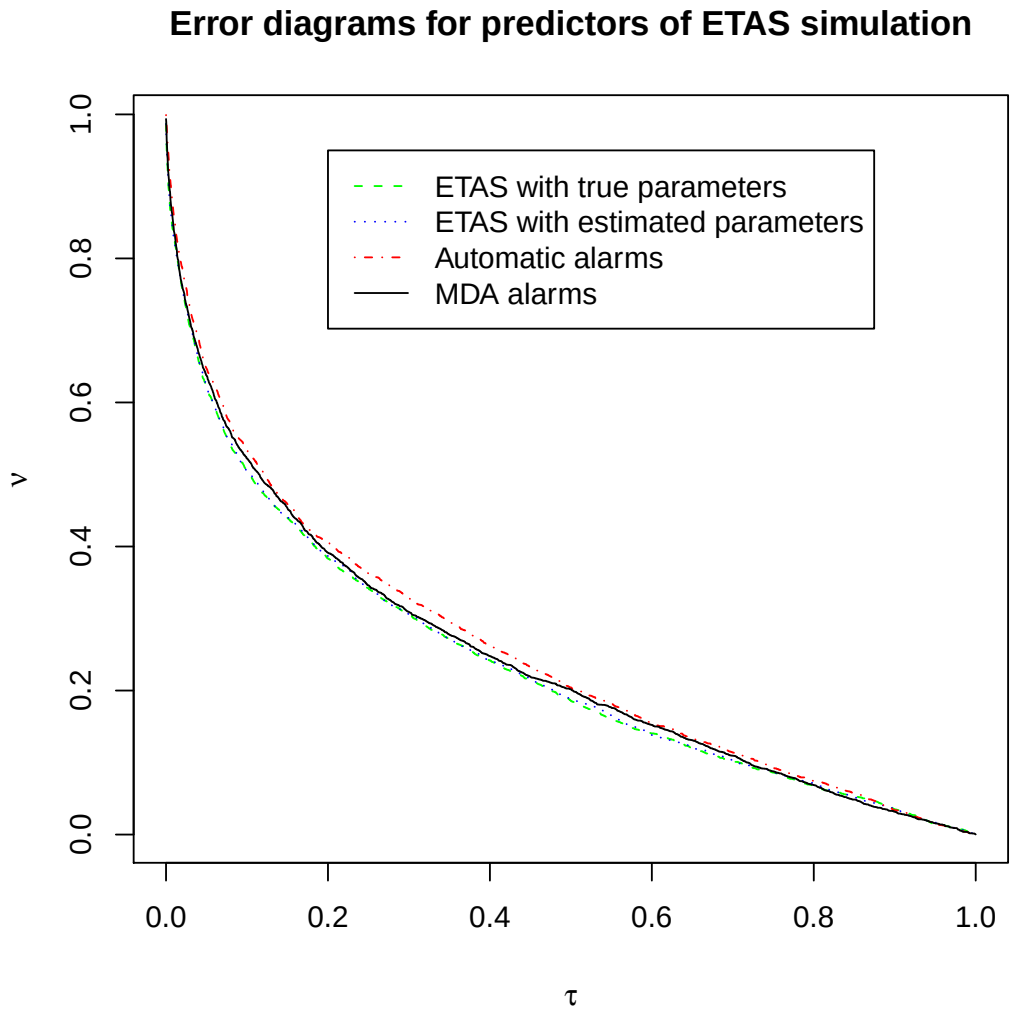

Figure 5.7: Error diagrams for predictors of a simulated temporal ETAS sequence. The parameters used in the simulation were those estimated for Southern Californian seismicity: $m_0 = 3, \mu = 0.1687, K = 0.04225, \alpha = 0.4491, c = 0.1922, p = 1.222$. Models were fitted to a 20-year training set and assessed on a 10-year test set. The ETAS conditional intensity predictor with the true parameters (green dashed line) performs very similarly to the ETAS conditional intensity predictor with estimated parameters (blue dotted line). The magnitude-dependent automatic alarms have parameter $u = 3.70$, chosen to minimise area under the error diagram in the training set. In the test set (solid black line), they perform slightly better than automatic alarms (red dotted-dashed line) and slightly worse than the ETAS conditional intensity predictors. No single strategy dominated any other single strategy.

Set of parameters	m_0	Auto alarm success rate	Optimal alarm success rate
Tokachi 1926-45	5	39.7%	41.6%
Izu Peninsula	2.5	86.5%	87.4%
Izu Islands	4	65.8%	67.2%
Tokachi 1952-61	5	36.5%	39.4%
Matsuhiro swarm	3.9	55.8%	55.1%
East of Izu	2.9	79.9%	85.2%

Table 5.2: Success of alarms for ETAS simulations that are on 10% of the times. The column "Set of parameters" names a catalog for which Ogata [3, 1] fitted temporal ETAS models. The parameter estimates for these catalogs are given in Table 5.1, while the column "m_0" gives the catalog minimum magnitude. The third and fourth columns give the percentages of events in simulations that fall within simple automatic and optimal conditional intensity alarms respectively.

parameters (0.236). The optimal predictor performs only very slightly better than a conditional intensity predictor that uses estimated parameters (0.237). In fact, Table 5.4 suggests that predictions using estimated parameters are nearly optimal, even when the training set is small.

5.5 Predicting Southern Californian seismicity

Figure 5.8 shows the performance of predictors on a training set. The parameter u in the MDA alarms is chosen to minimise the area under the error diagram. (Other selection criteria are possible—we could minimise ν for a fixed value of τ, or minimise the area under the error diagram for a limited τ range, or use a measure of entropy such as Kagan's information score [117].) In fact, a conditional intensity predictor using Veen and Schoenberg's estimated space-time ETAS parameters and a conditional intensity predictor using temporal ETAS estimated parameters do comparably well. (See Table 4.2 for parameter estimates.) A conditional intensity predictor using the typical parameters in Table 4.3 performs similarly. The value $u = 5.8$ minimises the area under the error diagram for an MDA alarm strategy. This strategy performs almost as well as the conditional intensity predictors, while simple automatic alarms are notably worse. However, the more complex strategies have more degrees of freedom.

It is better to measure predictive performance on a test set separate from the data from which parameters were estimated. We do not quite keep training and test data separate—although we find parameter estimates for an ETAS model from the

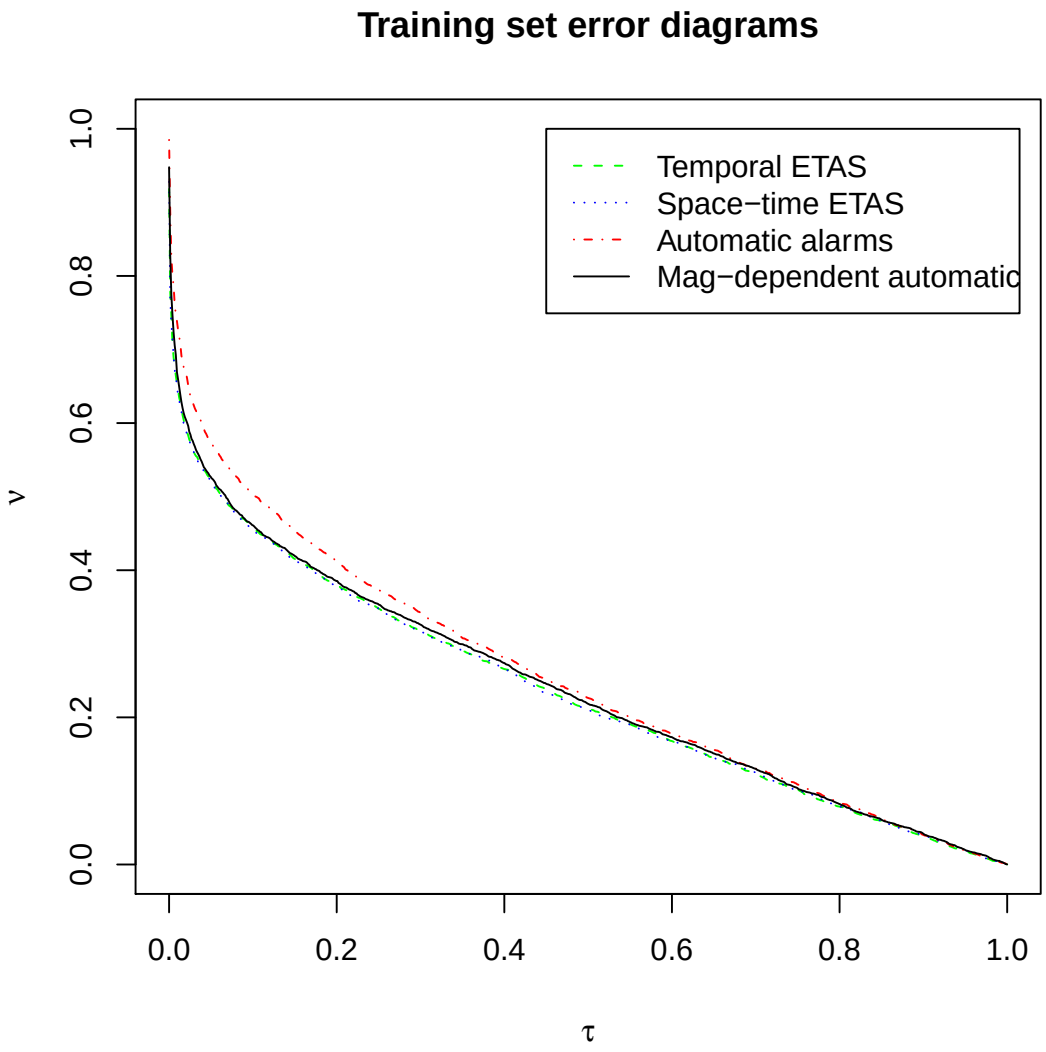

Figure 5.8: Error diagrams for predictors of Southern Californian seismicity on a training set of data. The catalog is the SCEC catalog of $M \geq 3$ earthquakes from January 1st, 1984 to June 17th, 2004. The estimated ETAS models and the MDA alarm strategy (with parameter chosen to minimise the area under the curve) all perform comparably well, and outperform a simple automatic alarm strategy for most values of $\hat{\tau}$.

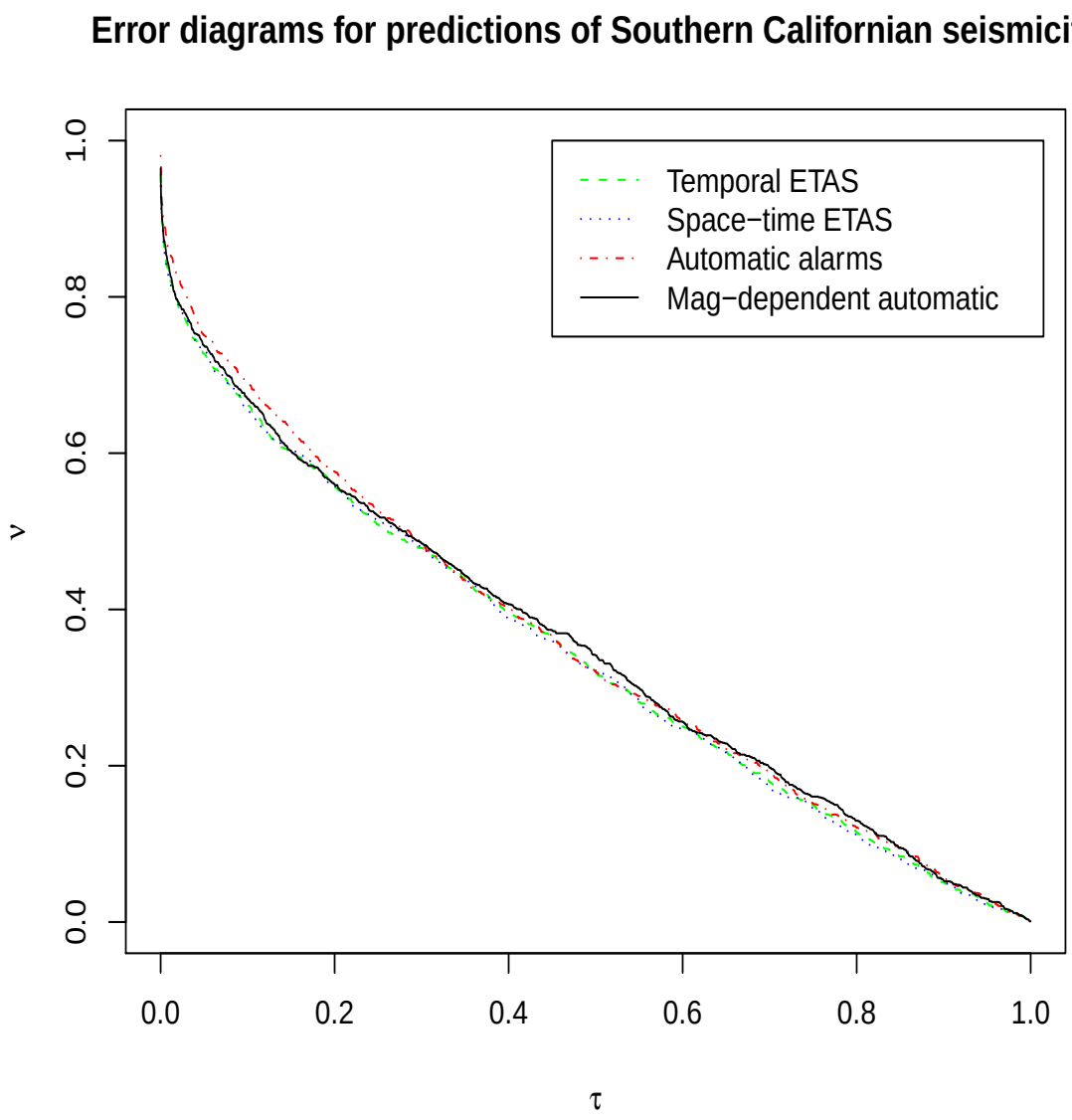

Figure 5.9: Error diagrams for predictors of Southern Californian seismicity. The predictors were fitted to the SCEC catalog from January 1st, 1984 to June 17th, 2004, and tested on the SCEC catalog from June 18th, 2004 to December 31st, 2009. For low values of $\hat{\tau}$, simple automatic alarms do not perform as well as the ETAS predictors. For high values of $\hat{\tau}$, MDA alarms do not perform as well as the ETAS predictors. Note that although success rates are determined for the test set only, predictors used both training and test data to determine times since past events (for simple automatic and MDA alarms) and conditional intensity (for ETAS predictors).

Predictor	Area under error diagram	Area under left quarter
Optimal ETAS	0.236	0.127
Estimated ETAS	0.237	0.128
Typical ETAS	0.241	0.130
MDA, $u = 3.7$	0.242	0.130
MDA, $u = 2$	0.246	0.131
Simple auto	0.252	0.133

Table 5.3: Success of several predictors of a simulated ETAS sequence. Predictors are trained on a 20-year simulated catalog, and tested on a subsequent 10-year simulated catalog. The simulation parameters are $m_0 = 3, \mu = 0.1687, K = 0.04225, \alpha = 0.4491, c = 0.1922, p = 1.222$. The measures of success are area under the error diagram, and area under the left quarter of the error diagram (since alarms that are on less often are more attractive). "Optimal ETAS" is a conditional intensity predictor using the simulation parameters, given in the "VS spatial estimate" column of Table 4.2. "Estimated ETAS" uses parameters estimated from a training set. "Typical ETAS" uses the parameters in Table 4.3. "MDA, $u = 3.7$" is a magnitude-dependent automatic alarm strategy with base determined by fitting alarms to a test set. "MDA, $u = 2$" is an MDA alarm strategy with base 2. "Simple auto" is a simple automatic alarm strategy.

training set and not the test set, we calculate conditional intensity given these estimates by summing contributions from events in both the training and test sets (to avoid inaccuracy at the beginning of the test set). Figure 5.9 shows error diagrams for the predictors fitted to the above training set, tested on the SCEC catalog from June 18th, 2004 to December 31st, 2009. There is very little difference in performance between conditional intensity alarms using the parameters estimated by Veen and Schoenberg for space-time ETAS, and conditional intensity alarms using the parameters we estimated for temporal ETAS—even though the parameters are very different, the alarm times are similar. For example, if thresholds are set so that both strategies turn on alarms for 10% of the study period, then for 9.4% of the study period, both strategies have their alarms on. A conditional intensity predictor based on the "typical" parameters in Table 4.3 gives a similar error diagram to those of the estimated conditional intensity predictors. (For $\hat{\tau} = 10\%$, all three strategies have alarms on simultaneously 8.7% of the time.) The areas under the curve are 0.340 for Veen-Schoenberg parameter estimates, 0.341 for temporal ETAS estimates, and 0.345 for typical parameters. (Recall that perfect prediction gives area 0 and random guessing gives expected area 0.5.)

ETAS conditional intensity predictors outperform the simple automatic alarm strategy by a small but clear margin for most values of $\hat{\tau}$. The difference is most

Training years	Optimal ETAS	Estimated ETAS	Typical ETAS	MDA $w = 2$	Simple automatic
1	0.250	0.251	0.254	0.259	0.264
2	0.245	0.246	0.248	0.252	0.257
5	0.240	0.241	0.242	0.250	0.256
10	0.261	0.262	0.263	0.271	0.275
20	0.262	0.264	0.262	0.273	0.279
50	0.257	0.259	0.259	0.265	0.270

Table 5.4: Effect of length of training set on accuracy of prediction. The column "Training years" gives the length of a simulated ETAS training set in years. The simulation parameters are $m_0 = 3, \mu = 0.1687, K = 0.04225, \alpha = 0.4491, c = 0.1922, p = 1.222$. An ETAS model was estimated from the training set, then the parameter estimates were used to calculate a conditional intensity predictor for a 10-year test set. (Event in both the training and test sets were included in the ETAS conditional intensity calculations.) The column "Estimated ETAS" gives the area under the error diagram for this predictor. Other columns give areas under the error diagram for other predictors as comparisons. For each length of training set, all predictors were assessed on the same set. In each case, the estimated ETAS predictor does slightly worse than the optimal ETAS predictor. Training set length has little effect on the accuracy of predictions from the estimated ETAS model. Note that predictions were better for two years of training data than for 50 years—this is the result of sampling variability.

Predictor	Training area	Test area	LQ test area
Space-time ETAS	0.234	0.340	0.161
Temporal ETAS	0.235	0.341	0.161
Typical ETAS	0.236	0.345	0.161
MDA, $u = 2$	0.253	0.348	0.165
MDA, $u = 5.8$	0.240	0.351	0.163
Simple auto	0.254	0.352	0.168

Table 5.5: Success of several predictors of Southern Californian earthquakes of magnitude $M \geq 3$. The predictors are fitted to a training set of data (the SCEC catalog from January 1st, 1984 to June 17th, 2004) and assessed on a test set of data (the catalog from June 18th, 2004 to December 31st, 2009). The predictors have parameters estimated on the training set, but may use times and magnitudes of training events in the test. The measures of success are area under the training set error diagram, area under the test set error diagram, and area under the left quarter of the test set error diagram. "Space-time ETAS" is a conditional intensity predictor using Veen and Schoenberg's space-time parameter estimates, given in the "VS spatial estimate" column of Table 4.2. "Temporal ETAS" uses parameters estimated using a temporal ETAS model, given in the "Temporal estimate" column of Table 4.2. "Typical ETAS" uses the parameters in Table 4.3. "MDA, $u = 2$" is a magnitude-dependent automatic alarm strategy with base 2. "MDA, $u = 5.8$" is an MDA alarm strategy alarm strategy with base determined by fitting alarms to a test set. "Simple auto" is a simple automatic alarm strategy.

pronounced for small values of $\hat{\tau}$. The MDA alarms with fitted parameter $u = 5.8$ perform comparably to the conditional intensity predictors for small values of $\hat{\tau}$. For large values of $\hat{\tau}$, they perform slightly worse than both conditional intensity predictors and simple automatic alarms. For the MDA alarm to capture half of events, the alarm would have to be on 28% of the time. In comparison, an alarm based on conditional intensity estimated from a temporal ETAS model would have to be on 26% of the time to capture half of events. In both cases, the observed predictive success is far from that required for operational earthquake prediction.

Table 5.5 gives the area under the training and test error diagrams for a number of predictors. No predictor dominated any other predictor for all values of $\hat{\tau}$. In fact, each predictor was uniquely best for at least some values of $\hat{\tau}$. The predictor based on Veen-Schoenberg estimates was best most often (outright best for 47% of $\hat{\tau}$ values, equal best for a further 10% of $\hat{\tau}$ values). We would like to perform similar analyses on other geographic areas, as well as on subregions of Southern California, to see if we obtain similar results.

5.6 Discussion

The error diagram (Molchan diagram) allows evaluation of earthquake alarm strategies. Empirical error diagrams display performance on data sets, while expected error diagrams show the theoretical predictability of stochastic point processes. If seismicity actually followed a stochastic process, an optimal alarm would be "on" when the conditional intensity of the process is above some threshold. The family of optimal alarms determines an optimal error diagram. We can compare the performance of optimal alarms for complex models to that of automatic alarms to study the value of the extra complexity.

In a renewal process, conditional intensity is a function of the time since the last event, so this time determines whether an optimal alarm is off or on. When seismicity is a realisation of a renewal process with decreasing hazard, simple automatic alarms are theoretically optimal. This is the case when the inter-event distribution is gamma with shape $\kappa \leq 1$. In practice, renewal models fitted to real seismicity generally have decreasing hazard because of clustering.

Molchan and Keilis-Borok [135] previously found that the ETAS model was broadly as predictable as realistic renewal process models for seismicity. We examined the success of simple and magnitude-dependent automatic alarm strategies at predicting simulations from the ETAS model. We found that MDA alarms performed slightly worse than conditional intensity predictors, and simple automatic alarms performed slightly worse than MDA alarms. No strategy was dominated. Over a variety of training set lengths, whether parameters were known or estimated had little effect on the conditional intensity predictor—despite the estimated parameters being very different from the true parameters.

In a test on real seismicity (the SCEC catalog from June 18th, 2004 to December 31st, 2009), an ETAS model with "typical" parameters predicted about as successfully as one with estimated parameters. Generally, both did slightly better than magnitude-dependent automatic alarms, but MDA alarms were not dominated. (Optimising the parameter u in the MDA alarms did not result in better prediction than fixing $u = 2$.) MDA alarms, in turn, outperformed simple automatic alarms by a small amount, but did not dominate them.

The ETAS model has some predictive power over and above MDA alarms—the success of ETAS is not straightforwardly explained by the fact that large earthquakes are frequently followed by aftershocks. What the value is in the small increase in predictive success is up for debate. That ETAS parameter values have a weak effect on predictive accuracy is comforting, as parameters generally cannot be estimated accurately. Though different sets of ETAS parameters give very different conditional intensities, they may give roughly the same topology of level sets. (For instance, the parameter μ affects the conditional intensity, but does not change its level sets.) The correct approach to ETAS would seem to be to use it as a comparison to physics-

based models: just as stochastic models have no incremental value if they do not give better predictions than an automatic alarm strategy, physics-based prediction have no incremental value if they do not outperform ETAS. Properties of ETAS that are sensitive to the parameters, such as the branching ratio and conditional intensity, should be treated with scepticism.

In the future, we wish to examine the space-time aspect of prediction in more detail. We would like to apply space-time MDA alarms, which we briefly described in section 5.2.4, to real seismicity to establish a baseline against which to compare more complicated methods.

Chapter 6

Conclusion

6.1 Assessing models and predictions

All stochastic models of seismicity are wrong, but some are more wrong than others. For instance, seismicity is not a stationary process over geological time scales, but stationarity may be a reasonable assumption over human time scales, albeit a difficult assumption to test. The assumption that seismicity is a branching process—that every earthquake is triggered by no or one preceding event—is untrue, but may be useful in earthquake modelling. If such assumptions are made, the models should not be taken literally—a branching ratio, giving the expected number of events directly triggered by every event in the process, is an artifact of the model, not something that physically exists.

Stochastic models for seismicity have often been used as null hypotheses in statistical tests. Failure to reject a null hypothesis does not mean the null is true. The test may not have sufficient power to detect departures from the null. Different tests may have power against different alternatives, so it may be preferable to use multiple tests, combining them using Bonferroni's inequality. For example, it has been claimed that declustered catalogs are Poisson. Evidence for this claim is that a chi-square test does not reject the hypothesis that the times of the declustered catalog are a realisation of a temporal Poisson process. However, the claim is not true in time, as shown by a Kolmogorov-Smirnov test, and it is not true in space-time, since, unlike in a spatially heterogeneous, temporally homogeneous Poisson process, events in a declustered catalog cannot occur arbitrarily close in space-time. We tested a weaker space-time hypothesis: that conditional on the locations and times of the events in a declustered catalog, all permutations of the times are equally likely. For three declustering methods, we obtained P-values of 0.003, 0.005, and 0.069, rejecting the null in two of three cases and casting doubt on the hypothesis of exchangeable times.

Furthermore, because no stochastic model for seismicity is true, tests that compare earthquake models or predictions to a null hypothesis with random seismicity have

limited interpretations. The null hypothesis may be rejected because the seismicity model is wrong, and not because the test model or predictions are good. In addition, tests must recognise that if the seismicity were different, predictions would be different. We showed that a test of a naive predictor that held predictions constant while allowing seismicity to vary under the null gave a statistically significant result—not because the predictions were good, but because they took history into account.

Instead of performing tests assuming random seismicity, it is better to assess the successes of predictions and forecasts by comparing them to simple predictors such as automatic alarms. Automatic alarms are easy to fit to data. Simple automatic alarms are optimal for many unmarked renewal processes, while magnitude-dependent automatic alarms are a straightforward generalisation. The success of automatic alarms should be taken be taken as empirical, and not as indicating anything causal, or that the clustering structure is simple.

It is difficult to use automatic alarms as part of a null hypothesis in a probabilistic test. There is no inherent randomness in automatic alarms—unless seismicity is considered random, which is unwise for testing—and attempting to add randomness through a semi-automatic alarm strategy can result in a loss of optimality, and a lower threshold for "success" than that provided by automatic alarms. Testing predictions is a difficult problem that we do not claim to have resolved. But outperforming a random predictor by a statistically significant margin is not impressive—any predictor that exploits clustering should be able to do this. On the other hand, outperforming magnitude-dependent automatic alarms over a long test set requires at least some skill.

6.2 Building models and predictions

We believe that the idea behind declustering is backwards. The occurrence of earthquakes in space-time is very complex. The simplest observation we can make about the structure of earthquake catalogs is that earthquakes cluster in space and time. It is preferable to attempt to model this clustering, rather than attempt to remove clustering, then assert the remaining structure is simple. Firstly, clustered events occur, and can cause damage, whether or not we decluster the catalog. Secondly, though there may or may not be physical differences between "background" and "offspring" events, all existing methods to differentiate between them based on times, locations, and magnitudes are arbitrary. Thirdly, the declustered catalog cannot simply be assumed to be a realisation of a simple model such as the Poisson, or to have simple structure such as exchangeable times given locations, without rigorous and varied testing. Fourthly, if notable features such as inhomogeneity are found in the declustered catalog, it is difficult to tell if they are of physical interest or are artifacts of the declustering procedure. Fifthly, if the declustered catalog does have a simple structure, it may be that all features of interest have been removed from the

data.

Fitting even unsophisticated clustering models to clustered data is better than declustering, in part because unsophisticated clustering models can be made more complicated. The simple automatic alarm strategy exploits the fact that earthquakes cluster in space and time. We know that the rate of seismicity is elevated for longer after large earthquakes, so we can add complexity to automatic alarms by making them magnitude-dependent. Then we can compare the predictive performance of these MDA alarms to that of yet more complicated models, like ETAS.

The ETAS model has substantial complexity, and substantial problems. It does not provide a good fit to real data—for instance, short times between events are observed more often in real catalogs than in the model. The ETAS branching ratio, a key model property, is very sensitive to distributional assumptions and parameter values. In fact, whether the model is stationary can depend on whether the Gutenberg-Richter magnitude distribution is truncated. Parameter estimation for ETAS is poor even for long catalogs, so the branching ratio should taken as a statistical artifact, not something that reflects real seismicity.

Although real seismicity is substantially different from realisations of ETAS, the model is still of some use as a predictor. For a point process, the alarm strategy that minimises the expected fraction of events missed τ given an expected alarm fraction τ declares an alarm when the conditional intensity exceeds some threshold. We therefore obtain predictions from an ETAS model by turning on an alarm when the conditional intensity is high. Predictions are sensitive to the times and magnitudes of events in the catalog. This makes prediction at a lead time difficult. (One advantage of automatic alarm strategies is that they are readily adaptable to prediction with some lead time.) Predictions are surprisingly insensitive to exact parameter values.

We compared this conditional intensity predictor for ETAS to automatic alarm strategies. Both when the true process is ETAS and for real Southern Californian data, ETAS conditional intensity predictors outperform simple and magnitude-dependent automatic alarm strategies by a small margin for most values of $\hat{\tau}$. This indicates that the ETAS model has some value—as long as we do not take it too literally. On the other hand, the level of success we found for Southern Californian data (to capture 50% of events, an alarm based on an estimated temporal ETAS model would have to be on 26% of the time) was insufficient for operational earthquake prediction.

We have only compared the performance of ETAS conditional intensity predictors and automatic alarm strategies for one real catalog. In the near future, we wish to examine other catalogs. We do not know if parameter estimates will resemble those for Southern California. However, we expect relative success for ETAS conditional intensity predictors and automatic alarms broadly similar to what we observed for Southern Californian data.

Further effort should go into understanding the structure of earthquake clusters;

however, it seems unlikely that a model that allows realistic modelling of clusters using only a few parameters exists. Real seismicity is more complex than any stochastic model yet proposed. This does not mean that simple models have no value. Automatic alarm strategies have almost all of the predictive success of the ETAS model, while avoiding many of that model's drawbacks.

Bibliography

[1] Y. Ogata, J. Geophys. Res. **97**, 19845 (1992).

[2] J. Gardner and L. Knopoff, Bull. Seis. Soc. Am. **64**, 1363 (1974).

[3] Y. Ogata, Tectonophysics **169**, 157 (1989).

[4] A. Veen and F. P. Schoenberg, JASA **482**, 614 (2008).

[5] A. O. Öncel and O. Alptekin, Natural Hazards **19**, 1 (1999).

[6] C. R. Allen, Bull. Seis. Soc. Am. **66**, 2069 (1976).

[7] R. J. Geller, in *A Critical Review of VAN* (World Scientific, Singapore, 1996), pp. 155–238.

[8] T. L. Wright and T. C. Pierson, Technical report, U.S. Geological Survey Circular 1073 (unpublished).

[9] D. D. Jackson, in *The State of the Planet: Frontiers and Challenges in Geophysics* (American Geophysical Union, New York, 2004), pp. 335–348.

[10] P. Stark and D. Freedman, in *Earthquake Science and Seismic Risk Reduction*, Vol. 32 of *NATO Science Series IV: Earth and Environmental Sciences* (Kluwer, Dordrecht, The Netherlands, 2003), pp. 201–213.

[11] R. M. Allen, P. Gasparini, O. Kamigaichi, and M. Böse, Seismo. Res. Lett. **80**, 682 (2009).

[12] V. Kossobokov, L. Romashkova, V. Keilis-Borok, and J. Healy, Phys. Earth Planet. Inter. **111**, 187 (1999).

[13] M. C. Gerstenberger, S. Wiemer, L. M. Jones, and P. A. Reasenberg, Nature **435**, 323 (2005).

[14] P. Varotsos and M. Lazaridou, Tectonophysics **188**, 321 (1991).

[15] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. 2*, 2nd ed. (Wiley, New York, 1971).

[16] W. Feller, *An Introduction to Probability Theory and Its Applications*, 3rd ed. (John Wiley & Sons, Inc., New York, 1968), Vol. I, p. 509.

[17] G. E. P. Box and N. R. Draper, *Empirical Model-building and Response Surfaces* (Wiley, New York, 1987), p. 669.

[18] B. Luen and P. B. Stark, *IMS Lecture Notes—Monograph Series. Probability and Statistics: Essays in Honor of David A. Freedman* (Institute for Mathematical Statistics Press, Beachwood, 2008), pp. 302–315.

[19] R. Stothers, Acient Hist. Bull. **18**, 101 (2004).

[20] R. E. Buskirk, C. Frohlich, and G. V. Latham, Rev. Geophys. Space Phys. **19**, 247 (1981).

[21] M. Ikeya, *Earthquakes and Animals: From Folk Legends to Science* (World Scientific Publishing, Hackensack, NJ, 2004), p. 316.

[22] R. Kerr, Science **208**, 695 (1980).

[23] J. Kirschvink, Bull. Seis. Soc. Am. **90**, 312 (2000).

[24] D. Arabelos, G. Asteriadis, M. Contadakis, G. Zioutas, D. Xu, C. Zhang, and B. Zheng, Tectonophysics **338**, 315 (2001).

[25] F. Poitrasson, S. Dundas, J. Toutain, M. Munoz, and A. Rigo, Earth. Planet. Sci. Lett. **169**, 269 (1999).

[26] P. Varotsos, K. Alexopoulos, and K. Nomicos, Prakt. Akad. Athenon **56**, 277 (1981).

[27] T. Bleier and F. Freund, IEEE Spectrum **December**, 22 (2005).

[28] C. Y. Wang, R. E. Goodman, P. N. Sundaram, and H. F. Morrison, Geophys. Res. Lett. **2**, 525 (1975).

[29] F. Morgan, E. Williams, and T. Madden, J. Geophys. Res. **94**, 12,449 (1989).

[30] Z. Shou, Science and Utopia **64**, 53 (1999).

[31] V. I. Keilis-Borok and V. G. Kossobokov, Physics of the Earth and Planetary Interiors **61**, 73 (1990).

[32] B. Bodri, J. Geodynamics **32**, 289 (2001).

[33] F. F. Evison and D. A. Rhoades, New Zealand J. of Geol. and Geophys. **40**, 537 (1997).

[34] M. Gerstenberger, S. Wiemer, L. Jones, and P. Reasenberg, Nature **435**, 328 (2005).

[35] J. Holliday, J. Rundle, K. Tiampo, W. Klein, and A. Donnellan, Pure Appl. Geoph. **163**, 2433 (2006).

[36] R. Geller, A. Braginski, and W. Campbell, Earthquake Precursors or Background Noise?, http://spectrum.ieee.org/apr06/3275, 2006.

[37] F. F. Evison and D. A. Rhoades, New Zealand J. of Geol. and Geophys. **36**, 51 (1993).

[38] D. D. Jackson, Proc. Natl. Acad. Sci. **93**, 3772 (1996).

[39] Y. Y. Kagan and D. D. Jackson, J. Geophys. Res. **100**, 3943 (1995).

[40] Y. Shi, J. Liu, and G. Zhang, J. Appl. Probab. **38A**, 222 (2001).

[41] M. Wyss and R. Burford, Nature **329**, 323 (1987).

[42] V. Kossobokov, L. Romashkova, V. Keilis-Borok, and J. Healy, Phys. Earth Planet. Inter. **11**, 187 (1999).

[43] G. Zöller, S. Hainzl, and J. Kurths, J. Geophys. Res. **106**, 2167 (2001).

[44] P. B. Stark, Geophys. Res. Lett. **23**, 1399 (1996).

[45] P. Stark, Geophys. J. Int. **131**, 495 (1997).

[46] Y. Kagan, Geophys. Res. Lett. **23**, 1315 (1996).

[47] B. Bolt, *Earthquakes* (W.H. Freeman, New York, 1993).

[48] P. Stark and C. Frohlich, J. Geophys. Res. **90**, 1859 (1985).

[49] D. Freedman, Statistical Science **14**, 243 (1999).

[50] A. Udias and J. Rice, Bull. Seis. Soc. Am. **65**, 809 (1975).

[51] D. Vere-Jones, J. Roy. Stat. Soc., Ser. B **32**, 1 (1970).

[52] Y. Ogata, JASA **83**, 9 (1988).

[53] Y. Ogata, Pure Appl. Geophys. **155**, 471 (1999).

[54] R. Console, D. Pantosti, and G. D'Addezio, Ann. Geophys. **45**, 723 (2002).

[55] R. Console, Tectonophysics **338**, 261 (2001).

[56] F. Mulargia and P. Gasperini, Geophys. J. Int. **111**, 32 (1992).

[57] K. Riedel, Geophys. Res. Lett. **23**, 1407 (1996).

[58] P. Varotsos, K. Eftaxias, F. Vallianatos, and M. Lazaridou, Geophys. Res. Lett. **23**, 1295 (1996).

[59] P. Stark, SticiGui: Statistics tools for internet and classroom instruction with a graphical user interface, 1997-2007.

[60] D. Bowman, G. Ouillon, C. Sammis, A. Sornette, and D. Sornette, J. Geophys. Res. **103**, 24359 (1998).

[61] K. Nanjo, J. Holliday, C. c. Chen, J. Rundle, and D. Turcotte, Tectonophysics **424**, 351 (2006).

[62] K. Tiampo, J. Rundle, S. McGinnis, S. Gross, and W. Klein, Europhys. Lett. **60**, 481 (2002).

[63] A. Prozorov, Pure and Applied Geophysics **127**, 1 (1988).

[64] R. Aceves, S. Park, and D. Strauss, Geophys. Res. Lett. **23**, 1425 (1996).

[65] V. Keilis-Borok, L. Knopoff, and I. Rotvain, Nature **283**, 259 (1980).

[66] L. Knopoff, Proc. Natl. Acad. Sci. **97**, 11880 (2000).

[67] P. A. Reasenberg, J. Geophys. Res. **90**, 5479 (1985).

[68] Y. Honkura and N. Tanaka, Geophys. Res. Lett. **23**, 1417 (1996).

[69] Y. Kagan and D. Jackson, J. Geophys. Res. **99**, 13,685 (1994).

[70] A. Michael, Technical Report No. 96-67, USGS, Menlo Park, CA (unpublished).

[71] H. Utada, Geophys. Res. Lett. **23**, 1391 (1996).

[72] M. Wyss and A. Allmann, Geophys. Res. Lett. **23**, 1307 (1996).

[73] J. Holliday, J. Rundle, K. Tiampo, W. Klein, and A. Donnellan, Pure Appl. Geoph. **163**, 2433 (2006).

[74] D. Vere-Jones, J. Roy. Stat. Soc., Ser. B **32**, 1 (1970).

[75] L. Knopoff and J. Gardner., Geophys. J. Int. **28**, 311 (1972).

[76] J. Zhuang, Y. Ogata, and D. Vere-Jones, JASA **97**, 369 (2002).

[77] S. D. Davis and C. Frohlich, Geophys. J. Int. **104**, 289 (1991).

[78] S. Barani, G. Ferretti, M. Massa, and D. Spallarossa, Geophys. J. Int. **168**, 100 (2007).

[79] J. Romano, JASA **83**, 698 (1988).

[80] J. Romano, Ann. Stat. **17**, 141 (1989).

[81] J. E. Ebel, D. W. Chambers, A. L. Kafka, and J. A. Baglivo, Seismo. Res. Lett. **78**, 57 (2007).

[82] H. Kanamori and D. Anderson, Bull. Seis. Soc. Am. **65**, 1073 (1975).

[83] A. Helmstetter, Y. Y. Kagan, and D. D. Jackson, Seismo. Res. Lett. **78**, 78 (2007).

[84] D. L. Wells and K. J. Coppersmith, Bull. Seis. Soc. Am. **84**, 974 (1994).

[85] S. Shlien and M. Toksöz, Earthquake Notes **45**, 3 (1974).

[86] L. Knopoff, Y. Y. Kagan, and R. Knopoff, Bull. Seis. Soc. Am. **72**, 1663 (1982).

[87] Y. Ogata, Annals of the Institute of Statistical Mathematics **50**, 379 (1998).

[88] D. Marsan and O. Lengliné, Science **319**, 1076 (2008).

[89] D. Schorlemmer and M. C. Gerstenberger, Seismo. Res. Letters **78**, 30 (2007).

[90] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, *Springer Texts in Statistics*, 3rd ed. (Springer, New York, 2005).

[91] M. V. Matthews and P. A. Reasenberg, Pure Appl. Geoph. **126**, 357 (1988).

[92] P. A. Reasenberg and M. V. Matthews, Pure Appl. Geoph. **126**, 373 (1988).

[93] P. Massart, Ann. Prob. **18**, 1269 (1990).

[94] D. Sornette and S. Utkin, Phys. Rev. E **79**, 061110 (2009).

[95] T. Utsu, Y. Ogata, and M. Matsu'ura, J. Phys. Earth **43**, 1 (1995).

[96] S. Hainzl and D. Marsan, J. Geophys. Res. **113**, B10309 (2008).

[97] A. Helmstetter and D. Sornette, J. Geophys. Res. **107**, 2237 (2002).

[98] Y. Y. Kagan and L. Knopoff, J. Geophys. Res. **86**, 2853 (1981).

[99] Y. Y. Kagan and L. Knopoff, Science **236**, 1563 (1987).

[100] A. G. Hawkes, Biometrika **58**, 83 (1971).

[101] A. G. Hawkes, in *Stochastic Point Processes* (John Wiley, New York, 1972), pp. 261–271.

[102] A. G. Hawkes and D. Oakes, J. Appl. Prob. **11**, 493 (1974).

[103] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: 1 (Probability and Its Applications)*, 2003. corr. 2nd. ed. (Springer, Berlin, 2005).

[104] R. Crane and D. Sornette, Proc. Natl. Acad. Sci. **105**, 15649 (2008).

[105] D. Vere-Jones and T. Ozaki, Annals of the Institute of Statistical Mathematics **34**, 189 (1982).

[106] Y. Y. Kagan, Bull. Seis. Soc. Am. **92**, 641 (2002).

[107] Y. Y. Kagan and D. D. Jackson, Geophys. J. Int. **143**, 438 (2000).

[108] Y. Ogata, Transactions on Information Theory **IT-27**, 23 (1981).

[109] J. Møller and J. G. Rasmussen, Adv. Appl. Prob. **37**, 629 (2005).

[110] G. Bravaccino, P. Brémaud, and G. Nappo, Technical report, Department of Mathematics, Sapienza University of Rome (unpublished).

[111] A. Dempster, N. Laird, and D. Rubin, Journal of the Royal Statistical Society, Series B **39**, 1 (1977).

[112] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, Ann. Math. Statist. **41**, 164 (1970).

[113] J. Zhuang, Y. Ogata, and D. Vere-Jones, J. Geophys. Res. **109**, B05301 (2004).

[114] A. Veen, personal communication, 2009.

[115] A. Veen and F. P. Schoenberg, in *Case Studies in Spatial Point Process Models* (Springer, New York, 2005), pp. 293–306.

[116] T. Utsu, in *International handbook of earthquake & engineering seismology* (Academic Press, New York, 2002), Chap. 43, pp. 719–732.

[117] Y. Y. Kagan, Pure Appl. Geophys. **164**, 1947 (2007).

[118] D. Arabelos, G. Asteriadis, M. Contadakis, G. Zioutas, D. Xu, C. Zhang, and B. Zheng, Tectonophysics **338**, 315 (2001).

[119] Y. Kagan and D. Jackson, J. Geophys. Res. **96**, 21,419 (1991).

[120] G. M. Molchan, Phys. Earth Planet. Inter. **61**, 84 (1990).

[121] *Forecast Verification: a Practioner's Guide in Atmospheric Science*, edited by I. T. Jolliffe and D. B. Stephenson (J. Wiley, Chichester, 2003).

[122] V. I. Keilis-Borok, P. Shebalin, and I. Zaliapin, PNAS **99**, 16,562 (2002).

[123] G. M. Molchan, in *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction* (Springer, Heidelberg, 2003), Chap. 5, pp. 208–237.

[124] J. Neyman and E. S. Pearson, Phil. Trans. Roy. Soc. Ser. A **231**, 289 (1933).

[125] Y. Y. Kagan and D. D. Jackson, Geophys. Res. Lett. **23**, 1433 (1996).

[126] S. Asmussen, *Applied Probability and Queues* (Springer, New York, 2003).

[127] D. J. Daley and D. Vere-Jones, J. Appl. Probab. **41A**, 297 (2004).

[128] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth dover printing, tenth GPO printing ed. (Dover, New York, 1964).

[129] S. D. B. Goes, J. Geophys. Res. **101**, 5739 (1996).

[130] S. Nishenko and R. Buland, Bull. Seis. Soc. Am. **77**, 1382 (1987).

[131] I. A. Parvez and A. Ram, Pure Appl. Geophys. **149**, 731 (1997).

[132] T. Rikitake, Tectonophysics **23**, 299 (1974).

[133] T. Utsu, Bull. Earthq. Res. Inst. Univ. Tokyo **59**, 53 (1984).

[134] A. Helmstetter and D. Sornette, J. Geophys. Res. **108**, 2482 (2003).

[135] G. M. Molchan and V. I. Keilis-Borok, Geophys. J. Int. **173**, 1012 (2008).

[136] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, No. 38 in *Regional Conference Series in Applied Mathematics* (Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1982).

[137] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, *Mono. Stat. Appl. Probab.* (Chapman and Hall, London, 1993).

[138] V. N. Vapnik and A. Y. Chervonenkis, Theory Probab. Appl. **16**, 264 (1971).

[139] R. Beran and P. Millar, Ann. Statist. **15**, 1131 (1987).

[140] J. P. Romano, Technical report, Department of Statistics, Stanford University (unpublished).

Appendix A

Resampling and randomisation tests

Resampling techniques have become widely used since adequate computational power has become commonly available [136, 137]. In particular, tests of significance using randomisation or the bootstrap can often be used even when an analytic form for the distribution of the test statistic is not known. Romano [79, 80] developed methodology for bootstrap and randomisation tests of nonparametric hypotheses such as independence, symmetry, and exchangeability. Before describing his approach in detail, we first present a review of Vapnik-Chervonenkis theory.

A.1 Vapnik-Chervonenkis classes

Suppose the set $D \subset \chi$ has finite cardinality n. D is *shattered* by a collection of sets $\mathbf{V}$ if all 2^n subsets $d \subset D$ can be written $d = V \cap D$ for some $V \in \mathbf{V}$. The *Vapnik-Chervonenkis dimension* [138], or *VC dimension*, of a collection $\mathbf{V}$ is the cardinality of the largest set that can be shattered by $\mathbf{V}$.[1] If for all n there exists a set of cardinality n that can be shattered by $\mathbf{V}$, the VC dimension of $\mathbf{V}$ is infinite. $\mathbf{V}$ is a *VC class* if and only if its VC dimension is finite.

In $\mathbf{R}^n$, a lower-left quadrant is a set of the form

$$(-\infty, t_1] \times \ldots \times (-\infty, t_n] : t_1, \ldots, t_n \in \mathbf{R}. \tag{A.1}$$

The set of all lower-left quadrants on $\mathbf{R}^n$ is a VC class with VC dimension n. Note that the Cartesian product of VC classes is also a VC class.

[1] In some definitions, the VC dimension is the smallest n for which no set of cardinality n can be shattered. This definition gives a VC dimension one greater than does our definition.

A.2 Romano-type tests

Let Ω be a known collection of distributions on sample space χ. Suppose we observe a sample $\{X_j\}_{j=1}^n$ from a distribution $P \in \Omega$. Let Ω_0 be the set of distributions in Ω that are invariant under a given idempotent transformation $\tau : \Omega \to \Omega_0$; that is, $\tau(P_0) = P_0$ for all $P_0 \in \Omega_0 \subset \Omega$. Let our null hypothesis be that $P \in \Omega_0$.

Romano [79, 80] proposes hypothesis tests that rely on a seminorm that generalises Kolmogorov-Smirnov distance [79]. Let $\mathbf{V}$ be a VC class of subsets of χ. In particular, if the sample space is $\mathbf{R}^n$, then $\mathbf{V}$ may be the set of lower-left quadrants. Define $\delta_{\mathbf{V}} : \Omega \times \Omega \to \mathbf{R}^+$ as

$$\delta_{\mathbf{V}}(P, Q) \equiv \sup_{V \in \mathbf{V}} |P(V) - Q(V)|, \tag{A.2}$$

for $P, Q \in \Omega$. To perform a hypothesis test, select $\mathbf{V}$ and τ such that $\delta_{\mathbf{V}}(P, \tau P) = 0$ if and only if $P \in \Omega_0$. Let $\hat{P}_n = \hat{P}_n(X)$ be the empirical measure of $\{X_j\}_{j=1}^n$. Romano suggested the test statistic

$$T_n(X) \equiv \sqrt{n}\, \delta_{\mathbf{V}}\Big(\hat{P}_n, \tau(\hat{P}_n)\Big). \tag{A.3}$$

Informally, $\tau(\hat{P}_n)$ is like a "projection" of the empirical measure onto the set of measures that satisfy the null hypothesis. The null hypothesis is rejected if this test statistic is large; that is, when $\hat{P}_n$ and $\tau(\hat{P}_n)$ are distant in the seminorm (A.2).

We need to estimate the sampling distribution of T_n under the null hypothesis. Since $\tau(\hat{P}_n)$ is an element of the null set Ω_0, we could take bootstrap samples from $\tau(\hat{P}_n)$ and compute T_n for every one of these. The critical value of a level-α test will be approximately the $1 - \alpha$ quantile of the empirical distribution of the bootstrap sample test statistics.

An alternative to the bootstrap is randomisation. Suppose there is a known finite group of transformations $\mathbf{G}_n = \{g_{nj}\}$ of the sample space such that all $P_0 \in \Omega_0$ are invariant under elements of $\mathbf{G}_n$. The *orbit* $\mathcal{X}$ under $\mathbf{G}_n$ of a point $x \in \chi$ is the set

$$\mathcal{X} \equiv \{g(x) : g \in \mathbf{G}_n\}.$$

In a permutation test, $\mathbf{G}_n$ is some group of transformations. For example, if $\mathbf{G}_n$ is a permutation group, the orbit consists of permutations of the data. Every point in the orbit $\mathcal{X}$ is equally likely under the null hypothesis. To find a null distribution for the test statistic, compute the statistic $T_n(g(x))$ for every $g \in \mathbf{G}_n$. Then the one-tailed permutation test P-value, conditioning on the orbit, is the proportion of the statistics $T_n(g(x))$ that are greater than or equal to the observed test statistic $T_n(x)$. If a permutation test taking $T_n(g(x))$ to be the null sampling distribution of $T_n(x)$ is constructed to have level α given the orbit, it will also have level α unconditionally.

In both bootstrap and randomisation tests, we reject the null when the test statis-

tic (A.3) is large. The methods differ in how they estimate the null distribution, and hence in their critical values. Romano [79, 80] showed that under fairly general conditions, bootstrap and randomisation tests are asymptotically equivalent: for a given test statistic and sufficiently large samples, the methods have comparable power and critical value.

Calculating exact significance levels for randomisation tests may require huge amounts of computation. There are two main computational difficulties. Firstly, it is difficult to find the global supremum defined in (A.2) if the sample space is multi-dimensional. Pursuing Beran and Millar's idea of stochastic testing [139], Romano suggested that instead of searching through all sets in $\mathbf{V}$, one could randomly choose s search sets $\mathbf{V_s} \equiv \{V_1, \ldots, V_s\} \subset \mathbf{V}$ according to a probability on $\mathbf{V}$.[2] The supremum in (A.2) is then found over all sets in $\mathbf{V_s}$, rather than over all sets in $\mathbf{V}$. For the test to be consistent, the number of search sets must grow quickly with the sample size n [139]. The supremum over $\mathbf{V_s}$ is faster to find than the supremum over $\mathbf{V}$ but may be quite different, and the power of the test can be compromised.

Romano asserted that the search sets could be the same for all permutations, or could be selected independently for every permutation. We consider the former case here.[3] The distance seminorm becomes

$$\delta_{\mathbf{V_s}}(P, Q) \equiv \sup_{V \in \mathbf{V_s}} |P(V) - Q(V)|, \tag{A.4}$$

while the test statistic is

$$T_n \equiv \sqrt{n}\, \delta_{\mathbf{V_s}}\left(\hat{P}_n, \tau(\hat{P}_n)\right). \tag{A.5}$$

The test statistic for the original data is $T_n(x)$. Under the null hypothesis, every permutation $g(x) \in \mathcal{X}$ of the data is equally likely. So under the null and conditional on $X \in \mathcal{X}$, $T_n(x)$ is drawn uniformly at random from the set $\{T_n(g(x)), g(x) \in \mathcal{X}\}$ of test statistics evaluated on all permutations of the data. An exact test may thus be constructed by comparing $T_n(x)$ to this set.

A second computational issue is that the orbit is large. For instance, if $\mathbf{G}_n$ is the permutation group on n elements, there are $n!$ permutations. Instead of calculating the test statistic for all elements in the orbit, one may approximate the null distribution by randomly sampling r transformations from $\mathbf{G}_n$, then applying each to the data and calculating the test statistic for the transformed data. Suppose the sample is taken with replacement from the group. Let the test statistic evaluated on the original data be $T_n(x)$, and let the values of the statistic applied to the transformed

[2] Search sets chosen deterministically also give the correct level, but may miss regions of the search space.

[3] Romano [140] showed that for the bootstrap, either way of selecting search sets results in tests that have the correct asymptotic level under quite general conditions.

data be $(T_n(g_1(x)), \ldots, T_n(g_r(x)))$. If the null hypothesis is true, then $T_n(x)$ is drawn from the same distribution as the elements of $(T_n(g_1(x)), \ldots, T_n(g_r(x)))$. The choices of transformations, given n and r, are independent of the data and of each other. Thus the elements of the concatenated vector $(T_n(x), T_n(g_1(x)), \ldots, T_n(g_r(x)))$, conditional on $X \in \{x, g_1(x), \ldots, g_r(x)\}$, are independent and identically distributed if the null is true. If the sample of transformations is taken without replacement, the random vector is exchangeable. In either case, under the null hypothesis, the probability that $T_n(x)$ is one of the k largest elements of the vector is $k/(r+1)$ (provided there are no ties), since $r+1$ is the length of the vector. So if $T_n(x)$ is the kth-largest element of the vector, $k/(r+1)$ gives an approximate P-value for the test that uses all transformations in the group [80].

Appendix B

R code for test of exchangeability

```
# Input catalog with columns "longitude"; "latitude"; "times"

catalog = read.table(file.choose(),header=T)

# Find catalog length

n = nrow(catalog)

# Sort in time order

catalog=catalog[order(catalog$times),]

# Extract ranks (assume no ties)

x.rank = rank(catalog$longitude)
y.rank = rank(catalog$latitude)

# Find empirical distribution of spatial ranks

xy.upper = matrix(NA,n,n)
for(I in 1:n){
 for(J in 1:n){
    xy.upper[I,J] = sum((y.rank<=y.rank[I])*(x.rank<=x.rank[J]))
 }
}
# xy.upper[I,J] is the number of points
# with y <= y[i], x <= x[j]
# y is row, x is column
```

```
### Distance function

distfind <- function(x.rank,y.rank,xy.upper){
 n = length(x.rank)
 # Set some stuff to zero
 teststat = 0
 xyz.temp = matrix(0,n,n)
 # xyz.temp is the number of points
 # with y <= y[i], x <= x[j], z <= Z
 # i.e. empirical distribution at time Z
 # Now go through search space chronologically
 # update xyz.temp
 # find the max; check the min isn't close; if it is, look around
 # this is where we really want to minimise the number of operations
 for(Z in 1:n){
    xyz.temp = xyz.temp + (y.rank>=y.rank[Z])%*%t(x.rank>=x.rank[Z])
    dist.matrix = xyz.temp/n-xy.upper/n*Z/n
    teststat = max(teststat,abs(dist.matrix))
 }
 return(teststat)
}

teststat = distfind(x.rank,y.rank,xy.upper)

# Number of perms

N = 10000
permustat = rep(NA,N)

### It's permuting time

for(permu in 1:N){

 o=sample(n)
 x.perm = x.rank[o]
 y.perm = y.rank[o]
 xy.perm = xy.upper[o,o]

 permustat[permu] = distfind(x.perm,y.perm,xy.perm)
```

```
}

# P-value
mean(permustat>=teststat)
```

Appendix C

Proof of the optimal predictor lemma

This is a proof of the lemma in section 5.2.3.

Proof of (i): Existence

Let

$$\tau^-(\Lambda) = \mathbf{E}\left[\frac{1}{T}\int_0^T \mathbf{1}(\lambda(t) > \Lambda)dt\right].$$

This is non-increasing and right-continuous, with $0 \leq \tau^-(\Lambda) \leq 1$.

Let

$$\tau^+(\Lambda) = \mathbf{E}\left[\frac{1}{T}\int_0^T \mathbf{1}(\lambda(t) \geq \Lambda)dt\right].$$

This is non-increasing and left-continuous, with $0 \leq \tau^+(\Lambda) \leq 1$.

Given $0 \leq \tau \leq 1$, there is a unique Λ such that $\tau^-(\Lambda) \leq \tau \leq \tau^+(\Lambda)$. The following function satisfies (5.7):

$$H_t = \begin{cases} 1 & \text{when } \lambda(t) > \Lambda \\ \frac{\tau - \tau^-(\Lambda)}{\tau^+(\Lambda) - \tau^-(\Lambda)} & \text{when } \lambda(t) = \Lambda \\ 0 & \text{when } \lambda(t) < \Lambda. \end{cases} \tag{C.1}$$

The process H_t is previsible because λ is previsible.

Proof of (ii): Sufficiency

Suppose that H_t satisfies (5.7) and (5.9). Let H_t^* be some other previsible function satisfying (5.7). Let S^+ be the set for which $H > H^*$. For all $\{t, \omega\}$ in this set, $H_t > 0$ and $\lambda(t) \geq \Lambda$. Let S^- be the set for which $H < H^*$. For all $\{t, \omega\}$ in this set,

$H_t^*(\omega) > 0$ and $\lambda(t) \leq \Lambda$. Note that

$$\int_{S^+ \cup S^-} H_t \, d\mu = \int_{S^+ \cup S^-} H_t^* \, d\mu.$$

The difference in the expected number of events that occur during alarms is

$$\begin{aligned} \mathbf{E}\left[\int_0^T H_t \, dN(t)\right] - \mathbf{E}\left[\int_0^T H_t^* \, dN(t)\right] &= \mathbf{E}\left\{\int_0^T [H_t - H_t^*] \, dN(t)\right\} \\ &= \mathbf{E}\left\{\int_0^T [H_t - H_t^*] \, \lambda(t) dt\right\} \\ &= \int_{S^+ \cup S^-} [H_t - H_t^*] \, \lambda(t) d\mu \end{aligned}$$

Now,

$$\int_{S^+} [H_t - H_t^*] \, \lambda(t) d\mu \geq \Lambda \int_{S^+} [H_t - H_t^*] \, d\mu$$

and

$$\int_{S^-} [H_t^* - H_t] \, \lambda(t) d\mu \leq \Lambda \int_{S^-} [H_t^* - H_t] \, d\mu.$$

So

$$\begin{aligned} \int_{S^+ \cup S^-} [H_t - H_t^*] \, \lambda(t) d\mu &\geq \Lambda \int_{S^+ \cup S^-} [H_t - H_t^*] \, d\mu \\ &\geq 0. \end{aligned}$$

Proof of (iii): Necessity

Suppose that some previsible H_t^* maximises (5.6) subject to (5.7). Let H_t satisfy (5.7) and (5.9). Let S^+ be the set for which $H > H^*$ and S^- be the set for which $H < H^*$. Define the set S as

$$S = (S^+ \cup S^-) \cap \{t, \omega : \lambda(t) \neq \Lambda\}. \tag{C.2}$$

Then

$$\int_{S^+ \cup S^-} (H_t - H_t^*)(\lambda(t) - \Lambda) d\mu = \int_S (H_t - H_t^*)(\lambda(t) - \Lambda) d\mu. \tag{C.3}$$

Suppose that $\mu(S) > 0$. On S, when $H_t > H_t^*$, $\lambda(t) > \Lambda$; and when $H_t < H_t^*$, $\lambda(t) < \Lambda$. So $(H_t - H_t^*)(\lambda(t) - \Lambda)$ is strictly positive on S. Thus

$$\int_S (H_t - H_t^*)(\lambda(t) - \Lambda) d\mu > 0. \tag{C.4}$$

This implies

$$\int_{S^+ \cup S^-} H_t[\lambda(t) - \Lambda] d\mu > \int_{S^+ \cup S^-} H_t^*[\lambda(t) - \Lambda] d\mu \tag{C.5}$$

and hence that H_t gives a larger value of (5.6) than H_t^*, yielding a contradiction. Therefore $\mu(S) = 0$, and $H^* = H$ almost everywhere with respect to μ, except where $\lambda(t) = \Lambda$.