Statistical Models for Earthquakes Incorporating Ancillary Data

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

at Massey University, Palmerston North,
New Zealand.

Ting Wang

2010

In memory of Professor Ma Li

Abstract

This thesis consists of two parts. The first part proposes a new model – the Markov-modulated Hawkes process with stepwise decay (MMHPSD) to investigate the seismicity rate. The MMHPSD is a self-exciting process which switches among different states, in each of which the process has distinguishable background seismicity and decay rates. Parameter estimation is developed via the expectation maximization algorithm. The model is applied to data from the Landers earthquake sequence, demonstrating that it is useful for modelling changes in the temporal patterns of seismicity. The states in the model can capture the behavior of main shocks, large aftershocks, secondary aftershocks and a period of quiescence with different background rates and decay rates. The state transitions can then explain the seismicity rate changes and help indicate if there is any seismicity shadow or relative quiescence.

The second part of this thesis develops statistical methods to examine earthquake sequences possessing ancillary data, in this case groundwater level data or GPS measurements of deformation. For the former, signals from groundwater level data at Tangshan Well, China, are extracted for the period from 2002 to 2005 using a moving window method. A number of different statistical techniques are used to detect and quantify coseismic responses to P, S, Love and Rayleigh wave arrivals. The P phase arrivals appear to trigger identifiable oscillations in groundwater level, whereas the Rayleigh waves amplify the water level movement. Identifiable coseismic responses are found for approximately 40 percent of magnitude 6+ earthquakes worldwide. A threshold in the relationship between earthquake magnitude and well–epicenter distance is also found, satisfied by 97% of the identified coseismic responses, above which coseismic changes in groundwater level at Tangshan Well are most likely.

A non-linear filter measuring short-term deformation rate changes is introduced to extract signals from GPS data. For two case studies of a) deep earthquakes in central North Island, New Zealand, and b) shallow earthquakes in Southern California, a hidden Markov model (HMM) is fitted to the output from the filter. Mutual information analysis indicates that the state having the largest variation of deformation rate contains precursory information that indicates an elevated probability for earthquake occurrence.

Acknowledgements

I wish to acknowledge all those who have helped me through the three years of my Ph.D. study. First and foremost, I would like to express my gratitude to my supervisor Mark Bebbington. Three years ago, I followed my dream and came to Massey University to become a research professional with a special interest in probabilistic models for geophysical hazard but with little background knowledge about geophysics. Without Mark's kind guidance and motivation, nothing could have been accomplished. It's from Mark that I learnt first how to crawl, then how to walk and finally how to run in academic research. During the three years, I have never found a proper way to thank him. In completion of this thesis I would like to thank Mark for being patient when I was not making good progress, inspiring and motivating me when I got lost, kindly encouraging me when I was frustrated, making my research a top priority, and showing me the beauty of the New Zealand countryside.

I am very grateful to my co-supervisors David Harte and David Vere-Jones for their continuing support, guidance and encouragement, as well as all their helpful discussions for this research. David H always came to my rescue whenever I was faced to programming issues and patiently explained everything in detail. I also want to thank him for teaching me how to write an R package. David VJ was also inspirational for my undertaking of this adventure and provided insightful comments for this work. I would also like to thank David VJ for inviting me to the King Lear play, to the benefit of my cultural education, and encouraging me to enjoy the world beyond academia. My gratitude goes to all three supervisors for thoroughly reviewing and improving this manuscript.

My special thanks also go to two of the most important people in my life, Professor Ma Li and Jill Bebbington. If it is Mark who made me love statistics-related geophysics, it was Professor Ma Li who inspired my first interest in this field. She was always willing to ask for and provided the best for her students. For her having been truly inspirational, I would like to dedicate this thesis to her memory. By having Mark as my principal supervisor, I was actually very lucky to find myself two teachers. Jill taught me how to live a valuable life while Mark was trying to inspire and improve my research skills. If there was a Nobel Prize in cooking and raising a family, I wish Jill would be awarded one. My love also goes to my two little friends, Anna and Craig Bebbington for all the joy they have brought to me.

I also want to acknowledge all those who have provided insightful comments on my research or otherwise helped me, particularly Marco Brenna, Cheryl and Ray Brownrigg, Professor Huizhong

Chen, Robert Davies, Alistair Gray, Jianping Huang, Takaki Iwata, Professor Chin-Diew Lai, Wenjing Li, Nicholas Look, Dejun Luo, Tim O'Dea, Professor Yoshihiko Ogata, Patrick Rynhart, Professor Martha Kane Savage, Peter Thomson, Baojun Yin, Jiancang Zhuang and Walter Zucchini, and my apologies to all those whom I failed to mention. I would also like to express my appreciation to Takaki Iwata, Professor Chin-Diew Lai, and Roger Littlejohn for their thorough review of my thesis.

It was indeed delightful to work with a group of ever friendly statisticians, the Statistics Research Group at Massey University. I would like to thank all of them for all their help and encouragement. The postgraduate students in the Stats PostGrads Office AHA2.74 accompanied me through these three years in the office and made my student life much more enjoyable. I am grateful to all of them for cheering me up when I was down and offering help whenever I needed it.

My appreciation also goes to those friends who have always accompanied and supported me outside of the academic world, Valentine Borges, Boon Feei Chong, Beibei Feng, Kasey Gordon, Marissa Isidro, Wirya Khim, Xiaojing Peng, Bandeth Ros, Zhongwei Xing, Zhiwen Yu, and many others.

This research project was financially supported by Marsden Fund administrated by the Royal Society of New Zealand. I am grateful to the Tangshan Earthquake Administration for providing the well data.

Last but not the least, to my parents, my brother and my sister-in-law, goes my deepest love for always being there to support me.

Contents

List of Figures

List of Tables

Glossary

AMR	Accelerated moment release model
Craton	An old and stable part of the continental lithosphere
ETAS	Epidemic type aftershock sequence model
Declustered data	Earthquake sequences with aftershocks removed
Gutenberg-Richter law	The total number of earthquakes in a population that are larger than or equal to some magnitude M varies as 10^{bM}
HMM	Hidden Markov models
Hypocenter	The point within the earth where an earthquake rupture starts
LSRM	Linked stress release model
MMGLM	Markov-modulated generalized linear model
MMHPSD	Markov-modulated Hawkes process with stepwise decay
MMPP	Markov-modulated Poisson process
NHMM	Nonhomogeneous hidden Markov model
Relative quiescence	A significant decrease of earthquakes compared with the occurrence rate expected from a point-process model for ordinary seismic activity
Seismicity/stress shadows	Seismicity rate decreases
SRM	Stress release model

Chapter 1

Introduction

1.1 Motivation

1.1.1 Earthquake Cycles

The earthquake process is observed only through data that generally do not directly reflect the underlying dynamics. The observations usually consist of records of earthquake features such as magnitude, location and depth, while the underlying dynamics, for example the accumulation of stress, remain unobserved. Fedotov (1968) described the concept of a seismic cycle and suggested an increasing seismicity rate before the second main shock of the cycle. Note that the earthquake cycle considered in this thesis refers to a sequence of stages of seismicity, not to periodic or "characteristic" earthquake cycles. Mogi (1968) also suggested that prior to a main shock, there are concentrative occurrences of foreshocks in a region near the epicenter of the main rupture preceded by relative seismic quiescence. Bufe and Varnes (1993) and Bufe et al. (1994) reported an increase in cumulative regional energy release prior to a large earthquake and explored the accelerated moment release (AMR) model. The statistical version of this model was studied by Vere-Jones et al. (2001), Jaumé and Bebbington (2004) and Bebbington et al. (2010), all indicating that the AMR model only models a particular seismic cycle whereas in reality there may have been multiple such cycles (e.g., Bufe and Varnes, 1993). However, Hardebeck et al. (2008) found AMR present in data which were simulated without benefit of AMR. Jaumé and Sykes (1999) reviewed the evidences that many large earthquakes are preceded by a period of accelerating seismic activity of moderate-sized earthquakes, and pointed out that the changes in the level of driving stress (stress accumulation and release) during a seismic cycle were the cause of the observed seismicity changes. This is also

supported by the study of Sammis and Smith (1999). They showed a seismic cycle in which a large event is followed by a shadow period of quiescence and then a new approach back toward the critical state, in a scaling region at times close to which the events become larger and energy release increases. These, as well as many others, such as Bowman et al. (1998), Zhuang (2000), Lombardi et al. (2006) and Pievatolo and Rotondi (2008), indicate that seismicity may present a cyclic nature of some sort.

Hidden Markov models (HMMs) form a remarkably general statistical framework for modelling such partially observed systems, by assuming that the unobserved (or hidden) process is a Markov chain and the observations are conditionally independent given the hidden states. The HMM was initially introduced by Baum and Petrie (1966) as a probabilistic function of a discrete-time Markov chain with a finite number of states. There have been extensive developments of this model, including the methods of parameter estimation (Rabiner, 1989 and the references therein), extension to processes where the hidden process is a continuous-time Markov chain with discrete states and the observation is from a Poisson process such as Markov-modulated Poisson process (MMPP; Fischer and Meier-Hellstern, 1993 and the references therein), extension to processes where either the observed or the hidden process depends on another observed variable (Markov-modulated generalized linear models (MMGLM) for the former, Harte (2005); nonhomogeneous hidden Markov models (NHMM) for the later, Hughes (1993) and the references thereafter).

1.1.2 Clustering

To model earthquake processes, there should be more significant extensions to HMMs, as the hidden process is a continuous-time Markov chain and the observations of earthquake processes are point processes which often occur in a self-exciting way (a process is considered to be self-exciting if it depends on the entire or some of its history which affects the intensity function of the process). None of the existing discrete-time HMMs, MMPPs, MMGLMs and NHMMs is adequate to capture these features simultaneously.

Vere-Jones and Davies (1966) investigated the earthquake data from the main seismic region of New Zealand for the years 1942–1961. The analysis revealed a significant degree of dependence among earthquakes and suggested that the earthquake process is among the general class of ‘contagious processes’, in which the events already occurred often trigger new ones. Using the triggered model suggested by them, investigations of the earthquake data from New Zealand in Vere-Jones (1970) and from 14 regions listed in Table 1 in Hawkes and Adamopoulos (1973) further supported

such significant correlations between earthquakes. The extensive applications of the Epidemic Type Aftershock Sequence (ETAS) model (Ogata, 1988 and subsequent papers), in which each earthquake is able to produce offspring in the same way as a diseased people may pass the disease to other people in epidemics, demonstrate that earthquakes do present some self-exciting features.

Vere-Jones (1978) proposed the stress release model (SRM) for triggering between main sequence events, which incorporates a deterministic build-up of stress within a single region and its stochastic release through earthquakes. As reviewed in Bebbington and Harte (2001; 2003), earthquake interaction by means of stress triggering and stress shadows (across different regions at a distance, see also Zheng and Vere-Jones, 1994) has been well accepted. In order to account for this interaction, instead of using the SRM, they used the linked stress release model (LSRM) (Liu et al., 1998; Lu et al., 1999) to investigate seismicity in multiple regions. Both the SRM and LSRM were developed for triggering between main sequence events only (with aftershocks removed, cf., Bebbington and Harte, 2001). In order to formulate main shocks and aftershocks in one model, a two-node stress release/transfer model was presented by Borovkov and Bebbington (2003) and Bebbington (2008), which was also shown to be an alternative to the ETAS model for aftershock sequences. The two-node model is a two region version of the LSRM to describe main shocks by one region (or node) and aftershocks by the second node. Borovkov and Bebbington (2003) also pointed out that in sequences which include a secondary sequence triggered by main shocks, a third node might be added to the two-node model, i.e., a three-node model might be considered. These studies suggest that an assumption of independence of the events in an earthquake catalogue is inadequate for the analysis of evolution of seismicity.

A Markov-modulated Poisson process (MMPP) is a doubly stochastic Poisson process, the intensity of which is controlled by a finite, non-observable, continuous-time Markov chain. It is parameterized by the r-state continuous-time Markov chain with infinitesimal generator $Q = (q_{ij})_{r\times r}$ and the r Poisson arrival rates $\lambda_1, \lambda_2, \cdots, \lambda_r$, where

$$q_{ii} = -\sum_{\substack{j=1 \\ j\neq i}}^{r} q_{ij}.$$

The event occurrence rate of the process in one particular state is independent of the previous events and is thus not suitable here. It is necessary to develop an HMM type model in which the process switches among some finite states and the occurrence rate of the events in each state depends on the historical events. Instead of the well accepted stochastic models for seismicity analysis such as the

SRM or the ETAS model, the self-exciting Hawkes process with exponential decay rate (Hawkes, 1971) will be used to formulate the event occurrence rate in each hidden state. This is because this Hawkes process, which consists of two types of events, immigrants with a Poisson occurrence rate and offspring generated from an exponential decay rate, has the Markovian property and is thus mathematically tractable. For a seismic cycle, the relative quiescence can be considered as a state in which there are no conspicuous clusters, i.e., with a negligible background intensity function, while an immigration-branching rate with different background and decay rates can be assigned to each of the foreshock and aftershock periods. Note that this is in some sense a generalization of the two-node or three-node stress transfer model (Borovkov and Bebbington, 2003), in which the regions and nodes were pre-identified and fixed. The HMM type model, however, inherits the Markovian property, which automatically switches into a new regime or node when the underlying dynamics change.

1.1.3 Ancillary Data

Historically, there has been little data available beyond the observed record of the earthquake catalogue. A recent innovation has been the availability of related, or ancillary, data which are (hopefully) related to the process generating the hazard. More and more ancillary data, such as groundwater level oscillations and GPS measurements of deformation, are becoming available, which may either be dependent on the original observations (e.g., Matsumoto, 1992; Kitagawa and Matsumoto, 1996), or reflect the underlying dynamics (e.g., Roeloffs, 2006).

Earthquakes disrupt and change hydrologic processes. Oscillations or sustained changes are the two typical anomalous phenomena in groundwater levels in wells. The Tangshan Well in northern China has long been recognized by Chinese seismologists as sensitive to seismic disturbances. There exist more than thirty years of groundwater level observations at this well. Such a long record as that from Tangshan Well is nearly unique to the best of our knowledge (cf., Elkhoury et al., 2006). A great number of oscillations of the groundwater level are recorded at the Tangshan Well which may have been induced by large earthquakes from the global catalogue since the advent of 1-minute digital records in 2001. To date there has been little in the way of a coherent statistical examination of the possible link between earthquakes and hydrological responses. A better understanding of the whys and wherefores of occurrence or non-occurrence following an earthquake event is a necessary first step. The question also remains as to whether the oscillations occur (or are detectable) in all cases, and what factors affect the response, and by how much.

A great deal of effort has been put into analysis of anomalous changes in crustal deformation rates prior to large earthquakes in order to look for warnings of impending earthquakes. Motivated by the evidences of anomalous changes in deformation prior to the 1944 Tonankai earthquake (M_s=8.2) and the 1946 Nankaido earthquake (M_w=8.3), the Japanese government has intensively monitored aseismic deformation with the goal of issuing a warning prior to the next great earthquake (EERI Committee on the Anticipated Tokai Earthquake and Scawthorn, 1984). Roeloffs (2006) reviewed the evidences of preseismic deformation rate changes which have been found prior to 5 large earthquakes in Japan, 3 in the USA, 1 in Chile, 1 in Peru and 1 in Greece. However, preseismic deformation rate changes are relatively rare in the existing record, even after the wide deployment of Continuous GPS stations in the early 1990s. Some statistical analyses have been carried out on anomalies of GPS measurements from the apparent long-term pre-earthquake slip which may be related to earthquake occurrences (e.g., Ogata, 2007). As discussed in Roeloffs (2006), most of the time it is difficult to detect apparent long-term displacement. There may exist some anomalous changes of the deformation rate prior to large earthquakes, the detection of which, however, may require more sophisticated techniques.

As discussed above, for the groundwater level at Tangshan Well, the water level oscillations are buried in millions of data, while for the case of GPS measurements of deformation, there are very few obvious displacements. Therefore, methods for detection of groundwater level oscillation signals and extraction of subtle deformation rate changes in GPS measurements are necessary. The ancillary groundwater level and deformation data and the earthquake catalogue are observed independently. Moreover, an earthquake sequence has long been considered as a point process, whereas both the groundwater level and the GPS measurements of deformation are continuous time series (multivariate for the latter case). Hence, methods for identifying dependence between such geophysical processes are needed.

1.2 Overview

The two major objectives of this thesis are 1) to develop HMM type models in which the hidden process is governed by a discrete-state, continuous-time Markov chain, and the observations in each state are from a self-exciting point process, and 2) to use statistical methods to investigate earthquake data incorporating ancillary data. Hence, this thesis consists of two parts. The first part starts with a comprehensive review of HMMs including model definition and parameter estimation. It is then

followed by the development of HMM type models which account for the self-exciting feature of earthquake occurrences. The new HMM type model is called a Markov-modulated Hawkes process with stepwise decay. An application of this model is demonstrated using a real data set. The second part investigates methodologies of extracting signals from potential candidates for earthquake-related processes and detecting association between the resulting series of events. A brief overview of each chapter is as follows.

Chapter 2 provides an overview of HMMs. Given that the HMM framework contains two components, the observed part and the hidden (or unobserved) part, all three of the models mentioned above, the discrete-time HMM, the NHMM and the MMGLM, can be generalized in one model. This chapter provides the definition of this concise model and then reviews the three models in detail, including definitions, parameterizations, and parameter estimation.

Chapter 3 develops the Markov-modulated Hawkes process with stepwise decay (MMHPSD). An expectation maximization (EM) algorithm for the parameter estimation of this model is proposed, along with a detailed implementation algorithm. The method for evaluating the goodness-of-fit problem is also provided. Consistency of parameter estimation will be demonstrated by simulation.

Chapter 4 presents a check of the consistency of the parameter estimation algorithm by way of a simulation study of the MMHPSD, and an application of the model to a real data set from the Landers–Hector Mine sequence. The ETAS model is used to simulate an earthquake catalogue. The MMHPSD is then fitted to this catalogue in order to see what meaning can be attached to the hidden states the model identifies. The parameter estimates from this simulated catalogue are used to conduct the consistency test of the estimation algorithm as proposed in Chapter 3. A case study of the earthquake data around Landers is carried out for this model. The seismicity rate changes characterized by this model are compared with results from other investigations. Further modifications and possible improvements of this model for earthquake analysis are discussed.

Chapter 5 provides a detailed review of three statistical techniques characterizing the association between series of events, the coherence, mutual information and Lin–Lin model, highlighting their use in earthquake modelling and its ancillary data analysis. Coherence, which provides a measure of the degree of linear predictability of one process from another, is inadequate as a measure of general association for it may be identically 0 when two series are in fact related. However, such behavior does not occur for the coefficient of mutual information, which is a measure of the amount of information that one random variable contains about another. The Lin–Lin model, which describes the

influence of a point process input on a point process output, can identify linear causal relationships between one sequence of events and another. The three methods can be used to clarify association and identify causation.

Chapter 6 contains a systematic statistical analysis of earthquake-related groundwater level oscillations from the 1-minute digital records of Tangshan Well data since 2001. A method of extracting signals compatible with coseismic responses from the approximately two million such data during the four year period 2002–2005 is demonstrated, and their relation to 600 earthquakes of magnitude 6.0 or greater in the global catalogue during this same period is investigated. Identifiable coseismic responses are found for approximately 40 percent of the total number of such earthquakes. The association of the initial oscillations of the groundwater level with the arrivals of the earliest P phase, the earliest S phase, the Love and Rayleigh waves are examined using the coherence, mutual information and Lin–Lin model. The detection probabilities of different types of earthquakes are assessed in terms of the earthquake magnitudes, distances, depths and azimuths. The relation between the well signal characteristics (delay, length, maximum amplitude) and the earthquake magnitude, distance, depth and azimuth is also discussed.

Chapter 7 is devoted to an HMM based investigation of possible link between GPS measurements of deformation and earthquakes. An exploratory analysis of the earthquake occurrences around Taupo using HMMs, NHMMs, and MMGLMs with binary distribution for earthquake occurrences and the GPS measurements of deformation as ancillary variable is carried out, but none of the models seems to be informative for understanding the underlying dynamics of earthquake occurrences. Instead, a non-linear filter for the GPS process which serves as a smoothing tool is introduced. Essentially, it measures the short-term deformation rate ranges and is able to extract useful signals which are distinguishable from the majority of the data. For two case studies of a) deep earthquakes in the central North Island, New Zealand, and b) shallow earthquakes in Southern California, HMMs are fitted to the deformation rate ranges of the GPS measurements. The HMMs classify the deformation data into different patterns which form proxies for states of the earthquake cycle. Mutual information is then used to examine whether there is any correlation between these patterns, in particular the Viterbi path, and subsequent (or previous) earthquakes. The class of GPS movements identified by the HMM as having the largest range of deformation rate changes is examined in detail in terms of the precursory character for large earthquakes. The performance of possible earthquake forecast is assessed by comparing a decision rule (based on model characteristics) with the actual outcome.

Chapter 8 summarizes the conclusions of this thesis and provides detailed suggestions for future research involving the development of the HMM type models for geophysical data, NHMM and MMGLM analyses of earthquakes incorporating the filtered GPS measurements as covariate(s), and further test of HMM analysis of GPS data for earthquake forecasting.

Part I

HIDDEN MARKOV TYPE MODELS FOR EARTHQUAKES

Chapter 2

A General Class of Discrete-time Models

This chapter first introduces a class of general discrete-time hidden Markov models, and then reviews the concepts and parameter estimation issues for the three special discrete-time hidden Markov type models.

2.1 A General Formulation for Discrete-time Models

Consider a system which may be described at any time as being one of a set of r distinct states, $\{1, \cdots, r\}$. Denote the actual state at time t as S_t. The probability of the system in state j at time t given the predecessor states is

$$P(S_t = j \,|\, S_{t-1}, S_{t-2}, \cdots).$$

However, in reality, in most of the time this system is not observable, but instead, a set of observations $\mathbf{O}_t = \{O_{t1}, \cdots, O_{tn}\}$ can be produced at time t with the system in state $S_t \in \{1, \cdots, r\}$. Denote the observation sequence by $\mathbf{O} = (\mathbf{O}_1, \cdots, \mathbf{O}_T)$, and the hidden state sequence by $S = (S_1, \cdots, S_T)$. In practice, along with the observed process of the system, there might be another set of observed processes $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_T)$, which may not only strongly influence or directly reflect the hidden state, but also strongly associate with the observed sequence $\mathbf{O}$, where $\mathbf{X}_t = \{X_{t1}, \cdots, X_{tk}\}$ for $1 \leq t \leq T$. We refer to such processes as ancillary data or covariates. Here we should bear in mind that the process $\mathbf{O}$ is what we are interested in, and that the process $\mathbf{X}$ provides us ancillary information about $\mathbf{O}$. Then the probability of $\mathbf{O}$ being produced by the system

without assuming the knowledge of the state sequence in which it was generated is

$$\begin{aligned}
P(\mathbf{O} \mid \mathbf{X}, \theta) &= \sum_{\text{all } S} P(\mathbf{O}, S \mid \mathbf{X}, \theta) \\
&= \sum_{\text{all } S} P(\mathbf{O} \mid S, \mathbf{X}, \theta) P(S \mid \mathbf{X}, \theta) \\
&= \sum_{\text{all } S} P(\mathbf{O}_1, \cdots, \mathbf{O}_T \mid S_1, \cdots, S_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) P(S_1, \cdots, S_T \mid \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \\
&= \sum_{\text{all } S} P(\mathbf{O}_1 \mid S_1, \cdots, S_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) P(S_1 \mid \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \cdot \\
&\quad \prod_{t=2}^{T} P(\mathbf{O}_t \mid \mathbf{O}_1, \cdots, \mathbf{O}_{t-1}, S_1, \cdots, S_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \\
&\quad \times P(S_t \mid S_1, \cdots, S_{t-1}, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta), \qquad (2.1)
\end{aligned}$$

where θ is the set of parameters.

It is worth noting that the righthand side of the last step in (2.1) indicates that the model actually consists of two parts, the conditional probability of observations given the hidden state sequence and the covariates, and the transition probability given the covariates. This structure accounts for a variety of formulations including some well-known hidden Markov models. Discrete-time hidden Markov models (HMMs; see Rabiner, 1989 for a comprehensive tutorial), nonhomogeneous hidden Markov models (NHMMs; Hughes and Guttorp, 1994a) and Markov-modulated generalised linear models (MMGLMs; Harte, 2005) are all special cases of the model (2.1).

2.2 Models without Ancillary Data

These models are the ones described in Chapter 2 in MacDonald and Zucchini (1997). Let us assume that there is no ancillary variable available, and the observations at time t only depend on the current hidden state, i.e.

$$P(\mathbf{O}_t \mid \mathbf{O}_1, \cdots, \mathbf{O}_{t-1}, S_1, \cdots, S_T, \theta) = P(\mathbf{O}_t \mid S_t, \theta).$$

In addition, assume that the hidden state is a first-order Markov chain

$$P(S_t \mid S_1, \cdots, S_{t-1}, \theta) = P(S_t \mid S_{t-1}, \theta).$$

Then we have

$$P(\mathbf{O} \mid \theta) = \sum_{\text{all } S} P(\mathbf{O}_1 \mid S_1, \theta) P(S_1 \mid \theta) \prod_{t=2}^{T} P(\mathbf{O}_t \mid S_t, \theta) P(S_t \mid S_{t-1}, \theta),$$

which is an HMM.

The forward-backward procedure, a recursive algorithm, was developed to efficiently calculate the likelihood of a hidden Markov model. If the forward variable is defined as

$$\alpha_t(i) = P(\mathbf{O}_1, \cdots, \mathbf{O}_t, S_t = i \mid \theta), \tag{2.2}$$

and the backward variable is defined as

$$\beta_t(i) = P(\mathbf{O}_{t+1}, \cdots, \mathbf{O}_T \mid S_t = i, \theta), \tag{2.3}$$

the likelihood can then be calculated recursively by

$$L(\theta) = P(\mathbf{O}_1, \cdots, \mathbf{O}_T \mid \theta) = \sum_{i=1}^{r} \alpha_t(i) \beta_t(i). \tag{2.4}$$

Baum et al. (1970) developed an iterative algorithm to estimate the parameters for HMMs by considering the hidden states as missing data, which was shown to be equivalent to the expectation maximization (EM) algorithm of Dempster et al. (1977). The EM algorithm is to iteratively perform the expectation (E) step, which computes an expectation of the log likelihood with respect to the current estimate of the distribution for the missing variables, and the maximization (M) step, which computes the parameters that maximize the expected log likelihood found on the E step, until the log likelihood converges. To apply the EM algorithm to an HMM, the observations $\mathbf{O}$ and the hidden process S form the 'complete data'. The complete likelihood of an HMM is

$$\begin{aligned} \mathcal{L}^c(\theta; \mathbf{O}, S) &= \sum_{S_1, \cdots, S_T} P(\mathbf{O}_1, \cdots, \mathbf{O}_T, S_1, \cdots, S_T \mid \theta) \\ &= \sum_{S_1, \cdots, S_T} P(S_1 \mid \theta) P(\mathbf{O}_1 \mid S_1, \theta) \prod_{t=2}^{T} P(S_t \mid S_{t-1}, \theta) P(\mathbf{O}_t \mid S_t, \theta). \end{aligned} \tag{2.5}$$

If θ_0 is a given parameter estimate, then the E-step of the EM algorithm is calculating the expectation

$$\mathcal{Q}(\theta; \theta_0) = \mathbb{E}_{S, \theta_0}[\log \mathcal{L}^c(\theta; \mathbf{O}, S) \mid \mathbf{O}] \tag{2.6}$$

with respect to S and θ_0. The M-step is obtaining the new estimate

$$\widehat{\theta} = \arg\max_{\theta} \mathcal{Q}(\theta; \theta_0). \tag{2.7}$$

The expected complete log-likelihood $\mathcal{Q}(\theta; \theta_0)$ is obtained by taking the expectation of the complete likelihood over S, given the current parameter estimates θ_0, conditional on the observed data $\mathbf{O}$,

$$\begin{aligned}\mathcal{Q}(\theta; \theta_0) = &\sum_j P(S_1 = j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \theta_0) \log P(S_1 = j) \\ &+ \sum_{t=2}^{T} \sum_{i,j} P(S_{t-1} = i, S_t = j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \theta_0) \log P(S_t = j \,|\, S_{t-1} = i, \theta) \\ &+ \sum_{t=1}^{T} \sum_{j} P(S_t = j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \theta_0) \log P(\mathbf{O}_t \,|\, S_t = j, \theta). \end{aligned} \tag{2.8}$$

Then the E-step of the EM algorithm for the HMM is to compute

$$\begin{aligned} v_t(j) &= P(S_t = j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \theta_0), \\ w_t(i,j) &= P(S_{t-1} = i, S_t = j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \theta_0), \end{aligned}$$

and the M-step is to maximize the second and third terms in (2.8),

$$\Psi(\mathfrak{S} \,|\, \theta_0) = \sum_{t=2}^{T} \sum_{i,j} w_t(i,j) \log P(S_t = j \,|\, S_{t-1} = i, \mathfrak{S}) \tag{2.9}$$

and

$$\Psi(\mathfrak{O} \,|\, \theta_0) = \sum_{t=1}^{T} \sum_{j=1}^{r} v_t(j) \log P(\mathbf{O}_t \,|\, S_t = j, \mathfrak{O}) \tag{2.10}$$

as functions of $\theta = (\mathfrak{O}, \mathfrak{S})$, where $\mathfrak{O}$ represents the set of parameters involved in the observation distribution and $\mathfrak{S}$ is the set of parameters in the transition probability matrix.

For geophysical processes, it is common to have multivariate observations. Taking GPS measurements of deformation for example, there are three components of movement (north, east and up components) for each station. Under fairly general conditions the central limit theorem says that the sum of a large number of independent random variables has an approximately normal distribution. For this reason, the normal distribution is widely used in statistics, and in natural and social

sciences. Therefore the estimation procedure for an HMM with observations from a multivariate normal distribution will be presented here. Let $\mathbf{O}_t = \{O_{t1}, \cdots, O_{tn}\}$ represent the variable of measurement with observed values of $\mathbf{o}_t = \{o_{t1}, \cdots, o_{tn}\}$, where o_{ti} is the measurement of component i at time t. Suppose that given the current hidden state the measurement of each component is normally distributed and the measurements of each two components are independent. The process has the probability density function

$$f(\mathbf{o}_t \mid S_t = s) = \prod_{i=1}^{n} \frac{1}{\sigma_{is}\sqrt{2\pi}} \exp\left(-\frac{(o_{ti} - \mu_{is})^2}{2\sigma_{is}^2}\right) \tag{2.11}$$

where μ_{is} and σ_{is} are the mean and standard deviation for the observations of the ith component in state s. Assume that there are r states. Then the number of parameters to be estimated in this part of the model is $2nr$. Set the transition probability from state i to j as $P(S_t = j \mid S_{t-1} = i) = \gamma_{ij}$, and the initial distribution vector of the Markov chain as $\delta = (\delta_1, \cdots, \delta_r)$. The forward and backward probabilities, (2.2) and (2.3), are computed from the following equations,

$$\begin{aligned}
\alpha_1(i) &= P(\mathbf{O}_1, S_1 = i \mid \theta) = \delta_i f(\mathbf{o}_1 \mid S_1 = i), \\
\alpha_{t+1}(i) &= P(\mathbf{O}_1, \cdots, \mathbf{O}_{t+1}, S_t = i \mid \theta) \\
&= \sum_{k=1}^{r} \alpha_t(k)\gamma_{ki} f(\mathbf{o}_{t+1} \mid S_{t+1} = i),
\end{aligned}$$

and

$$\begin{aligned}
\beta_T(i) &= 1 \\
\beta_t(i) &= P(\mathbf{O}_{t+1}, \cdots, \mathbf{O}_T \mid S_t = i, \theta) \\
&= \sum_{j=1}^{r} \beta_{t+1}(j)\gamma_{ij} f(\mathbf{o}_{t+1} \mid S_{t+1} = j).
\end{aligned}$$

Given an initial guess of the parameter, θ_0, assuming that the initial distribution vector of the Markov chain is $\delta_0 = (\delta_{01}, \cdots, \delta_{0r})$, the forward and backward probabilities can be calculated using the above formula. The E-step will then be computing v and w given θ_0 via the following equations,

$$v_t(j) = P(S_t = j \mid \mathbf{O}_1, \cdots, \mathbf{O}_T, \theta_0) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{r} \alpha_t(i)\beta_t(i)},$$

and

$$\begin{aligned} w_t(i,j) &= P(S_{t-1}=i, S_t=j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \theta_0) \\ &= \frac{\alpha_{t-1}(i)\gamma_{ij} f(\mathbf{o}_t \,|\, S_t=j)\beta_t(j)}{\sum_{k=1}^{r} \alpha_t(k)\beta_t(k)}. \end{aligned}$$

The M-step is to maximize $\Psi(\mathfrak{S} \,|\, \theta_0)$ in (2.9) with respect to the parameters in the transition probability matrix and $\Psi(\mathfrak{O} \,|\, \theta_0)$ in (2.10) with respect to the parameters involved in the observation distribution. By maximizing

$$\begin{aligned} \Psi(\mathfrak{S} \,|\, \theta_0) &= \sum_{t=2}^{T} \sum_{i,j} w_t(i,j) \log P(S_t=j \,|\, S_{t-1}=i, \mathfrak{S}) \\ &= \sum_{t=2}^{T} \sum_{i,j} w_t(i,j) \log \gamma_{ij} \end{aligned}$$

we can get the estimates for the transition probabilities as

$$\widehat{\gamma}_{ij} = \frac{\sum_{t=2}^{T} w_t(i,j)}{\sum_{j=1}^{r} \sum_{t=2}^{T} w_t(i,j)}.$$

Then maximizing

$$\begin{aligned} \Psi(\mathfrak{O} \,|\, \theta_0) &= \sum_{t=1}^{T} \sum_{j=1}^{r} v_t(j) \log P(\mathbf{O}_t \,|\, S_t=j, \mathfrak{O}) \\ &= \sum_{t=1}^{T} \sum_{j=1}^{r} v_t(j) \log \left\{ \prod_{k=1}^{n} \frac{1}{\sigma_{kj}\sqrt{2\pi}} \exp\left(-\frac{(o_{tk}-\mu_{kj})^2}{2\sigma_{kj}^2} \right) \right\} \end{aligned}$$

has a simple solution given by

$$\widehat{\mu}_{kj} = \frac{\sum_{t=1}^{T} v_t(j) o_{tk}}{\sum_{t=1}^{T} v_t(j)}, \tag{2.12}$$

and

$$\widehat{\sigma}_{kj} = \sqrt{\frac{\sum_{t=1}^{T} v_t(j)(o_{tk}-\widehat{\mu}_{kj})^2}{\sum_{t=1}^{T} v_t(j)}}. \tag{2.13}$$

This forms one step of EM algorithm. To get the parameter estimates using this algorithm, we update the initial guess of the parameter θ_0 using the estimated parameter from the M-step and

then iterate the E-step and M-step until the log likelihood converges, or in other words, until the difference between the log likelihoods from the previous step and from the current step is less than a specified tolerance.

2.3 Transition Probabilities Depending on Ancillary Data

When only the transition probabilities change over time depending on the ancillary data, the model is called a nonhomogeneous hidden Markov model (Hughes and Guttorp, 1994a). Assume that the process of interest only depends on the current hidden state and is conditionally independent of all previous observations, i.e.

$$P(\mathbf{O}_t \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_{t-1}, S_1, \cdots, S_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = P(\mathbf{O}_t \,|\, S_t, \theta), \tag{2.14}$$

and that the ancillary data only indirectly affect the processes of interest through the hidden state. In addition, we assume that the hidden state is a nonhomogeneous first-order Markov chain,

$$P(S_t \,|\, S_1, \cdots, S_{t-1}, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = P(S_t \,|\, S_{t-1}, \mathbf{X}_t, \theta). \tag{2.15}$$

Note that the transition probability depends on the covariates $\mathbf{X}_t$. Model (2.1) becomes an NHMM.

Increasing interest has arisen in this model after the successful modelling of local and regional precipitation by incorporating synoptic atmospheric data via an NHMM framework (Hughes and Guttorp, 1994a). The model was first introduced to relate large-scale atmospheric data (covariates) to local hydrologic phenomena such as precipitation (observed process) (Hughes, 1993). In this class of models, the observed process is supposed to be related to a set of covariates, but the covariates only indirectly affect the observed process through another unobserved process (i.e. the hidden states). The unobserved states serve to automatically categorize the covariates into a few classes that are most associated with particular patterns of the observations of interest. The idea is to introduce an indirect covariate effect to the hidden Markov models. The observed process is assumed to be conditionally independent given the hidden state. However, the transition intensities of the hidden Markov process are allowed to depend on covariates.

Although the models were proposed to model precipitation, the general NHMM structure is potentially applicable to other settings. In application, in order to fit an NHMM to the data of interest, both the assumptions (2.14) and (2.15) need to be parameterized. Assumption (2.14) usually takes

the form of a conditional probability mass function (for discrete random variables) or a conditional probability density function (for continuous random variables) given the current state. It can be parameterized according to the empirical distribution of the observations. If the observed data are from a Gaussian population, one can consider Model (2.11) for this assumption. If a binary data set is under consideration, for example, earthquake occurrence processes in one or n independent regions, one can use the independence model (see also Hughes and Guttorp, 1994a),

$$P(\mathbf{O}_t = \mathbf{o}_t \,|\, S_t = s) = \prod_{i=1}^{n} \widetilde{p}_{is}^{\,o_{ti}} (1 - \widetilde{p}_{is})^{1-o_{ti}} \tag{2.16}$$

where $\widetilde{p}_{is}$ is the probability of event occurrence in region i in hidden state s, $o_{ti} = 1$ if the event occurs on day t at region i and 0 otherwise.

When the transition probability matrix depends on the covariates, the assumption (2.15) can be parameterized via some link functions which should ensure that $P(S_t \,|\, S_{t-1}, \mathbf{X}_t)$ is between 0 and 1, and $\sum_i P(S_t = i \,|\, S_{t-1}, \mathbf{X}_t) = 1$. For example, one can use a parameterization which makes use of the Bayes formula (Hughes and Guttorp, 1994a). Let $\gamma_{ij} = P(S_t = j \,|\, S_{t-1} = i)$ with constraint $\sum_j \gamma_{ij} = 1$, and

$$P(\mathbf{X}_t \,|\, S_{t-1}, S_t) \propto \exp(-(\mathbf{X}_t - \mu_{S_{t-1},S_t})\Sigma^{-1}(\mathbf{X}_t - \mu_{S_{t-1},S_t})'/2),$$

where μ_{S_{t-1},S_t} is the mean vector (assumed to vary according to the current and past states), and Σ is the variance-covariance matrix of the covariates. Assume that $\sum_j \mu_{ij} = \mu_i = 0$. Applying Bayes formula it then follows

$$\begin{aligned} P(S_t \,|\, S_{t-1}, \mathbf{X}_t) &= \frac{P(\mathbf{X}_t \,|\, S_{t-1}, S_t) P(S_t \,|\, S_{t-1})}{\sum_{S_t} P(\mathbf{X}_t \,|\, S_{t-1}, S_t) P(S_t \,|\, S_{t-1})} \\ &\propto \gamma_{S_{t-1},S_t} \exp(-(\mathbf{X}_t - \mu_{S_{t-1},S_t})\Sigma^{-1}(\mathbf{X}_t - \mu_{S_{t-1},S_t})'/2). \end{aligned} \tag{2.17}$$

The probabilities γ_{ij} serves as the baseline transition probabilities of the hidden state process. The exponential term characterizes the effect of the covariates on the baseline transition probabilities. This model can be applied to earthquake modelling incorporating GPS measurements of deformation as ancillary data. Normally, this model is easy to handle in terms of parameter estimation and explanation of hidden states.

Having outlined the formulation of the model, the next question is how to estimate the parameters from the observed data. If we let θ denote the model parameters, the likelihood function is

then

$$\begin{aligned} L(\theta) &= P(\mathbf{O}_1, \cdots, \mathbf{O}_T \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \\ &= \sum_{S_1, \cdots, S_T} P(\mathbf{O}_1, \cdots, \mathbf{O}_T, S_1, \cdots, S_T \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta). \end{aligned} \tag{2.18}$$

The likelihood function appears to be computationally intractable. The forward-backward procedure developed to solve the HMM can be extended to the NHMM and makes the calculation possible. Again, assume that the initial distribution vector of the Markov chain is $\delta = (\delta_1, \cdots, \delta_r)$. The forward variable is defined as

$$\begin{aligned} \alpha_1(i) &= P(\mathbf{O}_1, S_1 = i \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = \delta_i P(\mathbf{O}_1 = \mathbf{o}_1 \,|\, S_1 = i, \theta), \\ \alpha_{t+1}(i) &= P(\mathbf{O}_1, \cdots, \mathbf{O}_{t+1}, S_{t+1} = i \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \\ &= \sum_{k=1}^{r} \alpha_t(k) P(S_{t+1} = i \,|\, S_t = k, \mathbf{X}_{t+1}) P(\mathbf{O}_{t+1} = \mathbf{o}_{t+1} \,|\, S_{t+1} = i, \theta), \end{aligned}$$

it then follows that

$$L(\theta) = \sum_{i=1}^{r} P(\mathbf{O}_1, \cdots, \mathbf{O}_T, S_T = i \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = \sum_{i=1}^{r} \alpha_T(i). \tag{2.19}$$

Define the backward variable as

$$\begin{aligned} \beta_T(i) &= 1 \\ \beta_t(i) &= P(\mathbf{O}_{t+1}, \cdots, \mathbf{O}_T \,|\, S_t = i, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \\ &= \sum_{j=1}^{r} \beta_{t+1}(j) P(S_{t+1} = j \,|\, S_t = i, \mathbf{X}_{t+1}) P(\mathbf{O}_{t+1} = \mathbf{o}_{t+1} \,|\, S_{t+1} = j, \theta). \end{aligned}$$

The likelihood function can then be obtained by

$$L(\theta) = \sum_{i=1}^{r} P(\mathbf{O}_1, \cdots, \mathbf{O}_T, S_1 = i \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = \sum_{i=1}^{r} \beta_1(i) P(\mathbf{O}_1 \,|\, S_1 = i) \pi_i(x_1), \tag{2.20}$$

where $\pi_i(x_1) = P(S_1 = i \,|\, \mathbf{X}_1, \theta)$. Alternatively, the likelihood can be calculated recursively by

$$L(\theta) = \sum_{i=1}^{r} \alpha_t(i) \beta_t(i). \tag{2.21}$$

Note that an NHMM has the same structure as a standard HMM, both consist of two parts, an observation probability distribution and a transition probability distribution. The difference is that the transition probability distribution of an NHMM depends on an additional variable. This allows us to borrow the idea of the EM approach for the standard HMM to estimate the parameters involved in an NHMM. To apply the EM algorithm here, the observations $\mathbf{O}$, the hidden process S, together with the ancillary data $\mathbf{X}$ form the 'complete data'. The complete likelihood is

$$\begin{aligned}\mathcal{L}^c(\theta;\mathbf{O},S,\mathbf{X}) &= \sum_{S_1,\cdots,S_T} P(\mathbf{O}_1,\cdots,\mathbf{O}_T,S_1,\cdots,S_T \,|\, \mathbf{X}_1,\cdots,\mathbf{X}_T,\theta) \\ &= \sum_{S_1,\cdots,S_T} P(S_1 \,|\, \mathbf{X}_1,\theta)P(\mathbf{O}_1 \,|\, S_1,\theta)\prod_{t=2}^{T} P(S_t \,|\, S_{t-1},\mathbf{X}_t,\theta)P(\mathbf{O}_t \,|\, S_t,\theta).\end{aligned} \tag{2.22}$$

If θ_0 is a given parameter estimate, then the E-step of the EM algorithm is

$$\mathcal{Q}(\theta;\theta_0) = \mathbb{E}_{S,\theta_0}[\log \mathcal{L}^c(\theta;\mathbf{O},S,\mathbf{X}) \,|\, \mathbf{O},\mathbf{X}]. \tag{2.23}$$

By comparing to Equation (2.6), the distinguishing point is that the ancillary data $\mathbf{X}$ are affecting the complete likelihood through the transition probabilities in Equation (2.23). The M-step is obtaining the new estimate

$$\widehat{\theta} = \arg\max_{\theta} \mathcal{Q}(\theta;\theta_0). \tag{2.24}$$

The expected complete log-likelihood $\mathcal{Q}(\theta;\theta_0)$ is obtained by taking the expectation of the complete likelihood over S at the current parameter estimation θ_0, conditional on the observed data $\mathbf{O}$ and $\mathbf{X}$, which is

$$\begin{aligned}\mathcal{Q}(\theta;\theta_0) =& \sum_j P(S_1 = j \,|\, \mathbf{O}_1,\cdots,\mathbf{O}_T,\mathbf{X}_1,\cdots,\mathbf{X}_T,\theta_0)\log P(S_1 = j \,|\, \mathbf{X}_1) \\ &+ \sum_{t=2}^{T}\sum_{i,j} P(S_{t-1}=i, S_t = j \,|\, \mathbf{O}_1,\cdots,\mathbf{O}_T,\mathbf{X}_1,\cdots,\mathbf{X}_T,\theta_0) \\ &\quad \times \log P(S_t = j \,|\, S_{t-1}=i,\mathbf{X}_t,\theta) \\ &+ \sum_{t=1}^{T}\sum_{j} P(S_t = j \,|\, \mathbf{O}_1,\cdots,\mathbf{O}_T,\mathbf{X}_1,\cdots,\mathbf{X}_T,\theta_0)\log P(\mathbf{O}_t \,|\, S_t = j,\theta).\end{aligned} \tag{2.25}$$

Then the E-step of the EM algorithm for the NHMM is to compute

$$\begin{aligned} v_t(j) &= P(S_t = j \mid \mathbf{O}_1, \cdots, \mathbf{O}_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta_0) \\ &= \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{r} \alpha_t(j)\beta_t(j)} \\ w_t(i,j) &= P(S_{t-1} = i, S_t = j \mid \mathbf{O}_1, \cdots, \mathbf{O}_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta_0) \\ &= \frac{\alpha_{t-1}(i)P(S_t = j \mid S_{t-1} = i, \mathbf{X}_t = \mathbf{x}_t)P(\mathbf{O}_t \mid S_t = j)\beta_t(j)}{\sum_{i=1}^{r}\sum_{j=1}^{r} \alpha_{t-1}(i)P(S_t = j \mid S_{t-1} = i, \mathbf{X}_t = \mathbf{x}_t)P(\mathbf{O}_t \mid S_t = j)\beta_t(j)}. \end{aligned}$$

and the M-step is to maximize

$$\begin{aligned} \Psi(\mathfrak{O} \mid \theta_0) &= \sum_{t=1}^{T}\sum_{j} v_t(j) \log P(\mathbf{O}_t \mid S_t = j, \mathfrak{O}) \\ \Psi(\mathfrak{S} \mid \theta_0) &= \sum_{t=2}^{T}\sum_{i,j} w_t(i,j) \log P(S_t = j \mid S_{t-1} = i, \mathbf{X}_t, \mathfrak{S}) \end{aligned}$$

as functions of $\theta = (\mathfrak{O}, \mathfrak{S})$.

The difficult part of the parameter estimation for NHMM is in the M-step. The difference between the complete likelihoods of the HMM and NHMM derives mainly from the covariates. This makes it difficult to obtain an explicit solution for the parameter estimation involved in the transition probability matrix. In application, the maximization of $\Psi(\mathfrak{S} \mid \theta_0)$ always requires numerical optimization. For the maximization of $\Psi(\mathfrak{O} \mid \theta_0)$, one normally can get an explicit solution. Take Model (2.16) for example, maximizing $\Psi(\mathfrak{O} \mid \theta_0)$ has a simple solution given by

$$\widehat{\widehat{p}}_{is} = \frac{\sum_t v_t(s) o_{ti}}{\sum_t v_t(s)}.$$

However, when more complicated models are considered for the observations so that there is no explicit solution anymore, one might consider an alternative method of parameter estimation for the observation distribution part (cf., Hughes and Guttorp, 1994b).

2.4 Observation Mean Depending on Ancillary Data

When the mean of the observations depends on the ancillary data, the model is called a Markov-modulated generalized linear model (Harte, 2005). Assume that the ancillary data only have influ-

ence on the process of interest, i.e.

$$P(\mathbf{O}_t \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_{t-1}, S_1, \cdots, S_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = P(\mathbf{O}_t \,|\, S_t, \mathbf{X}_t, \theta),$$

and do not affect the hidden state. In addition, assume that the hidden state is a homogeneous first-order Markov chain, i.e.

$$P(S_t \,|\, S_1, \cdots, S_{t-1}, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = P(S_t \,|\, S_{t-1}, \theta).$$

If the ancillary data influence the observed processes through a generalized linear model (see McCullagh and Nelder, 1989 for a comprehensive discussion of the generalized linear models), the model (2.1) then becomes an MMGLM.

The MMGLM assumes that the distribution of the observed response variable $\mathbf{O}_t$ is indirectly dependent on the current state of the Markov chain through another observed covariate $\mathbf{X}_t$. It assumes that

$$\mathbf{E}(\mathbf{O}_t) = f\left(\zeta_0 + \sum_{i=1}^{k} \zeta_i X_{ti}\right),$$

where the values of ζ_0 and ζ_i vary according to the hidden Markov state, and $f(\cdot)$ is the inverse mapping of the link function commonly used in generalized linear models. Moreover, the responses are assumed to be conditionally independent given the value of the Markov chain. To make a clearer statement, take the density function of the response variable from the exponential family for example. Assume that the response variable $\mathbf{O}$ has a distribution in the exponential family, taking the form

$$P(\mathbf{O}_t = \mathbf{o}_t \,|\, S_t = s, \mathbf{X}_t, \xi, \phi) = \exp\{(\mathbf{o}_t \xi_{ts} - b(\xi_{ts}))/a(\phi) + c(\mathbf{o}_t, \phi)\}$$

for some specific functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$, where $\xi_{ts} = \zeta_{0s} + \sum_{i=1}^{k} \zeta_{is} x_{ti}$ is the so called canonical parameter which depends on the current hidden state s. For some very commonly used distributions in the exponential family, such as normal, binomial and Poisson distributions, we can write this state-dependent distribution for the observations as, for the normal distribution $N(\mu, \sigma^2)$,

$$\begin{aligned} P(\mathbf{O}_t = \mathbf{o}_t \,|\, S_t = s, \mathbf{X}_t, \xi, \phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\mathbf{o}_t - \mu_{ts})^2}{2\sigma^2}\right\} \\ &= \exp\{(\mathbf{o}_t \mu_{ts} - \mu_{ts}^2/2)/\sigma^2 - (\mathbf{o}_t^2/\sigma^2 + \log(2\pi\sigma^2))/2\} \end{aligned}$$

with $\xi_{ts} = \mu_{ts} = \zeta_{0s} + \sum_{i=1}^{k} \zeta_{is} x_{ti}$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\xi_{ts}) = \xi_{ts}^2/2$ and $c(\mathbf{o}_t, \phi) = -(\mathbf{o}_t^2/\sigma^2 + \log(2\pi\sigma^2))/2\}$; for the binomial distribution $B(m, \mu)$,

$$P(\mathbf{O}_t = \mathbf{o}_t \,|\, S_t = s, \mathbf{X}_t, \xi, \phi) = \exp\left\{\left(\mathbf{o}_t \log\left(\frac{\mu_{ts}}{1-\mu_{ts}}\right) - m\log\left(\frac{1}{1-\mu_{ts}}\right)\right) + \log\left(C_{\mathbf{o}_t}^m\right)\right\} \tag{2.26}$$

with $\xi_{ts} = \log\left(\mu_{ts}/(1-\mu_{ts})\right) = \zeta_{0s} + \sum_{i=1}^{k} \zeta_{is} x_{ti}$, $\phi = 1$, $a(\phi) = \phi$, $b(\xi_{ts}) = m\log(1+e^{\xi_{ts}})$ and $c(\mathbf{o}_t, \phi) = \log\left(C_{\mathbf{o}_t}^m\right)$, where C_k^m means the combination of k items from m; and for the Poisson distribution $P(\mu)$,

$$P(\mathbf{O}_t = \mathbf{o}_t \,|\, S_t = s, \mathbf{X}_t, \xi, \phi) = \exp\{(\mathbf{o}_t \log(\mu_{ts}) - \mu_{ts}) - \log(\mathbf{o}_t)\}$$

with $\xi_{ts} = \log(\mu_{ts}) = \zeta_{0s} + \sum_{i=1}^{k} \zeta_{is} x_{ti}$, $\phi = 1$, $a(\phi) = \phi$, $b(\xi_{ts}) = \exp(\xi_{ts})$ and $c(\mathbf{o}_t, \phi) = -\log(\mathbf{o}_t)$.

Set the transition probability from state i to j as $P(S_t = j \,|\, S_{t-1} = i) = \gamma_{ij}$, and the initial distribution vector of the Markov chain as $\delta = (\delta_1, \cdots, \delta_r)$. Let θ denote the model parameters. Similar to a discrete-time HMM, the likelihood function of an MMGLM

$$L(\theta) = P(\mathbf{O}_1, \cdots, \mathbf{O}_T \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \tag{2.27}$$

can be efficiently calculated from the forward-backward procedure. Define the forward probabilities as

$$\begin{aligned}
\alpha_1(i) &= P(\mathbf{O}_1, S_1 = i \,|\, \mathbf{X}_1, \theta) = \delta_i P(\mathbf{O}_1 = \mathbf{o}_1 \,|\, S_1 = i, \mathbf{X}_1, \theta), \\
\alpha_{t+1}(i) &= P(\mathbf{O}_1, \cdots, \mathbf{O}_{t+1}, S_t = i \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \\
&= \sum_{k=1}^{r} \alpha_t(k)\gamma_{ki} P(\mathbf{O}_{t+1} = \mathbf{o}_{t+1} \,|\, S_{t+1} = i, \mathbf{X}_{t+1}, \theta),
\end{aligned}$$

and the backward probabilities as

$$\begin{aligned}
\beta_T(i) &= P(\cdot \,|\, S_T = i, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = 1 \\
\beta_t(i) &= P(\mathbf{O}_{t+1}, \cdots, \mathbf{O}_T \,|\, S_t = i, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \\
&= \sum_{j=1}^{r} \beta_{t+1}(j)\gamma_{ij} P(\mathbf{O}_{t+1} = \mathbf{o}_{t+1} \,|\, S_{t+1} = j, \mathbf{X}_{t+1}, \theta).
\end{aligned}$$

The likelihood can then be calculated recursively by

$$L(\theta) = P(\mathbf{O}_1, \cdots, \mathbf{O}_T \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) = \sum_{i=1}^{r} \alpha_t(i)\beta_t(i). \tag{2.28}$$

According the structure of this model, the EM algorithm can also be applied for parameter estimation. The observations $\mathbf{O}$, $\mathbf{X}$ and the hidden process S form the 'complete data'. The complete likelihood of an MMGLM is

$$\begin{aligned}\mathcal{L}^c(\theta; \mathbf{O}, S, \mathbf{X}) &= \sum_{S_1,\cdots,S_T} P(\mathbf{O}_1, \cdots, \mathbf{O}_T, S_1, \cdots, S_T \,|\, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta) \\ &= \sum_{S_1,\cdots,S_T} P(S_1 \,|\, \theta) P(\mathbf{O}_1 \,|\, S_1, \mathbf{X}_1, \theta) \prod_{t=2}^{T} P(S_t \,|\, S_{t-1}, \theta) P(\mathbf{O}_t \,|\, S_t, \mathbf{X}_t, \theta).\end{aligned} \tag{2.29}$$

If θ_0 is an initial value of the parameter estimate, then the E-step of the EM algorithm is

$$\mathcal{Q}(\theta; \theta_0) = \mathbb{E}_{S,\theta_0}[\log \mathcal{L}^c(\theta; \mathbf{O}, S, \mathbf{X}) \,|\, \mathbf{O}, \mathbf{X}]. \tag{2.30}$$

The difference of Equation (2.30) from Equation (2.23) is that the ancillary data $\mathbf{X}$ are driving the observations in the former. The M-step is obtaining the new estimate

$$\widehat{\theta} = \arg\max_{\theta} \mathcal{Q}(\theta; \theta_0).$$

The expected complete log-likelihood $\mathcal{Q}(\theta; \theta_0)$ is obtained by taking the expectation of the complete likelihood over S at the current parameter estimation θ_0, conditional on the observed data $\mathbf{O}$ and $\mathbf{X}$,

$$\begin{aligned}&\mathcal{Q}(\theta; \theta_0) \\ &= \sum_j P(S_1 = j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta_0) \log P(S_1 = j) \\ &\quad + \sum_{t=2}^{T} \sum_{i,j} P(S_{t-1} = i, S_t = j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta_0) \log P(S_t = j \,|\, S_{t-1} = i, \theta) \\ &\quad + \sum_{t=1}^{T} \sum_{j} P(S_t = j \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta_0) \log P(\mathbf{O}_t \,|\, S_t = j, \mathbf{X}_t, \theta).\end{aligned} \tag{2.31}$$

Similar to the procedures in the previous sections, the E-step of the EM algorithm for the MMGLM

is to compute

$$\begin{aligned} v_t(j) &= P(S_t = j \mid \mathbf{O}_1, \cdots, \mathbf{O}_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta_0) \\ &= \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{r} \alpha_t(j)\beta_t(j)} \\ w_t(i,j) &= P(S_{t-1} = i, S_t = j \mid \mathbf{O}_1, \cdots, \mathbf{O}_T, \mathbf{X}_1, \cdots, \mathbf{X}_T, \theta_0) \\ &= \frac{\alpha_{t-1}(i)\gamma_{ij} P(\mathbf{O}_t \mid S_t = j, \mathbf{X}_t = \mathbf{x}_t)\beta_t(j)}{\sum_{i=1}^{r}\sum_{j=1}^{r} \alpha_{t-1}(i)\gamma_{ij} P(\mathbf{O}_t \mid S_t = j, \mathbf{X}_t = \mathbf{x}_t)\beta_t(j)}. \end{aligned}$$

and the M-step is to maximize

$$\begin{aligned} \Psi(\mathfrak{O} \mid \theta_0) &= \sum_{t=1}^{T} \sum_{j} v_t(j) \log P(\mathbf{O}_t \mid S_t = j, \mathbf{X}_t, \mathfrak{O}) \\ \Psi(\mathfrak{S} \mid \theta_0) &= \sum_{t=2}^{T} \sum_{i,j} w_t(i,j) \log P(S_t = j \mid S_{t-1} = i, \mathfrak{S}) \end{aligned}$$

as functions of $\theta = (\mathfrak{O}, \mathfrak{S})$. The estimation of the parameters in the transition probability matrix is the same as that for discrete-time HMM. The parameter estimation involved in the observation distribution part, $\Psi(\mathfrak{O} \mid \theta_0)$, is simply maximizing the sum of the individual log-likelihood contributions of the response variable weighted by the Markov state probabilities $v_t(j)$. This can be carried out in R using the "HiddenMarkov" package (Harte, 2005).

Chapter 3

Markov-modulated Hawkes Process with Stepwise Decay

3.1 Introduction

As mentioned in Chapter 1, when a Markov-modulated Poisson process (MMPP) is in one specific state, say s, the event occurrences follow a Poisson process with rate λ_s. For a detailed review, see Fischer and Meier-Hellstern (1993), Rydén (1994), and Rydén (1996). This process has been applied to model bursty point processes, especially in telecommunications (see, e.g., Heffes and Lucantoni, 1986). Yet, due to the fact that it has a constant event occurrence rate in any particular state, application of MMPPs to problems involving real data is limited.

Bebbington (2007) investigated volcanic inter-onset times of flank eruptions during 1600–2006 from Mount Etna using HMMs and MMPPs, where the volcano is demonstrated to have longish periods of Poissonian behavior. For a geophysical process, there usually exists some additional information apart from the arrival times. For example, an earthquake series often has both spatial and temporal locations, as well as magnitude. Lv (2009) extended the MMPP to a Markov-modulated Poisson process with state-dependent marks. He introduced additional information to the MMPP in the form of marks. It provides the potential to analyze spatial-temporal point patterns or multivariate processes. The ground process arrives at a stationary rate.

However, many geophysical processes often occur in a non-independent or self-exciting way. The events already occurred often trigger new ones. As pointed out in Section 1.1, the observable point process which is governed by the hidden process tends to cluster with certain period. An assumption of independence of the observations in a particular state is no longer suitable here. The

framework should be extended to the case in which the hidden state is controlled by a continuous-time Markov chain, and when sojourning in each state, the observed process follows a self-exciting point process.

Based on the above considerations, this chapter develops a Markov-modulated Hawkes process with stepwise decay, a model in which the process switches among a finite number of states with a self-exciting occurrence rate of events from a Hawkes process. A method of estimating the parameters involved in the model using the EM algorithm is then introduced. The detailed implementation algorithm for the parameter estimation is also provided. After that, how to evaluate the goodness-of-fit problem is discussed. A simulation algorithm for this process is also provided so that the simulated events can be used to examine how the EM algorithm works for the parameter estimation of this model.

3.2 Markov-modulated Hawkes Process with Stepwise Decay

3.2.1 Hawkes Process with Stepwise Decay

The Hawkes process has an extensive application in seismology (see, e.g., Hawkes and Adamopoulos, 1973), epidemiology, neurophysiology (see, e.g., Brémaud and Massoulié, 1996), and econometrics (see, e.g., Bowsher, 2007). It is a point-process analogue of the autoregressive model in time series. Let N be a simple point process on $\mathbb{R}$ with successive occurrence times $T_0 = t_0 = 0, T_1 = t_1, \cdots, T_n = t_n$. The history $\mathcal{H}_t$ of the process at time t is defined by $\mathcal{H}_t = \{t_l : t_l < t\}$. The Hawkes process was first introduced as a self-exciting point process with conditional intensity function taking the form

$$\lambda(t) \stackrel{\triangle}{=} \lambda(t \,|\, \mathcal{H}_t) = \lambda + \int_{-\infty}^{t} g(t-u)\mathrm{d}N(u),$$

where $\lambda \geq 0$, $g(u) \geq 0$ and $\int_0^\infty g(u)\mathrm{d}u < 1$ (Hawkes, 1971). The process consists of two types of events, immigrants without extant parents in the process and offspring generated by existing points. The immigrants arrive in a Poisson process of rate λ, and each immigrant generates offspring in a non-stationary Poisson stream of rate $g(t-u)$. If we let $g(t-u) = \nu\eta e^{-\eta(t-u)}$, the process will then have the Markovian property (Daley and Vere-Jones, 2003) and become easy to handle mathematically. This model will be imbedded into the hidden Markov model framework with further restriction of the intensity function to stepwise decay. The purpose of this restriction is purely for

the convenience of computation.

For a point process N with occurrence times $t_0 = 0$, $t_1, \cdots, t_n$, the intensity of this process changes after each event occurs, but stays as a constant between each two consecutive events, i.e. for $t_k < t < t_{k+1}$, $\lambda^*(t) = \lambda^*(t_{k+1})$, and

$$\lambda^*(t_{k+1}) = \lambda + \nu\eta \sum_{t_j < t_k} e^{-\eta(t_k - t_j)}.$$

The general conditional intensity function for the process can be written as

$$\lambda^*(t) = \lambda + \nu\eta \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\eta(\max\{t_l : t_l < t\} - t_j)}.$$

The process consists of a series of immigrants which arrive according to a Poisson process at a constant rate λ. Any immigrant at the point t_i generates descendants in $t > t_{i+1}$ according to a stepwise rate $\nu\eta e^{-\eta(\max\{t_l : t_l < t\} - t_i)}$. Note that the only difference of this process from a Hawkes process with exponential decay is how the immigrants produce offspring. In a Hawkes process with exponential decay, an immigrant at the point t_i generates descendants in $t > t_i$ in a non-stationary Poisson stream of rate $\nu\eta e^{-\eta(t - t_i)}$ (see Hawkes and Oakes, 1974; Oakes, 1975; Daley and Vere-Jones, 2003). This new process we will call a Hawkes process with stepwise decay. Denote the inter-event times of this process by $X_1 = T_1 - T_0, \cdots, X_n = T_n - T_{n-1}$. The inter-event time distribution for a point process provides in principle a means of predicting future events in terms of the history of the process (see Jowett and Vere-Jones, 1972). The inter-event time distribution of this process is

$$\begin{aligned} P(X_n \le x_n \,|\, \mathcal{H}_{t_{n-1}}) &= 1 - \exp\left\{ - \int_{t_{n-1}}^{t_{n-1}+x_n} \left(\lambda + \nu\eta \sum_{t_j < t_{n-1}} e^{-\eta(t_{n-1} - t_j)} \right) \mathrm{d}t \right\} \\ &= 1 - \exp\left\{ - \left(\lambda + \nu\eta \sum_{t_j < t_{n-1}} e^{-\eta(t_{n-1} - t_j)} \right) x_n \right\} \\ &= 1 - \exp\left\{ -\lambda^*(t_n) x_n \right\}. \end{aligned}$$

3.2.2 Markov-modulated Hawkes Process with Stepwise Decay

Consider a Hawkes process with stepwise decay the parameters of which vary according to an r-state irreducible Markov process. Denote the infinitesimal generator of the underlying Markov

process $\{Y(t)\}$ by $Q = (q_{ij})_{r\times r}$ and $q_i = -q_{ii}$, $i = 1, \cdots, r$. The observed Hawkes process with stepwise decay is characterized by $\Lambda^*(t)$, an $r \times r$ diagonal matrix with diagonal elements $\lambda_1^*(t), \cdots, \lambda_r^*(t)$, where

$$\lambda_i^*(t) = \lambda_i + \nu_i\eta_i \sum_{t_j<\max\{t_l:t_l<t\}} e^{-\eta_i(\max\{t_l:t_l<t\}-t_j)}, \tag{3.1}$$

t_k is the occurrence time of the kth event, and λ_i, ν_i and η_i, $i = 1, \cdots, r$, are parameters. This process is said to be in state i, $1 \le i \le r$, when the underlying Markov process is in state i. When this process is in state i at time t, events occur according to a Hawkes process with stepwise decay rate $\lambda_i^*(t)$ as defined in (3.1). This process is called a Markov-modulated Hawkes process with stepwise decay (MMHPSD) of order r. Note that in this process, for the ancestors in the history, the influencing factors, ν and η, by which they generate descendants, are ν_i and η_i, where i is the present state, regardless of which state the ancestor was generated in. This is for the sake of mathematical tractability. Let Y_0 be the state of the Markov process having generator Q at time $t = 0$, and let $X_0 = 0$. The kth event of the MMHPSD is associated with the corresponding state Y_k of the underlying Markov process as well as the time X_k, $k \ge 1$, between the $(k-1)$st and the kth event.

The state of the underlying Markov process at time t is $Y(t)$. Let $N_n(u)$ denote the number of events in the MMHPSD in $(t_{n-1}, t_{n-1}+u]$, where t_n denotes the nth event time. Given the history, the transition probability without arrival is

$$H_{ij}^{(n)}(u) = P\{Y(t_{n-1}+u) = j, N_n(u) = 0 \,|\, Y(t_{n-1}) = i, \mathcal{H}_{t_{n-1}}\}.$$

It follows that

$$\lim_{\Delta u\to 0} \frac{H_{ij}^{(n)}(\Delta u)}{\Delta u} = q_{ij}.$$

Then for $1 \le i, j \le r$ and for $\Delta u > 0$ we have

$$\begin{aligned}
&H_{ij}^{(n)}(u+\Delta u) \\
&\quad = P\left\{Y(t_{n-1}+u+\Delta u) = j, N_n(u+\Delta u) = 0 \,|\, Y(t_{n-1}) = i, \mathcal{H}_{t_{n-1}}\right\} \\
&\quad = \sum_{k=1}^{r} P\{Y(t_{n-1}+u+\Delta u) = j, N_n(u+\Delta u) = 0, Y(t_{n-1}+u) = k \,| \\
&\qquad\qquad Y(t_{n-1}) = i, \mathcal{H}_{t_{n-1}}\}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{r} P\{Y(t_{n-1}+u+\Delta u)=j, N_n(u)=0, N((t_{n-1}+u, t_{n-1}+u+\Delta u])=0, \\
&\qquad Y(t_{n-1}+u)=k \,|\, Y(t_{n-1})=i, \mathcal{H}_{t_{n-1}}\} \\
&= \sum_{k=1}^{r} P\{Y(t_{n-1}+u+\Delta u)=j, N((t_{n-1}+u, t_{n-1}+u+\Delta u])=0 \,| \\
&\qquad N_n(u)=0, Y(t_{n-1}+u)=k, Y(t_{n-1})=i, \mathcal{H}_{t_{n-1}}\} \\
&\quad \times P\{Y(t_{n-1}+u)=k, N_n(u)=0 \,|\, Y(t_{n-1})=i, \mathcal{H}_{t_{n-1}}\} \\
&= \sum_{k=1}^{r} P\{Y(t_{n-1}+\Delta u)=j, N_n(\Delta u)=0 \,|\, Y(t_{n-1})=k, \mathcal{H}_{t_{n-1}}\} H_{ik}^{(n)}(u) \\
&= \sum_{k=1}^{r} H_{kj}^{(n)}(\Delta u) H_{ik}^{(n)}(u).
\end{aligned}
$$

Note that the penultimate step requires the conditional intensity between each two consecutive events to be constant. The probability of the process remaining in state j without arrival is

$$
\begin{aligned}
H_{jj}^{(n)}(\Delta u) &= P\{Y(t_{n-1}+\Delta u)=j, N_n(\Delta u)=0 \,|\, Y(t_{n-1})=j, \mathcal{H}_{t_{n-1}}\} \\
&= P\{Y(t_{n-1}+\Delta u)=j \,|\, N_n(\Delta u)=0, Y(t_{n-1})=j, \mathcal{H}_{t_{n-1}}\} \\
&\quad \times P\{N_n(\Delta u)=0 \,|\, Y(t_{n-1})=j, \mathcal{H}_{t_{n-1}}\} \\
&= P\{Y(t_{n-1}+\Delta u)=j \,|\, Y(t_{n-1})=j\} \\
&\quad \times P\{N_n(\Delta u)=0 \,|\, Y(t_{n-1})=j, \mathcal{H}_{t_{n-1}}\} \\
&= \exp\{-q_j \Delta u\} \exp\{-\lambda_j^*(t_n)\Delta u\},
\end{aligned}
$$

where the last step holds due to the formulation of the intensity function of the Hawkes process with stepwise decay rate that for $t_k < t < t_{k+1}$, $\lambda^*(t) = \lambda^*(t_{k+1})$. Applying the Taylor series expansion we have that

$$
\begin{aligned}
&\exp\{-q_j \Delta u\} \exp\{-\lambda_j^*(t_n)\Delta u\} \\
&\qquad = \exp\{-q_j \Delta u - \lambda_j^*(t_n)\Delta u\} \\
&\qquad = 1 - q_j \Delta u - \lambda_j^*(t_n)\Delta u + o(\Delta u).
\end{aligned}
$$

It then follows that

$$
\lim_{\Delta u \to 0} \frac{H_{jj}^{(n)}(\Delta u) - 1}{\Delta u} = -q_j - \lambda_j^*(t_n),
$$

and therefore we have for $u > 0$,

$$\begin{aligned}
&\lim_{\Delta u \to 0} \frac{H_{ij}^{(n)}(u + \Delta u) - H_{ij}^{(n)}(u)}{\Delta u} \\
&\quad = \lim_{\Delta u \to 0} \left(\sum_{k=1, k \neq j}^{r} H_{ik}^{(n)}(u) \frac{H_{kj}^{(n)}(\Delta u)}{\Delta u} + H_{ij}^{(n)}(u) \frac{H_{jj}^{(n)}(\Delta u) - 1}{\Delta u} \right) \\
&\quad = \sum_{k=1, k \neq j}^{r} H_{ik}^{(n)}(u) q_{kj} - H_{ij}^{(n)}(u)(q_j + \lambda_j^*(t_n)),
\end{aligned}$$

which is

$$H_{ij}^{(n)'}(u) = -H_{ij}^{(n)}(u)(q_j + \lambda_j^*(t_n)) + \sum_{k=1, k \neq j}^{r} H_{ik}^{(n)}(u) q_{kj}, \quad u > 0,$$

and

$$H_{ij}^{(n)'}(0) = \delta_{ij},$$

where $\delta_{ij} = 1$ for $i = j$, and 0 otherwise. We express this in matrix format as

$$\begin{cases} H^{(n)'}(u) = H^{(n)}(u)(Q - \Lambda^*(t_n)) \\ H^{(n)'}(0) = I \end{cases} \tag{3.2}$$

where $H^{(n)}(u) = \{H_{ij}^{(n)}(u)\}$, which yields $H^{(n)}(u) = \exp\{(Q - \Lambda^*(t_n))u\}$ for $u \geq 0$.

The transition probability matrix of the MMHPSD is then given by

$$\begin{aligned}
F^{(n)}(x) &= \int_0^x H^{(n)}(u) \Lambda^*(t_{n-1} + u) \mathrm{d}u \\
&= \int_0^x \exp\{(Q - \Lambda^*(t_n))u\} \Lambda^*(t_{n-1} + u) \mathrm{d}u,
\end{aligned}$$

where the elements $F_{ij}^{(n)}(x) = P(Y_n = j, X_n \leq x \,|\, Y_{n-1} = i, \mathcal{H}_{t_{n-1}})$. Thus the transition density matrix is

$$f^{(n)}(x) = \exp\{(Q - \Lambda^*(t_n))x\} \Lambda^*(t_n).$$

Let $\lambda = (\lambda_1, \cdots, \lambda_r)$, $\nu = (\nu_1, \cdots, \nu_r)$, and $\eta = (\eta_1, \cdots, \eta_r)$. Denote $\Theta = \{Q; \lambda, \nu, \eta\}$. Let $\pi = (\pi_1, \cdots, \pi_r)$ denote the initial distribution vector of the Markov process. If $x_1, \cdots, x_n$ are the

observed inter-event times, then the likelihood of the parameter Θ is

$$\mathcal{L}(\Theta; x_1, \cdots, x_n) = \pi \left\{ \prod_{i=1}^{n} f^{(i)}(x_i; \Theta) \right\} \mathbf{1},$$

where $\mathbf{1}$ is an $r \times 1$ vector of ones.

Similar to a standard discrete time HMM problem, the forward and backward probabilities can be defined as follows. Define the forward probability as

$$\begin{aligned}
\alpha_t(i) &= P\{T_1 = t_1, \cdots, T_{N(t)} = t_{N(t)}, 0 < t_1 < \cdots < t_{N(t)} \le t < t_{N(t)+1}, Y(t) = i\} \\
&= \pi \exp\{(Q - \Lambda^*(t_1))t_1\}\Lambda^*(t_1) \cdots \\
&\quad \times \exp\{(Q - \Lambda^*(t_{N(t)}))(t_{N(t)} - t_{N(t)-1})\}\Lambda^*(t_{N(t)}) \\
&\quad \times \exp\{(Q - \Lambda^*(t_{N(t)+1}))(t - t_{N(t)})\}e_i,
\end{aligned}$$

where e_i is an $r \times 1$ vector, the elements of which are zeros except the ith entry assigned by 1. We have that $\mathcal{L} = \sum_{i=1}^{r} \alpha_T(i)$. The backward probability is defined as

$$\begin{aligned}
\beta_t(j) &= P\{T_{N(t)+1} = t_{N(t)+1}, \cdots, T_n = t_n, t_{N(t)} < t \le t_{N(t)+1} < \cdots < t_n, Y(t) = j\} \\
&= e_j' \exp\{(Q - \Lambda^*(t_{N(t)+1}))(t_{N(t)+1} - t)\}\Lambda^*(t_{N(t)+1}) \\
&\quad \times \exp\{(Q - \Lambda^*(t_{N(t)+2}))(t_{N(t)+2} - t_{N(t)+1})\}\Lambda^*(t_{N(t)+2}) \cdots \\
&\quad \times \exp\{(Q - \Lambda^*(t_n))(t_n - t_{n-1})\}\Lambda^*(t_n)\mathbf{1}.
\end{aligned}$$

The likelihood can also be expressed as

$$\mathcal{L} = \sum_{i=1}^{r} \alpha_t(i)\beta_t(i).$$

3.3 The Complete Likelihood

This section provides the complete likelihood of an MMHPSD process, which will be used in Section 3.4 to get the parameter estimation for the process. Suppose that the occurrence times $t_0 = 0, t_1, \cdots, t_n = T$ from an MMHPSD of order r are observed. Let $x_1 = t_1 - t_0, \cdots, x_n = t_n - t_{n-1}$, which are the inter-event times. Assume that the hidden Markov process $\{Y(t)\}$ has transitions at the time-points $0 < u_1 < u_2 < \cdots < u_m < T$. Define $u_0 = 0$ and $u_{m+1} = T$, and write $I_k = [u_{k-1}, u_k)$, and $\Delta u_k = u_k - u_{k-1}$, for $1 \le k \le m + 1$. Moreover, denote the state of

$\{Y(t)\}$ during I_k by s_k, let z_k be the number of events in I_k (do not count the event at $t = 0$), and let $N(t) = \#\{k : 0 < k \leq n, t_k \leq t\}$, i.e. the number of events up to time t. Let $\rho_0 = 0$ and $\rho_k = z_1 + \cdots + z_k$, for $1 \leq k \leq m + 1$. Then the complete likelihood of Θ can be written as

$$\begin{aligned}\mathcal{L}^c = \pi_{s_1} &\left\{ \prod_{k=1}^{m} q_{s_k} e^{-q_{s_k}\Delta u_k} \times \frac{q_{s_k,s_{k+1}}}{q_{s_k}} \right\} e^{-q_{s_{m+1}}\Delta u_{m+1}} \\ &\times \left\{ \prod_{k=1}^{m+1} \prod_{i=1}^{z_k} \lambda^*_{s_k}(t_{\rho_{k-1}+i}) \exp\left\{ - \int_{t_{\rho_{k-1}+i-1}}^{t_{\rho_{k-1}+i}} \lambda^*_{s_k}(t)\mathrm{d}t \right\} \right\},\end{aligned}$$

where $q_k = -q_{kk}$. Taking the logarithm of the complete likelihood and simplifying, it then follows that

$$\begin{aligned}\log \mathcal{L}^c =& \sum_{i=1}^{r} I\{Y(0) = i\} \log \pi_i - \sum_{i=1}^{r} D_i q_i + \sum_{i=1}^{r} \sum_{\substack{j=1 \\ j\neq i}}^{r} w_{ij} \log q_{ij} \\ &+ \sum_{i=1}^{r} \sum_{k=1}^{n} \log \lambda^*_i(t_k) I\{Y(t_k) = i\} - \sum_{i=1}^{r} \int_0^T \lambda^*_i(t) I\{Y(t) = i\}\mathrm{d}t \qquad (3.3) \\ =& L_1 + L_2,\end{aligned}$$

where for $i \neq j$, $w_{ij} = \#\{k : 1 \leq k \leq m, s_k = i, s_{k+1} = j\} = \#\{t : 0 < t \leq T, Y(t-) = i, Y(t) = j\}$ is the number of jumps of $Y(t)$ from state i to state j in $[0, T]$, $D_i = \sum_{\{k:1\leq k\leq m+1, s_k=i\}} \Delta u_k = \int_0^T I\{Y(t) = i\}\mathrm{d}t$ is the time $\{Y(t)\}$ spends in state i during $[0, T]$,

$$L_1 = \sum_{i=1}^{r} I\{Y(0) = i\} \log \pi_i - \sum_{i=1}^{r} D_i q_i + \sum_{i=1}^{r} \sum_{\substack{j=1 \\ j\neq i}}^{r} w_{ij} \log q_{ij} \qquad (3.4)$$

and

$$L_2 = \sum_{i=1}^{r} \sum_{k=1}^{n} \log \lambda^*_i(t_k) I\{Y(t_k) = i\} - \sum_{i=1}^{r} \int_0^T \lambda^*_i(t) I\{Y(t) = i\}\mathrm{d}t. \qquad (3.5)$$

3.4 An EM Algorithm for Parameter Estimation

Having obtained the complete likelihood of the process, the EM algorithm can then be used to estimate the parameters. If Θ_0 is a given parameter estimate, then the E-step of the EM algorithm

is to calculate the expectation

$$\mathcal{Q}(\Theta;\Theta_0) = \mathbb{E}_{Y,\Theta_0}[\log \mathcal{L}^c(\Theta; N(t), Y) \,|\, N(t)] \tag{3.6}$$

with respect to Y and Θ_0. The M-step will then be maximizing the $\mathcal{Q}$ function to obtain the new estimate

$$\widehat{\Theta} = \arg\max_{\Theta} \mathcal{Q}(\Theta;\Theta_0). \tag{3.7}$$

The expected complete log-likelihood $\mathcal{Q}(\Theta;\Theta_0)$ is obtained by taking the expectation of the complete likelihood over Y at the current parameter estimate Θ_0 conditional on the observed data $\{N(s), 0 \le s \le T\}$.

The parameters involved in the model can then be estimated by carrying out the following steps. The initial distribution of the Markov process, π_i, can be estimated using the forward and backward probabilities. Since we have that

$$\alpha_0(i)\beta_0(i) = P\{T_1 = t_1, \cdots, T_n = t_n, Y(0) = i\},$$

and $\alpha_0(i) = \pi_i$, given the observed events, the initial distribution of the Markov process can be estimated by

$$\widehat{\pi}_i = \frac{\alpha_0(i)\beta_0(i)}{\mathcal{L}}.$$

Given the observed sample path $\{N(s), 0 \le s \le T\}$, the conditional expectation of the statistics D_i and w_{ij} are

$$\begin{aligned}\widehat{D}_i &= \mathbb{E}\left\{ \int_0^T I\{Y(t) = i\}\mathrm{d}t \,\middle|\, N(s), 0 \le s \le T \right\} \\ &= \int_0^T \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^r \alpha_s(j)\beta_s(j)}\mathrm{d}t\end{aligned}$$

and

$$\begin{aligned}\widehat{w}_{ij} &= \mathbb{E}\left\{ \int_0^T P\{Y(t-) = i, Y(t) = j \,|\, N(s), 0 \le s \le T\}\mathrm{d}t \right\} \\ &= \int_0^T \frac{\alpha_t(i) q_{ij} \beta_t(j)}{\sum_{k=1}^r \alpha_s(k)\beta_s(k)}\mathrm{d}t.\end{aligned} \tag{3.8}$$

Under the constraint $q_i = \sum_{j \neq i} q_{ij}$ for each i, the transition probabilities can then be estimated by calculating

$$\widehat{q}_{ij} = \frac{\widehat{w}_{ij}}{\widehat{D}_i}, \quad i, j = 1, \cdots, r, i \neq j. \tag{3.9}$$

Then the remaining problem is to estimate the parameters in the intensity function of the Hawkes process with stepwise decay rate. This can be done by maximizing the conditional expectation of L_2 in (3.5) given the observations $\{N(s), 0 \leq s \leq T\}$. This conditional expectation can be expressed as

$$\begin{aligned}
&\mathcal{Q}_2(\Theta; \Theta_0) = \mathbb{E}\{L_2 \,|\, N(s), 0 \leq s \leq T\} \\
&= \sum_{i=1}^{r} \mathbb{E}\left\{ \sum_{k=1}^{n} \log \lambda_i^*(t_k) I\{Y(t_k) = i\} - \int_0^T \lambda_i^*(t) I\{Y(t) = i\} \mathrm{d}t \,\middle|\, N(s), 0 \leq s \leq T \right\} \\
&= \sum_{i=1}^{r} \sum_{k=1}^{n} \left\{ \frac{\alpha_{t_k}(i)\beta_{t_k}(i)}{\mathcal{L}} \log\left(\lambda_i^*(t_k)\right) - \int_{t_{k-1}}^{t_k} \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}} \lambda_i^*(t_k) \mathrm{d}t \right\}.
\end{aligned}$$

In order to get the estimates of the parameters involved in the intensity part, the forward and backward probabilities will be calculated in the E-step, and numerical optimization will then be carried out to maximize $\mathcal{Q}_2(\Theta; \Theta_0)$ in the M-step.

3.4.1 Implementation

This section will provide the implementation steps to carry out the parameter estimation. Given the observed sequence $\mathbf{x} = \{x_1, \cdots, x_n\}$ of $\mathbf{X} = \{X_1, \cdots, X_n\}$, the likelihood function of the sample path $\{N(s), 0 \leq s \leq T\}$ can be expressed as

$$P\{N(s), 0 \leq s \leq T\} = p(x^n) = \pi \prod_{k=1}^{n} f^{(k)}(x_k) \mathbf{1}, \tag{3.10}$$

and $p(x^n) = \prod_{k=1}^{n} p(x_k \,|\, x^{k-1})$, where $p(x_1 \,|\, x_0) = p(x_1)$. Let $c_k = p(x_k \,|\, x^{k-1}), k = 1, \cdots, n$. Equation (3.8) then becomes

$$\begin{aligned}
\widehat{w}_{ij} = &\sum_{k=1}^{n} \frac{q_{ij}}{c_k} \left(\pi \prod_{l=1}^{k-1} \frac{f^{(l)}(x_l)}{c_l} \right) \int_{t_{k-1}}^{t_k} H^{(k)}(t - t_{k-1}) e_i \\
&\times e_j^T \exp\{(Q - \Lambda^*(t_k))(t_k - t)\} \Lambda^*(t_k) \mathrm{d}t \left(\prod_{l=k+1}^{n} \frac{f^{(l)}(x_l)}{c_l} \mathbf{1} \right).
\end{aligned} \tag{3.11}$$

Set $L(0) = \pi$, $R(n+1) = \mathbf{1}$, and let

$$L(k) = \pi \prod_{l=1}^{k} \frac{f^{(l)}(x_l)}{c_l}, \quad k = 1, \cdots, n \tag{3.12}$$

$$R(k) = \prod_{l=k}^{n} \frac{f^{(l)}(x_l)}{c_l} \mathbf{1}, \quad k = n, \cdots, 1. \tag{3.13}$$

Since we have

$$p(x^k) = \prod_{l=1}^{k} c_l = \pi \prod_{l=1}^{k} f^{(l)}(x_l)\mathbf{1},$$

it then follows that

$$c_k = \pi \prod_{l=1}^{k-1} \frac{f^{(l)}(x_l)}{c_l} f^{(k)}(x_k)\mathbf{1} = L(k-1) f^{(k)}(x_k)\mathbf{1}.$$

Therefore, we get the following recursive expressions

$$L(k) = \frac{L(k-1) f^{(k)}(x_k)}{L(k-1) f^{(k)}(x_k)\mathbf{1}}, \tag{3.14}$$

$$R(k) = \frac{f^{(k)}(x_k) R(k+1)}{L(k-1) f^{(k)}(x_k)\mathbf{1}}. \tag{3.15}$$

From the definition of the forward and backward probabilities we also have the following result,

$$\frac{\alpha_{t_k}(i)\beta_{t_k}(i)}{\mathcal{L}} = L_i(k) R_i(k+1). \tag{3.16}$$

Let $\widehat{w} = (\widehat{w}_{ij})_{r \times r}$. Then we have

$$\begin{aligned} \widehat{w}^T = Q^T \odot \sum_{k=1}^{n} \frac{1}{c_k} \int_{t_{k-1}}^{t_k} & \exp\{(Q - \Lambda^*(t_k))(t_k - t)\}\Lambda^*(t_k) \\ & \times R(k+1) L(k-1) H^{(k)}(t - t_{k-1}) \mathrm{d}t, \end{aligned} \tag{3.17}$$

where $\odot$ denotes element-by-element multiplication of the two matrices. Using $x_k = t_k - t_{k-1}$, we have

$$\widehat{w} = Q \odot \sum_{k=1}^{n} \frac{\mathcal{I}_k^T}{c_k} \tag{3.18}$$

where

$$\mathcal{I}_k = \int_0^{x_k} \exp\{(Q - \Lambda^*(t_k))(x_k - x)\}\Lambda^*(t_k)R(k+1)L(k-1)\exp\{(Q - \Lambda^*(t_k))x\}\mathrm{d}x. \tag{3.19}$$

This integral can be efficiently evaluated by using the $2r \times 2r$ block-triangular matrix (Van Loan, 1978; Roberts et al., 2006)

$$\mathcal{C}_k = \begin{pmatrix} Q - \Lambda^*(t_k) & \Lambda^*(t_k)R(k+1)L(k-1) \\ 0 & Q - \Lambda^*(t_k) \end{pmatrix},$$

and $\mathcal{I}_k$ is the $r \times r$ upper-right block of the matrix $\mathrm{e}^{\mathcal{C}_k x_k}$. We also have that $\widehat{D}_i = \widehat{w}_{ii}/q_{ii}$, and the log-likelihood of the MMHPSD is given by

$$\log \mathcal{L} = \sum_{k=1}^{n} \log p(x_k \,|\, x^{k-1}) = \sum_{k=1}^{n} \log c_k = \sum_{k=1}^{n} \log(L(k-1)f^{(k)}(x_k)\mathbf{1}). \tag{3.20}$$

To simplify the integral in $\mathcal{Q}_2(\Theta; \Theta_0)$, we set

$$\begin{aligned} g_{ii} &= \sum_{k=1}^{n} \lambda_i^*(t_k) \int_{t_{k-1}}^{t_k} \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}}\mathrm{d}t \\ &= \sum_{k=1}^{n} \frac{\lambda_i^*(t_k)}{c_k} \left(\pi \prod_{l=1}^{k-1} \frac{f^{(l)}(x_l)}{c_l}\right) \int_{t_{k-1}}^{t_k} H^{(k)}(t - t_{k-1})e_i \\ &\qquad \times e_i^T \exp\{(Q - \Lambda^*(t_k))(t_k - t)\}\Lambda^*(t_k)\mathrm{d}t \left(\prod_{l=k+1}^{n} \frac{f^{(l)}(x_l)}{c_l}\mathbf{1}\right). \end{aligned}$$

Let $G = \mathrm{diag}(g_{ii})$. We can obtain that

$$\begin{aligned} G^T &= \sum_{k=1}^{n} \frac{\Lambda^*(t_k)}{c_k} \odot \int_{t_{k-1}}^{t_k} \exp\{(Q - \Lambda^*(t_k))(t_k - t)\}\Lambda^*(t_k) \\ &\qquad \times R(k+1)L(k-1)H^{(k)}(t - t_{k-1})\mathrm{d}t \\ &= \sum_{k=1}^{n} \frac{\Lambda^*(t_k) \odot \mathcal{I}_k}{c_k}, \end{aligned} \tag{3.21}$$

and it follows

$$G = \sum_{k=1}^{n} \frac{\Lambda^*(t_k) \odot \mathcal{I}_k^T}{c_k}, \tag{3.22}$$

and

$$\mathcal{Q}_2(\Theta;\Theta_0) = \sum_{i=1}^{r}\sum_{k=1}^{n}\left\{\frac{\alpha_{t_k}(i)\beta_{t_k}(i)}{\mathcal{L}}\log\left(\lambda_i^*(t_k)\right) - \int_{t_{k-1}}^{t_k}\frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}}\lambda_i^*(t_k)\mathrm{d}t\right\}$$
$$= \sum_{i=1}^{r}\sum_{k=1}^{n}\left\{\frac{\alpha_{t_k}(i)\beta_{t_k}(i)}{\mathcal{L}}\log\left(\lambda_i^*(t_k)\right)\right\} - \sum_{i=1}^{r} g_{ii}.$$

Algorithm 3.4.1 *Given the initial values* $\lambda_0 = (\lambda_{01},\cdots,\lambda_{0r})$, $\nu_0 = (\nu_{01},\cdots,\nu_{0r})$, *and* $\eta_0 = (\eta_{01},\cdots,\eta_{0r})$, $Q_0 = (q_{0ij})_{r\times r}$ *and* $\pi_0 = (\pi_{01},\cdots,\pi_{0r})$, *and given the observed occurrence times* $t_0 = 0, t_1,\cdots,t_n = T$ *from an MMHPSD of order* r, *with inter-event times* $x_1 = t_1 - t_0,\cdots,x_n = t_n - t_{n-1}$, *the EM algorithm for the MMHPSD can be carried out as follows.*

(1) Let $L(0) = \pi_0$, *and for* $k = 1,\cdots,n$, *let* $L(k) = L(k-1)f^{(k)}(x_k)/L(k-1)f^{(k)}(x_k)\mathbf{1}$.

(2) Let $R(n+1) = \mathbf{1}$, *and for* $k = n,\cdots,1$, *let* $R(k) = f^{(k)}(x_k)R(k+1)/L(k-1)f^{(k)}(x_k)\mathbf{1}$.

(3) For $k = 1,\cdots,n$, *let*

$$\mathcal{C}_k = \begin{pmatrix} Q-\Lambda(t_k) & \Lambda^*(t_k)R(k+1)L(k-1) \\ 0 & Q-\Lambda(t_k) \end{pmatrix}$$

and calculate the matrix $e^{\mathcal{C}_k x_k}$. *Then set* $\mathcal{I}_k$ *be the* $r\times r$ *upper-right block of this matrix.*

(4) Let

$$w = Q\odot\sum_{k=1}^{n}\frac{\mathcal{I}_k^T}{c_k}, \tag{3.23}$$

and $D_i = w_{ii}/q_{0ii}$.

(5) We have

$$\log\mathcal{L} = \sum_{k=1}^{n}\log(L(k-1)f^{(k)}(x_k)\mathbf{1})$$

and the new estimates

$$\widehat{q}_{ij} = \frac{w_{ij}}{D_i}, \quad i,j = 1,\cdots,r, i\neq j, \tag{3.24}$$

$$\widehat{q}_i = \sum_{\substack{j=1\\ j\neq i}}^{r}\widehat{q}_{ij}, \quad i = 1,\cdots,r \tag{3.25}$$

and

$$\widehat{\pi}_i = \pi_{0i} e_i^T R(1), \quad i = 1, \cdots, r. \tag{3.26}$$

(6) For $i = 1, \cdots, r$ and $k = 1, \cdots, n$, set

$$B_{ik} = L(k) e_i e_i^T R(k+1) = L_i(k) R_i(k+1). \tag{3.27}$$

(7) Let

$$G^T = \sum_{k=1}^{n} \frac{\Lambda^*(t_k) \odot \mathcal{I}_k^T}{c_k}.$$

Then numerically optimize

$$\begin{aligned} \mathcal{Q}_2(\Theta; \Theta_0) &= \sum_{i=1}^{r} \sum_{k=1}^{n} \left\{ \frac{\alpha_{t_k}(i) \beta_{t_k}(i)}{\mathcal{L}} \log\left(\lambda_i^*(t_k)\right) \right\} - \sum_{i=1}^{r} g_{ii} \\ &= \sum_{i=1}^{r} \sum_{k=1}^{n} B_{ik} \log\left(\lambda_i^*(t_k)\right) - \sum_{i=1}^{r} \sum_{k=1}^{n} \frac{\mathcal{I}_{ik}}{c_k} \lambda_i^*(t_k) \end{aligned}$$

and get estimation for the parameters λ, ν, and η.

(8) Calculate the log likelihood

$$LL_{old} = \sum_{k=1}^{n} \log(L(k-1) f^{(k)}(x_k) \mathbf{1}).$$

Then by using all the estimated parameters in step (1) with $L(0) = \widehat{\pi}$, and calculate the 'new' log likelihood

$$LL_{new} = \sum_{k=1}^{n} \log(L(k-1) f^{(k)}(x_k) \mathbf{1}).$$

(9) Iteratively use $(\widehat{Q}, \widehat{\mathbf{\Lambda}}, \widehat{\nu}, \widehat{\eta})$ in place of $(Q_0, \mathbf{\Lambda}_0, \nu_0, \eta_0)$ and repeat step (1) to step (8) until the difference in the log likelihoods becomes less than the terminative condition.

One problem is the parameter estimation involved in $\mathcal{Q}_2$ which requires some numerical optimization technique, such as the standard nonlinear optimization technique suggested in Fletcher and Powell (1963). To carry out this we need the gradients and Hessian of $\mathcal{Q}_2$, which are the partial first- and second-order derivatives of the function $\mathcal{Q}_2$ with respect to the parameters. To simplify

the notations in the derivatives, denote

$$\begin{aligned}
\xi_{ik} &= \eta_i \sum_{t_j<t_{k-1}} e^{-\eta_i(t_{k-1}-t_j)},\\
\zeta_{ik} &= \frac{1}{\lambda_i + \nu_i\eta_i \sum_{t_j<t_{k-1}} e^{-\eta_i(t_{k-1}-t_j)}},\\
\varphi_{ik} &= \sum_{t_j<t_{k-1}} e^{-\eta_i(t_{k-1}-t_j)} - \eta_i \sum_{t_j<t_{k-1}} e^{-\eta_i(t_{k-1}-t_j)}(t_{k-1}-t_j),\\
\psi_{ik} &= -2\nu_i \sum_{t_j<t_{k-1}} e^{-\eta_i(t_{k-1}-t_j)}(t_{k-1}-t_j) + \nu_i\eta_i \sum_{t_j<t_{k-1}} e^{-\eta_i(t_{k-1}-t_j)}(t_{k-1}-t_j)^2.
\end{aligned}$$

Since we already have that

$$\begin{aligned}
\mathcal{Q}_2(\Theta;\Theta_0) = &\sum_{i=1}^{r}\sum_{k=1}^{n} B_{ik}\log\left(\lambda_i + \nu_i\eta_i \sum_{t_j<t_{k-1}} e^{-\eta_i(t_{k-1}-t_j)}\right) -\\
&\sum_{i=1}^{r}\sum_{k=1}^{n} \frac{\mathcal{I}_{ik}}{c_k}\left(\lambda_i + \nu_i\eta_i \sum_{t_j<t_{k-1}} e^{-\eta_i(t_{k-1}-t_j)}\right),
\end{aligned}$$

the gradients of $\mathcal{Q}_2$ are given by

$$\begin{aligned}
\frac{\partial\mathcal{Q}_2}{\partial\lambda_i} &= \sum_{k=1}^{n} B_{ik}\zeta_{ik} - \sum_{k=1}^{n}\frac{\mathcal{I}_{ik}}{c_k},\\
\frac{\partial\mathcal{Q}_2}{\partial\nu_i} &= \sum_{k=1}^{n} B_{ik}\zeta_{ik}\xi_{ik} - \sum_{k=1}^{n}\frac{\mathcal{I}_{ik}}{c_k}\xi_{ik},\\
\frac{\partial\mathcal{Q}_2}{\partial\eta_i} &= \sum_{k=1}^{n} B_{ik}\zeta_{ik}\nu_i\varphi_{ik} - \sum_{k=1}^{n}\frac{\mathcal{I}_{ik}}{c_k}\nu_i\varphi_{ik}.
\end{aligned}$$

The Hessian of the function $\mathcal{Q}_2$ is given by

$$\begin{aligned}
\frac{\partial^2\mathcal{Q}_2}{\partial\lambda_i^2} &= -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2,\\
\frac{\partial^2\mathcal{Q}_2}{\partial\lambda_i\partial\nu_i} &= -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2\xi_{ik},\\
\frac{\partial^2\mathcal{Q}_2}{\partial\lambda_i\partial\eta_i} &= -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2\nu_i\varphi_{ik},
\end{aligned}$$

$$\frac{\partial^2 \mathcal{Q}_2}{\partial \nu_i^2} = -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2\xi_{ik}^2,$$
$$\frac{\partial^2 \mathcal{Q}_2}{\partial \nu_i \partial \eta_i} = -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2\nu_i\varphi_{ik}\xi_{ik} + \sum_{k=1}^{n} B_{ik}\zeta_{ik}\varphi_{ik} - \sum_{k=1}^{n}\frac{\mathcal{I}_{ik}}{c_k}\varphi_{ik},$$
$$\frac{\partial^2 \mathcal{Q}_2}{\partial \eta_i^2} = -\sum_{k=1}^{n} B_{ik}\zeta_{ik}^2\nu_i^2\varphi_{ik}^2 + \sum_{k=1}^{n} B_{ik}\zeta_{ik}\psi_{ik} - \sum_{k=1}^{n}\frac{\mathcal{I}_{ik}}{c_k}\psi_{ik}.$$

The nonlinear optimization technique of Davidon–Fletcher–Powell (see Dennis and Schnabel, 1983) can then be used to obtain the parameter estimates.

3.5 Goodness-of-fit

3.5.1 Estimated Intensity Function of the Observed Process

After fitting an MMHPSD to a set of observed data, the probability of the hidden state occupying a specific state at time t given the entire observed process will provide a clear illustration of the underlying process. As discussed in Zucchini and Guttorp (1991) and MacDonald and Zucchini (1997), this probability can be directly calculated using the forward and backward probabilities with the estimated parameters,

$$\begin{aligned} P(S_t = i \,|\, \mathcal{H}_T) &= \frac{P(T_1 = t_1, \cdots, T_n = t_n, Y(t) = i)}{P(T_1 = t_1, \cdots, T_n = t_n)} \\ &= \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}}. \end{aligned}$$

The estimated intensity function using the parameter estimation can be obtained by

$$\begin{aligned} \widehat{\lambda}^*(t) &= \sum_{i=1}^{r}\left(\widehat{\lambda}_i + \widehat{\nu}_i\widehat{\eta}_i \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\widehat{\eta}_i(\max\{t_l : t_l < t\} - t_j)}\right) P(S_t = i \,|\, \mathcal{H}_T) \\ &= \sum_{i=1}^{r}\left(\widehat{\lambda}_i + \widehat{\nu}_i\widehat{\eta}_i \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\widehat{\eta}_i(\max\{t_l : t_l < t\} - t_j)}\right) \frac{\alpha_t(i)\beta_t(i)}{\mathcal{L}}. \end{aligned} \tag{3.28}$$

The estimated intensity function of the MMHPSD process illustrates how the model performs.

Algorithm 3.5.1 *After fitting an MMHPSD to the data, the estimates of the parameters* $\widehat{\lambda} = (\widehat{\lambda}_1, \cdots, \widehat{\lambda}_r)$, $\widehat{\nu} = (\widehat{\nu}_1, \cdots, \widehat{\nu}_r)$, *and* $\widehat{\eta} = (\widehat{\eta}_1, \cdots, \widehat{\eta}_r)$, $\widehat{Q} = (\widehat{q}_{ij})_{r\times r}$ *and* $\widehat{\pi} = (\widehat{\pi}_1, \cdots, \widehat{\pi}_r)$ *are obtained. Given*

the observed occurrence times $t_0 = 0, t_1, \cdots, t_n = T$ *with inter-event times* $x_1 = t_1 - t_0, \cdots, x_n = t_n - t_{n-1}$*, the procedure of estimating the probability of the hidden state occupying a specific state at time* t *and the intensity function is as follows.*

(1) Let $L(0) = \widehat{\pi}$*, and for* $k = 1, \cdots, n$*, let* $L(k) = L(k-1)f^{(k)}(x_k)/L(k-1)f^{(k)}(x_k)\mathbf{1}$.

(2) Let $R(n+1) = \mathbf{1}$*, and for* $k = n, \cdots, 1$*, let* $R(k) = f^{(k)}(x_k)R(k+1)/L(k-1)f^{(k)}(x_k)\mathbf{1}$.

(3) For $i = 1, \cdots, r$ *and* $t_{k-1} \leq t < t_k$*, set*

$$\begin{aligned}\widehat{p}_i(t) =& L(k-1)\exp\{(Q - \Lambda^*(t_k))(t - t_{k-1})\}e_i \\ & \times e_i^{'}\exp\{(Q - \Lambda^*(t_k))(t_k - t)\}\Lambda^*(t_k)R(k+1)/c_k.\end{aligned}$$

Thus $\widehat{p}_i(t)$ *is the estimated probability of the hidden state occupying state* i *at time* t.

(4) For $i = 1, \cdots, r$ *and* $t_{k-1} \leq t < t_k$*, set*

$$\widehat{\lambda}^*(t) = \sum_{i=1}^{r}\left(\widehat{\lambda}_i + \widehat{\nu}_i\widehat{\eta}_i \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\widehat{\eta}_i(\max\{t_l : t_l < t\} - t_j)}\right)\widehat{p}_i(t),$$

which is the estimated intensity function of the MMHPSD.

3.5.2 Residual Analysis

Residual analysis (see, e.g., Ogata, 1988; Bebbington and Harte, 2001) can be used to assess the utility of the MMHPSD. According to Theorem 7.4.I in Daley and Vere-Jones (2003), if a point process in time with intensity function $\lambda(t)$ is rescaled using the random time change

$$\tau = \Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s, \tag{3.29}$$

then the rescaled time τ is a stationary Poisson process with unit rate. For an observed point process with occurrence times $t_1, t_2, \cdots, t_n$, the true model for the data is usually unknown. We can fit a stochastic model to the data and then examine whether the fitted model captures the main features of the observed point process, or in other words, whether it is a good approximation of the true model for the data. Residual analysis is different from other model selection criteria such as the log

likelihood ratio test, Akaike Information Criterion (AIC; Akaike, 1974) which is defined as

$$\text{AIC} = -2\log(\text{likelihood}) + 2k,$$

and Bayesian information criterion (BIC; Schwarz, 1978) which is defined as

$$\text{BIC} = -2\log(\text{likelihood}) + k\log(n),$$

where k is the number of parameters to be estimated and n is the number of observations. As pointed out in Bebbington and Harte (2001), the former is used to identify systematic deviation of the data from the fitted model. It would indicate a significant factor underlying the data which is not included in the model. Therefore, it provides an absolute evidence of a good fit. The latter three criteria, however, only suggest which model is relatively better and do not guarantee the model is a good approximation of the true model.

Assume that the estimated intensity function $\widehat{\lambda}(t)$ is a good approximation of the true intensity function of the data. The transformed time by the random time change

$$\tau_i = \Lambda(t_i) = \int_0^{t_i} \widehat{\lambda}(s)\mathrm{d}s,$$

which is called the residual point process, is then expected to be a stationary Poisson process with unit rate. However, if the transformed time deviates from the unit-rate Poisson process, it might suggest that the fitted model is not sufficient to explain the features of the data. In order to test whether the residual point process is a stationary Poisson process with unit rate, the following steps should be conducted.

The Kolmogorov–Smirnov test can be used to check whether the arrival times have a uniform distribution over the period of observation. For the fitted MMHPSD, the estimated intensity function (3.28) is used to get the transformed time

$$\Lambda(t) = \int_0^t \widehat{\lambda}^*(s)\mathrm{d}s = \int_0^t \sum_{i=1}^{r} \left(\widehat{\lambda}_i + \widehat{\nu}_i\widehat{\eta}_i \sum_{t_j < \max\{t_l : t_l < s\}} e^{-\widehat{\eta}_i(\max\{t_l : t_l < s\} - t_j)} \right) \widehat{p}_i(s)\mathrm{d}s.$$

Let

$$\widehat{\lambda}_i^*(s) = \widehat{\lambda}_i + \widehat{\nu}_i\widehat{\eta}_i \sum_{t_j < \max\{t_l : t_l < s\}} e^{-\widehat{\eta}_i(\max\{t_l : t_l < s\} - t_j)}.$$

For each occurrence time t_k, we then have

$$\begin{aligned}
\tau_k = \Lambda(t_k) &= \int_0^{t_k} \sum_{i=1}^{r} \widehat{\lambda}_i^*(s)\widehat{p}_i(s)\mathrm{d}s \\
&= \sum_{h=1}^{k} \int_{t_{h-1}}^{t_h} \sum_{i=1}^{r} \widehat{\lambda}_i^*(t_h)\widehat{p}_i(s)\mathrm{d}s \\
&= \sum_{h=1}^{k} \sum_{i=1}^{r} \widehat{\lambda}_i^*(t_h) \int_{t_{h-1}}^{t_h} \widehat{p}_i(s)\mathrm{d}s \\
&= \sum_{h=1}^{k} \sum_{i=1}^{r} \frac{\widehat{\lambda}_i^*(t_h)}{c_h} \left(\pi \prod_{l=1}^{h-1} \frac{f^{(l)}(x_l)}{c_l} \right) \int_{t_{h-1}}^{t_h} H^{(h)}(s - t_{h-1})e_i \\
&\qquad \times e_i^T \exp\{(Q - \Lambda^*(t_h))(t_h - s)\}\Lambda^*(t_h)\mathrm{d}s \left(\prod_{l=h+1}^{n} \frac{f^{(l)}(x_l)}{c_l} \mathbf{1} \right).
\end{aligned}$$

After the transformation procedure, the cumulative number of events versus the transformed times τ_i can be plotted with the 99% confidence limits of the Kolmogorov–Smirnov statistic under the null hypothesis of the uniform empirical distribution. If the cumulative curve exceeds either of the limits, it may suggest that there is some feature of the data not captured by the fitted model.

Let $E_k = \tau_k - \tau_{k-1} = \Lambda(t_k) - \Lambda(t_{k-1})$, $k = 1, \cdots, n$, and $U_k = 1 - \exp(-E_k)$. If E_k are independent and identically distributed exponential random variables, then U_k are independent and uniformly distributed on $[0, 1)$. Therefore, in order to examine whether the inter-arrival times of the residual process have an exponential distribution, the Kolmogorov–Smirnov statistic can be used to test whether U_k have a uniform distribution on $[0, 1)$. The method suggested by Hall (1991) can be adopted to verify the independence of the inter-arrival times and the t-test can be used to check the null hypothesis that the correlation coefficient between E_k and E_{k+1} equals zero. Note that independence implies zero correlation, however, absence of correlation does not necessarily mean that the two variables are independent. Therefore, after the t-test, the graphical illustration suggested by Berman (1983) can be used to check whether there is any pattern in the scatter plot of U_{k+1} against U_k.

3.6 Simulation Algorithm

To see how well the parameter estimation method works, an MMHPSD process can be simulated, and then the model can be fitted to the simulated data set. The method proposed above can be used to estimate the parameters. By comparing the estimated parameters with the true parameters, one

can see how the parameter estimation method performs.

Algorithm 3.6.1 *Given the initial state $y_1 = j$, the parameters Q, λ_l, ν_l and η_l, $l = 1, \cdots, r$, and the history data set $\mathcal{H} \subset \{t \leq 0\}$, the following steps can be carried out to generate n events from an MMHPSD process.*

1. *Set $i = s = 1$ and $t_i = t_s = 0$.*

2. *Use $r_i = q_j + \lambda_j + \nu_j \eta_j \sum_{t_l < t_i} e^{-\eta(t_i - t_l)}$ as the rate and generate an inter-event time τ_i from the exponential distribution.*

3. *Set $t_{s+1} = t_s + \tau_i$. Generate a uniform random variable $U \in (0, 1)$.*

   - *If $U > q_j / r_i$, then add $t_{i+1} = t_{s+1}$ into the history. Set $i = i + 1$, $s = s + 1$ and go to Step 2.*
   - *If $U \leq q_j / r_i$, then this point is a state transition point. Use $(q_{jk}/q_j)_{1 \leq k \leq r}$ to generate the next state y_2. Set $j = y_2$, $s = s + 1$ and go to Step 2.*

   If $i = n$, then stop.

4. *The sequence of times simulated is the generated MMHPSD process.*

Chapter 4

Simulation Study and an Application of the MMHPSD

4.1 Introduction

As many geophysical processes occur in a self-exciting way, in which the events already occurred often trigger new ones, and as the underlying dynamics for these processes might be represented as being governed by a Markov chain, Chapter 3 introduced the Markov-modulated Hawkes process with stepwise decay. In this model the hidden process switches among some finite states of a continuous Markov chain and in each state the observed events follow a self-exciting Hawkes process with a stepwise decay rate. A parameter estimation method is also developed by using the EM algorithm for this model. Before putting this into application, the parameter estimation algorithm needs to be validated. Once we have established that the parameter estimation from the EM algorithm performs reasonably well, an exploratory data analysis of the model on earthquake sequences can be carried out to study how this model captures seismicity rate changes.

There have been many investigations on seismicity rate changes before and after the magnitude 7.3 Landers earthquake on June 28, 1992, the Big Bear earthquake of magnitude 6.4 which occurred three hours after the Landers main shock, and the 1999 magnitude 7.1 Hector Mine earthquake thought to have been triggered by aftershocks of the Landers earthquake (Felzer et al., 2002). A significant seismicity rate increase following the Landers earthquake has been observed as far as 600km away from the Landers source region (Hill et al., 1993, 1995). Wyss and Wiemer (2000) investigated the seismicity rate changes for Landers using declustered data, comparing the data for the 12 years before Landers earthquake to the 7 years following. They concluded that the 1992 Lan-

ders earthquake shut off the production of small earthquakes in some regions (the volumes south of the future Hector Mine rupture and north of Big Bear) while increasing the seismicity in the neighboring regions (the volume surrounding the future Hector Mine hypocenter and north of Landers). They also detected that on average more small earthquakes were produced after this shock. Gomberg et al. (2001) detected an increase of seismicity rate following the Hector Mine earthquake within 250km from the main shock. Marsan (2003) observed seismicity shadows east of the Joshua Tree rupture, which occurred on April 22, 1992 with a magnitude of 6.1, following the Landers earthquake. This correlates well with the stress shadows modeled by King et al. (1994) and McCloskey et al. (2003). Ogata et al. (2003) used residual analysis of the Epidemic Type Aftershock Sequence (ETAS) model on the Landers aftershock sequence. This analysis revealed relative quiescence about 6 months after the main shock, which lasted nearly 7 years leading up to the Hector Mine earthquake. They also detected relative quiescence in the aftershock sequence of the Joshua Tree earthquake for a period leading up to the Landers rupture. Marsan and Nalbant (2005) observed seismicity shadows developing after a few days of the Landers earthquake in the region of the Joshua Tree earthquake, which are sometimes preceded by instances of early triggering. The MMHPSD will be applied to the sequence of data collected from Joshua Tree, Landers, Big Bear and Hector Mine to examine how this model captures the seismicity rate changes in the selected area.

In this chapter, first, the performance of the EM algorithm for the parameter estimation of the MMHPSD is evaluated. The simulation algorithm of this model was provided in Chapter 3. An arbitrary set of parameters is used to simulate 100 sequences of MMHPSD events, and the MMHPSD is then refitted to each of the simulated sequences. The parameters are estimated using the EM algorithm for the 100 sequences and the histogram of the parameter estimates is plotted to examine how the EM algorithm works for the parameter estimation of this model. Another simulation study is conducted via a simulated ETAS sequence. The estimated intensity function of the MMHPSD is compared with the true ETAS intensity to check how well the model captures the simulated data. The estimated parameters for this simulated earthquake catalogue are used to conduct a consistency test for the parameter estimation of the MMHPSD. After the simulation studies, a case study of the model is carried out using the earthquake data around Landers. A discussion section concludes the chapter.

4.2 Evaluation of Parameter Estimation Algorithm

Before the MMHPSD is used to investigate any real data, we need to first evaluate how the parameter estimation algorithm proposed in Chapter 3 works for the model. Some arbitrary parameters are used to generate 100 series of MMHPSD events, each with 5000 observations. Then an MMHPSD is fitted to each of the simulated series of events and the EM algorithm is used to estimate the parameters. The self-exciting feature in conjunction with the underlying Markov structure suggests that a burn-in period for the simulations might improve the consistency, but as we will see, it is not necessary.

The parameters $\lambda_1 = 0.1$, $\lambda_2 = 10$, $\nu_1 = 0.1$, $\nu_2 = 0.5$, $\eta_1 = 0.05$, $\eta_2 = 5$, $q_1 = 0.001$ and $q_2 = 0.1$ are used to simulate the 100 series of events. The cumulative curve of a typical sequence of events is illustrated in Figure 4.1. The obvious jumps correspond to event occurrence rate changes.

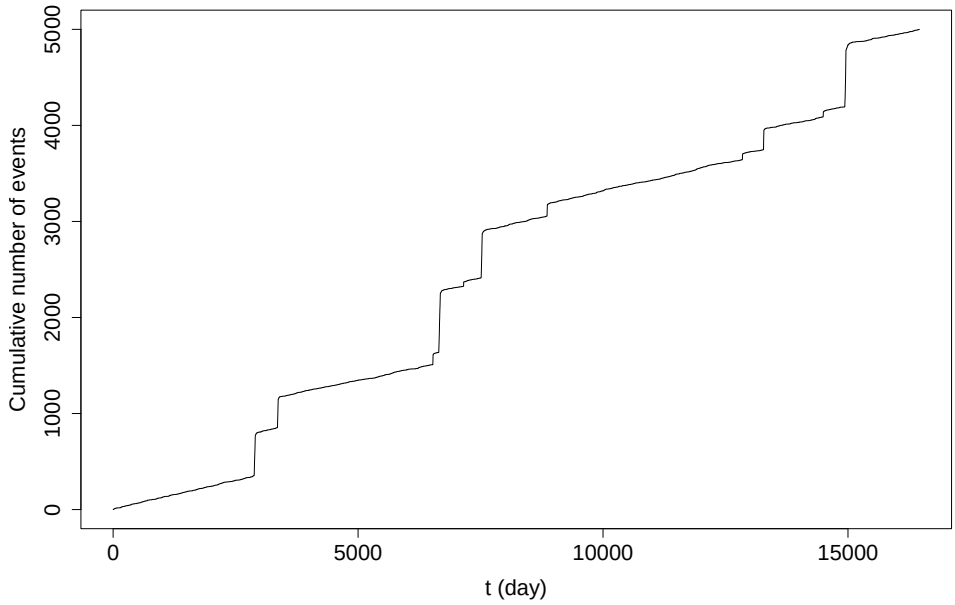


Figure 4.1: Cumulative curve of a simulated MMHPSD sequence with $\lambda_1 = 0.1$, $\lambda_2 = 10$, $\nu_1 = 0.1$, $\nu_2 = 0.5$, $\eta_1 = 0.05$, $\eta_2 = 5$, $q_1 = 0.001$ and $q_2 = 0.1$.

The EM algorithm is used to estimate the parameters. The histograms of the estimated parameters are shown in Figure 4.2. From the histograms, we can see that the estimated parameters appear to be centered around the true parameters, which were used to simulate the series of events. The sample means and standard deviations of the estimated parameters are listed in Table 4.1 for reference.

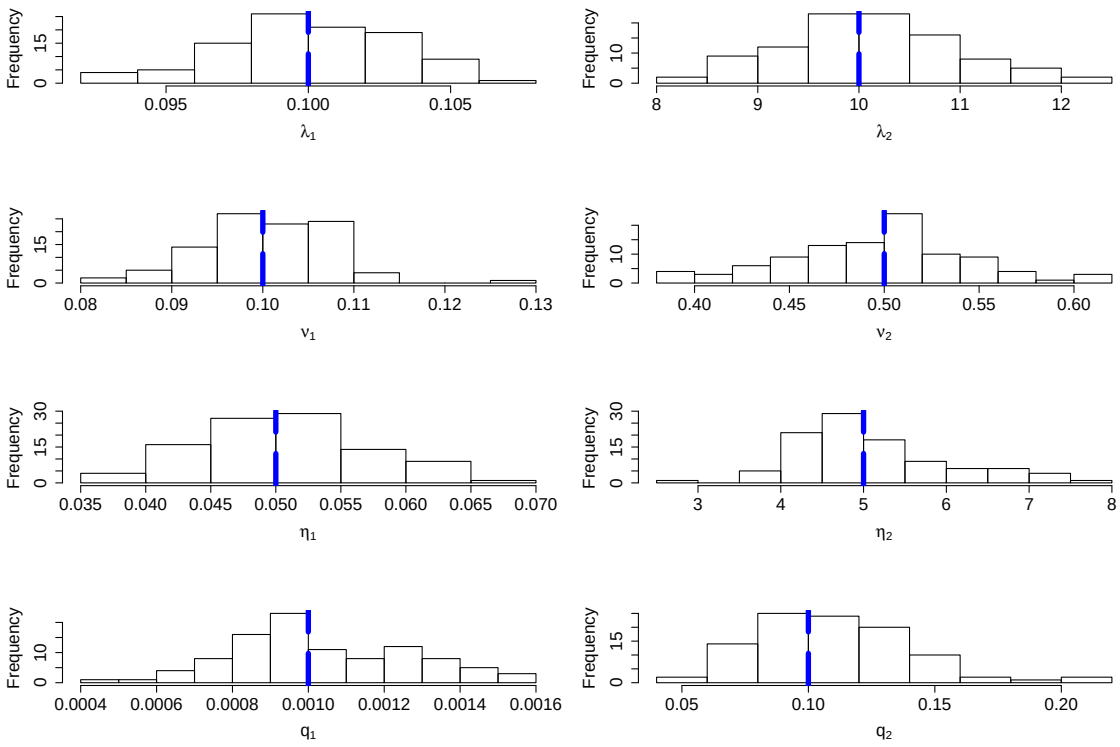


Figure 4.2: Histograms of the estimates $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\nu}_1$, $\hat{\nu}_2$, $\hat{\eta}_1$, $\hat{\eta}_2$, $\hat{q}_1$ and $\hat{q}_2$ of the parameters $\lambda_1 = 0.1$, $\lambda_2 = 10$, $\nu_1 = 0.1$, $\nu_2 = 0.5$, $\eta_1 = 0.05$, $\eta_2 = 5$, $q_1 = 0.001$ and $q_2 = 0.1$ which are indicated in the graphs using dashed lines.

Table 4.1: The sample means (Est mean) and standard deviations (Est s.d.) of the estimated parameters for the simulated MMHPSDs.

Parameter	True value	Est mean	Est s.d.
λ_1	0.1	0.100	0.003
λ_2	10	10.131	0.840
ν_1	0.1	0.100	0.007
ν_2	0.5	0.496	0.048
η_1	0.05	0.051	0.006
η_2	5	5.090	0.948
q_1	0.001	0.0010	0.0002
q_2	0.1	0.110	0.031

4.3 Simulation Study Using a Simulated ETAS Sequence

Having checked the performance of the parameter estimation algorithm for this model using some arbitrary parameters, we now investigate the performance of this model using a simulated earth-

quake catalogue. A well accepted stochastic model, which has been extensively applied to various aftershock sequences, will be used to simulate a long enough sequence of events. (A long enough sequence is to ensure that we get enough seismic cycles for the investigation of this Markov-modulated model, but not too long in order to control the computational cost.) MMHPSDs with different numbers of states are then fitted to the simulated earthquake catalogue. This provides insights that will lead to understanding the structure of the model. After this, the estimated parameters are used to simulate further catalogues of varying lengths to study the consistency of the parameter estimation using the EM algorithm.

4.3.1 Fitting MMHPSDs to Simulated ETAS Sequence

The Epidemic Type Aftershock Sequence (ETAS; Ogata, 1988) model and its extensions (Ogata, 1998 and references therein) are nowadays almost invariably used to investigate the characteristics of aftershock sequences. The ETAS model is a point process and was derived based on the following assumptions. The background rate of events within a given region is assumed as a constant, μ; each event including aftershocks can generate its own aftershocks; the rate of aftershock activity from an event decays with time following a modified Omori law, which describes the frequency of aftershocks per unit time interval in the form of $n(t) = K(t+c)^{-p}$ (Utsu, 1961), where K depends on the magnitude cutoff of the aftershocks counted in $n(t)$, p and c are independent of it; the average/mean number of aftershocks generated by an earthquake of magnitude M_i is proportional to $e^{\alpha(M_i - M_0)}$, where M_0 is the magnitude threshold. The conditional intensity function for ETAS model is given by

$$\lambda(t \,|\, \mathcal{H}_t) = \mu + \sum_{i:t_i<t} e^{\alpha(M_i - M_0)} \frac{K}{(t - t_i + c)^p},$$

where t_i are the event occurrence times. In this model, μ is the occurrence rate for the background seismic activity, i.e., the immigrants follow a stationary Poisson process with a constant rate μ. The parameter c is a time delay constant, which is used to correct for the fact that the power law becomes infinite as $t - t_i$ goes to 0. It is usually found to be on the order of a fraction of a day (e.g., Ogata, 1988; Reasenberg and Jones, 1989; Ogata, 1992; Felzer et al., 2003). The parameter p characterizes aftershock decay rate and is usually slightly larger than 1 (e.g., Reasenberg and Jones, 1989; Ogata, 1992). The parameter α provides a measure of the power of a shock generating its aftershocks. For instance, swarm-type activity has a smaller α value than that of ordinary main shock and aftershock

activity, and an aftershock sequence with no conspicuous secondary aftershocks has a large α value (Ogata, 1987). The parameter K can be explained as the productivity of events dependent on the magnitude threshold M_0.

Zhuang (2000) used the ETAS model to study the seismicity of the earthquake sequence occurred off Cape Palliser at the southeastern tip of North Island, New Zealand from January 1, 1978 to May 31, 1996 (996 events in total). The whole period was divided visually according to the magnitude–time plot into four stages: early background period, relatively quiescent period, main shock and aftershock sequence, and an active period of post-aftershocks.

We use the same study area as in Zhuang (2000). Some 920 events with minimum magnitude 2.0 are selected from the SSLib package (http://homepages.paradise.net.nz/david.harte/SSLib/, last accessed on September 22, 2009) within the cylinder centered at latitude 41.686S and longitude 175.508E with a radius of 36km and a depth of 40km from January 1, 1978 to May 31, 1996. Note that the catalogue has changed slightly since the year 2000. An ETAS model is fitted to this sequence with the maximum likelihood estimates of the parameters, $\hat{\mu} = 0.025$, $\hat{K} = 11.002$, $\hat{\alpha} = 1.468$, $\hat{c} = 0.004$ and $\hat{p} = 1.127$. The estimated parameters are then used to simulate a sequence of 3000 events. Figure 4.3 shows the cumulative curve of the simulated sequence of ETAS events. We can clearly see the big jumps in the cumulative curve which indicate seismicity rate changes.

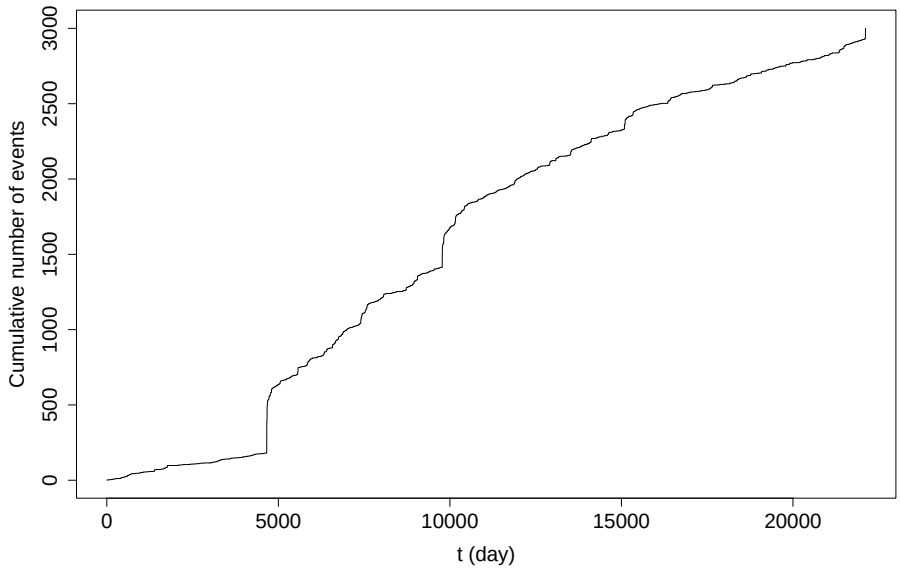


Figure 4.3: Cumulative curve of the simulated ETAS events.

Next, the MMHPSDs will be fitted to this simulated sequence. The purpose is to examine how

this model works and what kind of hidden states can be characterized by this model. MMHPSDs with two, three and four hidden states are fitted to the series of events. The estimated parameters for each model are listed in Table 4.2. Note that the states are ordered according to the decay rate

Table 4.2: Estimated parameters of the MMHPSDs fitted to the simulated ETAS events.

State	1	2	3	4
$\hat{\lambda}$	0.003	6.324		
$\hat{\nu}$	0.192	0.583		
$\hat{\eta}$	0.044	19.502		
$\hat{\lambda}$	0.005	0.449	31.108	
$\hat{\nu}$	0.175	0.446	0.738	
$\hat{\eta}$	0.019	1.569	51.219	
$\hat{\lambda}$	0.002	0.067	3.302	53.776
$\hat{\nu}$	0.173	0.256	0.889	0.647
$\hat{\eta}$	0.005	0.295	2.608	145.374

η. The estimated Q-matrices are, for the two-state model,

$$\hat{Q} = \begin{pmatrix} -0.083 & 0.083 \\ 8.036 & -8.036 \end{pmatrix}, \tag{4.1}$$

for the three-state model,

$$\hat{Q} = \begin{pmatrix} -0.061 & 0.026 & 0.035 \\ 1.770 & -3.330 & 1.560 \\ 1.086 & 49.878 & -50.964 \end{pmatrix}, \tag{4.2}$$

and for the four-state model,

$$\hat{Q} = \begin{pmatrix} -0.047 & 0.007 & 0.023 & 0.017 \\ 0.447 & -1.027 & 0.253 & 0.326 \\ 0.762 & 8.936 & -13.726 & 4.028 \\ 0.078 & 9.858 & 99.952 & -109.888 \end{pmatrix}. \tag{4.3}$$

Note that the parameter λ characterizes the background rate (or the immigrant rate) which is equivalent to the background rate μ in the ETAS model, η describes the aftershock decay rate, and ν is a parameter to ensure a finite total number of descendants of any immigrant with probability one.

The main difference between the ETAS model and the Hawkes process is in the term involving magnitude and the decay rate term. The ETAS model assumes that for each event, both the time and magnitude of the event influence the intensity. However, the conditional intensity function of the Hawkes process is only time-dependent. The decay rate of the ETAS model follows a power law, t^{-p}, while that of the Hawkes process is an exponential decay function, $e^{-\eta t}$. Let us compare the two decay functions disregarding the magnitude effect. At a fixed time t, if we want the two decay rates to be equal, we will have

$$t^{-p} = e^{-\eta t},$$

which provides a relationship between p and η as

$$\eta = \frac{p}{t} \log t.$$

For example, for $p = 1.127$, estimated from the Cape Palliser data set, and for a given $t = 5$ days, we have $\eta = 0.363$. For this p and η, the two decay functions, t^{-p} and $e^{-\eta t}$, versus time t are plotted in Figure 4.4. It is not difficult to notice that when we change η, only the slope of

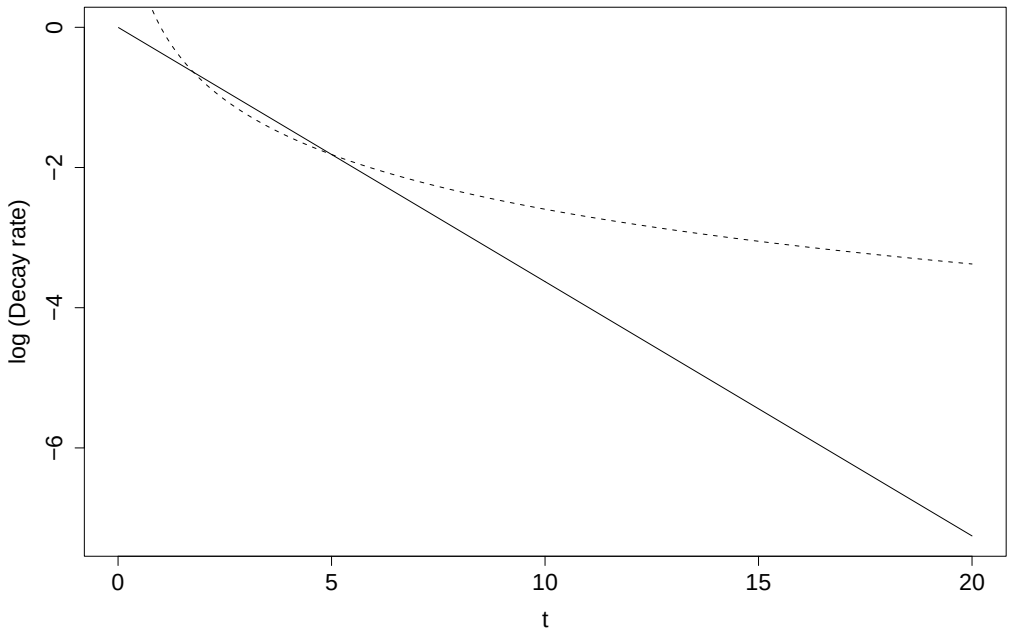


Figure 4.4: The log-scaled decay rate t^{-p} (dashed) and the log-scaled decay rate $e^{-\eta t}$ (solid).

the linear solid line varies, with the starting point of the line always at $(0, 0)$. This suggests that for very small and very large $t - t_i$, the ETAS model always has larger decay rate. Moreover, in

the ETAS model, the decay rate is multiplied by the exponential term of the magnitude (which is always larger than or equal to 1), which makes the ETAS intensity much larger than that of the Hawkes process. The relationship between ν in the Hawkes process and K in the ETAS model is not as obvious as the other parameters. Assume that the distribution of earthquake magnitudes M follows the Gutenberg–Richter distribution $b\log(10)10^{-b(M-M_0)}$. The average number of offspring (triggered per immigrant) predicted by the ETAS model is (cf., Helmstetter and Sornette, 2002)

$$a_E = \int_0^{+\infty} \mathrm{d}t \int_{M_0}^{+\infty} \mathrm{d}M b\log(10)10^{-b(M-M_0)} \frac{K10^{\alpha^*(M-M_0)}}{(t+c)^p} = n_0 \int_0^{+\infty} \frac{\mathrm{d}t}{(t+1)^p}$$

where $\alpha^* = \alpha/\log(10)$ and $n_0 = Kb/c^{p-1}(b-\alpha^*)$, and is

$$\int_0^{+\infty} \nu\eta e^{-\eta t}\mathrm{d}t = \nu$$

predicted by the Hawkes process. Given the same number of offspring, we have that

$$\nu = \frac{K}{c^{p-1}} \frac{b}{b-\alpha^*} \int_0^{+\infty} \frac{\mathrm{d}t}{(t+1)^p}.$$

For example, for $b = 1$, $\alpha = 1.468$, i.e. $\alpha^* = 0.638$, as discussed above, restraining $0 < \nu = a_E < 1$ ensures that each immigrant has a finite total number of offspring with probability one. However, in the ETAS model, there is no such constraint on the parameter K, i.e., there are cases when $a_E \geq 1$. For an aftershock sequence, the ETAS model may provide a good fit. However, for other sequences, it may put too much weight on the magnitude effect which would result in a higher intensity than the true intensity of the events.

For the three MMHPSDs with 2, 3 and 4 hidden states fitted to the simulated ETAS sequence, let us first compare the estimates of the event occurrence rates from Table 4.2, and then discuss the transition probability matrices from (4.1), (4.2) and (4.3). For State 1 in each of the three models, the background rate is smaller than that of the ETAS model, which suggests that the immigrants in State 1 occur in a very low rate. Therefore State 1 in each model is considered as the ground state. For the largest state in each of the three models, both the background and the decay rates increase with increasing number of states. For example, in the two-state model, the immigrants in State 2 occur at a rate of about 6 events per day; whereas in the four-state model, the immigrants in State 4 occur at a rate of about 54 events per day. Moreover, on average the offspring in State 4 of the four-state model have shorter inter-event times (i.e., larger decay rate) than that in State 2 of the

two-state model. Basically, the ground state, State 1, in each of the three models does not have too much difference from each other. The higher states, however, present some sort of cascading style. The high state in the two-state model is divided into the high and medium states in the three-state model. The latter two states are then separated into three states in the four-state model. It seems that the various states capture earthquakes with different ranges of magnitudes. State 2 in the two-state model is more related with large and medium events. State 3 in the three-state model and State 4 in the four-state model mainly correspond to large earthquakes. The estimated probability of staying in each state and the estimated intensity functions are illustrated in Figures 4.5 to 4.7.

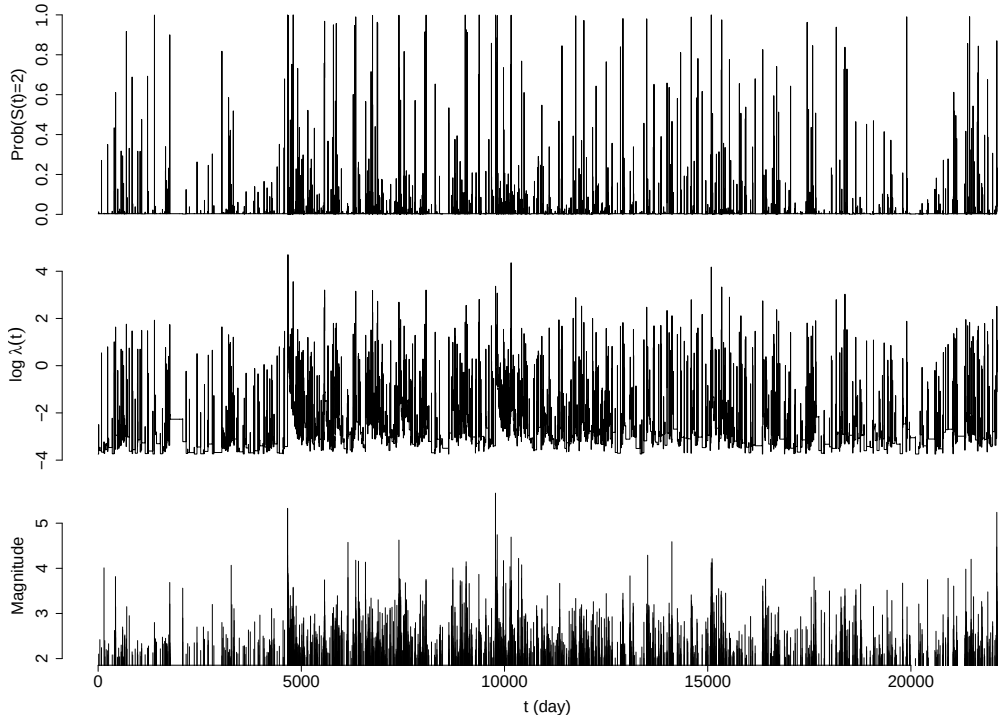


Figure 4.5: The estimated probability of the hidden state occupying State 2 (top), and the estimated intensity function (middle) for the two-state MMHPSD. The bottom one is the magnitude-time plot for the earthquake sequence.

For the three-state model, the transition intensity matrix suggests a quick switch of the hidden state out of State 3, which has the largest intensity representing the period of main shocks, mainly to State 2. It indicates that after large events, the seismicity rate increases. State 2 may characterize the aftershock state, and State 1 is a ground state which has very low seismicity rate.

According to the estimated transition intensity matrix, the state transition structure for the four-state model is that State 4 mostly jumps to State 3 which is more likely to make a transition to a lower intensity State 2, State 2 transits to every state, while State 1 mainly transits to State 3. If we

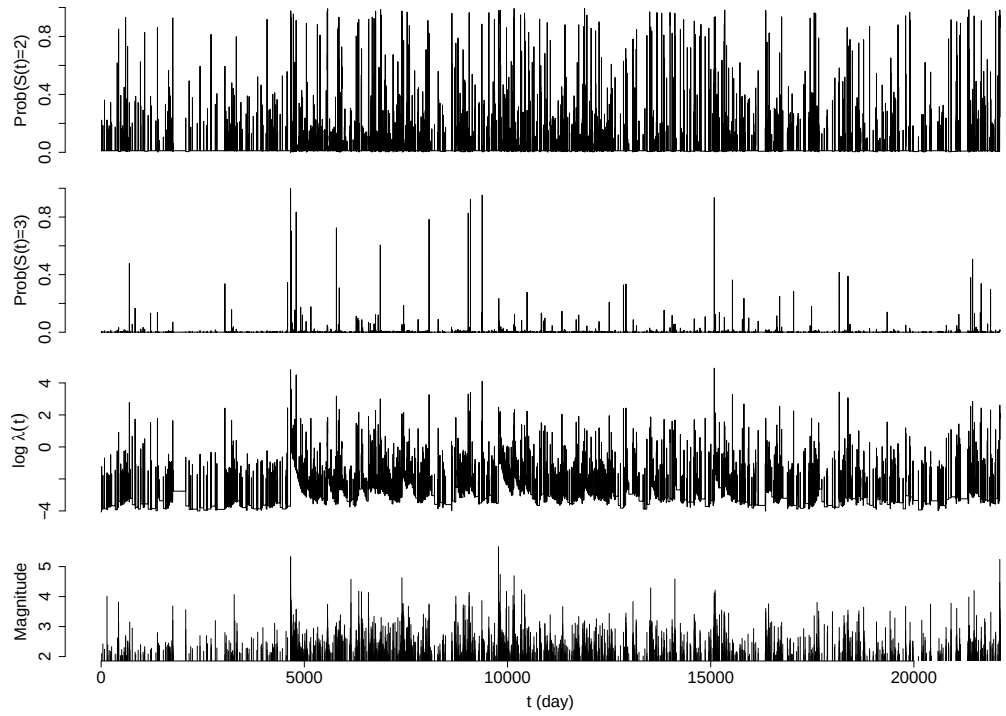


Figure 4.6: The estimated probability of the hidden state occupying each state (top two), and the estimated intensity function (third) for the three-state MMHPSD. The bottom one is the magnitude–time plot for the earthquake sequence.

consider possible foreshocks with their own aftershocks as the sequence State 3–2, then it seems that there is a favored transition pattern like State 1–(3–2)–(4–3–2)(the main sequence)–1. State 4 with the largest intensity represents the period of main shocks. The seismicity rate increases after the main shocks (State 2 preceding State 4 has lower seismicity rate than State 3 which follows State 4). State 3 accounts for the immediate aftershocks, while State 2 may correspond to swarms, or smaller aftershocks which are known as secondary aftershocks (Richter, 1958).

Now let us compare the true intensity function of the ETAS model with the estimated intensity functions for the fitted MMHPSDs with 2, 3 and 4 hidden states. On average, the intensity of each of the fitted MMHPSDs is smaller than that of the ETAS model. Figure 4.8 shows the differences between the log-scaled intensities of the MMHHSDs and the ETAS model normalized by the log-scaled intensity of the ETAS model. The big spikes indicate large differences between the intensities derived from the MMHPSDs and the ETAS model, especially around the large event occurrences. This is due to the observation above (cf., Figure 4.4) that the ETAS model has larger decay rates immediately following an event than the MMHPSDs, which are further multiplied by the exponential term of the magnitude and hence result in much larger ETAS intensities.

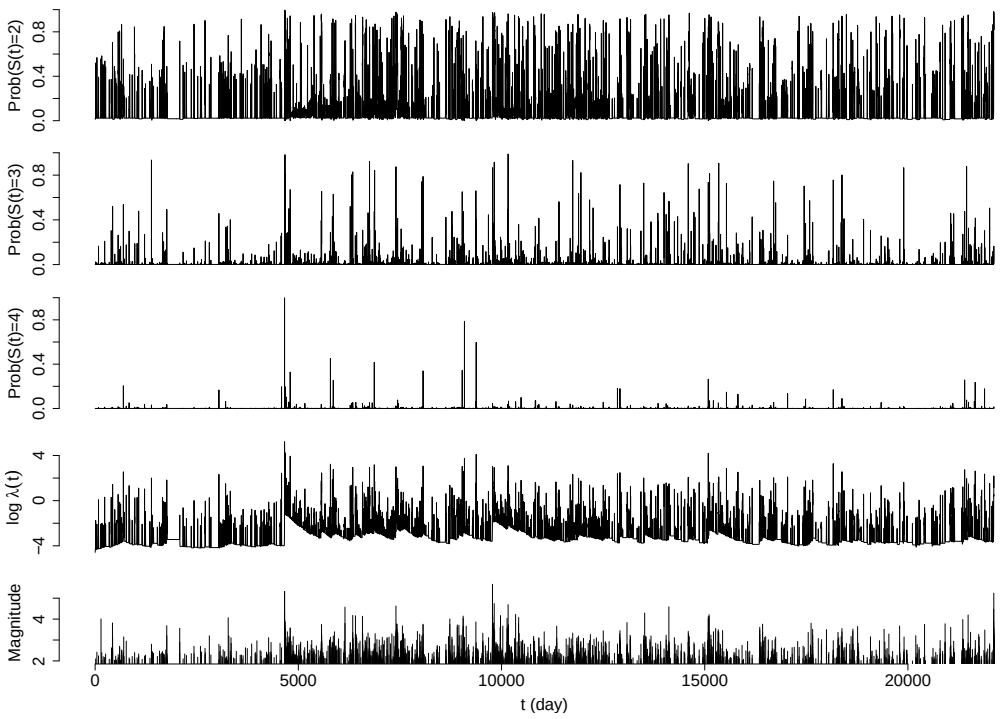


Figure 4.7: The estimated probability of the hidden state occupying each state (top three), and the estimated intensity function (fourth) for the four-state MMHPSD. The bottom one is the magnitude–time plot for the earthquake sequence.

4.3.2 Consistency of the Parameter Estimation

We now proceed to examine the consistency of the parameter estimation for the MMHPSDs. Instead of using some arbitrary parameters, the parameters of the two-state MMHPSD estimated from the simulated ETAS sequence will be used. The parameters $\lambda_1 = 0.003$, $\lambda_2 = 6.324$, $\nu_1 = 0.192$, $\nu_2 = 0.583$, $\eta_1 = 0.044$, $\eta_2 = 19.502$, $q_1 = 0.083$ and $q_2 = 8.036$ are used to simulate four groups of MMHPSD events. Each group consists of 100 sequences. Each sequence has 500 events in Group 1, 1000 events in Group 2, 2000 events in Group 3 and 5000 events in Group 4. Then the MMHPSD is refitted to each of the simulated sequences and the parameters are estimated.

The boxplot of the estimated parameters for each group is shown in Figure 4.9. We see that the mean of the parameter estimates of each of the parameters is very close to the true value. From Group 1 of which each sequence has 500 events to Group 4 of which each sequence has 5000 events, the variation of the parameter estimates consistently decreases. It suggests that with more events in the sequence, we get closer estimates to the true values, which indicates consistency.

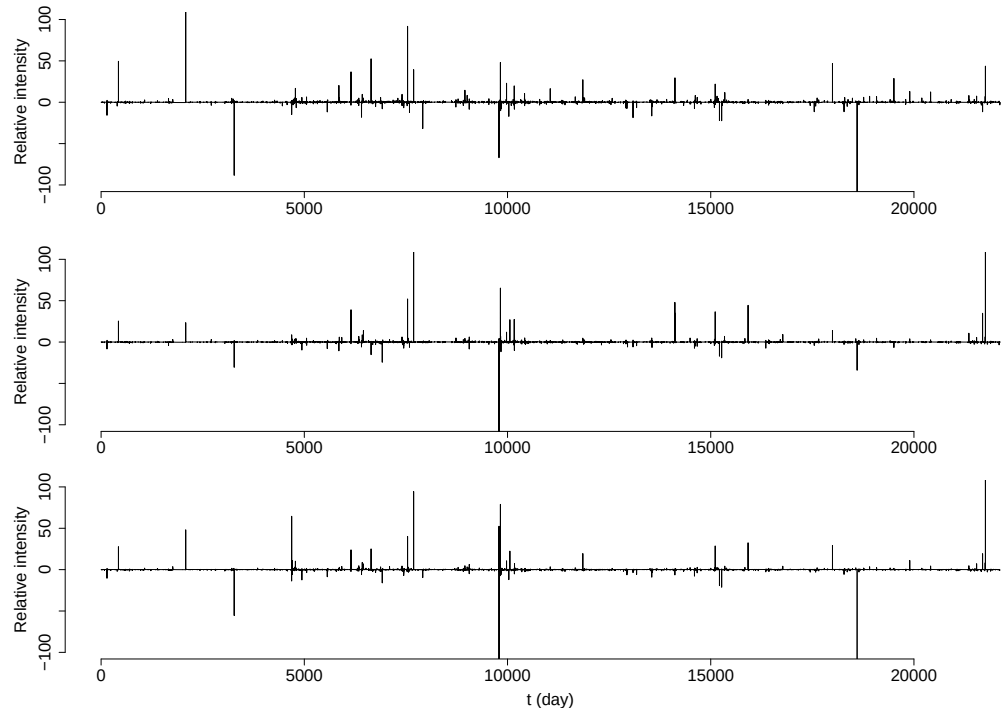


Figure 4.8: The estimated intensity function for the two-state (top), three-state (middle) and four-state (bottom) MMHPSDs (log-scaled) minus the true ETAS intensity function (log-scaled).

4.4 MMHPSD Investigation of Earthquake Data around Landers

4.4.1 Earthquake Data around Landers

The magnitude 7.3 Landers earthquake on June 28, 1992 is one of the most significant, and the most studied earthquakes in Southern California. It was preceded by the Joshua Tree earthquake of magnitude 6.1 on April 22, 1992, and followed by the magnitude 6.4 Big Bear earthquake three hours later, about 35km west of the Landers epicenter. The Landers aftershock sequence contains more than 10 large aftershocks of magnitude 5.0 or above. The 1999 magnitude 7.1 Hector Mine earthquake struck about 30 km northeast of the Landers source region 7 years later. There were three earlier earthquakes with magnitude larger than or equal to 5.0 near the Landers source region between 1976 and 1992, the March 15, 1979 $M = 5.2$ earthquake about 23km northwest of Joshua Tree, the July 8, 1986 $M = 5.7$ earthquake about 6km southwest of Morongo Valley, and the December 15, 1988 $M = 5.0$ earthquake about 12km northeast of Cabazon.

The seismicity in the region around the above events will be studied. In order to choose the magnitude threshold, the cutoff should ensure that the earthquakes above this magnitude follow the Gutenberg–Richter (G-R) law, because this is assumed in the ETAS formulation. Moreover, on the

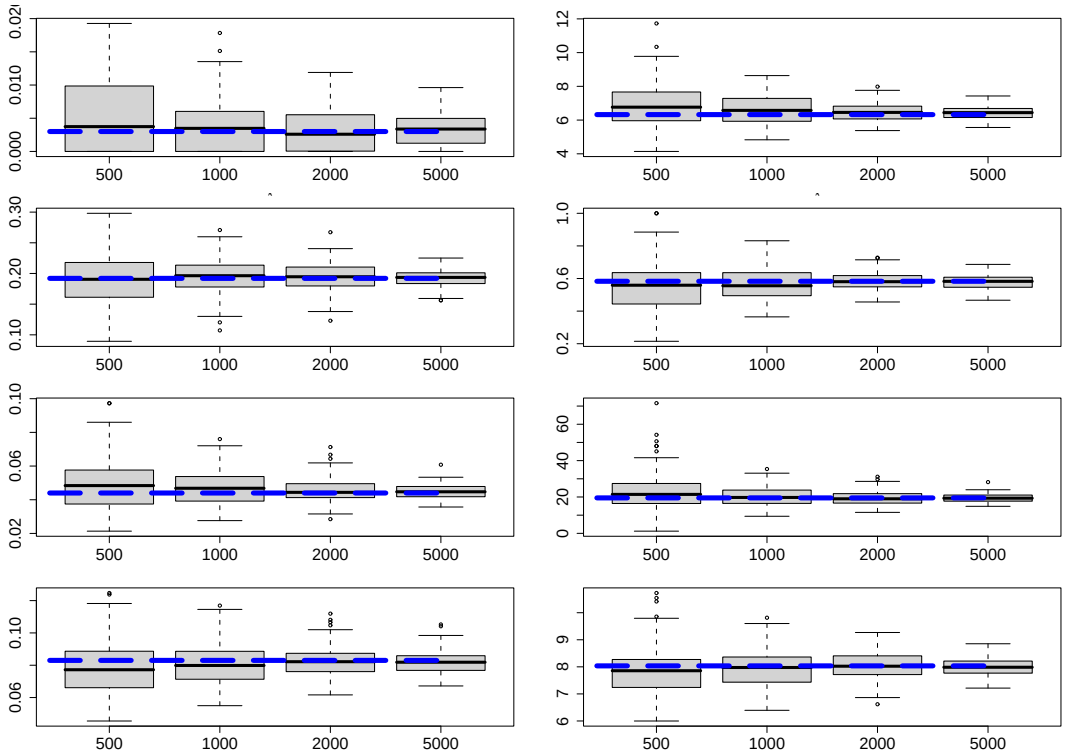


Figure 4.9: Boxplot of the estimated parameters $\hat{\lambda}_1$, $\hat{\lambda}_2$, $\hat{\nu}_1$, $\hat{\nu}_2$, $\hat{\eta}_1$, $\hat{\eta}_2$, $\hat{q}_1$ and $\hat{q}_2$ for each group. The true parameters are $\lambda_1 = 0.003$, $\lambda_2 = 6.324$, $\nu_1 = 0.192$, $\nu_2 = 0.583$, $\eta_1 = 0.044$, $\eta_2 = 19.502$, $q_1 = 0.083$ and $q_2 = 8.036$ which are indicated in the graphs using dashed lines.

one hand, a reasonably long enough data set is needed to make sure that there are several seismic cycles. On the other hand, the number of events should be less than 5000, otherwise the parameter estimation will be very time-consuming. Thus the earthquakes from Joshua Tree, Landers, Big Bear and Hector Mine sequences with minimum magnitude 3 from January 1, 1976 to December 31, 2008 are chosen. The area is between latitude from 33.8N to 34.8N, and longitude from 117.1W to 116.1W. The maximum depth of the selected earthquakes is 31km, and the majority of them are of local magnitude scale. The frequency–magnitude plot of the selected data is shown in Figure 4.10, which suggests that the selected earthquakes are complete according to the G-R law.

Some 2431 events are selected, including the magnitude 7.3 Landers earthquake which occurred on June 28, 1992 and the magnitude 7.1 Hector Mine earthquake on October 16, 1999. The data is from the SCSN catalogue which is available from the southern California earthquake data center website (http://www.data.scec.org/index.html, last accessed on September 19, 2009). The locations of the earthquakes are plotted in Figure 4.11. Figure 4.12 shows the cumulative number of earthquakes in the study area during the 33 years. We can clearly see several jumps which indicate seismicity rate changes. The MMHPSD will be fitted to this set of data and what kind of features

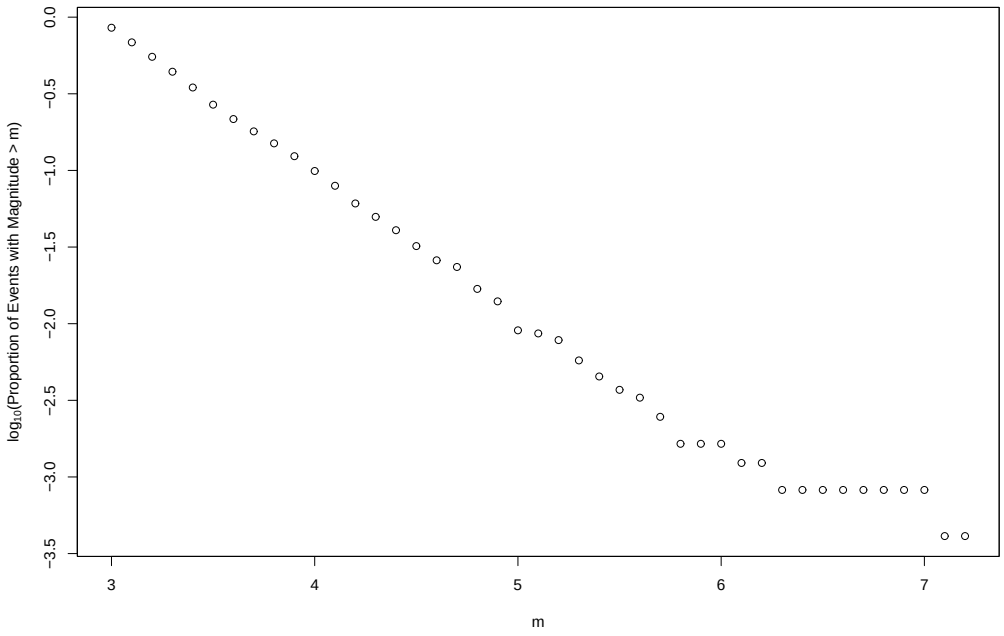


Figure 4.10: The frequency–magnitude plot of earthquakes with minimum magnitude 3.0 in the area between latitude 33.8N to 34.8N, and longitude 117.1W to 116.1W from January 1, 1976 to December 31, 2008.

this model captures will be discussed. After that, the data will be examined using the ETAS model with magnitude threshold $M_0 = 3.0$ for comparison.

4.4.2 Exploratory Data Analysis Using MMHPSDs

Starting from a two-state MMHPSD, one more state is added each time to investigate the seismicity rate changes of the study area from 1976 to the end of 2008, until the residual point process of the current model is a stationary Poisson process with unit rate. As mentioned in Section 3.5.2 in Chapter 3, if the residual point process is a stationary Poisson process with unit rate, then the model is a good approximation of the true model. For each model, the random time change $t \mapsto \tau$ is considered, i.e., for each occurrence time t_i, the transformed time

$$\tau_i = \Lambda(t_i) = \int_0^{t_i} \hat{\lambda}(s)\mathrm{d}s,$$

is calculated, which is called the residual process. The cumulative number of the residual process versus the transformed time $\{\tau_i\}$ is then plotted. The two-sided 95% and 99% confidence limits of the Kolmogorov–Smirnov statistics are indicated by the dashed lines. The cumulative number

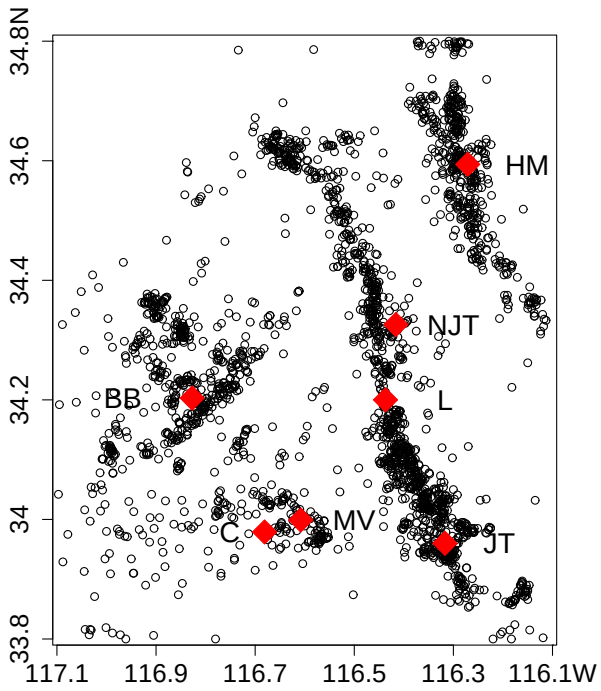


Figure 4.11: Location map of the earthquakes with minimum magnitude 3.0 in the area between latitude 33.8N to 34.8N, and longitude 117.1W to 116.1W from January 1, 1976 to December 31, 2008. The large symbols ♦ indicate the above mentioned earthquakes, the March 15, 1979 $M = 5.2$ earthquake (NJT), the July 8, 1986 $M = 5.7$ earthquake (MV), the December 15, 1988 $M = 5.0$ earthquake (C), the April 22, 1992 $M = 6.1$ Joshua Tree earthquake (JT), the June 28, 1992 $M = 7.3$ Landers earthquake (L), the June 28, 1992 $M = 6.4$ Big Bear (BB) and the October 16, 1999 $M = 7.1$ Hector Mine earthquake (HM).

versus the transformed time curves for the fitted MMHPSDs with 2 and 3 hidden states exceeded the 95% confidence limits not long before the Landers earthquake and then further exceeded the 99% confidence limits after the Big Bear earthquake (see, e.g., Figure 4.13 for the plot of the three-state MMHPSD). The residual process of the MMHPSD with 4 hidden states, however, is well approximated by the standard stationary Poisson process. Figure 4.14 shows that the curve of the cumulative number of events versus the transformed time lies close to the line of unit slope and between both the 95% and 99% confidence limits.

For the MMHPSD with 4 hidden states, for $k = 1, \cdots, n$, let

$$E_k = \tau_k - \tau_{k-1} = \Lambda(t_k) - \Lambda(t_{k-1}), \tag{4.4}$$

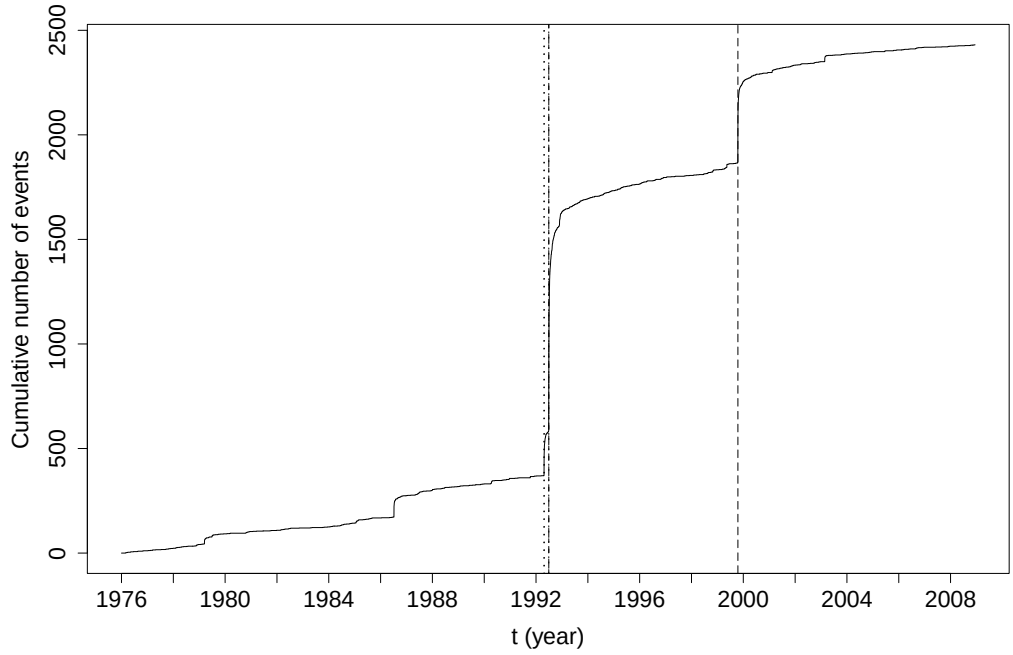


Figure 4.12: Cumulative number of earthquakes with minimum magnitude 3.0 in the area between latitude 33.8N to 34.8N, and longitude 117.1W to 116.1W from January 1, 1976 to December 31, 2008. The vertical lines from left to right respectively indicate the occurrence times of the magnitude 6.1 Joshua Tree earthquake on April 22, 1992, the magnitude 7.3 Landers earthquake on June 28, 1992, the magnitude 6.4 Big Bear earthquake on June 28, 1992, and the magnitude 7.1 Hector Mine earthquake on October 16, 1999.

and

$$U_k = 1 - \exp(-E_k). \tag{4.5}$$

We then use the Kolmogorov–Smirnov statistic to test whether U_k have a uniform distribution on $[0, 1)$. The empirical distribution of U_k is plotted in Figure 4.15. We see that the empirical distribution lies close to the theoretical uniform distribution on [0,1) and between both the 95% and 99% confidence limits. Therefore, the inter-arrival times are exponentially distributed.

The method suggested by Hall (1991) is adopted to verify the independence of the inter-arrival times. Figure 4.16 shows the plot of E_{k+1} versus E_k, which does not show any evidence of association between the two variables. Since the majority of the values of E_k are close to zero, we first take logarithm of them and then calculate the t-statistic of the hypothesis test with the null hypothesis that the correlation coefficient between $\log(E_k)$ and $\log(E_{k+1})$ equals zero. The t-statistic is -1.4514 with a P-value of 0.1468. Thus there is not enough evidence to reject the null hypothesis.

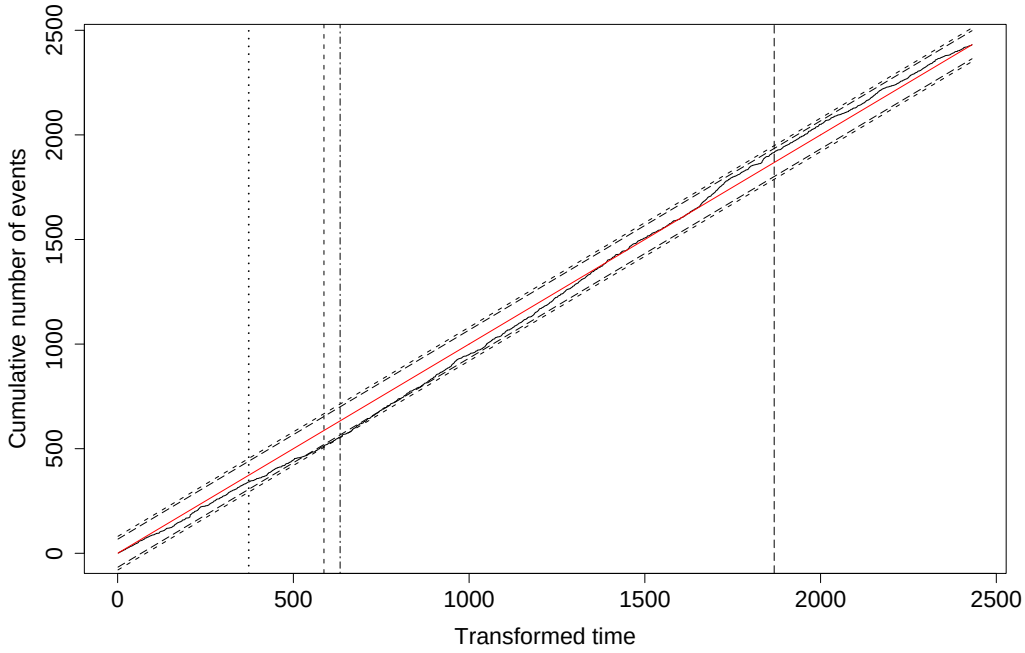


Figure 4.13: Cumulative number of the residual process versus the transformed time for the MMHPSD with 3 states fitted to the earthquakes around Landers. The solid straight line is of unit slope. The dashed lines indicate the two-sided 95% (longdash) and 99% (dashed) confidence limits of the Kolmogorov–Smirnov statistics. The vertical lines from left to right respectively indicate the occurrence times of the magnitude 6.1 Joshua Tree earthquake on April 22, 1992, the magnitude 7.3 Landers earthquake on June 28, 1992, the magnitude 6.4 Big Bear earthquake on June 28, 1992, and the magnitude 7.1 Hector Mine earthquake on October 16, 1999.

The scatter plot of U_{k+1} against U_k, as shown in Figure 4.17, suggests no particular pattern either. Therefore, we can say that the inter-arrival times are independent, and hence the fitted MMHPSD with 4 hidden states is a good approximation of the true intensity of the data.

The parameters estimated via the EM algorithm for the four-state MMHPSD are shown in Table 4.3, together with the parameter estimates for the fitted MMHPSDs with 2 and 3 hidden states for comparison. Note that the states are ordered according to the decay rate η. The estimated Q-matrices are, for the two-state model,

$$\hat{Q} = \begin{pmatrix} -0.0038 & 0.0038 \\ 1.6348 & -1.6348 \end{pmatrix}, \tag{4.6}$$

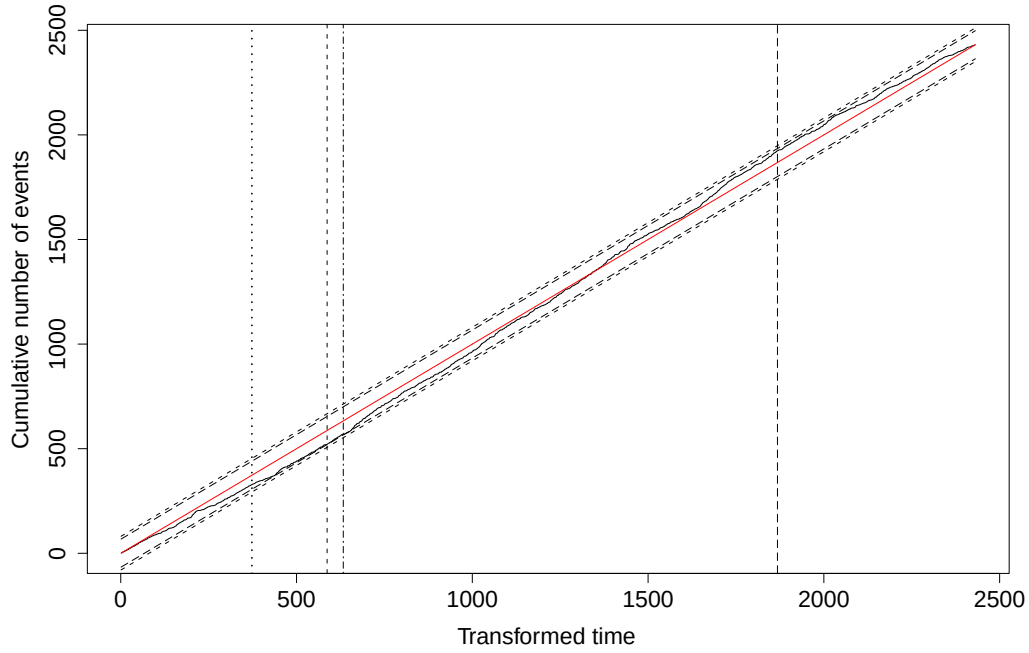


Figure 4.14: Cumulative number of the residual process versus the transformed time for the MMHPSD with 4 states fitted to the Earthquakes around Landers. The solid straight line is of unit slope. The dashed lines indicate the two-sided 95% (longdash) and 99% (dashed) confidence limits of the Kolmogorov–Smirnov statistics. The vertical lines are the same as those in Figure 4.13.

Table 4.3: Estimated parameters of the MMHPSDs fitted to the data around Landers.

State	1	2	3	4
$\hat{\lambda}$	0.025	9.325		
$\hat{\nu}$	0.380	0.818		
$\hat{\eta}$	0.069	21.049		
$\hat{\lambda}$	0.022	0.220	81.918	
$\hat{\nu}$	0.272	0.640	0.742	
$\hat{\eta}$	0.030	1.327	40.643	
$\hat{\lambda}$	0.022	0.000	0.783	154.098
$\hat{\nu}$	0.254	0.545	0.914	0.999
$\hat{\eta}$	0.026	0.521	19.286	188.787

for the three-state model,

$$\hat{Q} = \begin{pmatrix} -0.0051 & 0.0037 & 0.0014 \\ 0.1102 & -0.3499 & 0.2397 \\ 8.3578 & 16.1655 & -24.5233 \end{pmatrix}, \tag{4.7}$$

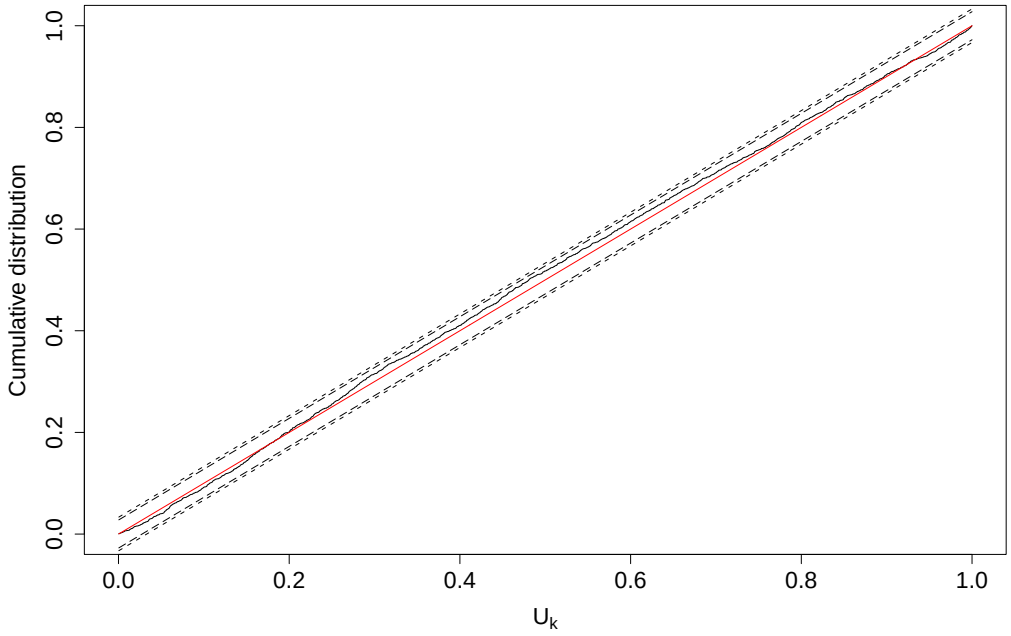


Figure 4.15: Empirical distribution of U_k (as in Equation (4.5)) for the fitted MMHPSD with 4 states to the Landers earthquakes. The solid straight line is of unit slope. The dashed lines indicate the two-sided 95% (longdash) and 99% (dashed) confidence limits of the Kolmogorov–Smirnov statistics.

and for the four-state model,

$$\hat{Q} = \begin{pmatrix} -0.0066 & 0.0000 & 0.0064 & 0.0002 \\ 0.0001 & -0.4677 & 0.0002 & 0.4674 \\ 0.0001 & 0.0643 & -6.1022 & 6.0378 \\ 58.0805 & 132.6127 & 121.2158 & -311.9090 \end{pmatrix}, \tag{4.8}$$

with stationary distribution $(0.9652, 0.0315, 0.0032, 0.0001)$. The estimated probability of the hidden state occupying each state and the estimated intensity function of each of the MMHPSDs are shown in Figures 4.18 to 4.20. Note that the residual analysis shows that the MMHPSD with 4 hidden states is the simplest MMHPSD model which captures the main features of the data. Hence it can be concluded that the four-state model is the 'best' in this sense. The log likelihood and BIC value for the fitted MMHPSDs with 2, 3 and 4 hidden states are listed in Table 4.4. Among the three models, the Bayesian information criterion suggests that the fitted MMHPSD with 4 hidden states is the best fit. Although the BIC may improve with additional states, these parameters cannot be

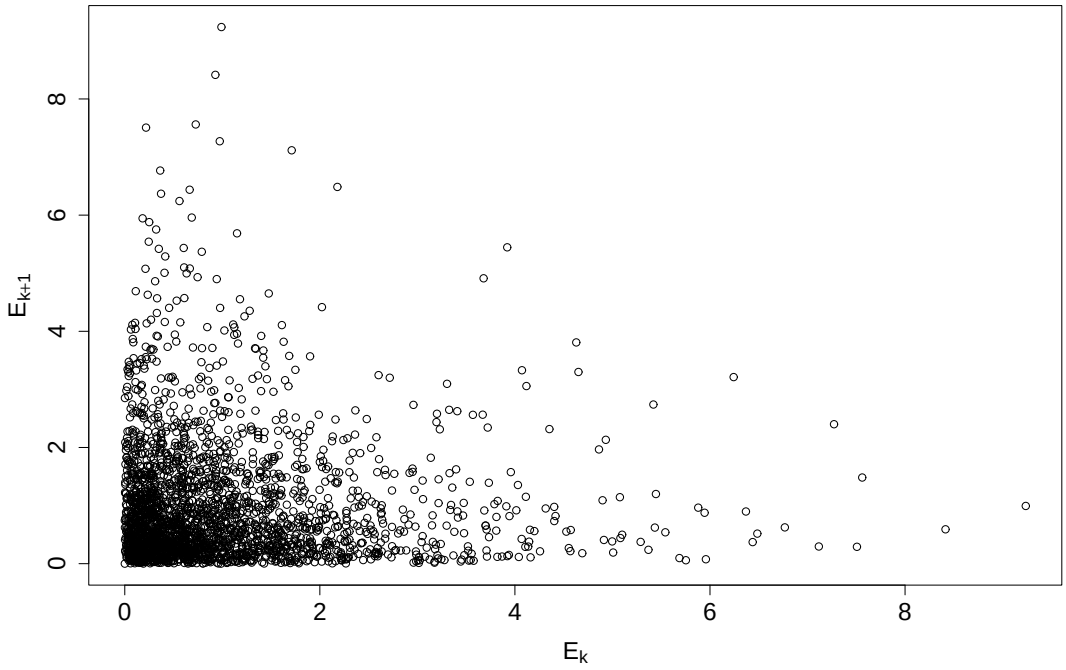


Figure 4.16: The scatter plot for E_{k+1} versus E_k (as in Equation (4.4)).

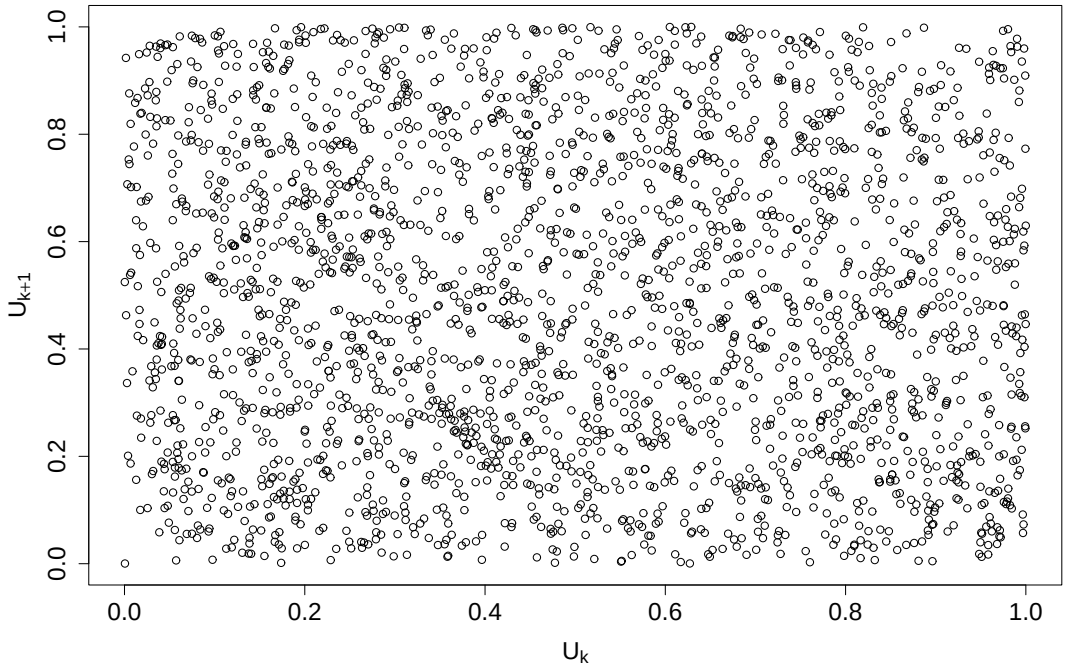


Figure 4.17: The scatter plot for U_{k+1} versus U_k (as in Equation (4.5)).

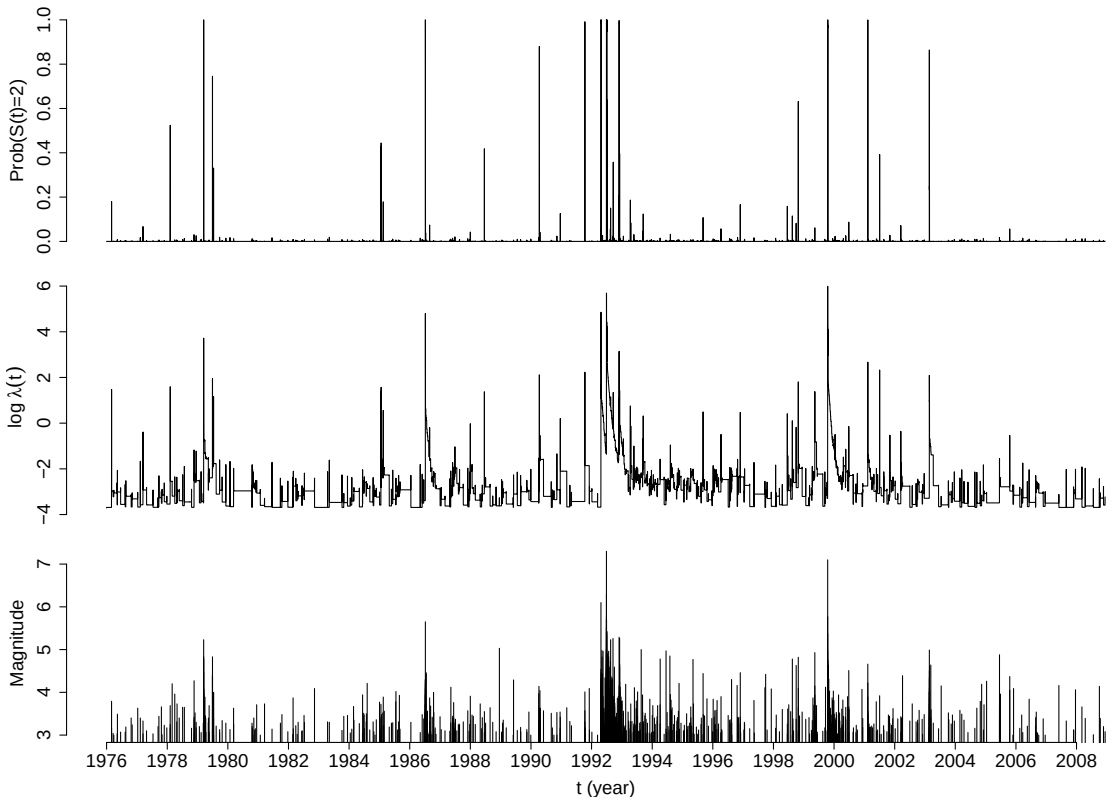


Figure 4.18: Illustration of the fitted two-state MMHPSD. Top: the estimated probability of the hidden state occupying state 2; middle: the estimated intensity function; bottom: the magnitude–time plot for the Earthquakes around Landers.

Table 4.4: The log likelihood (LL) and BIC for the fitted MMHPSDs respectively with 2 states (MMHP2S), 3 states (MMHP3S) and 4 states (MMHP4S), and the ETAS model.

	MMHP2S	MMHP3S	MMHP4S	ETAS
LL	1760.669	1942.465	2019.240	2020.587
BIC	-3458.966	-3767.983	-3851.365	-4002.192

justified by the residual analysis, which shows that the data is already explained by the four-state model.

The results for the fitted MMHPSDs with 2 and 3 hidden states are shown for comparison, though they fail to capture some main features of the data. They show a similar feature to that concluded from the analysis of the simulated ETAS sequence (cf., the pattern detected using MMHPSD for the ETAS sequence in Section 4.3.1). The ground state, State 1, in the MMHPSD with 2 hidden states remains in the MMHPSDs with 3 and 4 hidden states. However, both the background rate λ and the decay rate η for the largest state increase when we add more states in the model. We see

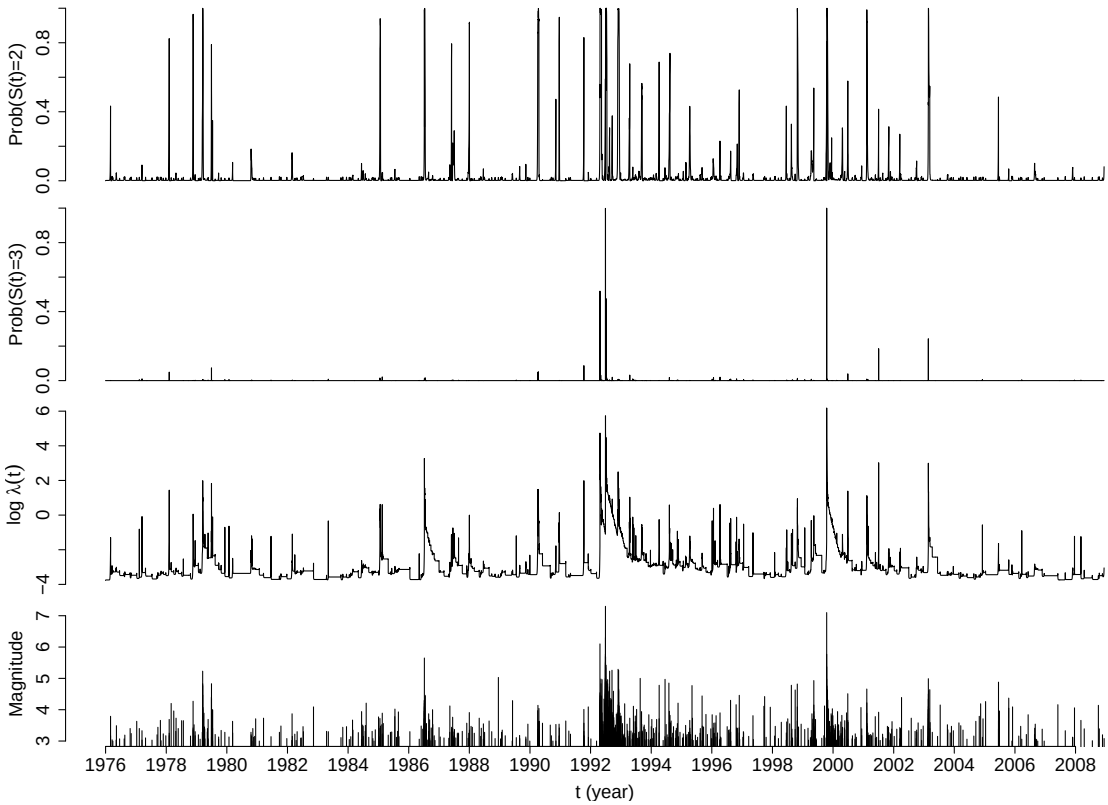


Figure 4.19: Illustration of the fitted three-state MMHPSD. Top two plots: the estimated probability of the hidden state occupying states 2 and 3; next two plots: the estimated intensity function; bottom: the magnitude–time plot for the Landers earthquakes.

from Table 4.3 and Figure 4.20 that State 4 corresponds to the Landers main shock and its immediate aftermath. In this state, there is a very high immigration rate λ, indicating that the events are not triggered by other events. The decay rate η is also very high, so these events do not themselves trigger other events, and $\nu \sim 1$, indicating a period of almost constant activity rate. This may relate to the multiple segment nature of the Landers event. State 3 appears to account for the other main sequence events, and for the major aftershocks, while State 2, with a negligible immigration rate and a small decay rate, is the principal aftershock state. Finally, State 1 has a low immigration rate, a very low decay rate, and a small value of ν. Coupled with the long sojourn time implied by (4.8), we see that this corresponds to a ground, or quiescent, state.

Further examining the transition matrix in (4.8), we see that State 4 has a very short sojourn time (approximately 5 minutes on average), and exits to one of the aftershock states. State 3 appears to possibly have some precursory properties for State 4, commensurate with the Joshua Tree – Landers triggering (Hill et al. 1993; 1995). This is further supported by the preferred transition from State 1 to State 3. In short, our seismic cycle is identified: State 1 (quiescent) – State 3 (precursory/main

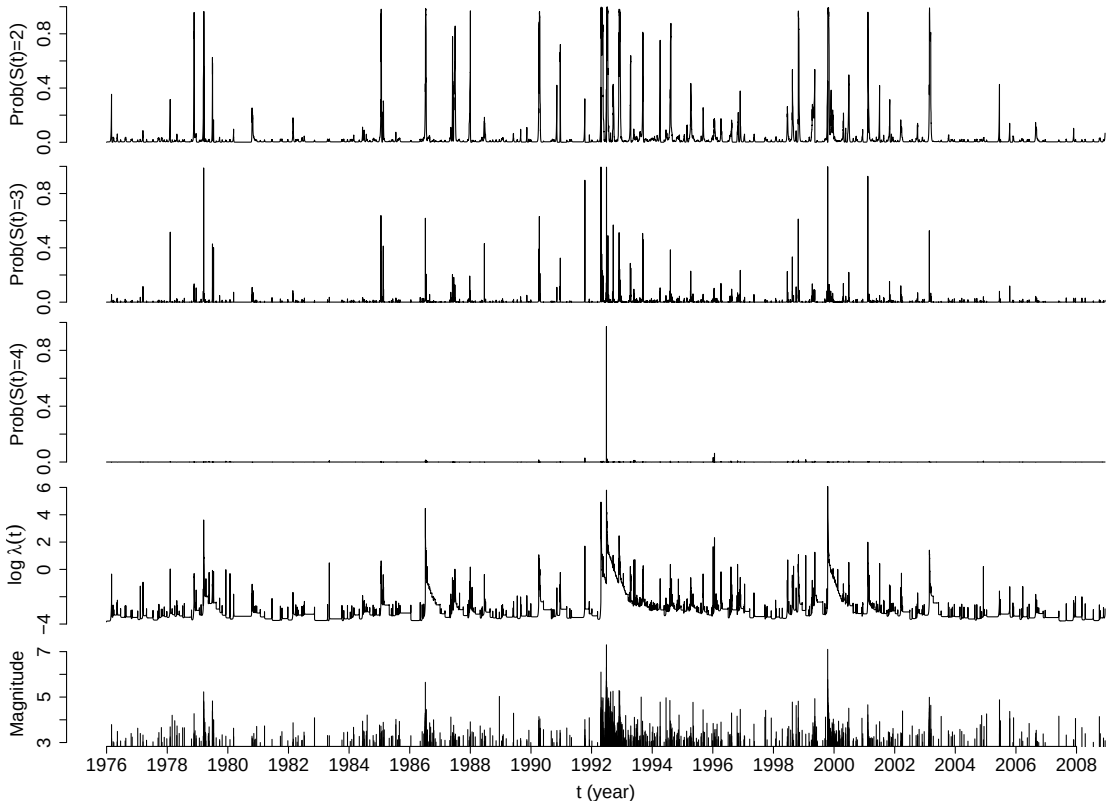


Figure 4.20: Illustration of the fitted four-state MMHPSD. Top three plots: the estimated probability of the hidden state occupying each state; next two plots: the estimated intensity function; bottom: the magnitude–time plot for the Landers earthquakes.

shock) – [State 4 (main shock) – State 3 (primary aftershocks)] – State 2 (aftershocks) – State 1, with the steps in brackets being optional, so to speak. A simulated sequence of 2500 events using the estimated parameters for the four-state MMHPSD is shown in Figure 4.21.

Again, we notice a similar cascading style to that discussed in Section 4.3.1. The conditional intensity function of a point process in time can be used to interpret the inter-event times. Larger intensities correspond to shorter inter-event times. From the above MMHPSD analysis of the data using different numbers of states, it is not difficult to notice that the inter-event times are highly related with magnitude. Take the fitted four-state model for example. State 4 with the highest conditional intensity function captures the feature of the largest earthquake; State 3, which has medium-large intensities, is related to the large-magnitude aftershocks; States 1 and 2 with low intensities correspond to smaller earthquakes. The inter-event times indirectly reflect the magnitude influence. Though we do not include the magnitude effect in our formulation of the conditional intensity function, the fitted models still account for the magnitude effect.

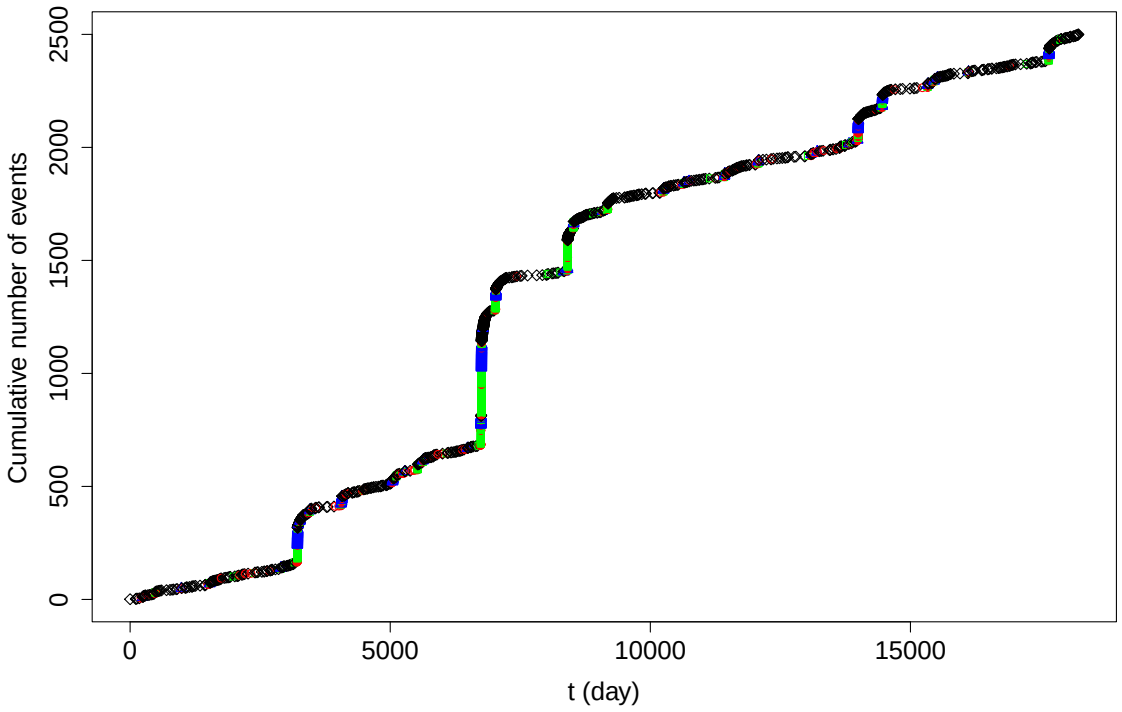


Figure 4.21: Cumulative curve of the simulated MMHPSD events using estimated parameters for the four-state MMHPSD fitted to the earthquakes around Landers. ◊: events in State 1, △: events in State 2, □: events in State 3, ○: events in State 4.

4.4.3 Comparison with the ETAS Model

Now let us examine the data using the ETAS model with magnitude threshold $M_0 = 3.0$. The ETAS model is fitted to the data and the maximum likelihood estimates of the parameters are listed in Table 4.5. The estimated intensity function as plotted in Figure 4.22 shows big spikes almost for

Table 4.5: Estimated parameters for the ETAS model fitted to the data around Landers.

Parameters	μ	K	α	c	p
Estimation	0.0208	1.4217	1.6265	0.0381	1.2230

each event, whereas the intensity functions of the MMHPSDs tend to have less and on average lower spikes. Recall Figure 4.4 which shows that for very small and very large $t - t_i$, the ETAS model always has larger decay rate than the Hawkes process. The decay rate of the ETAS model multiplied by the exponential term of the magnitude will therefore produce intensities very sensitive to event magnitude. We can see that not only do large events influence the ETAS conditional intensity

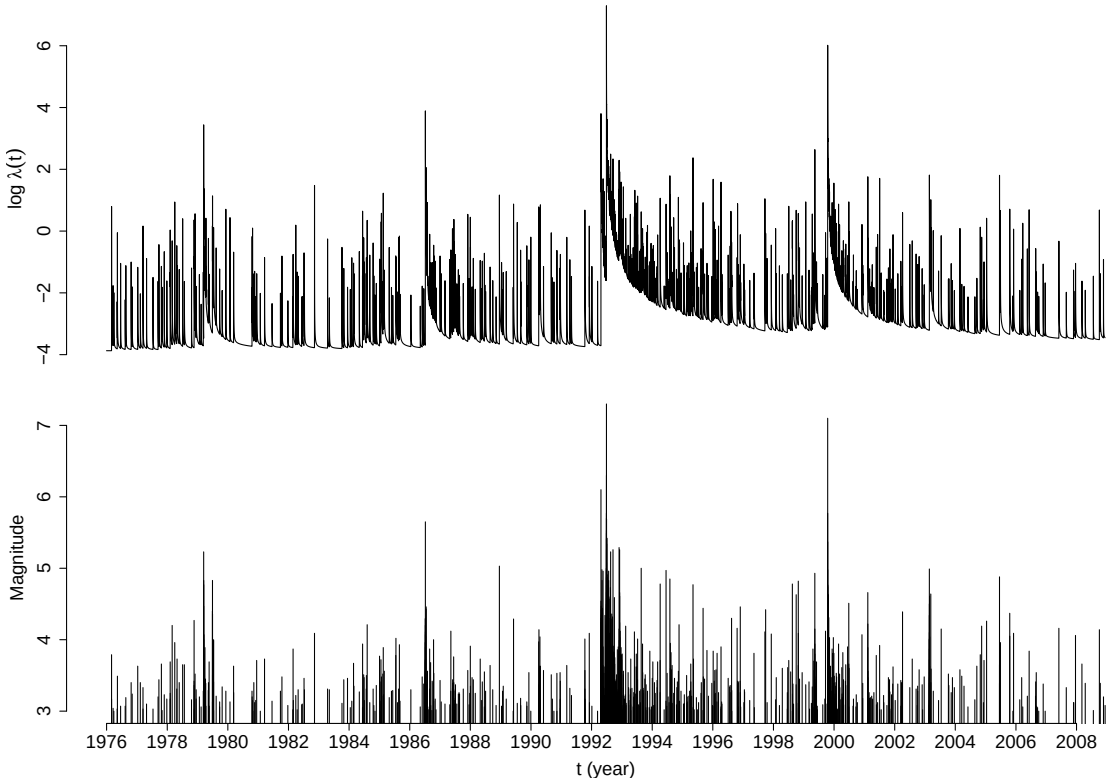


Figure 4.22: Illustration of the ETAS model fitted to the data around Landers.

function, events with small magnitudes also have strong effect on this intensity function. Comparing the parameter estimates of the MMHPSD with 4 hidden states and that of the ETAS model, we see that the very large background rates and decay rates of States 3 and 4 actually describe the fact that large magnitudes tend to result in short inter-event times.

The log likelihood and BIC value for the fitted ETAS model are listed in Table 4.4. We see that the fitted ETAS model has the largest log likelihood and the smallest BIC value among the four models. The BIC is a criterion for model selection among a class of parametric models with different numbers of parameters. However, for a point process, a smaller BIC does not guarantee that this model captures the main features of the data. In order to examine whether the intensity of the fitted ETAS model is a good approximation of the true intensity of the observed data, we examine the residual point process of the fitted ETAS model.

The cumulative number versus the transformed time plot for the fitted ETAS model as shown in Figure 4.23 lies above the line of unit slope and out of both the 95% and 99% confidence limits. The estimated distribution deviates significantly from the theoretical one. It suggests that the residual process of the ETAS fit is not a stationary Poisson process. Therefore, the ETAS intensity function

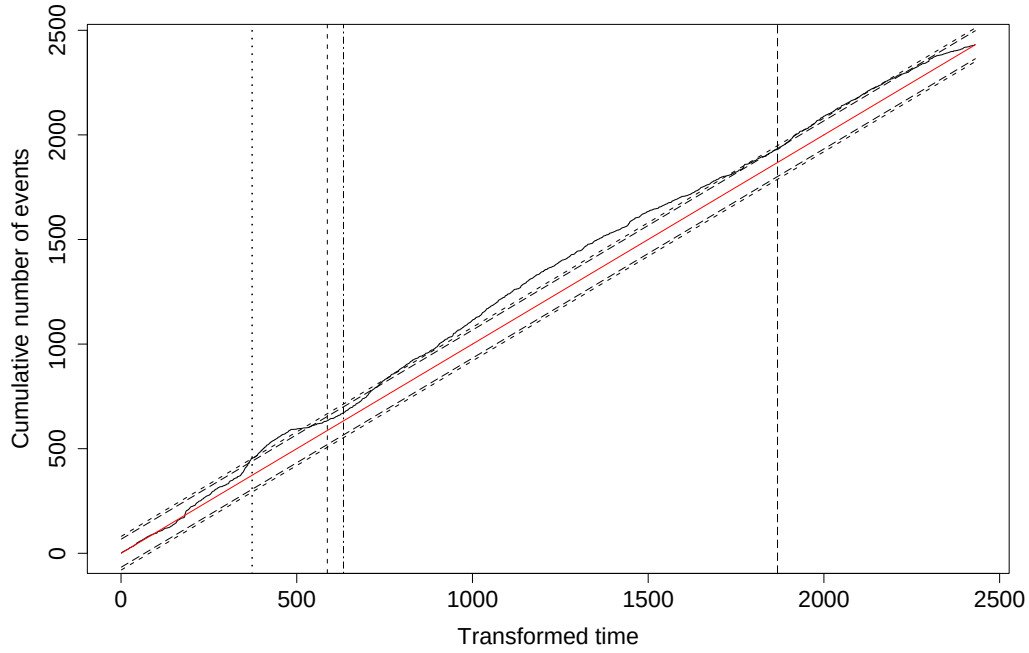


Figure 4.23: Cumulative number of the residual process versus the transformed time for the fitted ETAS model to the Landers earthquakes. The solid straight line is of unit slope. The dashed lines indicate the two-sided 95% (longdash) and 99% (dahsed) confidence limits of the Kolmogorov–Smirnov statistics. The vertical lines are the same as those in Figure 4.13.

does not capture the observed intensity well. Figure 4.23 clearly shows that after the Joshua Tree and Hector Mine earthquakes, there are sudden and short accelerations which are not well captured by the model. There is also a long acceleration following a period after the Big Bear earthquake. Such features may be corresponding to the activation of secondary aftershock clusters (cf., Marsan and Nalbant, 2005).

4.5 Conclusion and Discussion

In this chapter, the performance of the MMHPSD introduced in Chapter 3 is evaluated using both simulated and real earthquake catalogues. The simulation studies demonstrate that the EM algorithm performs reasonably well for the parameter estimation, though for a small data size (less than 1000), the parameter estimates may deviate from the true values. In general, the estimates seem to consistently capture the true parameters.

The simulation study using a simulated ETAS sequence suggests that different states of the MMHPSD seem to capture the magnitude effect parameterized as $e^{\alpha(M_i - M_0)}$ in the ETAS model.

The state with large intensities in the MMHPSD accounts for large earthquakes, while the state with low intensities corresponds to earthquakes with small magnitudes. The fitted MMHPSDs to this simulated sequence perform fairly well in modelling the seismicity of the data set. The comparison of the decay rate of the ETAS model with that of the Hawkes process suggests that for very small and very large time intervals, the ETAS model tends to produce larger intensities.

When we apply this model to the earthquakes from Joshua Tree, Landers, Big Bear and Hector Mine, this simple initial model demonstrates different states of seismicity rates. It provides an exploratory analysis of seismicity rate changes. The fitted four-state model suggests that before (some of) the large events, the March 15, 1979 $M = 5.2$ earthquake northwest of Joshua Tree, the July 8, 1986 $M = 5.7$ earthquake southwest of Morongo Valley, the December 15, 1988 $M = 5.0$ earthquake northeast of Cabazon, the April 22, 1992 $M = 6.1$ Joshua earthquake, the June 28, 1992 $M = 7.3$ Landers earthquake, the June 28, 1992 $M = 6.4$ Big Bear earthquake, and the October 16, 1999 $M = 7.1$ Hector Mine earthquake, there may exist relative quiescence. The seismicity rate increased after these events. The transition intensity matrix and the probability of staying in each state also suggest that there may be seismic shadows not long after the large events.

Note that in Marsan (2003) and Ogata et al. (2003), the seismicity rate changes were investigated for different subregions in the vicinity of Landers. Therefore, they provided temporal and spatial details of the seismic shadows. A possible improvement may be carried out for our analysis of the MMHPSD on the Landers sequences by dividing the whole area into different regions. We can then investigate the seismicity rate changes in each region. However, in order to get reliable estimates, the magnitude threshold may have to be lower than what we have chosen since there will be fewer events in smaller regions.

As discussed in Section 4.4.2, the MMHPSD captures the magnitude effect on the inter-event times, although we do not include a magnitude term in the intensity function. The state with larger intensities accounts for larger earthquakes or shorter inter-event times. The magnitude term in the ETAS model using an exponential function may be too strong along with the power law decay function. As a next step, a magnitude effect, $(M_i - M_0)^\alpha$, could be included in the intensity function of the MMHPSD. This may provide a better fit to the data and clearer state changes of the underlying Markov chain, given that the majority of the magnitude effect should be captured by the above introduced magnitude term.

Finally, we see that although the fitted ETAS model has the largest log likelihood and the smallest BIC value, it does not capture the main features of the events considered in this study (there

may still exist a better fit using the MMHPSD with more states, which has smaller BIC value than that of the ETAS model, but computational time precludes it from being included in this thesis). If the data is divided into several intervals each consisting of one main shock – aftershock sequence, and the ETAS model is then fitted to each sequence, we may get different sets of parameters for the sequences (e.g., Ogata et al., 2003). However, for the MMHPSD, each state acts as an attractor. If the features of the event occurrence pattern evolve towards a different attractor, the model will automatically switch into a different regime (or state), in a 'smooth' manner. We do not have to subjectively divide the entire sequence into different parts. Moreover, we see from Figure 4.23 that the most significant deviations of the curve for the cumulative number of the residual process versus the transformed time for the ETAS model from the line of unit slope are the two segments, the period after the Joshua Tree earthquake but before the Landers earthquake, and the period not long after the Big Bear earthquake. The two periods were shown to be relatively quiescent (Ogata et al., 2003; Marsan and Nalbant, 2005). The ETAS model fits well for the period immediately after the Landers earthquake. This suggests that the ETAS model captures well the features of an aftershock sequence, but may not be as good a fit for the quiescence period of a sequence.

This problem may be inherited from the assumption of the ETAS model. In the ETAS model, the frequency of the aftershocks triggered by an event with $\{t_i, M_i\}$ is assumed to be

$$\frac{K_i}{(t - t_i + c)^p}.$$

For different magnitudes M_i, K_i are different with

$$K_i = K e^{\alpha(M_i - M_0)}.$$

This suggests that for an event with magnitude M_0, the frequency of aftershocks triggered by the event is

$$\frac{K}{(t - t_i + c)^p};$$

and for a larger event with magnitude $M_i > M_0$, the frequency becomes

$$\frac{K e^{\alpha(M_i - M_0)}}{(t - t_i + c)^p},$$

which is $e^{\alpha(M_i - M_0)}$ times of that of an event with magnitude M_0. Note that the empirical formula $e^{\alpha(M_i - M_0)}$ was found for a sequence of aftershocks triggered by a main shock with magnitude M_i,

which describes the total number of aftershocks in the sequence. But in the ETAS model, this is assigned to all the events, including aftershocks and secondary aftershocks. This may result in a higher intensity than is consistent with the actual process. In the MMHPSD, however, we found distinguishable decay rates for different kinds of sequences, large for main shock activity, medium for aftershock activity, medium-small for secondary aftershocks and small for quiescence.

Part II

STATISTICAL ANALYSIS OF EARTHQUAKES WITH ANCILLARY DATA

Chapter 5

Statistics on Association Between Series of Events

5.1 Introduction

Correlation between two or more series of events is often an issue in statistics or other scientific subjects. For the case of earthquakes, the increasing availability of ancillary data relating to records of earthquake hazard processes, such as observations for groundwater level down wells and deformation measurements from GPS stations, has required techniques which can be used to examine if there is any association between these ancillary variables and earthquake occurrences. The ancillary data may provide insights into the study of earthquake hazard if there is indeed association between the two series of data. This chapter will review three very important and popular statistical methods for ascertaining the association between two point processes: coherence, mutual information and the Lin–Lin model. Previous literatures have compared the coherence with mutual information (e.g., Brillinger, 2002; Brillinger, 2003). However, to date, none have discussed the three approaches together. Note that the earthquake occurrences are point processes. The ancillary data, however, are time series, which can be transformed into 0-1 sequences, and hence point processes (cf., Section 6.4.3 in Chapter 6 and Section 7.2.1 in Chapter 7).

Coherence is a stationary process analog of the traditional correlation coefficient, taking on values between 0 and 1 at any given frequency. It is a measure of the time invariant linear dependence of the two processes at frequency λ, (Brillinger, 1975; Brillinger, 1994; Brillinger and Villa, 1997), and provides a measure of the degree of linear predictability of one process from another. When the coherence function is identically zero, one process is of no use in linearly predicting the other.

When it is identically one, one process gives a perfect linear prediction of the other (Brillinger, 1994; Iyengar, 2001). If it is significantly greater than zero over a limited frequency range, it implies association between the two processes over this frequency range.

Mutual information was originally introduced by Shannon (1948). It is a measure of the amount of information that one random variable contains about another (Cover and Thomas, 1991). The coherence is inadequate as a measure of general association for it may be identically 0 when two series are in fact related (Brillinger, 2003; Brillinger and Guha, 2007). Brillinger (2003) proved that such behavior does not occur for the coefficient of mutual information for random variables. The mutual information coefficient has the property of taking on the value 0 if and only if the variables are statistically independent. In this work, we transform the possible pairs of random variables, in the point process context, to pairs of intervals by considering the point process as a 0-1 time series, and then calculate the mutual information as a function of the time lag.

A significant cross correlation between two series of events, N_1 and N_2, doesn't mean that we can determine whether

(i) N_1 causes N_2;

(ii) N_2 causes N_1;

(iii) N_1 and N_2 cause each other; or

(iv) some other process causes both N_1 and N_2

(Ogata, 1999). To discriminate among the first three cases, as well as to test the significance, Ogata et al. (1982) suggested a parametric model based on the self-exciting and mutually exciting point processes introduced by Hawkes (1971). The model describes the influence of a point process input on a point process output, or in other words, it can identify linear causal relationships between one sequence of events and another (Ogata et al., 1982; Ma and Vere-Jones, 1997). Hence the model is named the Lin–Lin model. Unlike the previous two cases which can be defined very generally, this method applies specifically to point process models.

The data under consideration are functions, particularly realizations of stationary processes. In particular, two types of processes, 0-1 time series and point processes, are studied. In the following sections the three statistical methodologies, coherence, mutual information, and the Lin–Lin model, are briefly reviewed. They can be used to characterize the association between two or more series of events.

5.2 Coherence

Let (N_1, N_2) represent a bivariate point process. Let σ_m, $m = 0, \pm 1, \pm 2, \cdots$ and τ_n, $n = 0, \pm 1, \pm 2, \cdots$ denote the times of occurrences of the N_1 and N_2 events, respectively. For real-valued ω, let us set

$$d^T_{N_1}(\omega) = \int_0^T \exp\{-i\omega t\} \mathrm{d}N_1(t) = \sum_m e^{-i\omega\sigma_m} \tag{5.1}$$

$$d^T_{N_2}(\omega) = \int_0^T \exp\{-i\omega t\} \mathrm{d}N_2(t) = \sum_n e^{-i\omega\tau_n} \tag{5.2}$$

where T denotes the length of the time period of observation. Then one frequency domain measure of association which can be used to assess the linear dependency between the processes N_1 and N_2 is defined by

$$\lim_{T\to\infty} |\mathrm{Corr}\{d^T_{N_1}(\omega), d^T_{N_2}(\omega)\}|^2 = \lim_{T\to\infty} \frac{|\mathrm{Cov}\{d^T_{N_1}(\omega), d^T_{N_2}(\omega)\}|^2}{\mathrm{Var}\{d^T_{N_1}(\omega)\}\mathrm{Var}\{d^T_{N_2}(\omega)\}} = |R_{N_1N_2}(\omega)|^2, \tag{5.3}$$

which is called the coherence of the two processes at frequency ω, where

$$R_{N_1N_2}(\omega) = \lim_{T\to\infty} \mathrm{Corr}\{d^T_{N_1}(\omega), d^T_{N_2}(\omega)\}$$

is called the coherency at frequency ω. Note that the coherence is the coherency modulus-squared. This can be interpreted as the magnitude squared of the correlation between the finite Fourier transforms of processes N_1 and N_2. The definition of the correlation in variance and covariance terms,

$$\mathrm{corr}\{N_1, N_2\} = \mathrm{cov}\{N_1, N_2\}/\sqrt{\mathrm{var}\{N_1\}\mathrm{var}\{N_2\}},$$

leads to an alternative definition of the coherence as

$$|R_{N_1N_2}(\omega)|^2 = \frac{|f_{N_1N_2}(\omega)|^2}{f_{N_1N_1}(\omega) f_{N_2N_2}(\omega)}, \tag{5.4}$$

where $f_{N_1N_2}(\omega)$ is a hybrid cross-spectrum between the point processes N_1 and N_2, defined as

$$f_{N_1N_2}(\omega) = \lim_{T\to\infty} \frac{1}{2\pi T} \mathbf{E}\{d^T_{N_1}(\omega), \overline{d^T_{N_2}(\omega)}\}. \tag{5.5}$$

The auto-spectra $f_{N_1N_1}(\omega)$ and $f_{N_2N_2}(\omega)$ can be similarly defined.

Using the method of disjoint sections (Halliday et al., 1995), the complete record, denoted by T, is divided into L non-overlapping disjoint sections each of length K, where $T = LK$. The finite Fourier transform of the lth segment $(l = 1, \cdots, L)$ from process N_1 at frequency ω is defined as

$$d_{N_1}^K(\omega, l) = \int_{(l-1)K}^{lK} e^{-i\omega t} \mathrm{d}N_1(t) \approx \sum_{(l-1)K \leq \tau_n \leq lK} e^{-i\omega\sigma_m}. \tag{5.6}$$

A similar definition holds for $d_{N_2}^K(\omega, l)$. A consistent estimate of $f_{N_1N_2}(\omega)$ can be given by

$$\widehat{f}_{N_1N_2}(\omega) = \frac{1}{2\pi LK} \sum_{l=1}^{L} d_{N_1}^K(\omega, l)\overline{d_{N_2}^K(\omega, l)}, \tag{5.7}$$

with similar expressions for the estimation of the auto-spectra $\widehat{f}_{N_1N_1}(\omega)$ and $\widehat{f}_{N_2N_2}(\omega)$. For large K and $\omega \neq 0$, the estimated cross-spectrum $\widehat{f}_{N_1N_2}(\omega)$ can be interpreted as the covariance between the components, at frequency ω, of the processes N_1 and N_2. The estimated auto-spectrum, $\widehat{f}_{N_1N_1}(\omega)$, can be interpreted as the variance at frequency ω of the process N_1.

The coherence function can be estimated by direct substitution of the estimates of the appropriate spectra as

$$|\widehat{R}_{N_1N_2}(\omega)|^2 = \frac{|\widehat{f}_{N_1N_2}(\omega)|^2}{\widehat{f}_{N_1N_1}(\omega)\widehat{f}_{N_2N_2}(\omega)}. \tag{5.8}$$

Coherence functions provide a bounded and normative measure of association. In the case of independence, $|R_{N_1N_2}(\omega)|^2 = 0$, the distribution of $|R_{N_1N_2}(\omega)|^2$ can be evaluated in terms of the incomplete Beta function with parameters 1 and $(L-1)$ (Brillinger, 2001). The pointwise $100\alpha\%$ confidence limit is given by $1 - (1-\alpha)^{1/(L-1)}$. Thus we will include the level

$$1 - 0.05^{1/(L-1)} \tag{5.9}$$

in the following coherence plots as a benchmark of the upper 95% confidence limit under the hypothesis of independence. Estimated values of coherence lying below this line can be taken as evidence for the lack of a linear association between the two processes, i.e. that zero coherence is plausible at that frequency (Brillinger, et al., 1976; Halliday et al., 1995).

The second moment

$$d_{N_1}^T(\omega) = \int_0^T \exp\{-i\omega t\} \mathrm{d}N_1(t) \tag{5.10}$$

needs to be corrected by subtracting the mean $\bar{n}\mathrm{d}t$, (see, for example, Vere-Jones and Ozaki, 1982) where $\bar{n}$ is the average rate. Thus we can use

$$\begin{aligned} d^*_{N_1}(\omega) &= \int_0^T \exp\{-i\omega t\}(\mathrm{d}N_1(t) - \bar{n}\mathrm{d}t) \\ &= d^T_{N_1}(\omega) - \int_0^T \exp\{-i\omega t\}\bar{n}\mathrm{d}t \\ &= d^T_{N_1}(\omega) - \frac{\bar{n}i}{\omega}(\exp\{-i\omega T\} - 1). \end{aligned} \tag{5.11}$$

The estimates are

$$\widehat{d}^T_{N_1}(\omega) = \sum_n e^{-i\omega\sigma_m} \quad \text{and} \quad \widehat{\bar{n}} = \frac{N(T)}{T}, \tag{5.12}$$

where $N(T) = \#(0, T)$. Therefore we obtain

$$\widehat{d}^*_{N_1}(\omega) = \sum_n e^{-i\omega\sigma_m} - \frac{iN(T)}{\omega T}(\exp\{-i\omega T\} - 1). \tag{5.13}$$

For the method of disjoint sections, we use

$$\widehat{d}^*_{N_1}(\omega, l) = \sum_{(l-1)K<\sigma_m\leq lK} e^{-i\omega\sigma_m} - \frac{iN(K)}{\omega K}(\exp\{-i\omega lK\} - \exp\{-i\omega(l-1)K\}), \tag{5.14}$$

where $N(K) = \#((l-1)K, lK]$.

5.3 Mutual Information

The mutual information of a bivariate random variable (U, V) is defined as

$$I_{UV} = \mathbf{E}\left\{\log_2\left(\frac{p_{UV}(u,v)}{p_U(u)p_V(v)}\right)\right\},$$

where $p_{UV}(u, v)$ is the joint probability mass function, and $p_U(u)$ and $p_V(v)$ are the marginal probability mass functions. The mutual information is non-negative and measures the strength of dependence in that $I_{UV} = 0$ if and only if U and V are independent, $I_{UV_1} \leq I_{UV_2}$ if U is independent of V_1 given V_2, and for the continuous case, $I_{UV} = \infty$ if $V = g(U)$.

When a bivariate variable (U, V) has a continuous distribution, the mutual information is defined

as

$$I_{UV} = \int\int \log_2 \left(\frac{p_{UV}(u,v)}{p_U(u)p_V(v)} \right) p_{UV}(u,v) \mathrm{d}u\mathrm{d}v.$$

One popular mutual information estimator is obtained by substituting suitable density estimators $\widehat{p}_{UV}(u,v)$, $\widehat{p}_U(u)$ and $\widehat{p}_V(v)$ into the above formula (Strong et al., 1998; Antos and Kontoyiannis, 2001), which takes the form

$$\widehat{I}_{UV} = \int\int \log_2 \left(\frac{\widehat{p}_{UV}(u,v)}{\widehat{p}_U(u)\widehat{p}_V(v)} \right) \widehat{p}_{UV}(u,v) \mathrm{d}u\mathrm{d}v.$$

We can use either parametric density estimators (Brillinger, 2004) or nonparametric ones (either histogram-based, Moddemeijer, 1989; or kernel-based, Mars and van Aragon, 1982; Joe, 1989; Granger and Lin, 1994; Moon et al., 1995).

For a bivariate discrete variable (U,V) with U taking on the values $1,\cdots,K$ and V taking on $1,\cdots,J$ and

$$P\{U=k, V=j\} = p_{kj}.$$

The marginals are then

$$p_{k+} = P\{U=k\} = \sum_{j=1}^{J} p_{kj}, \quad p_{+j} = P\{V=j\} = \sum_{k=1}^{K} p_{kj},$$

and the mutual information becomes

$$I_{UV} = \sum_{k,j} p_{kj} \log_2 \frac{p_{kj}}{p_{k+}p_{+j}},$$

assuming that $p_{jk} \neq 0$. Let $w = \{w_{kj}\}$, with

$$w_{kj} = \begin{cases} 1, & U=k, V=j, \\ 0, & \text{otherwise.} \end{cases}$$

Suppose that there are n independent realizations, $\{w_{kjl}, l=1,\cdots,n\}$, of w. The maximum likelihood estimates of p_{kj} are $\widehat{p}_{kj} = \sum_l w_{kjl}/n$ and the plug-in estimate of the mutual information (Brillinger, 2004; Brillinger and Guha, 2007) is

$$\widehat{I}_{UV} = \sum_{k,j} \widehat{p}_{kj} \log_2 \frac{\widehat{p}_{kj}}{\widehat{p}_{k+}\widehat{p}_{+j}}, \tag{5.15}$$

where $\widehat{p}_{k+} = \sum_j \widehat{p}_{kj}$ and $\widehat{p}_{+j} = \sum_k \widehat{p}_{kj}$. Note that a point process can be transformed into a 0-1 time series, with 1 at the occurrence times (of the process) and 0 otherwise. The likelihood ratio test statistic of the null hypothesis that the two variables U and V are independent is

$$G^2 = 2n \sum_{k,j} \widehat{p}_{kj} \log_2 \frac{\widehat{p}_{kj}}{\widehat{p}_{k+}\widehat{p}_{+j}}$$

(Christensen, 1997; Brillinger, 2004; Brillinger and Guha, 2007). The asymptotic null distribution of G^2 under the hypothesis of the independence of U and V is $\mathcal{X}^2_{(J-1)(K-1)}$. Noticing the proportional relationship between the estimate $\widehat{I}_{UV}$ in (5.15) and G^2, we have that under the null hypothesis that the two variables U and V are independent, the large sample distribution of the estimate (5.15) is $\mathcal{X}^2_{(J-1)(K-1)}/2n$.

5.4 Ogata's Lin–Lin Model

Consider a point process $\{N_t\}$ with intensity function

$$\lambda(t) = \mu + \int_0^t g(t-s)\mathrm{d}N_s + \int_0^t h(t-s)\mathrm{d}X_s, \tag{5.16}$$

where $\{X_t\}$, the input process, may be either a point process or a cumulative process

$$X_t = \int_0^t x(s)\mathrm{d}s$$

of a stochastic process $x(t)$. The self-exciting term $g(t)$ describes the nature of the point process, while the transfer term $h(t)$ indicates the strength of the causal relation between the input process X_t and the output N_t. When the function $h(t) \equiv 0$, it means there is no causal relationship between the input and output processes, whereas when the function $g(t) \equiv 0$ and $h(t) \neq 0$, it means that the output process is a doubly stochastic Poisson process whose intensity is modulated only by the input process (Ogata et al., 1982). Ogata et al. (1982) proposed the following parameterization of the two response functions

$$g(t) = \sum_{k=1}^{K} a_k t^{k-1} e^{-ct}, \tag{5.17}$$

$$h(t) = \sum_{k=1}^{L} b_k t^{k-1} e^{-ct}. \tag{5.18}$$

Given the occurrence times of two sequences of events $\{t_i : i = 1, \cdots, I\}$ (output) and $\{\tau_m : m = 1, \cdots, M\}$ (input) over the time interval $[0, T]$, the parameters are estimated by maximizing the partial log likelihood

$$\log L_T(\theta) = \sum_{i:0 \le t_i < T} \log \lambda_\theta(t_i) - \int_0^T \lambda_\theta(t) \mathrm{d}t,$$

where θ denotes $(\mu, c, a_1, \cdots, a_K, b_1, \cdots, b_L)$. To determine the orders K and L of the response functions, we use Akaike's information criteria (AIC; Akaike, 1974)

$$\mathrm{AIC}(K, L) = -2 \max_\theta (\log L_T(\theta)) + 2(K + L + 2),$$

and choose K and L which minimize the AIC.

Ogata (1983) suggested the following model to examine case (iv),

$$\lambda(t) = a_0 + P_J(t) + C_K(t) + \sum_{t_i < t} g_N(t - t_i) + \sum_{\tau_m < t} h_X(t - \tau_m).$$

The second term on the right-hand side

$$P_J(t) = \sum_{j=1}^{J} a_j \phi_j(t/T), \quad 0 < t < T,$$

represents the evolutionary trend. The total length of the observed interval is T, and $\phi_j(\cdot)$ is a polynomial of order j. The third term

$$C_K(t) = \sum_{k=1}^{K} \{b_{2k-1} \cos(2k\pi t/T_0) + b_{2k} \sin(2k\pi t/T_0)\}$$

is the Fourier expansion for cyclic effects with a given fixed cycle length T_0. The fourth term is the clustering effect, i.e., a response function of the output process. The last term describes the causal relation from the input process to the output process. The last two terms are parameterized as in (5.17) and (5.18), but with different exponential decay rates for the input and output processes (see e.g. Ma and Vere-Jones, 1997). If there is no causal relation from $\{\tau_m\}$ to the conditional intensity function $\lambda(t)$, or to the occurrence of $\{t_i\}$, then $h_X(x) = 0$ is expected. Maximum likelihood estimation can be used to estimate the parameters and the AIC can be used to determine the order(s).

Wang et al. (2010) presented a comparative study of the three statistics using the groundwater level data at Tangshan Well in northern China and earthquakes from the global catalogue. A strong association between the groundwater level oscillations and the earthquake occurrences was found, in which the earthquake occurrences appear to trigger water level oscillations with a time delay. The time delay may be related to the seismic wave travel times from earthquakes to the well. The next chapter will provide a detailed investigation of the linkage and causal relationship between the groundwater level oscillations and some prominent seismic phases using the three statistics.

Chapter 6

Transient Coseismic Responses at Tangshan Well

6.1 Introduction

Earthquakes disrupt and change hydrologic processes. Drastic hydrologic anomalies have been observed after earthquakes, such as oscillations or sustained changes in groundwater levels in wells (e.g., Montgomery and Manga, 2003, and references therein).

Sustained water level changes in distant wells have been demonstrated by, for example, Coble (1965), Bower and Heaton (1978), Igarashi et al. (1992), and Roeloffs (1998). Using larger samples, King et al. (1999) and Wang et al. (2004) showed that these changes could be from a number of physically different classes. However, statistical analysis (Matsumoto, 1992; Kitagawa and Matsumoto, 1996; Matsumoto et al., 2003) showed that such sustained changes occur as the consequence of a very small proportion of earthquakes (approximately one in 500 and one in 2500 in the latter two studies). They also suggested that there may exist a threshold combining magnitude and well–epicenter distance above which such changes are likely.

However, in addition to sustained changes in groundwater level, earthquakes can induce, via the passage of seismic waves, oscillations in water level with duration of up to a few hours. While sustained changes in groundwater level appear to be rare, transient oscillations of groundwater level are considered much more common. The amplitude, and hence detection, of such oscillations appears to be a function of the characteristics of the well–aquifer system (Cooper et al., 1965; Kunugi et al., 2000). The existing literature mainly discusses groundwater level changes at one well related to seismic waves of a few distant large earthquakes (Liu et al., 1989; Ohno et al., 1997; Brodsky et al.,

2003), or at several wells related to one or two earthquakes (Kitagawa et al., 2006; Ramana et al., 2007; Chia et al., 2008). Although Cooper et al. (1965) (see also King et al., 1999; Kitagawa et al., 2006) noted large variability in the response of different wells to a single earthquake, there appears to have been no systematic investigation of the variation in the response of a single well to multiple earthquakes. No statistical investigation in the spirit of those by Kitagawa and Matsumoto appears to have been conducted to investigate the variability of responses, what seismic wave arrivals they can be associated with, and what characteristics of the earthquake influence them.

In this chapter a statistical examination of transient coseismic responses to distant earthquakes will be conducted using four years of groundwater level measurements (sampling frequency per minute, approximately two million data) and the global catalogue of earthquakes of minimum magnitude 6.0 during the same period (600 events). The aim is to first construct an algorithm to automatically detect oscillations in groundwater level, and extract these as 'signals'. Secondly, the statistical methodologies reviewed in Chapter 5 will be adopted to investigate the relationship between these signals and the arrival times of various seismic waves, in particular the earliest P phase, S phase, Love Wave and Rayleigh Wave arrivals. Once the candidate wave which is most likely related to the initiation of groundwater oscillations has been found, it will be used to identify those oscillations that appear to be coseismics in nature, as opposed to deriving from other causes. The former will then provide the basis for investigating the detection probability, and the well signal characteristics, as a (statistical) function of the earthquake characteristics.

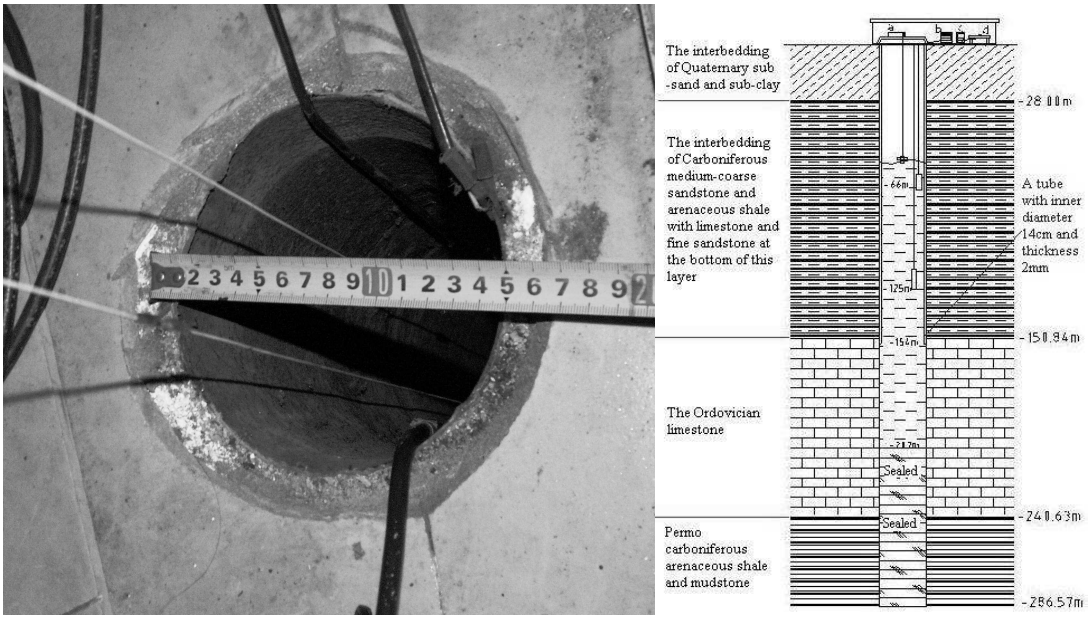
The earthquake characteristics are chosen rather than the amplitudes of seismic waves in the Tangshan area for two reasons. The relationship between groundwater level movements and seismic waves is well understood (e.g., Cooper et al., 1965; Bower and Heaton, 1978; Liu et al., 1989; Kunugi et al., 2000). Hence the question of interest is rather the relationship between the earthquake, and its manifestation in the distant region of the well, be that measured by seismic waves or groundwater level movement. The analysis of the latter, while still very data-intensive, is more feasible than examining thousands of seismograms, recorded at much higher frequency.

In the next section, the data and the pre-processing techniques used to 'clean' it prior to the analysis proper will be described. Following is the discussion of how to utilize statistical analysis to extract signals from the groundwater level record. Using the signals thus extracted, in Section 6.4, a dispersion test of Poisson process for both the well signal series and the earthquake series are used to examine the cluster properties of the two sequences. Then the coherence and mutual information are adopted to calculate the correlation between the well signals and the various seismic wave arrivals,

and the Lin–Lin model, a point process technique, is applied to identify which arrival is the most probable initiator of the oscillations. The detection threshold is examined both directly, and via logistic regression, in Section 6.5. A meta-analysis using multiple regression is conducted in the penultimate section in an attempt to quantify possible relations between earthquake and well signal characteristics. This chapter concludes with a discussion of the results.

6.2 Data

There are more than thirty years of groundwater level observations at the Tangshan Well, a well located in DaZhao Park in Tangshan City, 100km southeast of Beijing, China, with longitude 118.18E, latitude 39.62N and altitude 23.43m above sea level, which is located along the Tangshan fault. The well (see Figure 6.1(a)) was drilled by the Geology Department of Tangshan Mining, Kailuan Mining Administration in 1969 for the purpose of hydrologic observation, with the alternative name Shanxi Water 2. The borehole drilling terminated on October 30, 1969, at a depth of 286.57m.



(a) (b)

Figure 6.1: Tangshan Well (a) and its columnar section (b) (Li Ma, personal communication).

The alluvium thickness is 10.27m. After drilling, the borehole was sealed at 207.00m depth, which was the water depth. A tube with inner diameter 14cm and thickness 2mm was set in the borehole from the ground surface to 154.00m depth. For a detailed geological column of the well, see Figure 6.1(b). The natural period of the well–aquifer system was 21.9s in 2002 (Zhang et al., 2002),

but this varies with the depth of the well water.

Such a long record as that from Tangshan Well is rare to the best of our knowledge (cf. Elkhoury et al., 2006). To date there has been no coherent statistical examination of the possible link between earthquakes and the hydrological response of a well with potentially hundreds of responses, and little if any statistical analysis has been done for the data at Tangshan Well.

The observation frequency has increased over time, beginning with daily observations of water level on January 1, 1974. From January 1, 1981, analog observations with hourly sampling have been made using monitoring equipment SW40-1 with 1mm precision. Digital monitoring equipment LN3 for the water level began operating on September 16, 2001. These data have precision 0.0005m with a sampling rate of one observation per minute, and the available record extends to the end of 2005. The time series analysis (state-space model) methods (Kitagawa and Gersch, 1984; Kitagawa and Matsumoto, 1996) are not feasible as, after decomposing the data into air pressure, rainfall and earth tide responses, the dominant component of the residual water level appears to be related to groundwater pumping for industrial use, of which there is no detailed record. While we are hence unable to investigate whether or not Tangshan Well experiences sustained changes in water level, the question of transient changes remains open. From a sampling inspection of the water level data, including immediately following the largest and closest earthquakes, water level oscillations in the well appear to last for at most a few hours. This implies that hourly sampled data will not be sufficient for detecting and examining the oscillations at this well. However, the availability of data observed at minute intervals provides the opportunity to examine coseismic responses which are transient, rather than steps, in the water level, over the period 2002–2005 (in total 2,103,840 data, with 88,897 missing). Missing data is due to a number of causes – electrical failure, sensor adjustments, and malfunctions such as sensor, water-level recorder or data communication unit failure (Baojun Yin, personal communication).

The four year record of water level y_t is plotted in Figure 6.2, including the occurrence times of the largest earthquakes during the period, as listed in Table 6.1. Features worthy of remark include that the 27 earthquakes with minimum magnitude 7.5 during this period appear to have no sustained effect on water level, and little in the way of trend or seasonal variation. Note the break in the record in September 2002, where there is no record for a 10 day period (14,400 minutes). The erratic behavior in February and March 2003, and in October 2005 is due to sensor adjustments and malfunctions.

The clock time at Tangshan Well is synchronized every week to the same GPS time by which

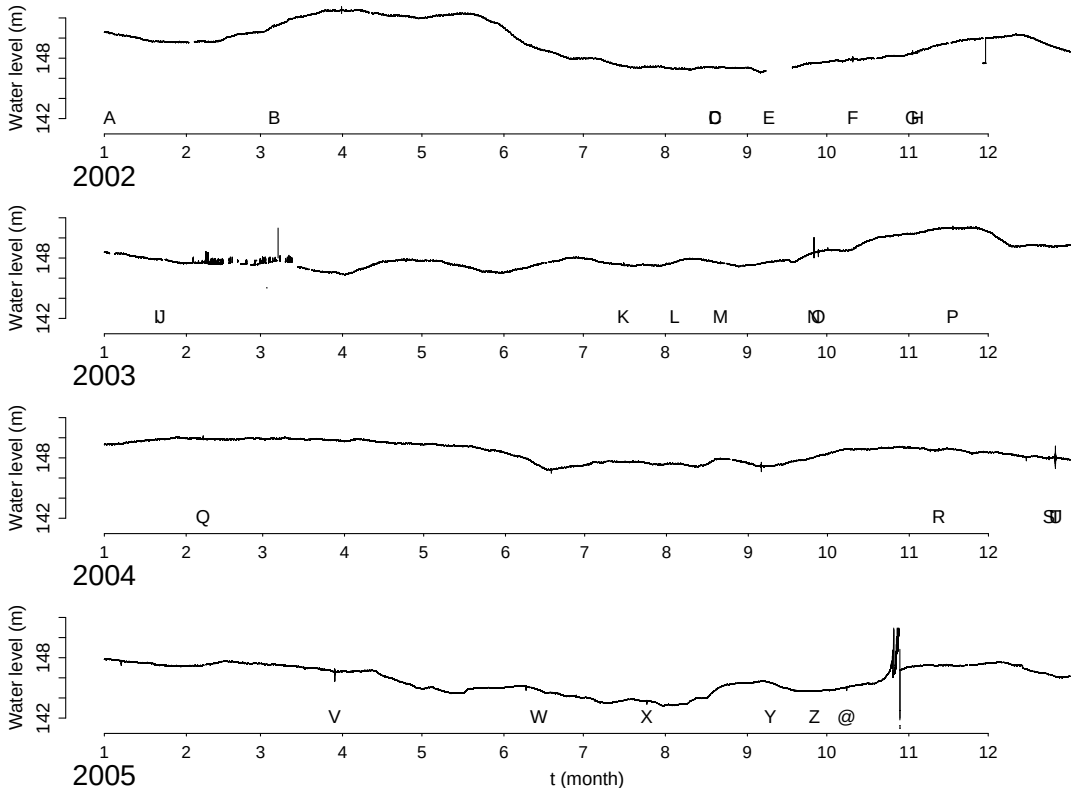


Figure 6.2: Groundwater level record at Tangshan Well from January 1, 2002 to December st, 2005. The symbols A–Z and @ indicate the occurrence of earthquakes with magnitude greater than or equal to 7.5 (details in Table 6.1).

earthquake origin times are measured in the NEIC catalogue. Between adjustments the clock time can lag behind the GPS time, with an error always less than 60 seconds, usually less than 30 seconds (Baojun Yin, personal communication). Thus the average time reading error of groundwater level record at this well is less than 30 seconds. Since we are considering data sampled by minute, we chop the earthquake occurrence times to minutes (i.e., omit the seconds), with a consequent average time reading error for earthquakes of approximately 30 seconds. The well error will lead to water level movements occurring 'early', i.e., the seismic waves arriving 'late', whereas the chopping of earthquake times makes the seismic waves arrive 'early'. Thus the two errors tend to cancel, and the maximum error in either direction is 1 minute. To test whether the sum of these time reading errors affects our analysis, we will perform sensitivity tests by moving the water level record 1 minute forward or 1 minute backward, and repeating all analyses. The results will be compared with that from the original water level data.

Table 6.1: Earthquakes obtained from the USGS-NEIC catalogue (http://neic.usgs.gov/neis/epic/epic_global.html) with minimum magnitude 7.5 from January 1, 2002 to December 31, 2005, corresponding to the symbols A–Z and @ in Figure 6.2.

Symbol	Date yyyymmdd	Time	Latitude	Longitude	Depth (km)	Mag	Region	Dis to well (km)	Azimuth (°)
A	20020102	17:22:48.76	-17.60	167.86	21	7.50	Vanuatu Islands	8184	229.1
B	20020305	21:16:09.13	6.03	124.25	31	7.50	Mindanao, Philippines	3784	190.7
C	20020819	11:01:01.19	-21.70	-179.51	580	7.70	Fiji Region	9389	235.4
D	20020819	11:08:24.31	-23.88	178.49	675	7.70	South of the Fiji Islands	9428	232.6
E	20020908	18:44:23.71	-3.30	142.95	13	7.80	Near North Coast of New Guinea, P.N.G.	5402	213.7
F	20021010	10:50:20.57	-1.76	134.30	10	7.70	Irian Jaya Region, Indonesia	4885	203.4
G	20021102	01:26:10.70	2.82	96.08	30	7.60	Northern Sumatra, Indonesia	4660	145.5
H	20021103	22:12:41.00	63.52	-147.44	4	8.50	Central Alaska	6338	327.7
I	20030120	08:43:06.07	-10.49	160.77	33	7.80	Solomon Islands	7094	227.4
J	20030122	02:06:34.61	18.77	-104.10	24	7.60	Offshore Colima, Mexico	12179	317.4
K	20030715	20:27:50.53	-2.60	68.38	10	7.60	Carlsberg Ridge	6906	119.7
L	20030804	04:37:20.13	-60.53	-43.41	10	7.60	Scotia Sea	17365	156.9
M	20030821	12:12:49.79	-45.10	167.14	28	7.50	South Island of New Zealand	10611	211.9
N	20030925	19:50:06.36	41.81	143.91	27	8.30	Hokkaido, Japan Region	2173	284.7
O	20030927	11:33:25.08	50.04	87.81	16	7.50	Southwestern Siberia, Russia	2635	53.8
P	20031117	06:43:06.80	51.15	178.65	33	7.80	Rat Islands, Aleutian Islands, Alaska	4749	305.9
Q	20040207	02:42:35.21	-4.00	135.02	10	7.50	Near the South Coast of Papua, Indonesia	5147	203.4
R	20041111	21:26:41.15	-8.15	124.87	10	7.50	Kepulauan Alor, Indonesia	5356	188.5
S	20041223	14:59:04.41	-49.31	161.35	10	8.10	North of Macquarie Island	10756	206.2
T	20041226	00:58:53.45	3.30	95.98	30	9.00	Off the West Coast of Northern Sumatra	4616	144.9
U	20041226	04:21:29.81	6.91	92.96	39	7.50	Nicobar Islands, India Region	4423	138.0
V	20050328	16:09:36.53	2.09	97.11	30	8.60	Northern Sumatra, Indonesia	4685	147.2
W	20050613	22:44:33.90	-19.99	-69.20	115	7.80	Tarapaca, Chile	17721	339.7
X	20050724	15:42:06.21	7.92	92.19	16	7.50	Nicobar Islands, India Region	4371	136.3
Y	20050909	07:26:43.73	-4.54	153.47	90	7.70	New Ireland Region, P.N.G.	6094	224.5
Z	20050926	01:55:37.67	-5.68	-76.40	115	7.50	Northern Peru	15968	334.6
@	20051008	03:50:40.80	34.54	73.59	26	7.70	Pakistan	3954	83.6

6.2.1 Well Data Levelling

Groundwater level changes are caused not only by tectonic factors, such as earth crust deformation related to earthquakes; but also by non-tectonic factors, such as rainfall, air pressure changes, earth tides, exploitation of mines and underground water. Although procedures for removing tidal and air pressure effects exist (e.g., Igarashi and Wakita, 1991; Kitagawa and Matsumoto, 1996), the available rainfall data is on quite a different time scale, and there is no record of industrial pumping. To avoid difficulty with the non-tectonic factors causing changes in water level, it is much simpler to just consider the first differences $y_{t+1} - y_t$ of the groundwater level data, which measures the changes of the groundwater level relative to the previous minute. This eliminates 'slow' factors such as air pressure, rainfall, earth tides and pumping, without introducing possible errors from additional data, leaving only the oscillations, as exemplified in Figure 6.3. Figure 6.4(a) shows the

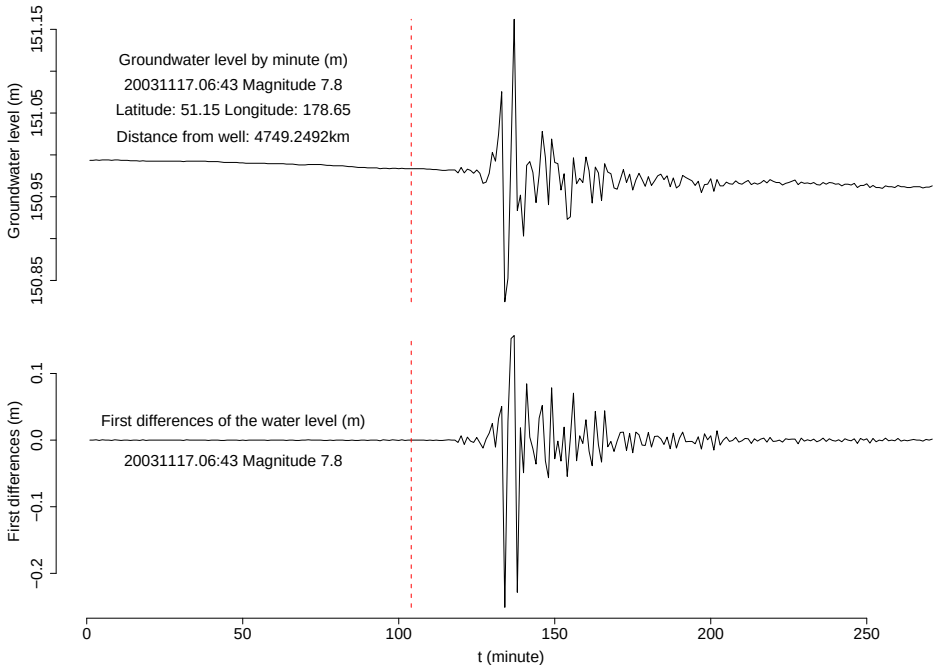


Figure 6.3: An example of the oscillations of the groundwater level at Tangshan Well, and their reduction to first differences. The dashed line indicates the origin time of the magnitude 7.8 Rat Islands, Alaska, earthquake of November 17, 2003.

distribution of the first differences of the groundwater level from 2002 to 2005. The 88,897 missing data translate to a total of 91,477 missing data in the first differences of the groundwater level. The 41 first differences of absolute value larger than or equal to 1m were individually examined. It appears that eight of them may have been caused by large earthquakes. The other 33 spikes appear to have been caused by sensor adjustments and malfunctions, which are irregular and always present

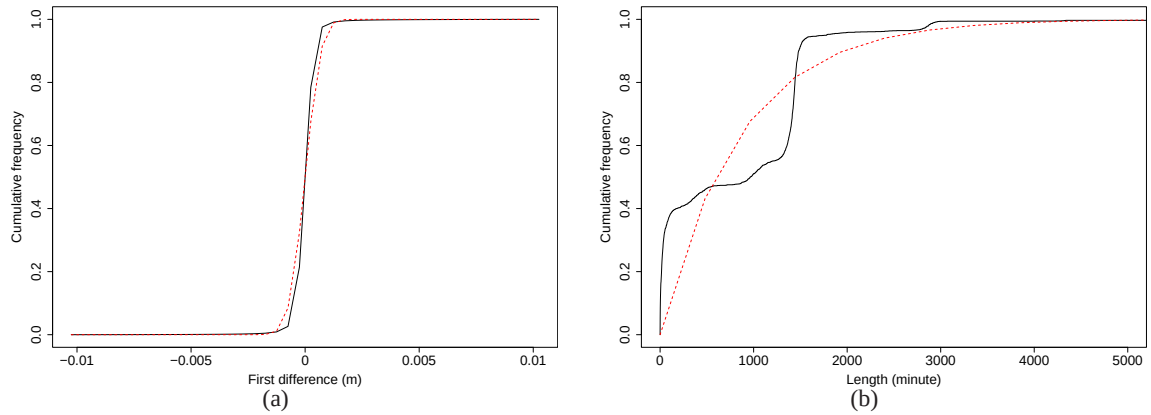


Figure 6.4: (a): The cumulative distribution of the first differences of the groundwater level (solid), with dashed line (for comparison) showing the fitted Gaussian cumulative distribution function with mean and variance calculated from the first difference data ($N(0.00025, 1.1969)$). (b): The cumulative distribution of the lengths of the non-missing intervals (solid), with dashed line (for comparison) indicating an exponential cumulative distribution function with rate estimated from the observed intervals ($1 - \exp(-x/851)$).

as either single jumps, or sustained oscillations which always reach their maximum amplitude at the end of the oscillations (Baojun Yin, personal communication). In this case we adopted the following filtering procedure: if the spike was due to sensor adjustment and the sensor was changed back to the previous level within two hours, we subtracted the adjustment from the original data; if the sensor was not changed back to the previous level within two hours, the adjustment was simply subtracted from the relevant first difference. For those spikes caused by malfunctions, the abnormal data was set to be missing. The data in October 2005, as shown in Figure 6.2, included a great number of sensor adjustments and malfunctions, which could not sensibly be reconciled. Hence, only the data from January 2002 to September 2005 are used here.

After this cleaning procedure, 1,971,360 data from January 2002 to September 2005 remain to be investigated, including 88,335 missing data, in the data set of the first differences of groundwater level. The missing data separate the first difference data into 2,213 intervals in which there is no missing datum. The longest length of these intervals is 23,478, and the shortest is 1. There are 97 intervals with length longer than 1,920, i.e. 32 hours. The distribution of the lengths of the non-missing intervals is shown in Figure 6.4(b). From this graph we can see that the breaks do not occur randomly, as the exponential distribution is a poor fit. There appear to be a large number of breaks at less than 2-hour intervals and at intervals of approximately 24 hours, which are almost certainly related to sensor adjustments and malfunctions. This may affect our analysis in the following sections.

6.2.2 Earthquake Data

The earthquakes are taken from the USGS-NEIC catalogue (http://neic.usgs.gov/neis/epic/epic_global.html). This catalogue is considered complete at magnitude 5.0. Hence a minimum magnitude cutoff of 6.0 will be used in this analysis, in order to have at least a minimal estimated amplitude from the seismic waves at the well. Almost all of the selected earthquake magnitudes are moment magnitudes (M_w). Rather than earthquake occurrence times, the arrival times of the earliest P phase, the earliest S phase, and the earliest Love and Rayleigh wave arrivals will be used, which were calculated based on the travel time table provided by Kennett and Engdahl (1991).

6.3 Identifying Signals in Well Data

In order to analyze the correlation between the water level oscillations and the earthquake occurrences, first the well signals have to be identified, which means the times when the oscillations begin (and end). Given that there are about two million data, it is hardly possible to do it manually. A method to automatically detect the occurrence times of oscillations probably due to earthquakes is needed.

6.3.1 Exponential Decay of Well Oscillations

After examining the oscillations of first differences individually, it appears that each set of oscillations may possess an exponential decay rate. Investigations of coseismic and noncoseismic water level changes of another well–aquifer system near Ito City, Shizuoka Prefecture, central Japan suggested that each water level change is followed by a damped harmonic oscillation (Kunugi et al., 2000). Therefore, an exponential decay function in conjunction with a cosine function can be fitted to the oscillations in order to get a reasonable decay rate and frequency for the oscillations of the groundwater level. This will then provide a suitable basis for an algorithm to detect induced changes in water level.

By choosing the point with the largest amplitude of each set of oscillations as a starting point, in order to model the oscillation frequency, a cosine modulated exponential function can be fitted to a cluster of cumulative sums C_t of first differences (i.e., to the values of the original, undifferenced values, subtracting a constant). The procedure is as follows. Set $t = 1$ at the location of the largest amplitude. Let $\delta_t = y_{t+1} - y_t$, which is the first differences of the groundwater level, $C_1 = \delta_1$ and

$C_t = \sum_{1 \le s \le t} \delta_s$ for $t > 1$. Then we fit the sets of oscillations using the equation

$$C_t = Ae^{-a(t+\phi)} \cos\left(\frac{2\pi(t+\phi)}{\omega}\right) + \varepsilon_t, \tag{6.1}$$

where A is the amplitude, a is the cluster decay rate, ϕ is the phase which allows for the origin to be somewhat between $t = 0$ and $t = 1$, ω is the frequency and ε_t is the residual. Model (6.1) involves four unknown parameters of this well. The real difficulty is that the sampling rate at this well is one observation per minute, approximately three times the natural period, which is now 20.99s, (Baojun Yin, personal communication, cf., Section 6.2). In order to get the estimated decay rate of the clusters in the presence of this aliasing effect we need to minimize the sum of the squared residuals $\sum_t g(t, A, a, \phi, \omega)$, where

$$g(t, A, a, \phi, \omega) = \left(Ae^{-a(t+\phi)} \cos\left(\frac{2\pi(t+\phi)}{\omega}\right) - C_t\right)^2.$$

However, in practice, at each point t, we use the minimum of $g(t, \theta)$, and the fitted value at ± 1 second, which allows for the stated precision in the timing of the water level measurements. A variety of oscillations corresponding to earthquakes with different ranges of magnitudes, depths and well–epicenter distances are chosen and then Model (6.1) is fitted to each set of the oscillations. The estimated frequency $\omega \approx 0.33 \pm 0.1\text{min}^{-1}$ (equivalent to 20 ± 6s period) and decay rate $a \approx 0.05 \pm 0.03$ from this model are consistent with the empirical values of the well parameters. Two examples with the fitted results are shown in Figure 6.5.

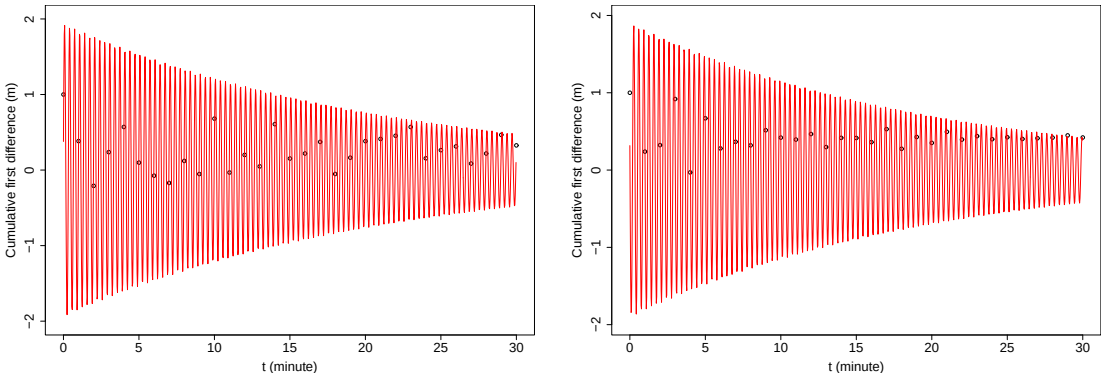


Figure 6.5: Two examples of cumulative first difference data, and hypothesized exponential damping showing the aliasing problem between sampling interval and well response period.

6.3.2 A Moving-window Detection Algorithm

Recalling the sampling rate of one per minute, a set of oscillations is most easily identified as an increase in the variability of the data. Hence, in order to identify where a set of water level oscillations begins and ends, a moving average method will be used to calculate a series of weighted variances of the first differences, and then those which are above a certain cutoff are considered to be signals.

The procedure is that the exponential decay rate estimated above is utilized to calculate a weighted variance of the first differences. Let $\{\delta'_t = (y_{t+1} - y_t)/0.0005 : 1 \leq t \leq N\}$ be the first difference sequence of the water level, measured in unit of 0.0005mm (the measurement precision) for ease of calculation. Set $t = 1$ at 00:00, January 1, 2002, and correspondingly, in the following sections, unless otherwise specified, the time origins (for time in minutes or hours) of both the water level and earthquake sequences are at this time point. Let l be the length of the moving window. After some experimentation, $l = 10$ was chosen, but the results do not appear to be sensitive to this choice. For each point t, we calculate two samples of size l, which are the weighted first differences for a window either preceding (backward) or following (forward) the point t, $X_{t1} = \{e^{-a(t-s)}\delta'_s : t - l + 1 \leq s \leq t\}$ and $X_{t2} = \{e^{-a(s-t)}\delta'_s : t \leq s \leq t + l - 1\}$. The exponential decay e^{-at} acts to down-weight the influence of the points further away from the current point. Then calculate the variances V_{t1} (backward variance) and V_{t2} (forward variance) of X_{t1} and X_{t2},

$$V_{t1} = (1/l) \sum_{s=t-l+1}^{t} \left(e^{-a(t-s)}\delta'_s - (1/l) \sum_{s=t-l+1}^{t} e^{-a(t-s)}\delta'_s \right)^2$$

and

$$V_{t2} = (1/l) \sum_{s=t}^{t+l-1} \left(e^{-a(s-t)}\delta'_s - (1/l) \sum_{s=t}^{t+l-1} e^{-a(s-t)}\delta'_s \right)^2,$$

respectively (as illustrated in Figure 6.6). The missing data in first differences is set to be 0 in order that the algorithm be conservative (i.e., we do not introduce any false positives). This has little effect when the intervals contain at most two missing data. However, when there are three or more missing data in an interval, oscillations may go undetected.

For a given quantile threshold V_c (c is the percentile) of the variances, a cluster within which every point has a weighted variance larger than or equal to V_c is considered as a signal, i.e., some

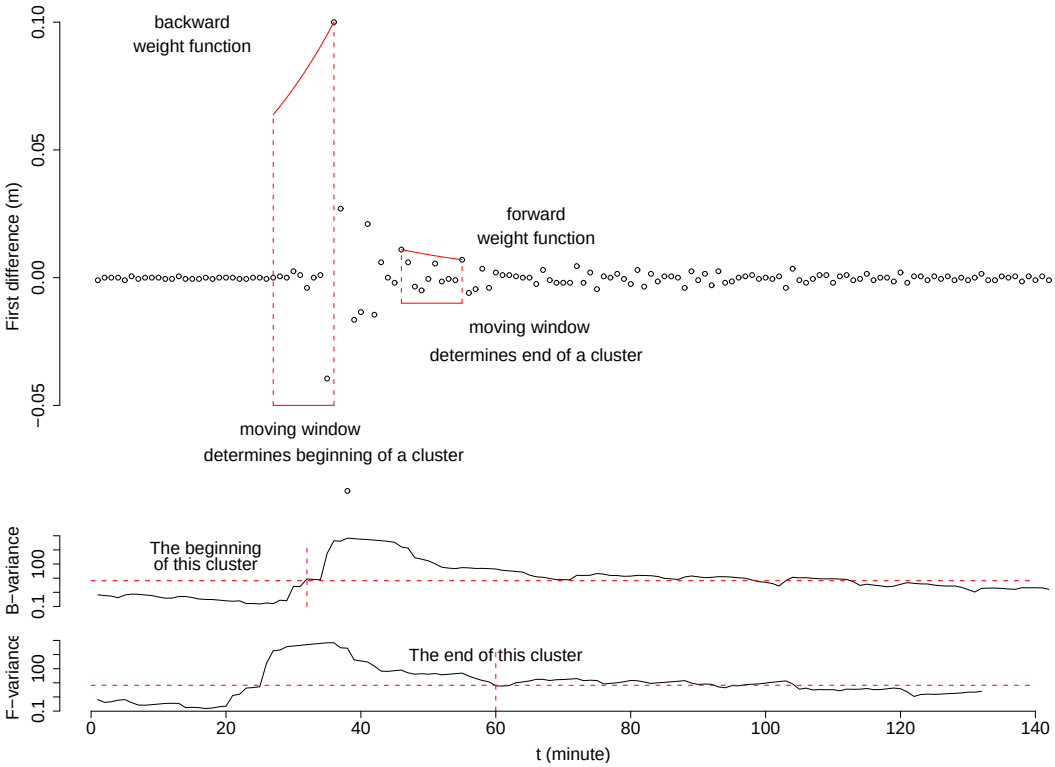


Figure 6.6: An illustration of calculating the weighted variance. B-variance: backward variance; F-variance: forward variance. The earthquake is the May 26, 2003, M 7.0, earthquake off the east coast of Honshu, Japan.

c percent of the data are treated as signals. Starting at $t = 1$, let t_b be the first point at which $V_{t_b1} \geq V_c$, and let $t_e = \min\{t > t_b : V_{t2} < V_c\}$. Call $\{\delta_t : t \in [t_b, t_e]\}$ a signal. Then repeat the above procedure starting at t_e until the end of the data. In the following sections, unless otherwise specified, a signal will be identified by its initiation time δ_{t_b}.

This procedure identifies some signals in the period from February 2003 to March 2003 that appear to have their maximum amplitude at the very last point of the signal. This nonphysical behavior is related to sensor malfunctions and adjustments, as mentioned above, and so these signals were eliminated from subsequent analysis. Readjustment of the sensor can cause apparent changes in groundwater level, which manifest in the first difference data as signals of length 1, while transcription errors (of single points) in the data create the appearance of signals of length 2. Hence signals of less than 3 minutes duration were likewise deleted.

6.3.3 Determining the Detection Threshold

In order to determine the threshold parameter c, we will borrow some ideas from prediction analysis. Kagan (1997) discussed prediction evaluation, and considered various measures of prediction efficiency in earthquake prediction, in which three basic ratios were used to evaluate the prediction: (a) the ratio of the total volume of alarm zones to the total volume of the region; (b) the ratio of missed qualified (should have been predicted but failed) earthquakes to the total number of such events; and (c) the ratio of the number of unsuccessful alarms to the total number of alarms. Since identifying the coseismic responses differs from predictions, and there is no alarm zone involved, we are going to use a criterion equivalent to class (c) of Kagan (1997). By determining the parameter c, we hope to maximize

$$R_1 = \text{percentage of earthquakes for which there are coseismic responses detected},$$

and minimize

$$R_2 = \text{percentage of signals which have no corresponding earthquake},$$

which we can formulate as maximizing the skill score R defined by

$$R = \frac{R_1}{R_2}. \tag{6.2}$$

We use the arrival times of the earliest P phase, the earliest S phase, the Love and the Rayleigh waves to define the coseismic responses. We suppose that a coseismic response is a well signal following an earthquake within a certain time interval L minutes after the arrival of each seismic phase. Note that R increases monotonically with L as R_1 (R_2) increases (decreases) monotonically with L. If a signal has no seismic phase arrival ahead of it within this time interval, we consider it as a 'false alarm'. The presence of missing data within the first differences of the water level is a complicating factor, which we will deal with by ignoring, in both the numerator and denominator of the calculation (6.2), earthquakes whose seismic wave arrivals lie within the interval of interest prior to any missing data.

For $c \leq 5$, we use the moving-window detection algorithm to identify well signals for each c, using $l = 10$ and $a = 0.05$. We restricted c to less than five percent of the data as five percent of 2 million data is equal to 1,667 hours of signals, which is easily more than any reasonable amount

of coseismic signals actually contained in the well data. Then for a fixed c and the identified well signals, we calculate R respectively for the earliest P phase, the earliest S phase, the Love and the Rayleigh wave arrivals of the earthquake series. The mean, median and mode of the c values at which the R scores reach the maximum are the same, $c = 1.5$, i.e. 1.5 percent of water level data to be considered as well signals, which we will thus use as our threshold. The parameter c selected in this way appears to be independent of the time interval L which we use to define the coseismic responses. The number of well signals isolated using these selected parameters is 754.

6.4 Tests of Dependence Between Earthquakes and Well Responses

Having extracted signals from our well data, we will utilize a dispersion test of the Poisson process for both the well signal series and the earthquake series to examine the cluster properties of the two sequences. After that we will discuss how to clarify the association between this series of events and the earthquakes. Various statistical methods have been adopted to characterize the association between two or more series of events, among which coherence, mutual information, and the Lin–Lin model are the most useful for our case.

6.4.1 Clustering Tendencies in Well Signals and the Earthquake Catalogue

Before investigating the correlation between the well signal and earthquake series, we first examine their properties separately. In this section we will see how they compare to a Poisson process using the dispersion test from Cox and Lewis (1966, see Appendix A).

To use the dispersion test statistic to test whether a series of events is a Poisson process, we let N be a point process of length T. Divide the observed sequence of N into k equal intervals of length t. Let $n_1, n_2, \cdots, n_k$ denote the number of events in the successive intervals and define

$$\bar{n} = \frac{n_1 + \cdots + n_k}{k}. \tag{6.3}$$

Then we calculate the statistic d of $n_1, n_2, \cdots, n_k$,

$$d = \sum_{i=1}^{k} \frac{(n_i - \bar{n})^2}{\bar{n}}. \tag{6.4}$$

The ratio $d/(k-1)$ has the expected value of unity for a Poisson process (Cox and Lewis, 1966), and for a large number of observations the distribution of d is asymptotically proportional to a $\mathcal{X}^2$

distribution with $(k-1)$ degrees of freedom (Vere-Jones and Davies, 1966). To test that the series of events is a Poisson process, we compare the statistic $d/(k-1)$, which is the ratio of the estimated variance of N to the estimated mean of N (Cox and Lewis, 1966), with unity.

The test results show that both the earthquake sequence and the series of 754 well signals are clustered, although the former is slightly less clustered than the latter (Figure 6.7). The histograms of

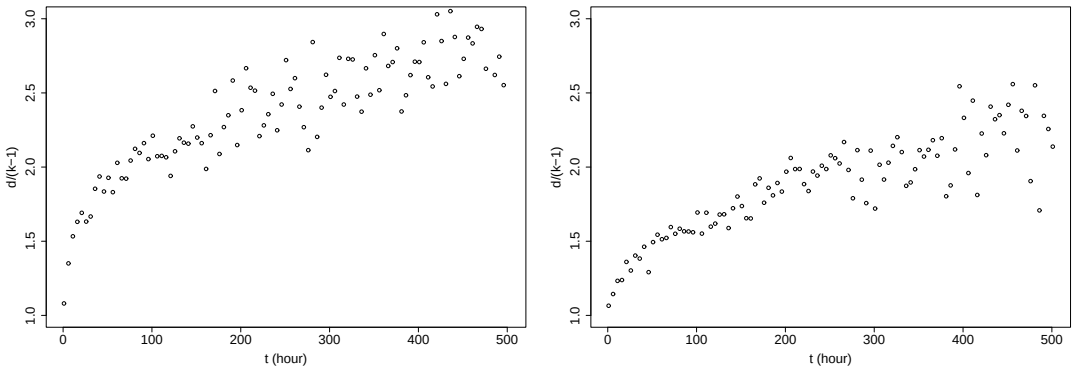


Figure 6.7: Variance against mean plot for well signals (left) and earthquakes (right).

the earthquake occurrence times, well signal beginning times, and the survivor functions of the well signal inter-event times and well signal lengths are shown in Figure 6.8. The histograms indicate that the earthquakes and the identified well signals have similar statistical properties. The survival

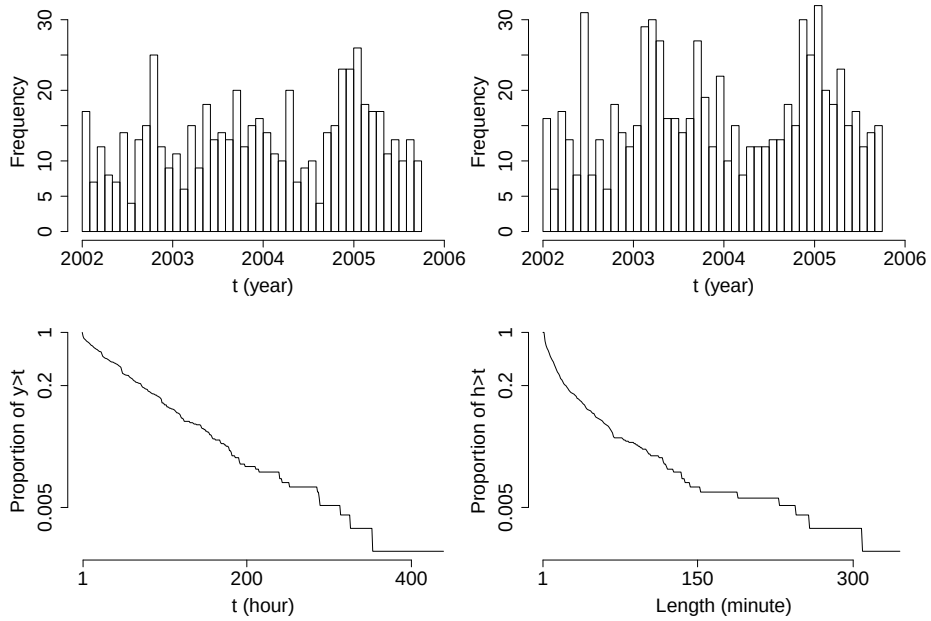


Figure 6.8: Time-occurrence histograms of the earthquake occurrence times (top left), well signals (top right), and the survival functions of the well signal inter-event time (bottom left) and well signal length (bottom right).

function of the well-signal inter-event times is not compatible with an exponential decay, confirming that there is some structure present, neither is the survival function of the well-signal lengths.

6.4.2 Correlations Between Earthquakes and Lagged Well Responses: (a) Coherence Analysis

Missing intervals of length 1 in the water level data set will cause missing intervals of length 2 in the first difference data set. Accordingly, here in this section we set the first differences in the missing interval of length 2 as 0. The resulting number of non-missing intervals with length longer than 1,920, i.e. 32 hours is 310. Figure 6.9 shows the coherence between the well sequence and each of the earliest P phase, the earliest S phase, Love wave and Rayleigh wave arrivals, using disjoint intervals of length 32 hours, respectively. We corrected the second moment by subtracting the mean (see

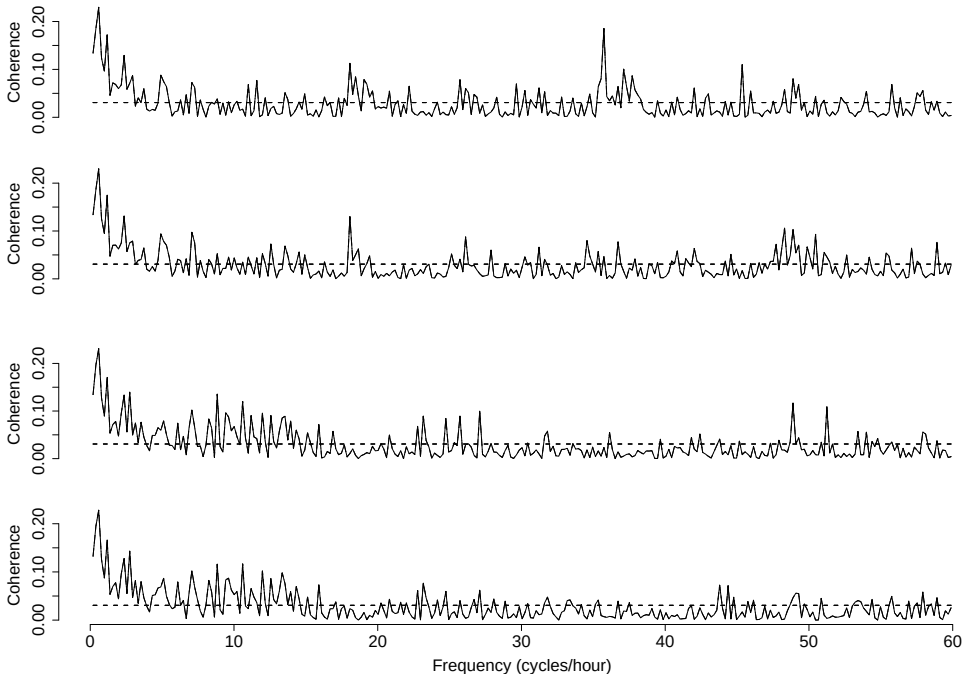


Figure 6.9: Coherence plots for well signals and the earliest P phase (top), the earliest S phase (second row), Love wave (third row) and Rayleigh wave (bottom) arrival times with disjoint intervals each having a length of 32 hours. Solid lines: coherence for well signals and the seismic phase arrivals; dashed lines: the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent.

Chapter 5 for details). Recall that the possible frequencies are limited by the discrete (per minute) nature of the data, and are not continuous. The graphs show strong association between each of the

four seismic wave arrivals and the first difference signals at low frequencies ω, apparent association up to about 6hr^{-1} or more than 10 minutes, i.e., the well signals are strongly correlated with the earthquake occurrences more than 10 minutes after the seismic wave arrivals. The sensitivity tests by moving the water level record 1 minute forward and 1 minute backward suggest that the coherence results for the perturbed data are not visibly different from the original plots.

6.4.3 Correlations Between Earthquakes and Lagged Well Responses: (b) Mutual Information Analysis

Let us consider the mutual information between well signals, denoted by U, and the arrival times of each of the seismic waves, denoted by V. We first transform the well signal series and the earthquake series into 0-1 sequences, with 1 indicating the beginning of the well signal, 0 otherwise. The earthquake series have a 1 in the minute of the wave arrival, 0 otherwise. We examine the mutual information between the two series $\{U_{t+u}\}$ and $\{V_t\}$, where u is the time lag. The mutual information results, as shown in Figure 6.10, suggest strong association between the seismic waves and the well signals. The figure indicates that while the P phase arrival precedes the signal onset,

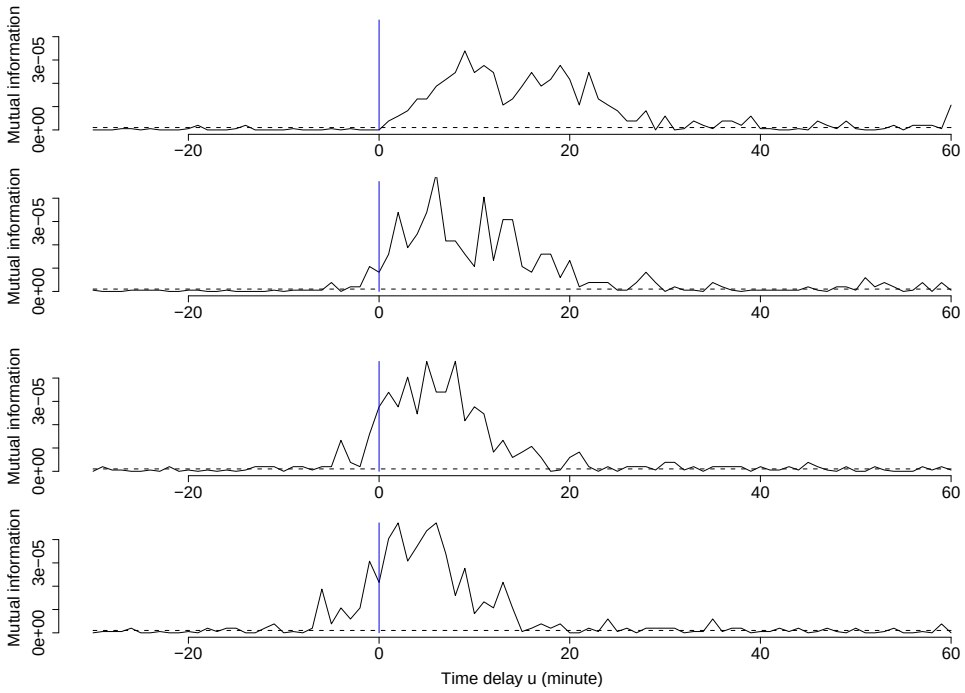


Figure 6.10: Mutual information for well signals and the earliest P phase (top), the earliest S phase (second top), earliest Love wave (third from top) and earliest Rayleigh wave (bottom) arrival times. The dashed lines show the upper level of the approximate 95% confidence interval under the null hypothesis that the two processes are independent.

the S phase, Love and Rayleigh wave arrivals appear to occur after some of the signal onsets.

The mutual information results for the well signals identified by using the water level recording moving 1 minute forward (or backward) can be read directly from Figure 6.10 by moving the plots 1 minute forward (or backward), and show that the conclusion is not sensitive to possible errors in timing.

An example of the first differences together with seismographs from the nearest seismic station, Douhe station, is shown in Figure 6.11. The Douhe seismic station is operated by Tangshan Earth-

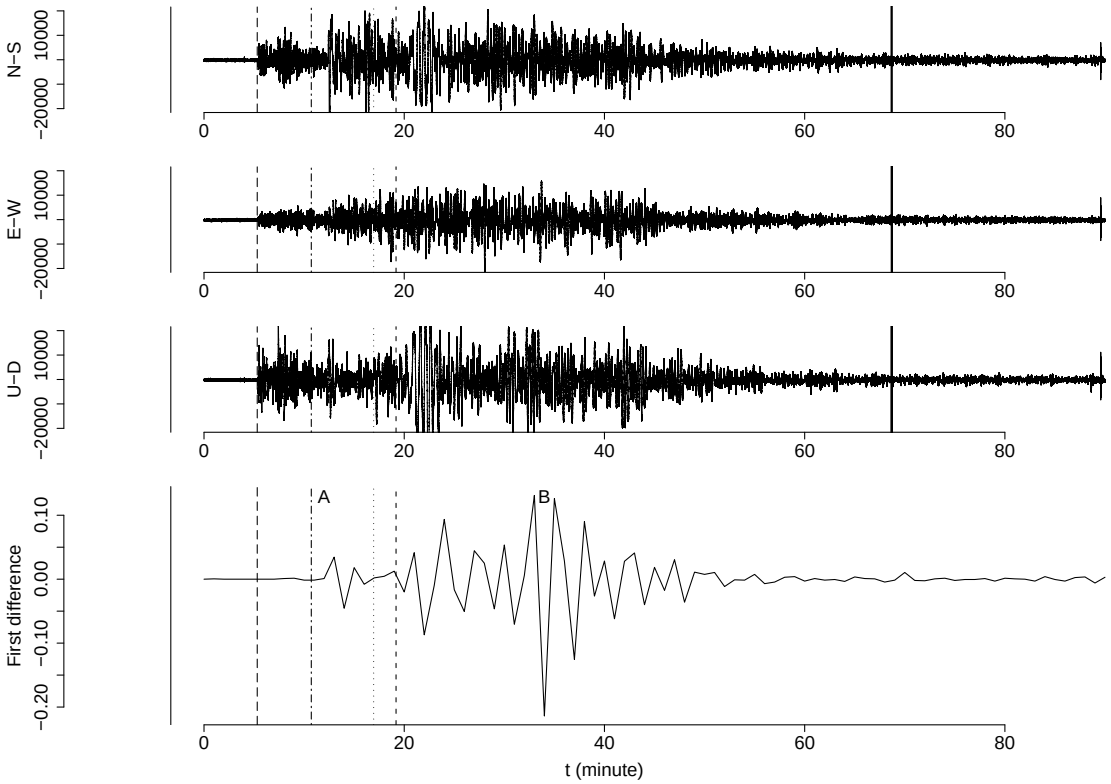


Figure 6.11: Example of an S phase arrival initiating well oscillations. The top three plots are seismographs from Douhe seismic station on the same time scale as the first difference plot of the groundwater level at the bottom. The solid line on the left-hand side in the figure is the earthquake occurrence time for the magnitude 7.5 earthquake in Kepulauan Alor, Indonesia on November 11, 2004 (latitude: 8.15S, longitude: 124.87E, well–epicenter distance: 5356.313km). The longdash, dot-dash, dotted and dashed lines indicate the corresponding earliest P phase, S phase, Love and Rayleigh wave arrival times for the earthquake, respectively. The letter'A' and 'B' in the first difference plot indicate respectively the times when the identified well signal starts and when the signal reaches its maximum amplitude.

quake Administration, and is located at latitude 39.74N, longitude 118.29E and altitude 55.42m above sea level, which is 16.3km from Tangshan Well. EDAS-C24 equipment is used to record the seismographs with a sample rate at 50Hz. The timing is synchronized with GPS time. Figure 6.11

shows an S phase arrival initiating the well response with the Rayleigh wave amplifying the oscillations. This is consistent with the conclusion from Cooper et al. (1965) that any type of earthquake wave that produces dilatation of the aquifer or vertical vibration of the well–aquifer system can cause the water level in a well to fluctuate, and that Rayleigh waves cause larger fluctuations in wells than any other wave that has been identified. Figure 8 in Brodsky et al. (2003) also shows that there is some energy that shows up before the Rayleigh wave arrival.

6.4.4 Earthquakes as an Explanatory Variable for Well Responses

Since we are interested in whether the earthquakes have induced the water level oscillations, we treat the well signal sequence as the output process and the wave arrival times as the input process(es). The estimated response functions for the self-exciting part, and the mutually exciting (to the seismic wave arrivals input) part are shown in Figure 6.12. We see that for the earliest P phase, the self-

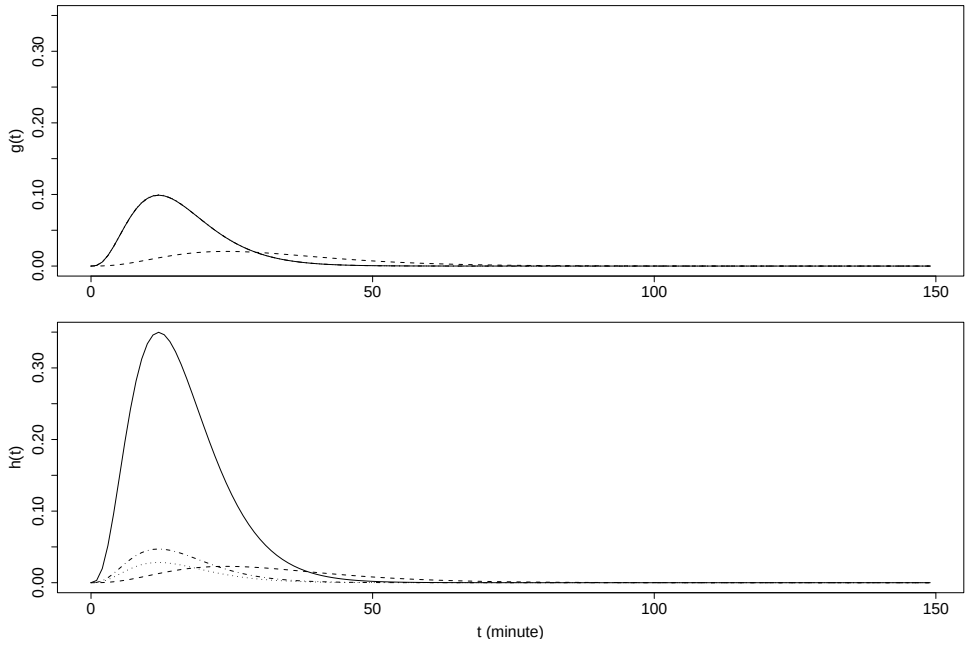


Figure 6.12: Estimated Lin–Lin response functions of well signals with the seismic phase arrival times: Earliest P phase (solid line), earliest S phase (dashed line), earliest Love wave arrival (dot-dashed line) and earliest Rayleigh wave arrival (dotted line). Top: Self-exciting response functions (P, L and R curves coincide). Bottom: Response function from the seismic wave arrival.

exciting part is small compared to the mutual exciting part, which clearly shows that the P phase arrival times do play a part in inducing the water level oscillations at Tangshan Well. The estimated response functions for the S phase arrivals are very similar to each other, and thus we can not

conclude whether or not the S phase arrival times play a part in inducing the water level oscillations at Tangshan Well. The response functions for the Love and the Rayleigh wave arrivals both have much larger self-exciting parts than mutual exciting parts, indicating that the Love and the Rayleigh wave arrival times do not have a clear causal relation with the initiation of water level oscillations at Tangshan Well. A sensitivity analysis moving the entire record 1 minute forward or backward produces almost identical plots, and thus does not alter the conclusions.

6.5 Coseismic Responses and Detection Probability

Having established that there appears to be some correlation between the seismic wave arrivals and the well signals, we can now examine the detection probability. First the detected and undetected events will be examined directly. Then a logistic regression analysis is used to examine the association of the probability that a coseismic response is detected with the earthquake statistics (magnitude, depth, azimuth and well–epicenter distance).

6.5.1 A Magnitude-distance Threshold for Detection

Kitagawa and Matsumoto (1996), and Matsumoto et al. (2003), in investigating sustained changes in groundwater level due to distant earthquakes, explained its occurrence as a function of magnitude and distance. A threshold combining earthquake magnitude and well–epicenter distance was thus proposed, above which earthquakes cause coseismic changes. Let us now investigate whether a similar form of threshold exists for the transient oscillations in the Tangshan Well data, and if so, how the slope differs from those proposed above.

We have seen above that the P phase arrival is the most strongly associated with the onset of well oscillations. Moreover, the mutual information for well signals and P phase arrival times indicates a strong association between the two series up to 40 minutes time delay. For P phase arrival times, we calculate the delay times to the start of the well signals from the arrival times. We will define those well signals as coseismic responses if the corresponding delay times are less than 40 minutes. This accounts for some 237 earthquakes (approximately 40 percent of the total number), and approximately 31 percent of the total number of well signals. Some 230 of these 237 earthquakes satisfy the relationship $M \geq 3.02 \log_{10} D - 5.50$, where M and D respectively denote magnitude and well–epicenter distance (km). This threshold line is determined such that the sum of misclassification of earthquakes below the line, for which there are coseismic responses detected,

and the misclassification of signals above the line, which have no corresponding earthquakes, is minimized. This classification by magnitude and distance is shown in Figure 6.13. Note that there

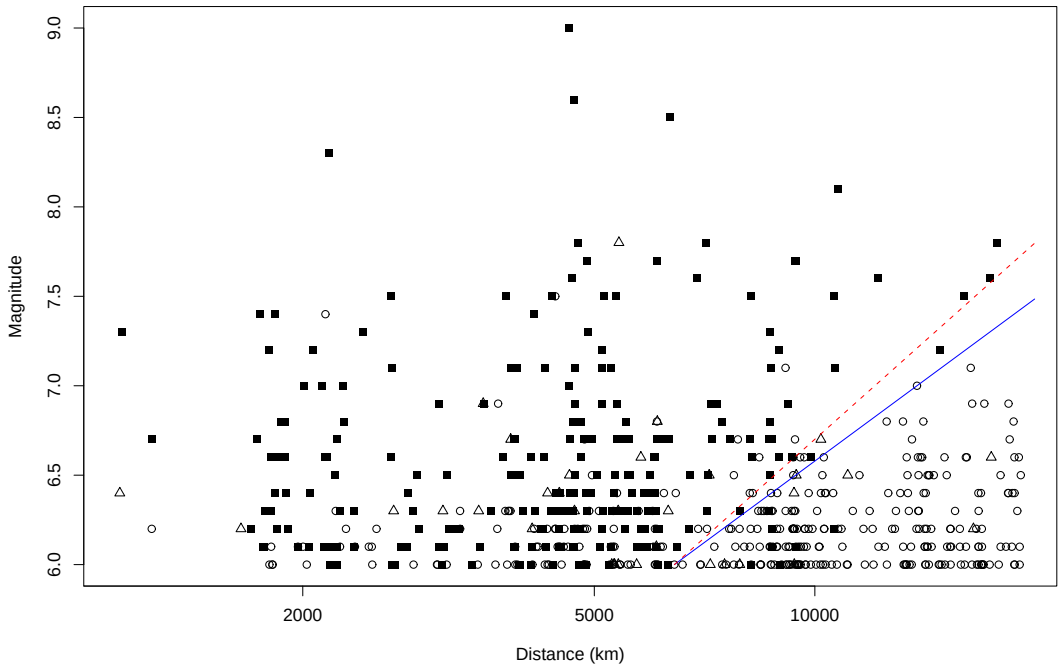


Figure 6.13: Magnitude versus well–epicenter distance plot. The solid line indicates $M = 3.02\log_{10}(D) - 5.50$. Some 230 of 237 coseismic responses satisfy the relationship $M \geq 3.02\log_{10}(D) - 5.50$. ■: the 237 earthquakes which have coseismic responses in the well data (i.e. there are signals detected within 40 minutes after the P phases arrived); ◯: the 329 earthquakes for which there is no coseismic response detected; △: those 34 earthquakes having missing well data in the 40 minutes following the P phase arrival. The dashed line $M = 3.69\log_{10}(D) - 8.07$ indicates the threshold using the interval of 30 minutes. The threshold using an interval of 50 minutes is the same as that for 40 minute interval.

were missing data in the well record in the 40 minutes following the earliest P wave arrival for 34 of the earthquakes. Hence these earthquakes are indicated separately, as detection was inconclusive.

It is instructive to examine the cases of misclassification. Of the 109 earthquakes denoted by circles above the threshold line (no signal detected), detailed visual inspection of the well record reveals that most of these are in fact followed by oscillations (which may or may not be coseismic responses) in groundwater level. These have not been picked up by the algorithm because the amplitudes of the oscillations are below the detection threshold. Recall that this threshold is a balance between failure to identify, and false identifications. The exceptions were the 30 earthquakes categorized in Table 6.2. Four of the seven events denoted by squares under the threshold line were

Table 6.2: The 30 events denoted in Figure 6.13 by circles above the line $M = 3.02\log_{10}(D) - 5.50$ for which there are no coseismic responses identified.

Number	Description
3	very deep earthquakes with depth larger than 200km
5	deep earthquakes with depth larger than 50km
3	with well signals beginning more than 40 minutes after the P wave arrived
1	with a well signal that begins earlier (!?) than the earthquake occurrence time, which would have obscured any response
15	immediately preceded by larger earthquakes (the smaller earthquake's response is merged together with the larger's one)
3	no apparent responses

very shallow earthquakes, while the remaining three earthquakes were followed by well signals which may have actually been initiated by other earthquakes which occurred a few minutes earlier elsewhere.

The map in Figure 6.14 (inset, Figure 6.15) shows the entire earthquake data set of the 600 events, with symbols, as in Figure 6.13 indicating whether or not the earthquake was detected in the well level. We see that there is no obvious effect of azimuth on whether or not a coseismic response is detected.

We can now consider in further detail how each seismic wave appears to affect the well signals. Table 6.3 categorizes the 237 earthquakes with coseismic responses, according to where the beginning of the well signal, and its maximum amplitude, fall in relation to the seismic wave arrivals. The

Table 6.3: Chronology of well signals and seismic wave arrivals.

	between P and S	between S and Love	between Love and Rayleigh	after Rayleigh
signal initiation	18	32	16	171
maximum signal amplitude	3	5	2	227

relationships with magnitude and well–epicenter distance are shown in Figure 6.16 and Figure 6.17. It appears that earthquakes with larger magnitudes tend to have detected coseismic responses before the Rayleigh wave arrivals, although some earthquakes with magnitudes less than 7.0 at long distances tend to induce well signals earlier than the Rayleigh wave arrivals. While the Rayleigh

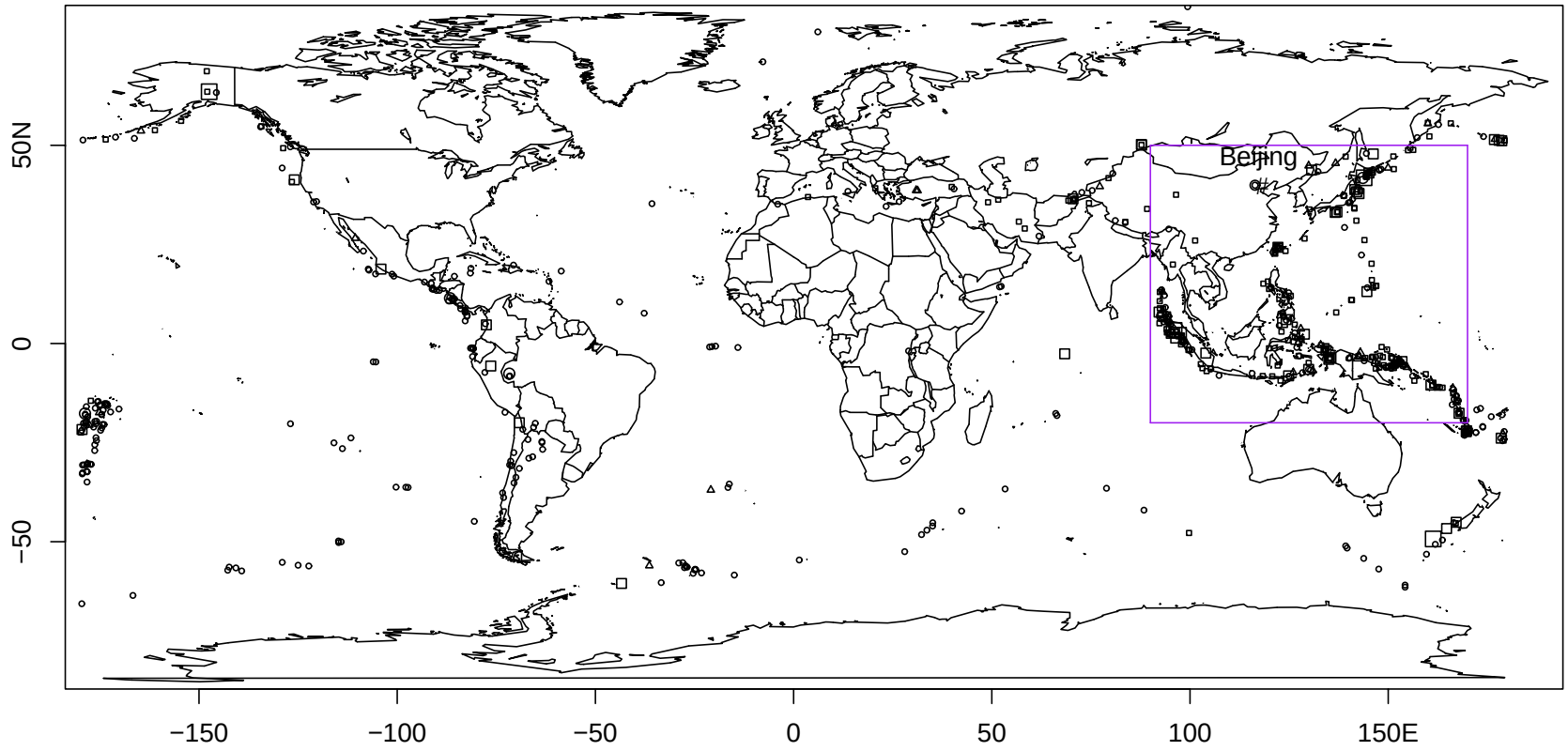


Figure 6.14: The global map and earthquakes with minimum magnitude 6.0 from January 1, 2002 to September 30, 2005. #: Tangshan Well; □: earthquakes for which there are coseismic responses detected; ◯: earthquakes for which there is no coseismic response detected; △: 34 earthquakes having missing well data in the 40 minutes following the P wave arrivals. The smallest symbol size is for earthquakes with magnitude larger than 6 and smaller than 7; the medium size for earthquakes with magnitude larger than 7 and smaller than 8; and the largest size for earthquakes with magnitude larger than 8. Inset is Figure 6.15.

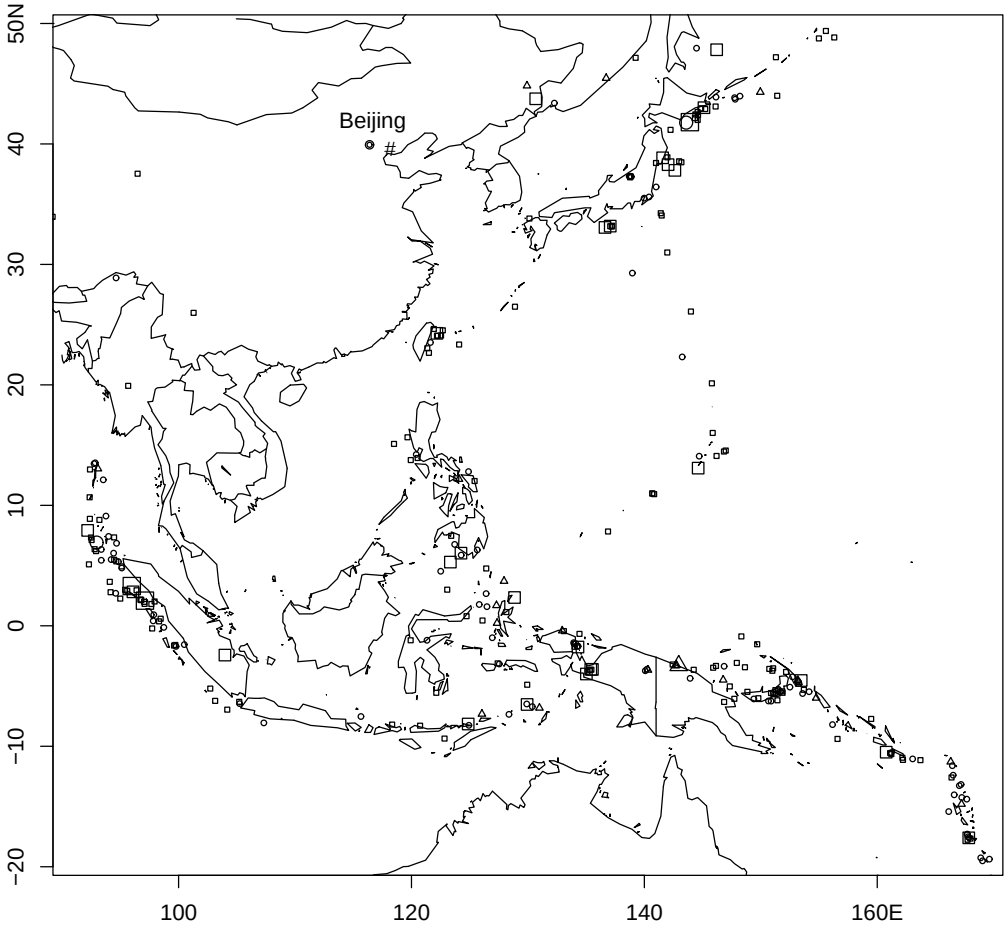


Figure 6.15: The enlarged version of the squared region in Figure 6.14. The symbols denote the same as in Figure 6.14.

waves are certainly the cause of most of the water level movement, in this well at least, arrivals from earlier phases are noticeable in many cases.

A sensitivity test for dependence of the results on the time interval used to define a coseismic response was conducted, using 30 minutes, and 50 minutes, instead of 40 minutes. For a 30 minute interval, 224 earthquakes have coseismic responses, 341 do not, and 35 had missing well data, while for a 50 minute interval, 241 have coseismic responses, 328 do not and 31 had missing data. While the threshold lines for the 40-minute and 50-minute intervals were the same, that for the 30-minute interval differed, with fewer distant earthquakes having coseismic responses. These appear to be smaller events where the oscillations are not detectable by the algorithm until the arrival of the amplifying surface waves.

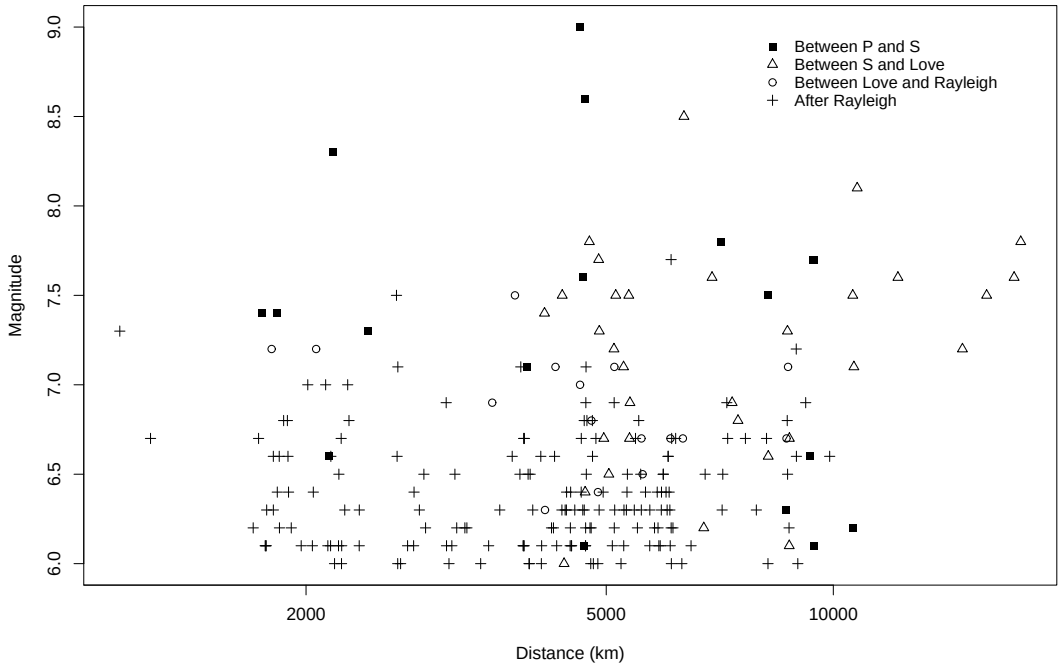


Figure 6.16: Magnitude versus well–epicenter distance plot. ■: earthquakes for which there are coseismic responses detected between the P and the S phase arrivals; △: earthquakes for which there are coseismic responses detected between the S phase and the Love wave arrivals; ◯: earthquakes for which there are coseismic responses detected between the Love and the Rayleigh wave arrivals; $+$: earthquakes for which there are coseismic responses detected after the Rayleigh wave arrivals.

6.5.2 Quantifying the Effects of Magnitude, Distance, Depth and Azimuth

A more quantitative analysis of how the detection probability p depends on the earthquake characteristics is possible using logistic regression (see Appendix B for statistical details). Besides magnitude, M, well–epicenter distance D (km), and depth H (km), we will use the azimuth angle ζ ($0^\circ \leq \zeta < 360^\circ$) from an earthquake to the well, determined as

$$\cos\zeta = \frac{1}{\sin\Delta}\left(\cos\theta_S \sin\theta_E - \sin\theta_S \cos\theta_E \cos(\phi_S - \phi_E)\right)$$
$$\sin\zeta = \frac{1}{\sin\Delta}\sin\theta_S \sin(\phi_S - \phi_E),$$

(Stein and Wysession, 2003), where θ_E is the colatitude of an earthquake, equal to $90^\circ -$latitude, ϕ_E is the longitude of an earthquake; θ_S and ϕ_S are the colatitude and longitude of the station or well similarly defined; and Δ is the source-to-receiver distance (in radians). See Figure 6.18 for

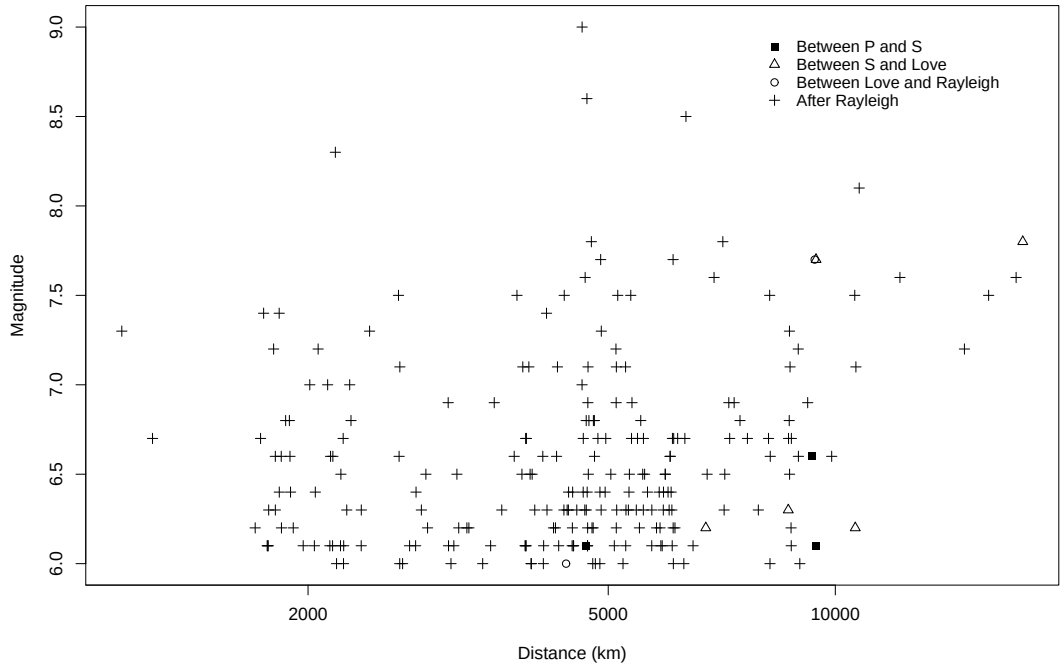


Figure 6.17: Magnitude versus well–epicenter distance plot. ■: earthquakes for which there are coseismic responses detected, which reach the maximum amplitudes between the P and the S phase arrivals; △: earthquakes for which there are coseismic responses detected, which reach the maximum amplitudes between the S phase and the Love wave arrivals; ◯: earthquakes for which there are coseismic responses detected, which reach the maximum amplitudes between the Love and the Rayleigh wave arrivals; +: earthquakes for which there are coseismic responses detected, which reach the maximum amplitudes after the Rayleigh wave arrivals.

a schematic illustration of the azimuth angle ζ ($0°$, $90°$, $180°$, $270°$) from an earthquake to the well. Here, and in the following section, we separate out well–epicenter distance and depth, rather than using well–hypocenter distance, as the seismic wave propagation effects of shallow and deep earthquakes differ considerably. We fit model (A-4) in Appendix B to the data and use maximum likelihood to obtain parameter estimates for b_0, b_i and b_{1i}. The resulting model is

$$\log\left(\frac{p}{1-p}\right) = -33.1 + 5.0M - 11.4(\log_{10} D - 3.43)^2 + 2.6 \times 10^{-5}(H - 384.6)^2 + \varepsilon, \quad (6.5)$$

where all the included terms are significant at a level of 0.001. The effect of the azimuth is not significant. The relative strengths of the effects can be gauged by noting that, in the data, $30 \leq 5.0M \leq 45$, $-8.26 \leq -11.4(\log_{10} D - 3.43)^2 \leq 0$ and $0 < 2.6 \times 10^{-5}(H - 384.6)^2 < 3.85$. The residual ε has standard deviation 0.33, small relative to the other terms, indicating quite a good fit.

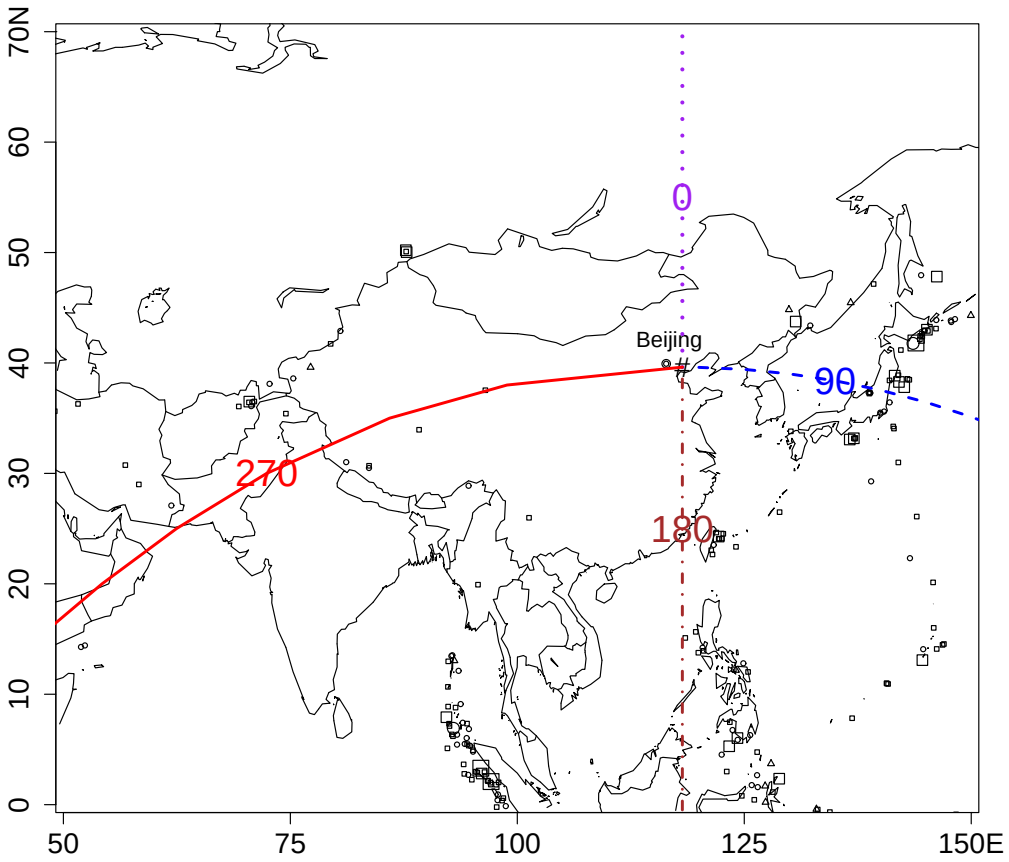


Figure 6.18: A schematic illustration of the azimuth angle ζ ($0°$, $90°$, $180°$, $270°$) from an earthquake to the well.

The fitted results by factor are shown in Figure 6.19. Due to the quadratic terms used in the analysis, and the dearth of events close to the well, we see that Model (6.5) is only physically meaningful when the distance is longer than 2690km ($\log_{10}(D) - 3.43 \geq 0$) and the depth is less than 384.6km ($H - 384.6 \leq 0$) as the sign of effect changes at these lengths.

In principle, Model (6.5) allows one to read off the effect on the detection probability of a change in any one of the significant factors. For example, we see that a increase of 1 unit in magnitude will result in a percentage increase in the probability of detection of

$$\frac{\delta p}{p} = \frac{1-p}{1/(e^5-1)+p},$$

where δp is the change in p. Similar calculations are possible for the effects of well–epicenter distance and depth.

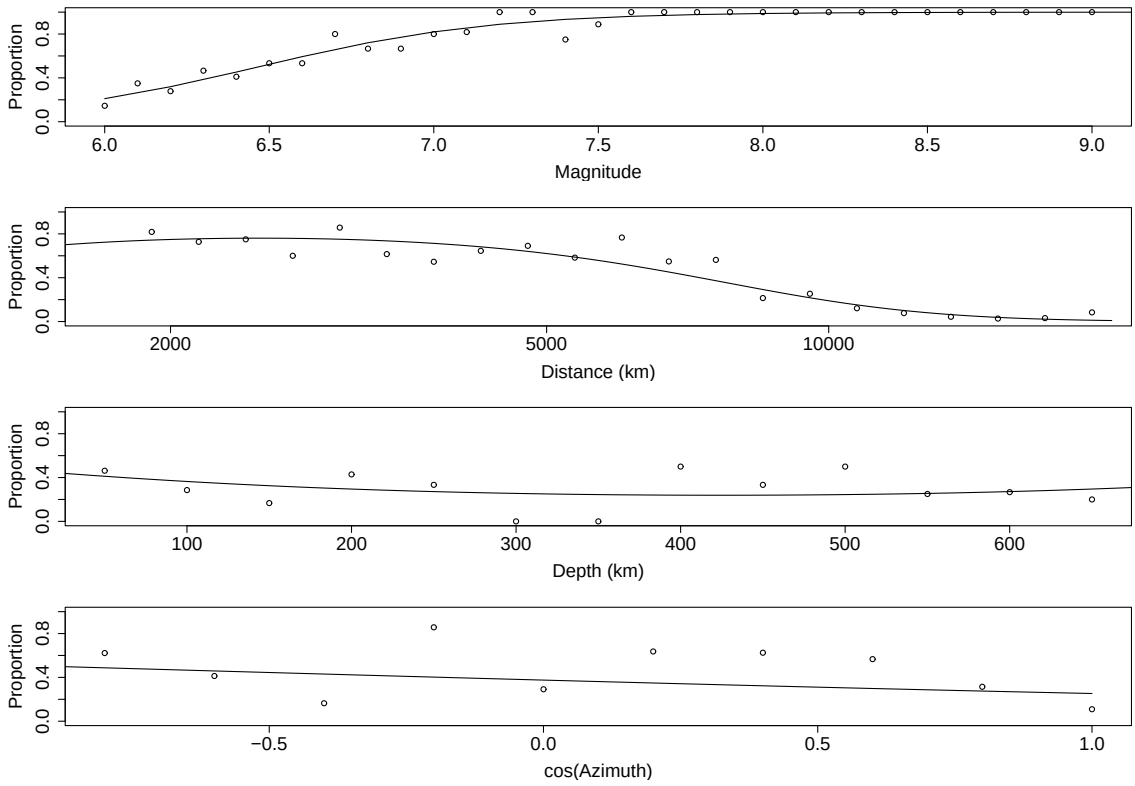


Figure 6.19: The proportion of earthquakes with coseismic responses (points) and the fitted probability that an earthquake produces a coseismic response (solid lines).

6.6 An Exploratory Analysis of Earthquake–well Signal Interaction

The well signals can be quantified by a set of characteristics (delays, amplitudes and lengths of well signals), as can the earthquakes (magnitudes, well–epicenter distances, depths and azimuths). In this section we will conduct some exploratory data analysis to investigate possible structural relationships between the two sets of characteristics. Another reason for the statistical analysis is to examine the effect of azimuth, as Tangshan Well is located in an ancient craton, and so there may be significant deviation from the Kennett and Engdahl (1991) radially symmetric travel-time model.

For P phase arrival times, we calculate Δ_P as the delay, in minutes, from the arrival of the earliest P phase to the start of the well signal. A well signal is defined as a coseismic response if Δ_P is less than 40 minutes, as above. For those earthquakes for which a coseismic response is defined, we denote similarly defined delay times for the earliest S phase, and the earliest Love and Rayleigh wave arrivals by Δ_S, Δ_L and Δ_R, respectively. The corresponding maximum amplitude and length of these well signals are denoted by A_w (m) and L_w (minutes), respectively. Figure 6.20 shows an

example of the scatter plots for the lengths L_w of the well signals versus the earthquake magnitudes and well–epicenter distances.

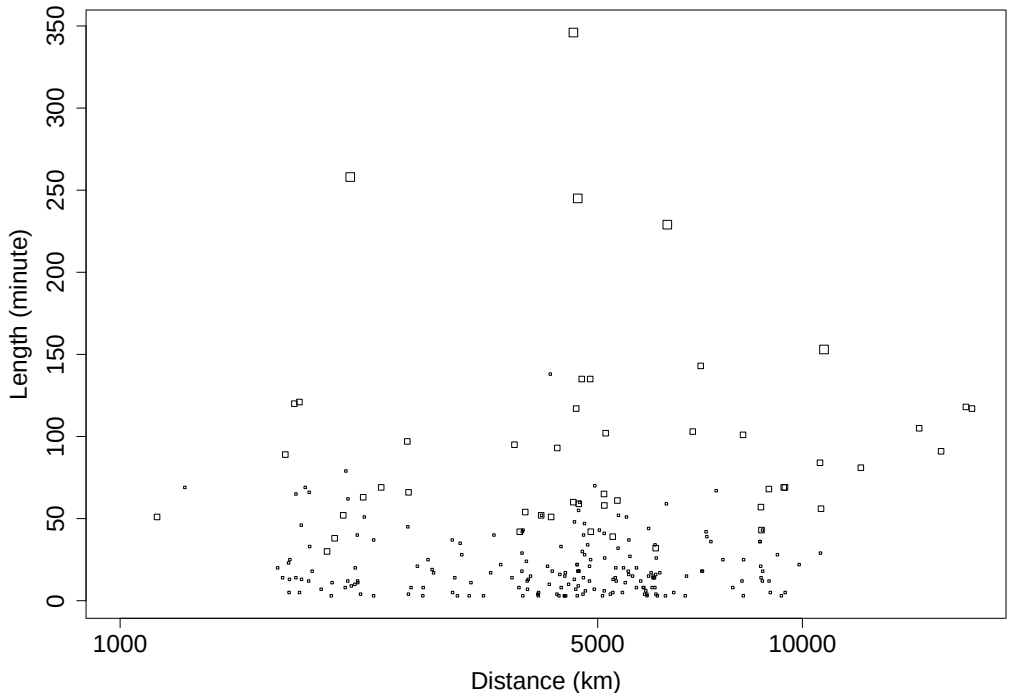


Figure 6.20: Well signals' lengths versus well–epicenter distances. The smallest symbols indicate earthquakes with M larger than 6 and smaller than 7; the medium symbols indicate earthquakes with M larger than 7 and smaller than 8; and the largest symbols indicate earthquakes with M larger than 8.

6.6.1 Correlations Between Earthquake and Well Signal Characteristics

In order to further validate our signal extraction algorithm we will calculate the correlation coefficients between the delay times, and the magnitude, well–epicenter distance, depth and azimuth. We will also calculate the correlation coefficients between amplitude, and length of the well signal, and the earthquake characteristics. The results are shown in Table 6.4. We see that earthquake magnitudes have the strongest overall effects on the well signal characteristics, negatively correlated with the delays and positively correlated with the amplitudes and lengths of well signals, all correlations being significant at a level of 0.001. Depth is positively, although less strongly (significant at a level of 0.05), correlated with the delay. The negative correlations with the maximum amplitudes and lengths of well signals are not significant. The positive correlation of well–epicenter distance with the P and S phase delays, and negative correlation with the Love and the Rayleigh wave delays is

Table 6.4: Correlation coefficients, where M is the earthquake magnitude, D is the well–epicenter distance (km), H is the earthquake depth (km), C_ζ is $\cos$(azimuth), S_ζ is $\sin$(azimuth), Δ_P is the P-phase delay, which equals to the time from the earliest P phase arrival to the well signal onset (minutes), Δ_S, Δ_L and Δ_R are respectively the S phase delay, Love wave delay and Rayleigh wave delay similarly defined, A_w is the amplitude of a well signal (m), and L_w is the length of a well signal (minutes).

	M	$\log(D)$	H	C_ζ	S_ζ
Δ_P	-0.340	0.447	-0.260	0.017	0.076
Δ_S	-0.401	0.316	-0.290	0.044	0.086
Δ_L	-0.578	-0.390	-0.324	-0.011	0.107
Δ_R	-0.578	-0.487	-0.311	-0.008	0.106
A_w	0.609	-0.109	-0.064	0.011	0.040
$\log(L_w)$	0.803	0.151	-0.023	0.079	-0.015

because Love and Rayleigh waves tend to arrive after the onset of the oscillations and so have minus delays. Thus the negative correlation for the latter means that longer distances have shorter 'minus' delays, and so the delays themselves are larger. These correlations are all significant at a level 0.05, as are those between well–epicenter distance and signal length. The remaining correlations are not significant, including all those involving the azimuth.

The correlations give the direction of any effect. For example, larger magnitudes lead to shorter delays, larger and longer signals. None of which are particularly surprising, which means that there is no reason to reject our algorithm. It is of more interest to try and quantify the size of any effect. For those earthquakes with identified coseismic responses, we can use multivariate regression (see Matsumoto, 1992, for an earlier application to coseismic water level changes), including interactions between the earthquake characteristics, to investigate the relationship between these characteristics and the well signal characteristics. See Appendix C for the statistical background. A very similar approach is the regression metamodel in simulation (Kleijnen, 1979), which likewise seeks a structural model between the inputs and outputs of an imperfectly understood system.

6.6.2 Metamodels for Delay, Amplitude and Length of Well Signals

First let us look at the relationship between the P phase delays and the earthquake statistics. The final fitted model for Δ_P is

$$\sqrt{\Delta_P} = 12.95 - 1.14M + 3.98\log_{10}(D) - 7.57 \times 10^{-6}(H - 1453.1)^2$$

$$+ 0.35C_\zeta + 0.002MH - 0.01\log_{10}(D)H - 0.004HC_\zeta + \varepsilon, \tag{6.6}$$

(in $\sqrt{\text{minutes}}$), where all the included terms are significant at a level of 0.05. The relative sizes of the effects can be gauged by noting that in the data, $-9.72 \leq -1.14M + 0.002MH \leq 1.62$, $-11.00 \leq (3.98 - 0.01H)\log_{10}(D) \leq 16.45$, and $-0.90 \leq (0.35 - 0.004H)C_\zeta \leq 1.85$. The residual ε has standard deviation 0.65, commensurate with the other terms, indicating considerable unexplained variation. Again, due to the quadratic terms in the model, Equation (6.6) is only physically valid for earthquakes with depths less than 570km. The equation (6.6) again provides a 'plug-in' formula from which one can extract the effect on delay of a change in one of the significant factors. For example, an increase of one unit in magnitude results in a decrease of the delay time of $(1.14 - 0.002H)^2 - 2(1.14 - 0.002H)\sqrt{\Delta_P}$, where $\sqrt{\Delta_P}$ is given by (6.6).

We see that using the radially symmetric travel time table of Kennett and Engdahl (1991) has led to a systematic error here, as the variation in delay with azimuth is significant, although the size of the effect is small, compared to the other factors. The quantitative analyses in this section can deal with azimuth effects directly, and so this systematic error will only affect the results if the azimuth has an effect on the identification of the seismic wave arrival immediately preceding the onset of well oscillations. However, a chi-squared test on the contingency table (Table 6.5) cross referencing azimuth (actually the sign of C_ζ) and seismic wave indicates that there is no dependence between the two. As the well signals appear predominately to be initiated 10 or more minutes (see above)

Table 6.5: Contingency table, where n_{PS} is the number of earthquakes with detected coseismic responses occurring between the P and the S phase arrivals; n_{SL} for coseismic responses detected between the S phase and the Love wave arrivals; n_{LR} for coseismic responses detected between the Love and the Rayleigh wave arrivals; and n_R for coseismic responses detected after the Rayleigh wave arrivals.

	n_{PS}	n_{SL}	n_{LR}	n_R
azimuth in the range $(-\pi/2, \pi/2)$	2	8	5	49
azimuth in the range $(\pi/2, 3\pi/2)$	16	24	11	122

after the P phase arrival, the size of the azimuth effect indicates that we are unlikely to be missing any coseismic responses because of the travel time error.

The multivariate regression analysis suggests a relationship between the maximum amplitudes

of the well signals and the earthquake characteristics of

$$\begin{aligned}\log_{10} A_w =& 5.52 - 0.09(M - 12.44)^2 - 1.07 \log_{10} D \\ &+ 7.02 \times 10^{-6} H^2 + 0.07 S_\zeta - 0.0007 MH + \varepsilon, \end{aligned} \tag{6.7}$$

(in units of $\log_{10}$m). All the fitted terms are significant at a level of 0.05. From the data, we have $-3.73 \leq -0.09(M - 12.44)^2 \leq -1.07$, $-4.55 \leq -1.07 \log_{10} D \leq -3.28$, $-0.87 \leq 10^{-6} H^2 - 0.0007 MH \leq -0.02$, and $-0.07 \leq 0.07 S_\zeta \leq 0.07$, from which we can see that the azimuth effect is negligible compared to the other effects. The residual ε has standard deviation 0.30, small relative to the magnitude and distance effects. Again, Equation (6.7) provides a formula from which the relative effects of the factors can be calculated. For example, at a depth of 100km, a magnitude 8.0 earthquake produces a signal amplitude approximately 6.5 times as that of a magnitude 7.0 earthquake. An earthquake at distance 12,000km will have a signal amplitude approximately half that of one at 6,000km, and so on.

For the well signal lengths, the following model is obtained

$$\log_{10} L_w = 3.07 - 0.21(M - 8.9)^2 - 0.27 \log_{10} D + 3.67 \times 10^{-6} (H - 408.7)^2 + \varepsilon, \tag{6.8}$$

(in $\log_{10}$ minutes), where all the terms are significant at a level of 0.001. The azimuth effect is not significant. The relative sizes of the effects in the data are $-1.77 \leq -0.21(M - 8.9)^2 \leq 0$, $-1.14 \leq -0.27 \log_{10} D \leq -0.82$, and $0 \leq 3.67 \times 10^{-6} (H - 408.7)^2 \leq 0.60$, respectively, while Equation (6.8) is obviously only valid for $M \leq 8.9$ and $H \leq 408.7$. The residual ε has standard deviation 0.25, small relative to distance effect, but possibly indicating some unexplained variation particularly around large and/or deep events. We can easily read off effects from Equation (6.8), for example a magnitude 8.0 earthquake will have a signal approximately 3.4 times as long as that of a magnitude 7.0 earthquake, while increasing the distance from 6,000km to 12,000km shortens the signal by about 17%.

6.7 Discussion

Having been acknowledged as one of the most sensitive wells to seismic disturbances in China, it appears as if the Tangshan Well acts as the recorder for a natural seismometer formed by the aquifer system into which the well penetrates. We have proposed an algorithm for automatically detecting

anomalous transient changes in groundwater level based on a moving average of the variance of the first differences in water level. The R-score and dispersion analyses were used to identify the best well signal series. Analysis of the resulting well signal catalog from the recent data indicates that the algorithm performs reasonably, although improvement is probably possible, especially if given a data set of more even quality than used here. A number of different statistical techniques were then used to detect and quantify coseismic responses to the earliest P phase arrivals.

As noted in Wang et al. (2010), the three statistics, coherence, mutual information and the Lin–Lin model, examine different types of association. When the problem of whether or not we can predict an output series from some input series through a linear relation is considered, coherence may be a very useful approach to use. It plays a diagnostic role in detecting association between two processes at certain frequency range. The estimation of coherence using the method of disjoint sections can work around missing data, which appears to be an advantage of using coherence.

However, the coherence may be identically 0 when two processes are actually correlated. The coefficient of mutual information overcomes this drawback and takes on the value 0 if and only if the two processes are statistically independent and thus forms a test of dependence. The mutual information presents a detailed conclusion about the time lag of the association between the two processes, as can be seen from Figure 6.10, for example, where the P phases and the well signals are strongly correlated with the P phases leading the well signals up to 40 minutes. Another important advantage of mutual information is that it does not really pay any attention to the values of the processes. In addition to telling us whether the hypothesis of independence is rejected, mutual information measures the strength of the correlation between two series of events. The drawback of the mutual information may be that it does not set up naturally for point processes, as we can see from the formulation as well as the application to the data. But a point process can be transformed into a time series which is suitable or convenient for mutual information analysis.

When we are interested in more than whether the two processes are correlated or not, or in other words, when we want to make clear which process is the driven process, the coherence and mutual information analyses become inadequate. The Lin–Lin model can identify cases in which either one of the two processes causes the other, or they cause each other, or some other process causes both of them. Unlike the generally defined coherence and mutual information, this model was designed specifically for point processes. The mutual-exciting term being identically 0 implies that there is no causal relationship between the two processes; whereas the self-exciting term being identically 0 while the mutual-exciting term being none zero means that the intensity of the output process is

modulated only by the input process.

The dispersion tests reveal that the well signal series and the earthquake series are themselves serially dependent. The coherence analysis concludes that the well signal series and the arrivals of the seismic waves have strong linear correlation, especially at lower frequencies, which may suggest that the earthquake occurrence times can provide a better linear prediction of the water level oscillations at lower frequencies. The mutual information analysis suggests that the oscillations of the groundwater level are strongly associated with the seismic wave arrivals, particularly the earliest P phase, which appears to initially trigger the fluctuations. The results from the Lin–Lin model indicate that the P phase arrivals are the most likely to be inducing the initial water level oscillations at Tangshan Well. The role of the S phase arrivals is unclear as the estimated self-exciting and mutual exciting response functions are similar. This is of interest, as Doan and Cornet (2007) described a hydraulic transient triggered near an active fault by seismic waves generated more than 10,000km away during the 2003 $M_w = 7.8$ Rat Island earthquake, where the pressure drop occurred simultaneously with the arrival of the S phase seismic waves and not during the passage of the Rayleigh waves. Brodsky et al. (2003) considered the Rayleigh wave, which generates the most significant dilatation, as being the likely cause of sustained changes in well water level, although their data also suggested that there may have been some detectable effect prior to the arrival of these waves in at least one case. The Love and Rayleigh waves do not have a clear causal relation with the initiation of water level oscillations at this well, although there appears to be a strong relation with the time of maximum amplitude, consistent with the results of Eaton and Takasaki (1959).

The signal detection algorithm detects coseismic responses for approximately 40% of global earthquakes with minimum magnitude 6.0. This is certainly a lower bound, as the performance of the algorithm balances this against the false detection probability. A threshold is also found in the relationship between earthquake magnitude and epicentral distance satisfied by 230 of 237 earthquakes with identified coseismic responses, $M \geq 3.02 \log_{10} D - 5.50$, above which coseismic changes in groundwater level at Tangshan Well mainly occur. This differs considerably from the thresholds $M = 2.62 \log_{10} D$ and $M = 2.62 \log_{10} D + 1$, (Kitagawa and Matsumoto, 1996) and $M = 2.45 \log_{10} D + 0.45$, (Matsumoto et al., 2003) which were obtained for groundwater level at wells in Japan. However, this threshold still supports the interpretation that the size of the responses is strongly correlated with the "distance-corrected magnitude". The differences are largely due to the consideration of transient, rather than sustained, water level changes, to which we can attribute the difference in intercepts, although the earthquake catalogue limits also differ. We see from the

larger slope in the present work that the furthest events, not investigated in Kitagawa and Matsumoto (1996) and Matsumoto et al. (2003), contribute coseismic responses at a rate below that one might expect based on the earlier thresholds.

The logistic regression analysis confirms, and quantifies, that earthquake magnitude and well–epicenter distance have very strong influence on the probability of coseismic detection at Tangshan Well, while the depth and azimuth of an earthquake have little or negligible effect. Similar formulae for the effects on the delays, and signal amplitude and length have been obtained via multivariate regression. The conclusions, firstly that earthquakes with larger magnitudes, and/or shorter well–epicenter distances are more likely to induce water oscillations at this well and secondly, that such earthquakes and/or those at shallower depths tend to have smaller delays before the initiation of detectable water level oscillations, and oscillations with larger amplitudes and longer duration, are hardly surprising. More importantly, the fitted formulae provide a means of estimating, or comparing, the observed signal characteristics, and identifying possible anomalies. Most importantly of all, the non-negligible residuals, particularly in the case of the delay function, indicate that there are factors controlling the rate of oscillation initiation in the aquifer that are not accounted for in the regression meta model. As there are no events of magnitude greater than 6.0 within 1125km of the well during the study period, earthquake induced changes in the well–aquifer system (Rojstaczer et al., 1995) do not appear to provide the reason. Instead they suggest the presence of large-scale local inhomogeneities in the well–aquifer system (cf. Igarashi et al., 1992).

The travel-time table of Kennett and Engdahl (1991) was used in the calculations to obtain the arrival times of the seismic waves. This travel-time table is of course radially symmetric, and the Tangshan Well is located in an ancient craton where seismic velocities may be significantly different from those in a spherical model down to a depth of several hundred kilometers. Our analysis identified, for events close to azimuth 0° or 180° from the well, a significant effect on the travel time delays. However, as there is no such effect on the detection probability, and there appears to be no effect on which seismic wave arrival immediately precedes the initial response, our conclusions are not affected. The effect of azimuth on the delay can be calculated using the regression model if required, while there was negligible effect on the maximum amplitudes and durations of the water level oscillations. Through a series of sensitivity analyses, we have also shown that all the results are also robust to possible ± 1 minute time reading errors in the well data.

Chapter 7

HMM and Mutual Information on GPS Measurements and Earthquakes

7.1 Introduction

Global Positioning System (GPS) was developed by the United States Department of Defense in the 1970s and 1980s as a means of global navigation primarily for military users. Several hardware and data processing developments (Blewitt, 1993) and the fact that GPS could potentially be used to observe ground motions led it to become a geodetic tool. It provides accurate real-time range measurements for point positioning. A constellation of 24 satellites were launched into near-circular orbits. Range measurements from 4 satellites are sufficient to solve the 3-dimensional (north, east, up) location of the receiver. A number of methods have been developed which can improve the accuracy of GPS measurements, such as Wide Area Augmentation System, Differential GPS, Inertial Navigation System and Assisted GPS. This appears to have caused the heteroscedasticity of the data mentioned in Section 7.4.

Geodetic anomalies preceding large earthquakes have long been of interest due to well documented pre-earthquake deformation rate changes observed before continuous GPS stations were widely deployed in the early 1990s (Roeloffs, 2006). The GPS measurements provide a good opportunity for scientists to further investigate pre-, co- and post-seismic deformation anomalies, but there is much 'noise' that needs to be filtered out of the observations. There has been considerable interest in whether GPS measurements have any predictive power for earthquake occurrences. Roeloffs (2006) reviewed the evidence for aseismic deformation rate changes prior to earthquakes and listed at least ten credible examples of tectonic earthquakes preceded by deformation rate changes. Ogata

(2007) detected slow slip during the three years period leading up to the October 2004 magnitude 6.6 earthquake in the Chuetsu area, central Japan, which was also exhibited in GPS observations around the rupture source. These anomalies were mostly from the apparent long-term pre-earthquake slip. However, it is very difficult to detect apparent displacement when there are anomalous changes of the geodetic measurements, and therefore techniques which can detect or extract subtle changes in GPS measurements, which may be related to earthquakes, are necessary.

Hashimoto et al. (2009) applied an inversion method based on Bayesian modelling (Matsu'ura et al., 2007) to horizontal and vertical velocities from GPS data and concluded that the slip-deficit zones identified with this method are potential source regions of large earthquakes. Granat and Donnellan (2002) and Granat (2003) introduced a hidden Markov model (HMM) based method which was applied to the GPS data from the southern California region. The application to the daily displacement time series collected in the city of Claremont, California clearly separated the states before and after the Hector Mine earthquake in October 1999. Granat (2006) applied this method to the daily GPS data from more than 100 stations of the Southern California Integrated Geodetic Network. The results revealed that approximately 70 out of the 127 stations had state changes on the day when the Hector Mine earthquake occurred. However, the different states are clearly dominated by the long-term trends of each component of the data, and the states are entered and existed only once. For example, see Figure 7.1, which corresponds to Figure 5 in Granat (2006). Thus the method is not suitable for predictive purposes.

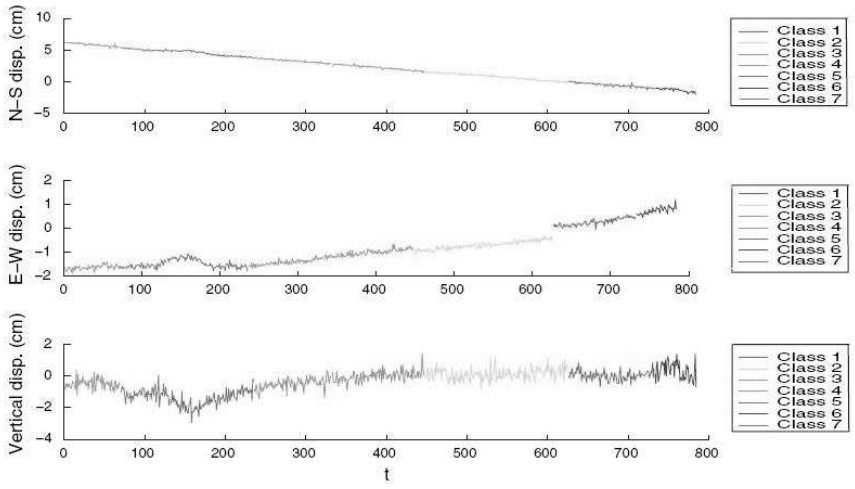


Figure 7.1: (Figure 5 in Granat, 2006) Classification results for a seven-state HMM applied to the GPS data set collected in the city of Claremont, California. (Granted with permission from Birkhäuser Verlag AG.)

Figure 7.2 shows the raw daily time series of the GPS measurements of deformation from a tectonically quiescent area, Alice Springs in Australia. We compare this daily time series with

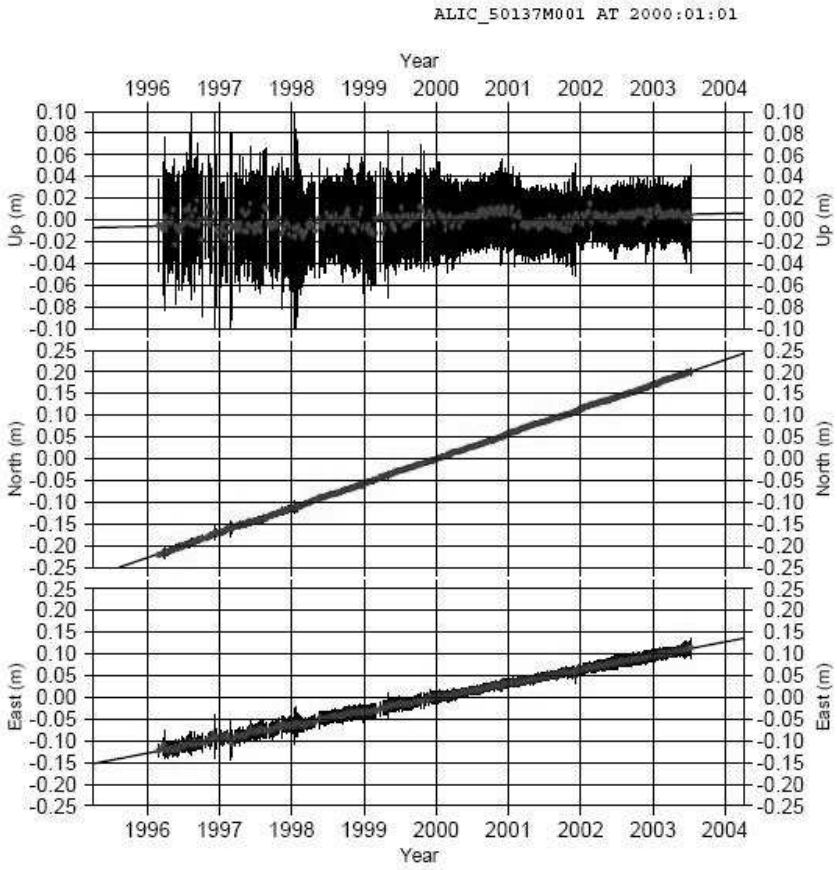


Figure 7.2: GPS measurements of deformation at Alice Springs, Australia.

the GPS movements from two tectonically active areas: Taupo (see Figure 7.7 for the average of the GPS measurements from the three stations, HAMT in Hamilton, NPLY in New Plymouth, and WANG in Wanganui, minus that from Taupo) and Southern California (see Figure 7.33 for the baselines between the Chilao Flats (CHIL) and the station LBC2 in Long Beach). Note that the time series from the latter two cases have the reference frame errors canceled by averaging and leveling (cf., Section 7.3.2 and Section 7.4.1). The daily time series from Alice Springs are evenly distributed around the linear trend line, whereas for the time series around Taupo and in Southern California, irregular spikes, occasional step jumps and trend changes are quite obvious from all three components, north, east and up. This suggests that the data from the tectonically active areas around Taupo and in Southern California may behave differently from that from a tectonically quiescent area, and that the trend of each of the GPS components from the tectonically active areas varies. The changes in the trend may be related to active earthquake occurrences (See also Section 7.2.1,

the choice of the number of hidden states for HMMs).

Therefore a non-linear filter is introduced for the GPS process which serves as a smoothing tool. It is able to extract useful signals which are distinguishable from the majority of the data. This non-linear filter is the range of the trends estimated for the GPS measurements in the previous 10 days. Essentially, the trend measures the short-term deformation rate. Consequently, this nonlinear filter calculates the maximum deformation rate changes in the previous 10 days, and reveals anomalous spikes in the deformation rate changes. For comparison, Ogata (2007) fitted a straight line to the time series of the baseline distance between each two stations around the 2004 Chuetsu earthquake during the period from 1997 towards the end of 2000. He then extrapolated the linear trends until 23 October 2004 and detected deviation of the time series from the predicted linear trends. The cumulated slip preceding the 2004 Chuetsu earthquake is estimated to be roughly equivalent to $M_w 6$, which is very close to the magnitude of the Chuetsu earthquake, $M_w 6.6$. This deviation is actually a long-term deformation rate change, whereas the variable considered in this chapter is the maximum short-term deformation rate changes in an interval of fixed length.

Studies concerning earthquake genesis indicate that seismicity may have a cyclic nature of some sort (Sammis and Smith, 1999; Jaumé and Sykes, 1999; Vere-Jones et al., 2001; Jaumé and Bebbington, 2004). GPS measurements of deformation may indirectly reflect the underlying dynamics (the unobservable or hidden states) for earthquake occurrences with large variations relating to a period of accelerating moment release and small variations corresponding to a period of quiescence (see also Section 7.2.1, the choice of the number of hidden states for HMMs). Let us assume that the hidden states switch among some phases in a seismic cycle. The transitions between phases or rather states may be assumed to be governed by a Markov chain. The variation of the deformation rate depends on the unobservable phases. On the basis of this assumption, HMMs can be used to extract signals from GPS data which may be related to earthquakes.

The nonlinear filter, which helps to extract subtle changes from the noise in the GPS data, and the underlying dynamics, which drives the observed process, form an HMM framework. The HMM can be used on the filtered GPS measurements to categorize the data into different states and then investigate the possible link between each category and the earthquakes. Finding a region which contains both long GPS data and enough earthquakes is not easy. Two data sets are considered, one from central North Island, New Zealand and the other from Southern California. First the nonlinear filter is introduced, which is a functional of the GPS data, and then HMMs are fitted to this functional, assuming the underlying dynamics switch among a finite number of phases in a seismic

cycle. The model will classify the data into different states, each state suggesting particular dynamics. The Viterbi algorithm (Viterbi, 1967; Forney, 1973) is thus used to track the most probable sequence of states from the GPS data. After that, the mutual information (MI) between each state from the most likely state sequence and the earthquake occurrences is calculated to examine if there is any association. A cross validation by dividing the entire data set into two parts is conducted for the data around Taupo after this examination to verify the declared association. A possible way of declaring a "Time of Increased Probability" (TIP) for the considered region is then discussed. Consequently, probability forecasts in time are investigated based on a Logistic regression model.

7.2 Methodology

7.2.1 Hidden Markov Model and Mutual Information Analysis

More than one GPS station will be used and the baselines between stations are calculated in order to cancel the reference frame errors. This provides us three series of GPS measurements, north, east and up. For the three series, the changes of the trend of GPS movements in the previous 10 days are considered. Take the north component N_t for example. At each time point t, a line is fitted to the data $\{N_s : t-9 \leq s \leq t\}$ using linear regression and the slope of this line is denoted as T_t. Then for each t, calculate

$$R_t = \max\{T_s : t-9 \leq s \leq t\} - \min\{T_s : t-9 \leq s \leq t\}, \tag{7.1}$$

which is the range of deformation rate changes in the previous 10 days, in other words, variation of deformation rate in the previous 10 days. Note that actually 20 days of data are used for calculating the variable R_t at t. As a result, the first 19 days during the selected time period will be excluded from the following study.

This non-linear filter for the GPS process serves as a smoothing tool. As defined in Equation (7.1), this nonlinear filter is the range of the trends, T_t, in the previous 10 days. Essentially, the trend T_t measures the short-term deformation rate. Therefore, this nonlinear filter calculates the maximum deformation rate change in the previous 10 days. Note that this filter is related to the maximum acceleration, $a_{\max}$, in the previous 10 days through the formula

$$R_t = \frac{T_{\max} - T_{\min}}{\Delta t_{\max\,-\,\min}} \cdot \Delta t_{\max\,-\,\min} = a_{\max} \cdot \Delta t_{\max\,-\,\min},$$

where $T_{\max} = \max\{T_s : t-9 \leq s \leq t\}$, $T_{\min} = \min\{T_s : t-9 \leq s \leq t\}$, and $\Delta t_{\max-\min} = |\arg\max_{t-9\leq s\leq t}\{T_s : t-9 \leq s \leq t\} - \arg\min_{t-9\leq s\leq t}\{T_s : t-9 \leq s \leq t\}|$. The purpose of this filter is to identify anomalous spikes in the deformation rate.

After calculating the range of deformation rate change for each component, north, east and up, a hidden Markov model (HMM) is used to fit this multivariate data set. A multivariate normal distribution is adopted for the observations (which are the ranges of deformation rates as calculated in Equation (7.1)). After fitting the model and estimating the parameters, we want to choose a corresponding state sequence which best "explains" the observations. The Viterbi algorithm (Viterbi, 1967; Forney, 1973) can be adopted to find the single best state sequence for the given GPS observations, i.e., to track the most probable sequence of states for the observed data. The algorithm finds the sequence of states with maximum probability of occurrence, which is different from maximizing the individual likelihood of being in state S_i at time t given the observation sequence and the fitted model. This algorithm will classify the GPS data into several different categories which can be shown in a 3-dimensional (north, east, up) plot.

For the problem of how many states for the HMM we shall use, a comparison study via simulation experiment was carried out by Bebbington (2007) to investigate the performances of the four model selection criteria, Akaike Information Criterion (Akaike, 1974), Bayes Information Criterion (Schwarz, 1978), the corrected AIC (Hurvich and Tsai, 1989) and the penalized minimum-distance (MacKay, 2002), for selecting the best number of hidden states for HMMs. The simulation experiment showed that the AIC most consistently estimated the correct number of hidden states, for the small sample sizes common in volcanology, but for the larger sample sizes here, the asymptotic properties of BIC are more appropriate. For this analysis of the GPS measurements, although the sample sizes are large, the number of states will be determined by a different criterion. It will be selected in order to minimize the number of parameters used, by avoiding detailed modelling of the 'noise', while extracting as much 'signal' as possible. More details are provided in the first example.

After determining which model to use, mutual information is used to examine whether there is any association between each state and the earthquake occurrences. In order to calculate the mutual information between the hidden states, denoted by U, and the earthquake occurrence times, denoted by V, the hidden states are transformed into 0-1 sequences, with 1 indicating the state we are interested in, 0 otherwise. For example, if we are interested in a state, say s, of a hidden Markov model, then all the days in which the HMM occupies state s will be set as 1, and the days the HMM spends in the other states will be set as 0. In a similar way, the earthquake series have a 1 in the

day of the earthquake occurrence, 0 otherwise. The mutual information between two series $\{U_{t+u}\}$ and $\{V_t\}$ is examined, where u is the time lag. The mutual information will then be plotted and compared with the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent. Significant association at negative time lags u's suggests that the current earthquake occurrence is strongly related with the GPS movements in that state some u days preceding the occurrence. If the significant association shows up at positive time lags, it means that the earthquakes have significant influence on the GPS movements which are observed some u days following the earthquake occurrences.

Note that the approximate 95% confidence interval is calculated for two processes which are not autocorrelated. However, the Viterbi path is autocorrelated being a Markov chain. In order to check whether this will affect the confidence interval band, we use the estimated transition probability matrix of the model to simulate 1000 series of autocorrelated sequences of Markov chains and use the number of earthquakes divided by the total number of days as the rate to generate 1000 series of Poisson processes. We then calculate the mutual information between the 1000 pairs of each Markov chain and Poisson process and get the simulated upper 95% confidence level for comparison.

7.2.2 Probability Forecast Using a Logistic Probability Model

A 'Time of Increased Probability' (TIP) for large earthquakes in a specified 'region of investigation' was first introduced by Keilis-Borok and Kossobokov (1990) when they developed the M8 algorithm to explore the use of pattern recognition methods to distinguish regions or time periods of heightened earthquake risk. Harte et al. (2007) used the Critical Series developed by Harte et al. (2003) to declare a TIP. They then utilized the series as a predictor variable to produce probability forecasts of a target event in the current time interval. From the mutual information analysis, if some state shows significant association with the earthquake occurrences at negative time lag u, it means that the GPS movements in this state are correlated with the earthquakes occurring some u days later. We therefore consider this state a precursory state. We can then define a TIP as follows. We divide the entire time period into non-overlapping intervals J_i with equal length of l days. For each interval J_i, if there is any day in which the HMM visits the precursory state, we consider the following interval J_{i+1} a TIP. The interval length l will be determined according to the mutual information results.

The problem of interest is how the GPS deformation rate ranges influence the probability of declaring a TIP. The objective here is to determine how the explanatory variable, some variable of the GPS measurements, in the interval $\{t : t \in J_i\}$ influences the probability, $P(J_{i+1})$, that a

large earthquake occurs in the interval J_{i+1}. The continuous explanatory variable suggests that a regression analysis may be suitable. The response variable, either a large earthquake occurrence in the interval J_{i+1} or no earthquake occurrence in this interval, which takes on the value 1 or 0, is binary. Instead of using a regression directly for the success probability, one usually transforms the probability scale from the range $(0, 1)$ to $(-\infty, \infty)$, and then uses a regression for the transformed values. Logistic transformation is usually recommended because it is more convenient. Moreover, it provides a direct interpretation in terms of the logarithm of the odds of success, which is defined to be the ratio of the probability of a success to the probability of a failure, $P(J_i)/(1 - P(J_i))$. Hence we can use logistic regression with binomial errors (see Collett, 1991 for example). Let M_i be some variable of the GPS measurements in the interval J_i. The variable M_i is used as a linear predictor of a target event in the time interval J_{i+1} in a Logistic linear regression

$$f(J_{i+1}) = \beta_0 + \beta_1 M_i,$$

where

$$f(J_{i+1}) = \log\left(\frac{P(\text{at least one earthquake in } J_{i+1})}{1 - P(\text{at least one earthquake in } J_{i+1})}\right)$$

is the log-odds ratio. We obtain the probability of a target event in the time interval J_{i+1} through

$$P(\text{at least one earthquake in } J_{i+1}) = \frac{\exp\{f(J_{i+1})\}}{1 + \exp\{f(J_{i+1})\}}.$$

We will fit this model to the data and examine whether there is any significant influence of the GPS deformation rate ranges on the probability of large earthquake occurrences. The parameters in this model will be estimated using the glm function in R package.

7.3 Data from Central North Island, New Zealand

We consider data from central North Island, New Zealand, located near the boundary of the Australian tectonic plate. According to DeMets et al. (1994), the Pacific and Australian tectonic plates are converging obliquely at about 42 mm/yr, accommodated by subduction of the Pacific plate and deformation of the overlying Australian plate (see Figure 7.3, which corresponds to Figure 1 in Reyners et al., 2006). The figure also outlines the Taupo Volcanic Zone (TVZ), an active continental rift in the central North Island. Wallace et al. (2004) showed that the TVZ region has a total

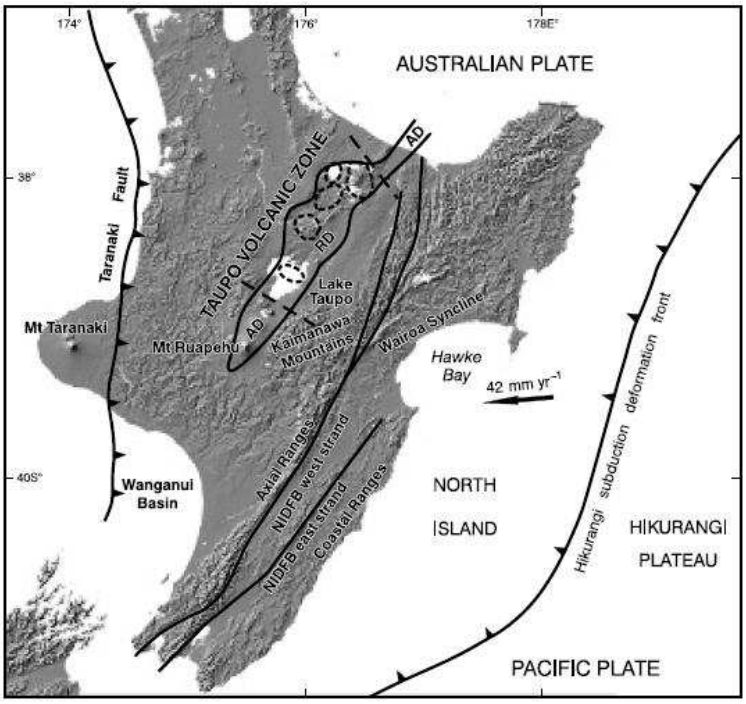


Figure 7.3: (Figure 1 in Reyners et al., 2006) Digital elevation map of the central North Island, New Zealand and its tectonic settings. The arrow indicates the velocity of the Pacific plate relative to the Australian plate (DeMets et al. 1994). The Taupo Volcanic Zone is outlined and subdivided into three distinct segments: andesite-dominant (AD) cone volcanoes in the northern and southern segments, and rhyolite-dominant (RD) caldera volcanoes (shown dashed) in the central segment (Wilson et al. 1995). NIDFB denotes the North Island Dextral Fault Belt. (Granted with permission from Wiley-Blackwell.)

predicted extension rate of about 15 mm/yr at the Bay of Plenty, decreasing to < 5 mm/yr near the southwestern termination of the TVZ. They also showed evidence of clockwise rotation of the forearc east of the TVZ, which is accommodated in the backarc by extension of the TVZ in the north, and by compression of the Wanganui Basin in the south (Reyners et al., 2006). Reyners (2009) suggested that the stresses driving the opening of the TVZ continue to be active at its southern termination which is marked by extensive earthquake activity throughout the crust to depths in excess of 40 km.

The GPS data can be obtained from the New Zealand GeoNet website (http://www.geonet.org.-nz/resources/gps/, last accessed on March 31, 2009). The daily estimates of relative coordinates are determined through the GPS processing using Bernese v5.0 software developed by the Astronomical Institute of the University of Bern. The coordinates and their formal uncertainties are converted to three dimensional displacement, north, east and up, in millimeters from an initial point. GeoNet removed some dubious data which are very different from the neighboring data points. We exam-

ine the resulting 'raw' daily GPS time series around Taupo. The continuous GPS stations around Lake Taupo which have the longest records available online for GPS measurements are HAMT in Hamilton, NPLY in New Plymouth, and WANG in Wanganui which are located on the west side of the TVZ; DNVK in Dannevirke, GISB in Gisborne and HAST in Hastings which are located on the east side of the TVZ; and TAUP in Taupo. See Figure 7.4 for the locations of the stations. The information on the GPS measurements at these stations are summarized in Table 7.1. The

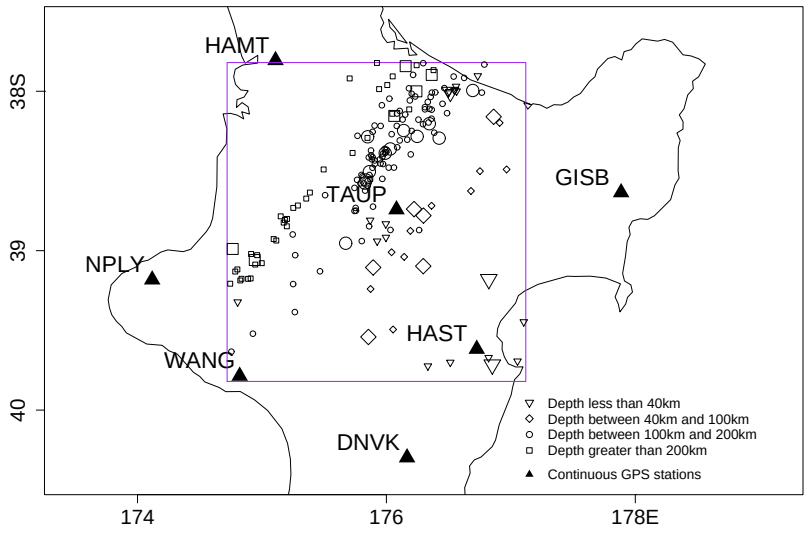


Figure 7.4: Location map for the GPS stations and earthquakes around Lake Taupo. The symbol ▲ indicates the location of a GPS station. The earthquakes are selected from the rectangular area. The small size of the symbols for earthquakes is for earthquakes with magnitude larger than 4.2 and smaller than 5.1; the large size for earthquakes with magnitude larger than 5.1. The maximum magnitude of the earthquakes in this area is 5.9.

missing data information during the time period from 2003.06.02 to 2008.12.31 is shown in Table 7.1 as well. The longest missing interval for NPLY is 33 days which occurred in early 2004 (2004.01.29—2004.03.01), other than that, the three stations on the west side of the Taupo Volcanic Zone have less missing data than the other three. So we will use the GPS measurements from HAMT, NPLY and WANG from 2004.03.02 to 2008.12.31 (1766 days) to avoid the 33 days missing data referred to above. Other missing data are interpolated by setting each of them as the mean of the non-missing data within the 10 days ahead of and 10 days following the missing point. Note that this is conservative in the model defined in (7.1).

Earthquakes in a rectangular area around Lake Taupo are chosen between latitude 37.82S and

Table 7.1: Information of the GPS measurements around Taupo, where N_m is the number of missing data from 2003.06.02 to 2007.12.31, L_{m1} and L_{m2} are the numbers of missing data in the intervals which have the most and the second most amount of missing data for each station, and 'Direction of movement' means the direction of long-term movement of each station relative to TAUP.

Station	Available record	N_m (day)	L_{m1} (day)	L_{m2} (day)	Direction of movement
DNVK	2002.06.05 —	37	4	4	southwest, *down*
GISB	2002.05.14 —	58	15	10	southwest, *down*
HAMT	2003.05.13 —	14	3	3	northwest, *down*
HAST	2002.09.19 —	46	12	5	southwest, *down*
NPLY	2003.03.20 —	88	33	10	northwest, *down*
TAUP	2002.03.22 —	16	5	3	N/A
WANG	2003.04.10 —	12	3	3	northwest, *down*

39.82S and between longitude 174.72E and 177.12E, from 2004.03.02 to 2008.12.31. The Gutenberg-Richter law states that the total number of earthquakes in a population that are larger than or equal to some magnitude M varies as 10^{bM} (Ishimoto and Iida, 1939; Gutenberg and Richter, 1944). It suggests that $\log_{10}(\text{Proportion of events with magnitude} > M)$ is proportional to bM. The frequency–magnitude plot is shown in Figure 7.5. We can see that there is a change of the slope at magnitude 4.2 and another change at magnitude 5.1. We will therefore consider two cases: earthquakes with

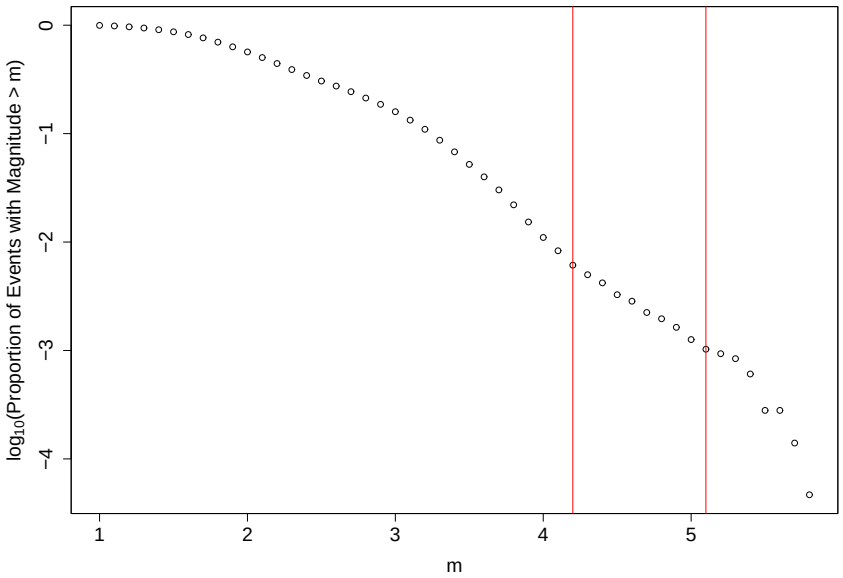


Figure 7.5: Frequency–magnitude plot for the earthquakes during the period 2004.03.02 to 2007.12.31 in the area between latitude 37.82S and 39.82S and longitude 174.72E and 177.12E.

magnitude larger than or equal to 4.2 (178 events in total) and those larger than or equal to 5.1 (27

events in total). See Figure 7.4 for the location map of the earthquakes.

7.3.1 NHMM and MMGLM Analyses of the Earthquake Data with Ancillary GPS Measurements

As a preliminary look to see whether incorporating the original non-filtered GPS baselines into an HMM framework will help classify the underlying dynamics into different regimes, an exploratory analysis is carried out using the three different HMMs reviewed in Chapter 2. Firstly, a discrete-time HMM with binary distribution is fitted to the earthquake occurrences around Taupo. Secondly, an NHMM is used with binary earthquake occurrences as observations and the first differences of the GPS measurements (three components, north, east and up) as covariates which influence the transition probabilities via the model (2.17). Thirdly, the data are examined via an MMGLM with binary earthquake occurrences as observations and the distance of the first differences of the GPS measurements to the origin (0,0,0) as a covariate which influences the observed earthquake sequence via a generalized linear model with Bernoulli family and logit link (2.26). We take a look at an example with three hidden states for each of the three models. The estimated most likely sequence of states for each model is tracked via the Viterbi algorithm. The earthquake occurrences and the Viterbi path for each of the models are plotted in Figure 7.6. We can see that the state changes in the HMMs appear to be following the earthquakes rather than preceding them, and thus incorporating the GPS measurements into the HMM framework directly does not seem to provide any precursory information. The three models are very similar in outcome.

As mentioned in Section 7.1, the irregular spikes in the GPS measurements might be related to earthquake activity. However, using the original data (first differences) do not seem to help extract any signals. Therefore, we need to consider some other methods to extract the potential signals, such as the non-linear filter presented in the previous section.

7.3.2 Hidden Markov Model and Mutual Information Analyses

Analysis of the Entire Data

The GPS measurements of deformation at the three stations, HAMT, NPLY and WANG, are averaged and then subtracted from the measurements at TAUP station, to cancel the reference frame errors. We get three series of GPS measurements, north, east and up as shown in Figure 7.7. For the three series, consider the variable R_t, defined in (7.1), of GPS movements, i.e., the range of

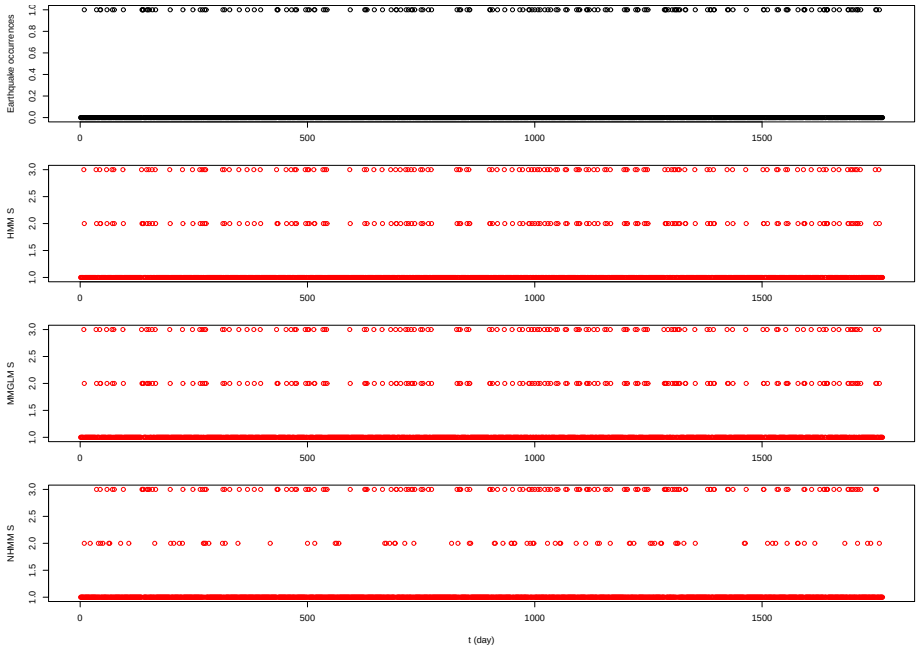


Figure 7.6: Earthquake occurrences and Viterbi paths from the HMM, NHMM and MMGLM analyses.

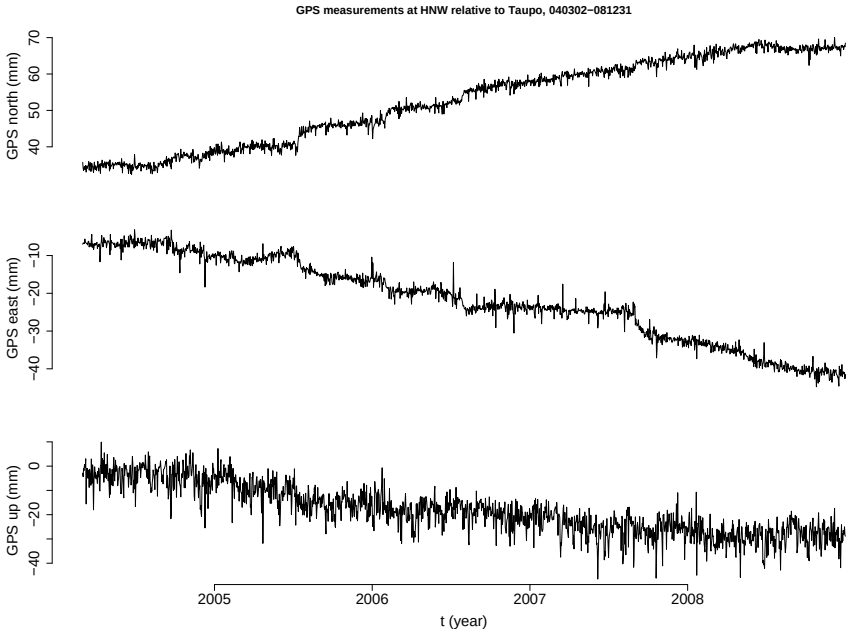


Figure 7.7: GPS movements of HAMT, NPLY and WANG relative to TAUP.

deformation rate observed in the previous 10 days. HMMs starting with 3 hidden states are fitted to the data. For the sake of easy comparison between models with different numbers of states, the states are ordered according to the Euclidean distance of the estimated means in each state from the origin (0,0,0). The log likelihood and the BIC value of each model are listed in Table 7.2. We can

Table 7.2: The log likelihood (LL) and BIC values of the fitted models to the entire data around Taupo and in Southern California (CA), where '# of S' is short for the number of states in the model.

Taupo					
# of S	LL	BIC	# of S	LL	BIC
3	1660.57	-3127.03	8	2872.05	-4915.41
4	2013.22	-3735.28	9	3047.08	-5093.76
5	2302.17	-4201.20	10	3166.47	-5145.91
6	2549.70	-4569.34	11	3220.44	-5052.27
7	2744.21	-4816.51			
CA					
# of S	LL	BIC	# of S	LL	BIC
3	528.01	-841.60	8	3336.89	-5758.41
4	1337.08	-2352.55	9	3573.67	-6042.29
5	2023.04	-3600.77	10	3787.20	-6263.17
6	2586.31	-4587.11	11	4122.24	-6710.59
7	2999.45	-5256.69	12	4178.39	-6583.75

clearly see that the log likelihood is monotonically increasing and the BIC reaches its maximum when 10 hidden states are considered which implies 159 parameters for less than 2000 data, too many for stability.

The purpose of this exploratory GPS data analysis using HMMs is to look for precursory signals. Let us compare the precursory information captured by each of the models. For each of the models with up to 10 hidden states, the Viterbi algorithm is used to track the most probable hidden state sequence corresponding to the GPS measurements. It categorizes the data into different classes each having the mean and standard deviation vectors from the corresponding state. As discussed in Section 7.1, the trend of each of the GPS components from tectonically active areas varies largely compared to that from a tectonically quiescent area. This suggests that the large variations in the trend may be related to active seismicity. As the aim is to examine whether a large variation of the trend has any predictive power for large earthquakes, consequently the largest state, the state which is the furthest from the origin, in each model will be examined. This state, on average, has the largest variations in the deformation rate. Then, for each model, firstly, the hidden states, denoted by U, and the earthquake occurrence times with minimum magnitude 5.1, denoted by V, are transformed into 0-1 sequences as described in Section 7.2.1; secondly, the mutual information between the two series $\{U_{t+u}\}$ and $\{V_t\}$ are calculated with a time lag u. The mutual information between the largest state of each of the models and earthquakes with minimum magnitude 5.1 is shown in Figure 7.8.

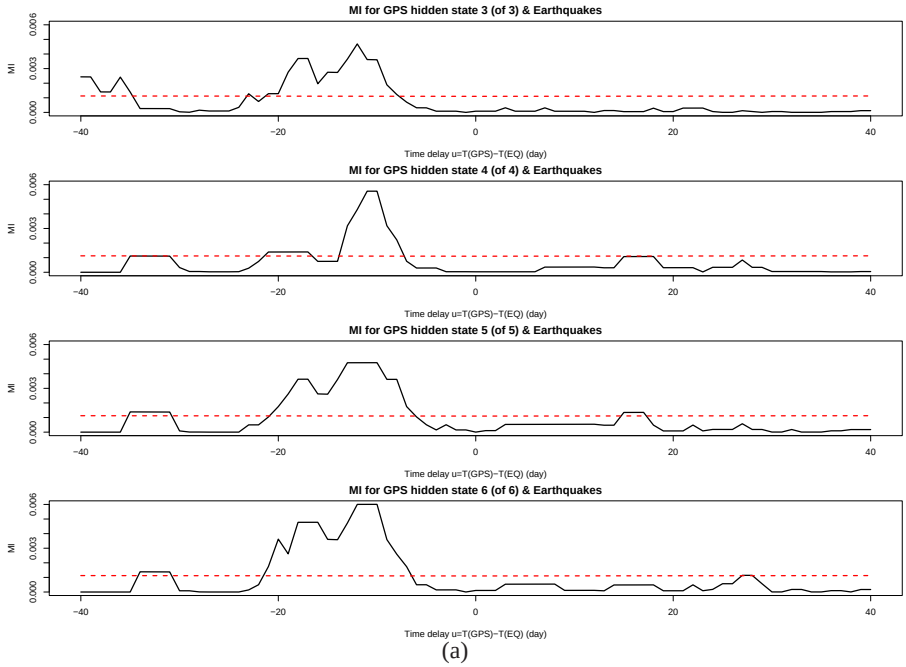
Figure 7.9 shows the mutual information between the state which mainly accounts for the north-component movement of each of the HMMs with 7, 8, 9 and 10 hidden states and the earthquakes. We can see that both the 'largest' state with the greatest variation in the deformation rate, and the state which mainly accounts for the north-component movement present preseismic information. Table 7.3 lists the number of precursory states in each of the model with up to 10 hidden states and the proportion of data in the precursory state(s). Examining the precursory states for the HMM with 10 hidden states down to 6 or 5 hidden states, we can clearly see that the two precursory states meld into one state in the 5 or 6 state model. Moreover, we are interested in extracting signals from the GPS measurements, rather than modelling the noise perfectly which is what the model selected by the BIC tries to do. We want the concentrated information in one state, and thus will concentrate on the class of movements with the greatest variation. There is a subtle distinction, and thus we will try and select the number of states on a slightly different criterion.

Table 7.3: The proportion of data in the precursory state(s) for data around Taupo, where '# of S' is short for the number of states in the model, '# of P' is short for the number of precursory states in the model, and 'proportion' is the proportion of data in the precursory state(s).

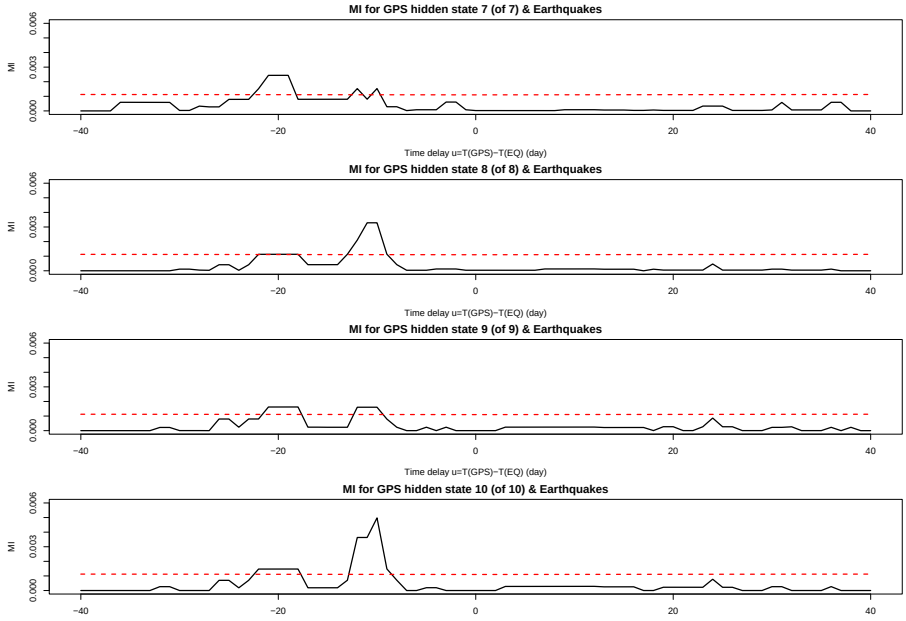
# of S	# of P	proportion	# of S	# of P	proportion
3	1	0.273	7	2	0.187
4	1	0.139	8	2	0.163
5	1	0.150	9	2	0.154
6	1	0.151	10	2	0.163

Beginning from a three-state model, we add more states one at a time. When the models start to consistently capture this state with very close means and variances (the distance from the mean μ_{kj} in (2.12) in Section 2.2 of this state in the s state model to that of this state in the $s-1$ state model is less than 0.1mm; for each component the difference of the standard deviations σ_{kj} in (2.13) in Section 2.2 between the two models is less than 0.05mm) and consisting of nearly the same amount of the data, say a% of the data (i.e., the proportion of time in this state is constant with ± 1%, which is the stationary distribution of the Markov chain occupying this state as defined in Appendix D), we stop adding more states. The model with the smallest number of parameters among those whose largest states consist of a% of the data is then chosen.

The state transitions of the fitted 3, 4, 5 and 6 state models are shown in Figure 7.10. We see that the proportion of time that the four models spend in the state furthest from the origin (0,0,0) are respectively 0.273, 0.139, 0.150 and 0.151. The fitted 5 and 6 state models consistently capture the



(a)



(b)

Figure 7.8: For data around Taupo. (a): Mutual information between the largest state in each of the HMMs with 3, 4, 5 and 6 hidden states and earthquakes of minimum magnitude 5.1 with a time lag u. (b): Mutual information between the largest state in each of the HMMs with 7, 8, 9 and 10 hidden states and earthquakes of minimum magnitude 5.1 with a time lag u. The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent.

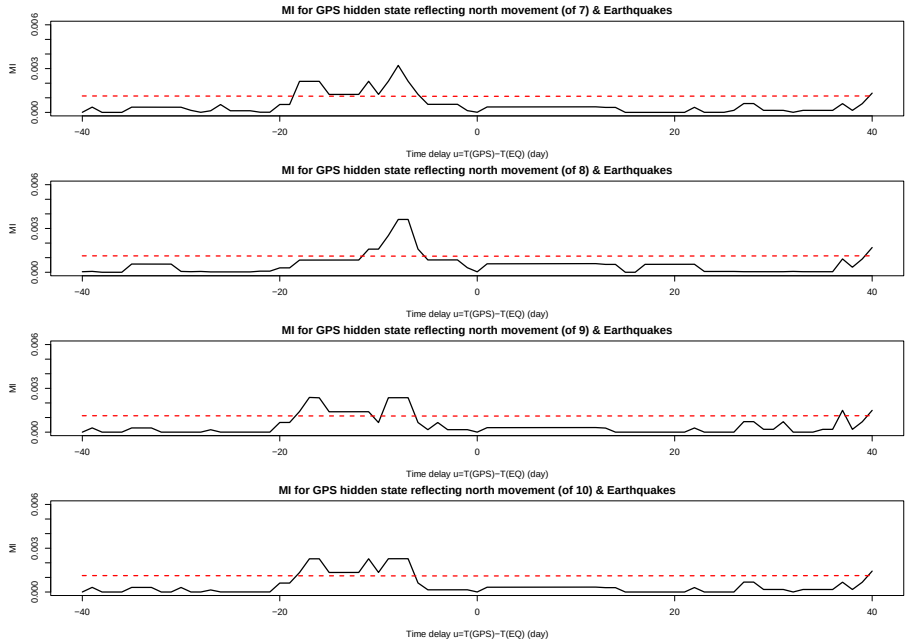


Figure 7.9: For data around Taupo. Mutual information between the state which accounts for the north-component movement in each of the HMMs with 7, 8, 9 and 10 hidden states and earthquakes of minimum magnitude 5.1 with a time lag u. The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent.

Table 7.4: The parameter estimates of the fitted five-state HMM for data around Taupo. The normalized means are obtained by subtracting the mean vector of state 1 from the mean vector of each state and then dividing the result by the standard deviation vector of state 1.

	State	1	2	3	4	5
Estimated	north	0.171	0.338	0.240	0.224	0.514
means	east	0.254	0.261	0.248	0.579	0.511
	up	1.030	1.140	1.840	1.810	1.960
Estimated	north	0.044	0.066	0.066	0.057	0.138
standard	east	0.094	0.088	0.072	0.190	0.233
deviations	up	0.260	0.310	0.300	0.680	0.742
Estimated	state 1	0.911	0.042	0.035	0.012	0.000
transition	state 2	0.037	0.892	0.030	0.004	0.037
probability	state 3	0.053	0.027	0.889	0.017	0.014
matrix	state 4	0.033	0.002	0.027	0.924	0.014
	state 5	0.000	0.040	0.015	0.032	0.914
Normalized	north	0	3.8	1.6	1.2	7.8
means	east	0	0.1	-0.1	3.5	2.7
	up	0	0.4	3.1	3.0	3.6

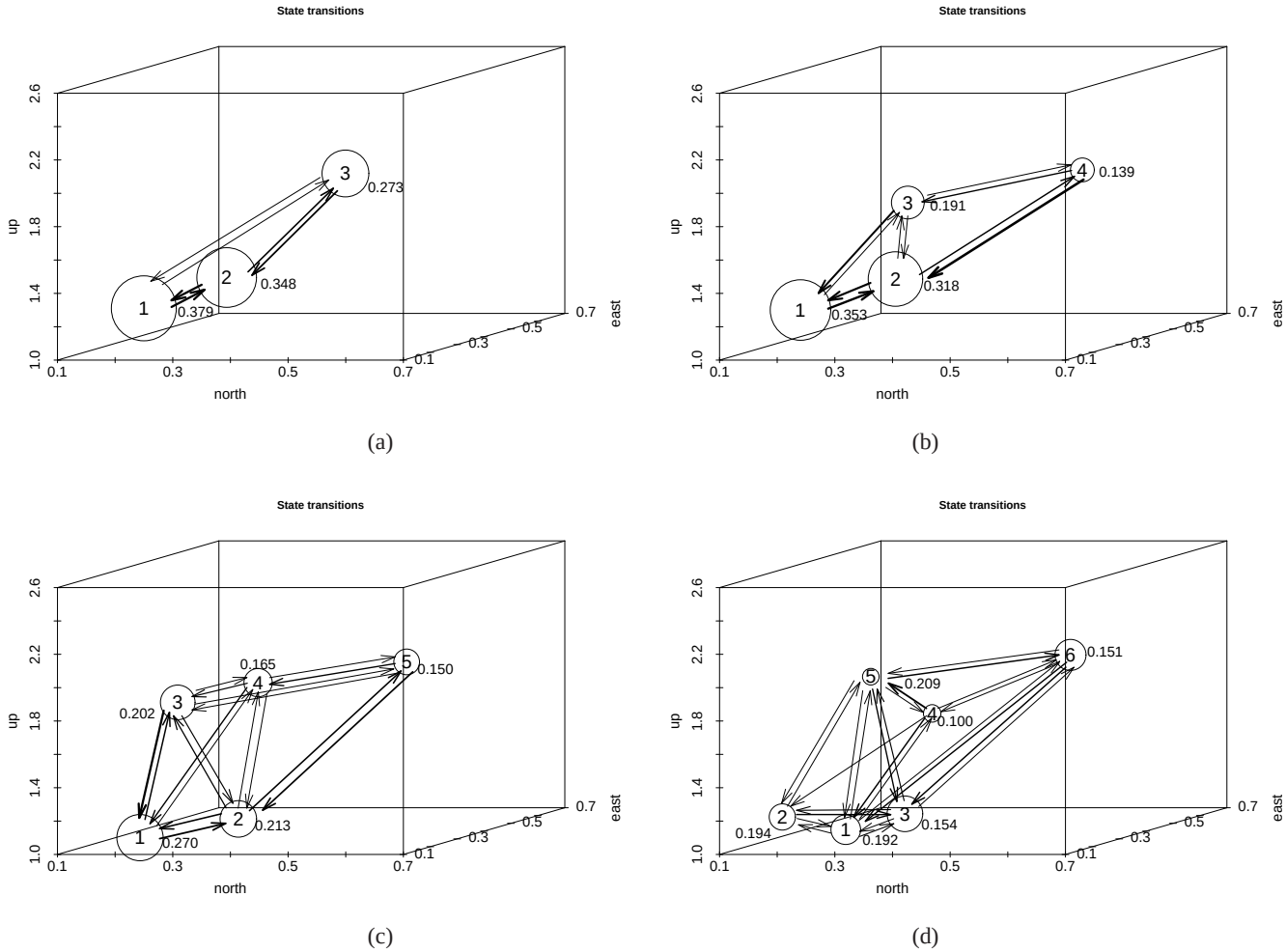


Figure 7.10: Illustration of state transitions for HMMs with 3 states (a), 4 states (b), 5 states (c) and 6 states (d) for data around Taupo. Circle sizes are proportional to the stationary distribution of the hidden Markov chain, i.e., the proportion of time in each state, which is shown as the number besides each circle. The thickness of the arrows is proportional to the transition probabilities. The numbers inside the circles indicate the states and are located at the means of the states estimated from the fitted model.

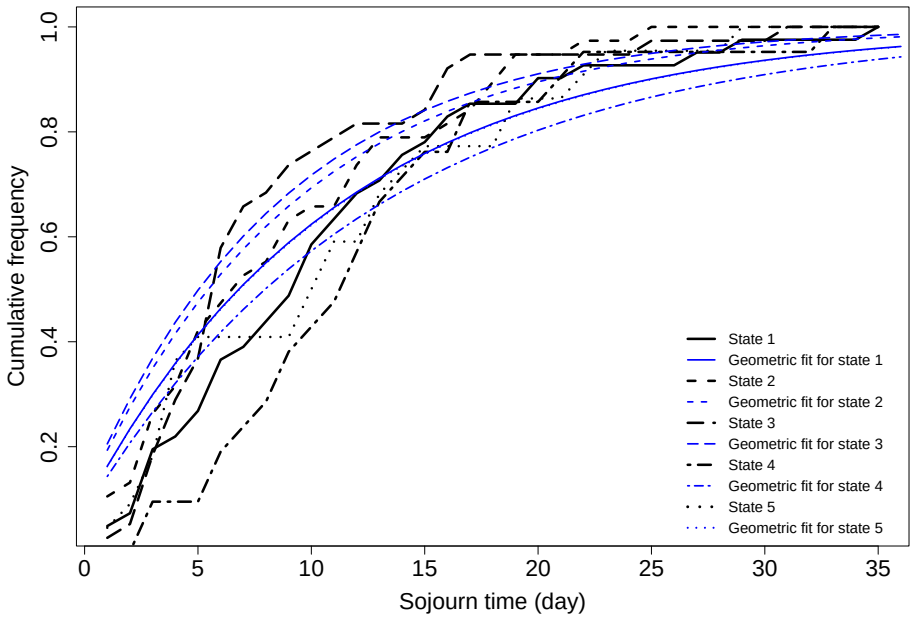


Figure 7.11: Sojourn time distribution for the fitted five-state HMM (thick lines) to the entire data, with the thin lines (for comparison) indicating Geometric distributions, each with mean calculated from the sample mean of the sojourn time in each state (Taupo).

furthest state with nearly the same mean and standard deviation for each GPS component, consisting of about 15% of the entire data. We therefore consider the 5 state hidden Markov model on the ground of the criterion discussed above. Given that the sojourn time in any state of a homogeneous discrete-time Markov chain has a Geometric distribution, we conduct a K-S test for the sojourn time distributions of the fitted five-state model. The P-values of the tests for State 1 to State 5 are, respectively, 0.1340, 0.2211, 0.0356, 0.1628, and 0.5242, which are not significantly different from the null hypothesis, given the multiple comparison issue. The sojourn time distribution for each state is shown in Figure 7.11.

The fitted five-state HMM results are shown in Table 7.4. State 5, the furthest from the origin, has the largest variances for all three components, and apart from the east component, the largest mean. State 5 does not transit to State 1 and vice versa. The stationary distribution for the Markov chain, i.e. the proportion of time in each state, is shown in Figure 7.10(c). The Markov chain spends the least amount of time in State 5 and mostly occupies State 1. State 1 has the least variation in trend ranges which we consider as a ground state. If we normalize the five states, by subtracting the mean vector of state 1 from the mean vector of each state and then dividing the result by the standard deviation vector of State 1, we see how each state moves relative to the ground state. The resulting normalized means are shown in Table 7.4. State 2 deviates from the ground state largely in

the north component. The main deviation of State 3 from the ground state is in the up component. Both the 4th and 5th states appear to deviate in the east and up components, while the 5th state also has the largest deviation in the north component.

In theory, the standard errors may be calculated using the approximate Hessian of the negative log likelihood at the minimum. However, this is only valid asymptotically, and the standard errors obtained this way will be very unstable if the sample size is not large enough (McLachan and Peel, 2000) for the number of parameters. With 55 parameters for less than 2000 data, the results will not be reliable, and the errors will be highly correlated. Instead, to see how well the HMM fits the data, we compare the histogram of the deformation rate ranges of each GPS component (north, east and up) and the mixed density of the five states,

$$\sum_{i=1}^{5} \pi_i f_i(x, \mu_{ki}, \sigma_{ki}),$$

where $\pi = (\pi_1, \cdots, \pi_5)$ is the stationary distribution of the estimated transition probability matrix, $f(\cdot, \cdot, \cdot)$ is the normal density function, k indicates either north, east or up component, μ_{ki} is the estimated mean of component k in State i, and σ_{ki} is the estimated standard deviation of component k in State i. Figure 7.12 shows the plots of the histograms of the deformation rate ranges of all three components along with the estimated mixed density. We can see that the fitted five-state model captures very well the main features of the data.

The deformation rate changes of GPS east component versus north, up versus north, and up versus east are plotted in Figures 7.13 to 7.15, with five symbols indicating the different classes (or states) in which the points are categorized (using the Viterbi algorithm). We see that the ellipse which illustrates the projected standard deviations of State 5 on the north-east surface centered at the estimated means does not overlap with that of the other states on that surface. Neither does that on the north-up surface. However, a large part of the standard deviation ellipse of State 5 on the east-up surface is covered by that of State 4 (Figure 7.15). It is obvious that State 5 (or the data in the 5th class) is clearly separated from the other states, except on the east-up surface, where it is mixed with State 4. The most likely state sequence tracked using the Viterbi algorithm and the trend ranges of the GPS movements as calculated in Equation (7.1) are plotted in Figure 7.16. The occurrence times of the earthquakes with magnitude larger than or equal to 5.1 are also shown in this figure. We can clearly see spikes in the trend ranges before some of the earthquake occurrences. Notice that when the HMM visits State 5, a large earthquake often follows. In order to examine

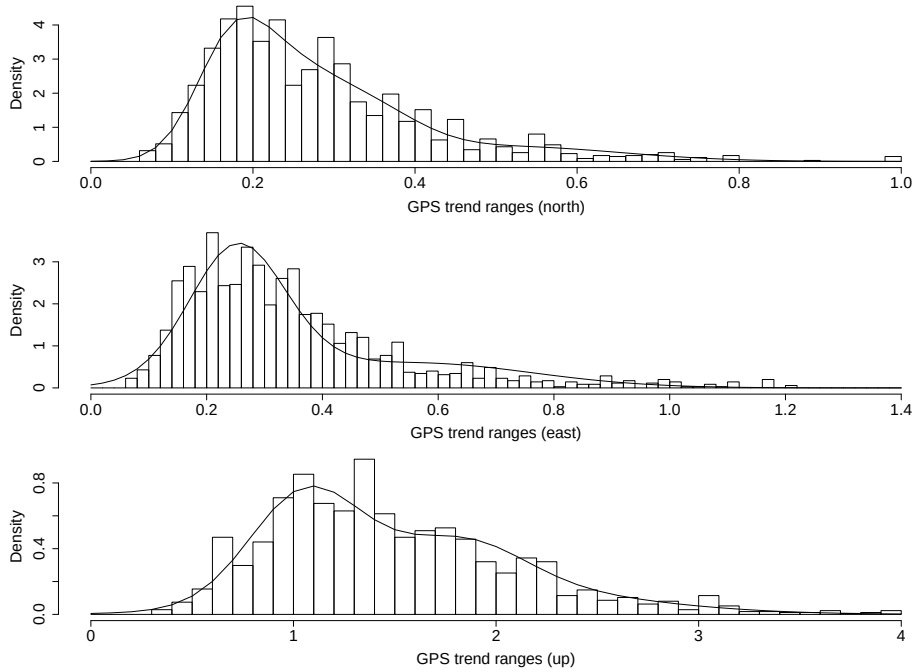


Figure 7.12: Histograms of the deformation rate ranges along with the estimated mixed normal density (Taupo).

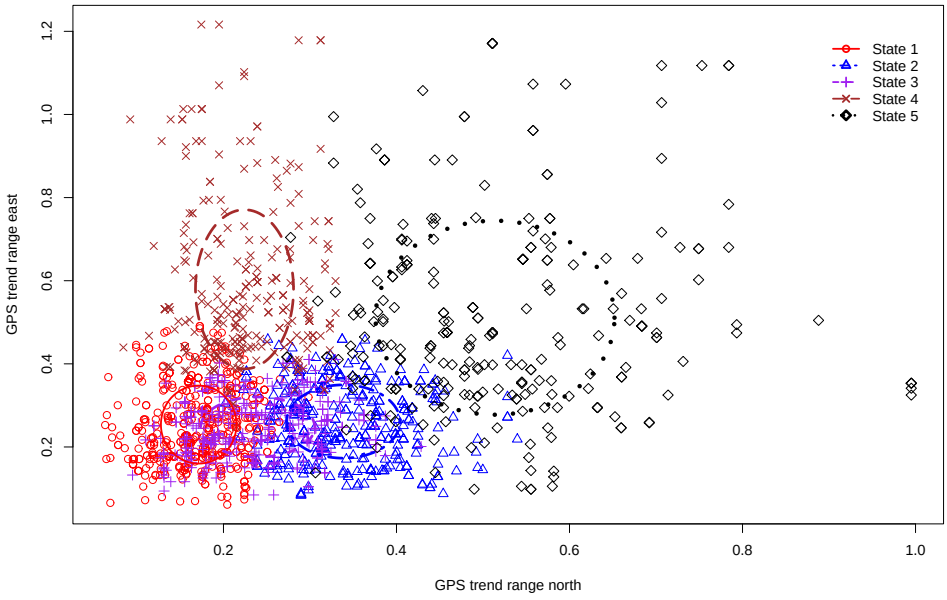


Figure 7.13: Scatter plot of the deformation rate ranges of the east component versus that of the north component. The ellipses illustrate the projected standard deviations for the north and east directions centered at the estimated means (Taupo).

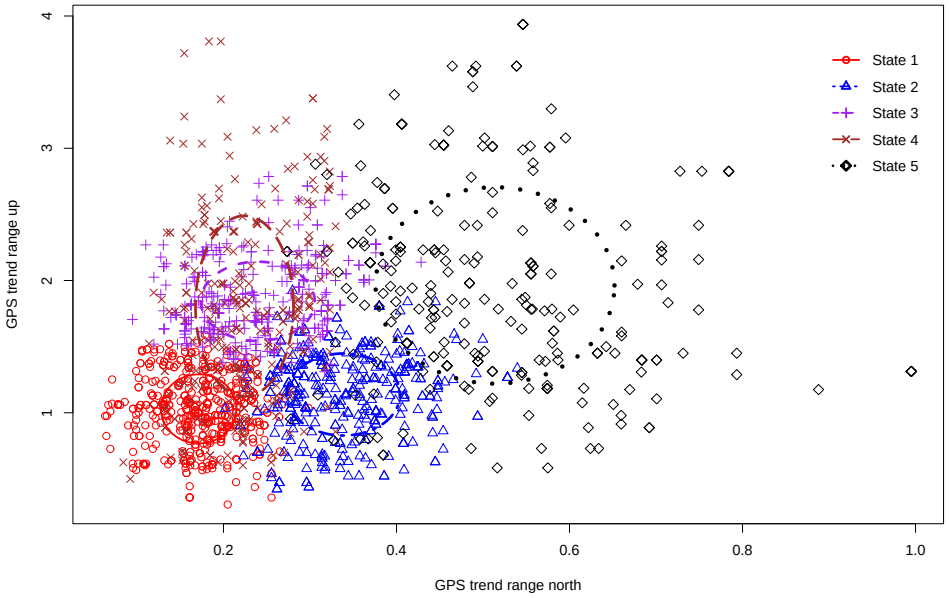


Figure 7.14: Scatter plot of the deformation rate ranges of the up component versus that of the north component. The ellipses illustrate the projected standard deviations for the north and up directions centered at the estimated means (Taupo).

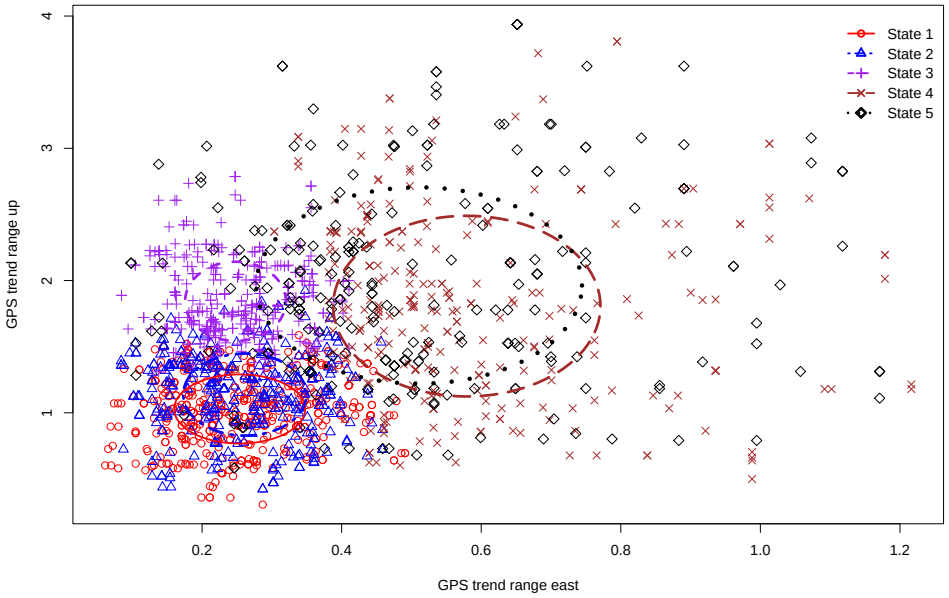


Figure 7.15: Scatter plot of the deformation rate ranges of the up component versus that of the east component. The ellipses illustrate the projected standard deviations for the east and up directions centered at the estimated means (Taupo).

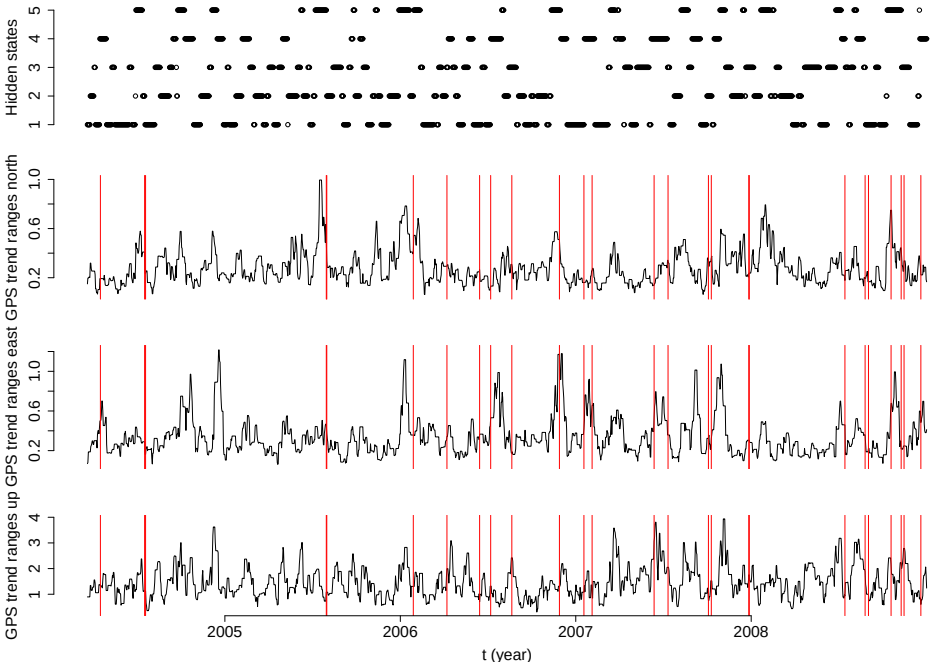


Figure 7.16: The Viterbi path and the trend changes of the GPS movements in the previous 10 days (Taupo), with vertical lines indicating the earthquake occurrence times with magnitude larger than or equal to 5.1.

whether this association is statistically significant, the mutual information between each state and the earthquake occurrences will be calculated.

The mutual information results for the five-state model are shown in Figure 7.17(a) for earthquakes with minimum magnitude 4.2, and in Figure 7.17(b) for earthquakes with minimum magnitude 5.1. For earthquakes with minimum magnitude 4.2, the mutual information suggests that States 3 and 5 show some preseismic information and State 2 may have some postseismic information, while the other states show little association with the earthquakes. When we use a larger threshold, magnitude larger than or equal to 5.1, the preseismic information in State 5 becomes stronger, while the other states appear to show little association with the earthquakes. State 5 is thus the 'precursory state'. The preseismic information in State 3, which accounts for the up component movement, weakens when increasing the magnitude threshold. This information may be related to the cluster effect either of the foreshocks or due to the previous large earthquakes. The postseismic state, State 2, appears to be mainly associated with small earthquakes. The mutual information between State 1 and the earthquakes for both magnitude thresholds shows little association. Moreover, State 1 is the closest to the origin. It may be the background state which suggests little movement of the underlying dynamics. As for State 4, it is more like an intermediate state between States 2

and 5, or between mainshock and aftershocks.

In order to validate the significance of the association, the simulated 95% confidence interval is calculated. Note that there is no observable clustering in this earthquake data set. We simulate 1000 series of Poisson processes with intensity equal to the number of earthquakes divided by the total number of days, and 1000 Markov chains using the estimated transition probability matrix from the fitted 5 state model. The mutual information between the 1000 pairs of processes is calculated and the simulated upper 95% confidence level can then be obtained. The dot-dash lines in Figure 7.17 indicate the simulated confidence levels for comparison. The preseismic information of State 5 is very clearly shown in the plot above the dot-dash line. This verifies that the association is not arbitrary, and confirms State 5 as a precursory state for the fitted five-state model. Note that by using 6 hidden states, the analysis results in a very similar conclusion to the five-state model in terms of precursory information. However, the five-state model contains less parameters and hence is more stable.

Cross validation

In order to verify this association, a cross validation is carried out. A five-state HMM is fitted to the first half of the GPS data to get the parameter estimates for the model. Then these estimated parameters are used to track the Viterbi path for the second half of the data. After that the mutual information between the estimated Viterbi path and the earthquakes occurring during this time period is calculated. The estimated parameters are listed in Table 7.5. Consistent with the result for the entire GPS data, we can see that the furthest state from the origin, State 5, which has the largest variances for all three components, and State 1 do not transit to each other. The mutual information between the estimated Viterbi path for the second half and the corresponding earthquakes are shown in Figure 7.18(a) for earthquakes with minimum magnitude 4.2, and in Figure 7.18(b) for earthquakes with minimum magnitude 5.1. The result suggests little association between the Viterbi path and earthquakes with minimum magnitude 4.2. However, the preseismic information remains in State 5, which, as shown above, is equivalent to the precursory State 5 for the entire data, when we use the magnitude threshold 5.1.

Conversely, the HMM is fitted to the second half and calculate the mutual information for the first half. The estimated parameters for the second half are listed in Table 7.6. State 5 has the largest trend ranges and the largest variances for the three components. This state, equivalent to the precursory State 5 for the entire data, does not transit to State 1 and vice versa. This result is also

Table 7.5: The parameter estimates of the fitted five-state HMM to the first half of GPS data (Taupo).

	State	1	2	3	4	5
Estimated	north	0.177	0.347	0.193	0.246	0.554
means	east	0.255	0.266	0.664	0.306	0.503
	up	1.029	1.151	1.371	1.891	1.893
Estimated	north	0.045	0.070	0.049	0.065	0.154
standard	east	0.096	0.081	0.208	0.103	0.226
deviations	up	0.268	0.230	0.525	0.361	0.733
Estimated	state 1	0.913	0.036	0.011	0.039	0.000
transition	state 2	0.035	0.888	0.000	0.055	0.023
probability	state 3	0.053	0.000	0.933	0.000	0.014
matrix	state 4	0.049	0.035	0.000	0.900	0.016
	state 5	0.000	0.055	0.018	0.009	0.918

Table 7.6: The parameter estimates of the fitted five-state HMM to the second half of GPS data (Taupo).

	State	1	2	3	4	5
Estimated	north	0.152	0.230	0.333	0.242	0.498
means	east	0.187	0.325	0.200	0.530	0.593
	up	1.280	1.060	1.580	1.958	1.978
Estimated	north	0.035	0.057	0.083	0.064	0.114
standard	east	0.041	0.060	0.061	0.171	0.238
deviations	up	0.441	0.324	0.513	0.651	0.806
Estimated	state 1	0.934	0.024	0.030	0.012	0.000
transition	state 2	0.023	0.919	0.032	0.025	0.000
probability	state 3	0.016	0.036	0.917	0.008	0.023
matrix	state 4	0.019	0.024	0.010	0.927	0.021
	state 5	0.000	0.000	0.017	0.054	0.929

consistent with the entire data set. The mutual information for the first half, shown in Figure 7.19(a) for earthquakes with minimum magnitude 4.2, shows some association between the Viterbi path and earthquakes with minimum magnitude 4.2. The preseismic information remains in State 5 and becomes clearer for larger magnitude threshold (Figure 7.19(b)).

For each of the above models (the HMM fitted to the entire data, the HMM fitted to the first half, and the one fitted to the second half), the precursory state (State 5) does not transit to State 1. In addition, State 1 has the smallest mean and standard deviation for the north component in all three

models. This state shows little association with the earthquake occurrence. It may be the quiescent state. This further justifies the selection of the number of states in the HMMs.

Analysis Using Only North and East Components

The Euclidean distance of the variable R_t from the origin $(0, 0, 0)$ is also calculated, which is

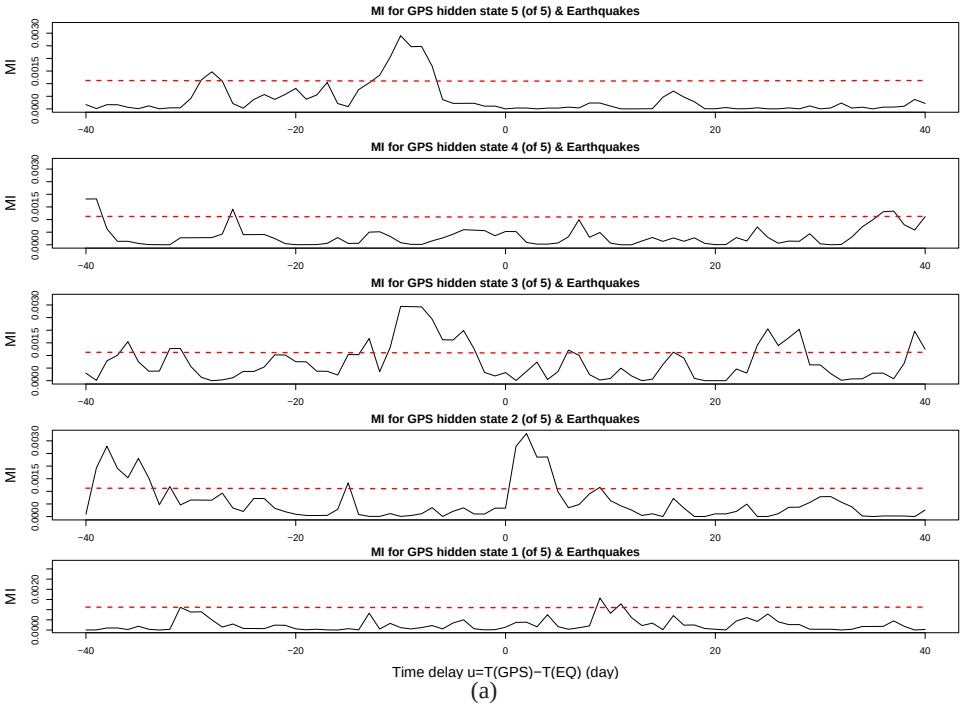
$$\sqrt{R_{Nt}^2 + R_{Et}^2 + R_{Ut}^2},$$

and then HMMs are fitted to the data. However, due to the fact that the up component is approximately 4 times as large as the north or east components and has about 3 times the standard deviation of the other two, it becomes the dominant factor for the Euclidean distance. The Euclidean distance down-weights the north and east component effects. The HMM analysis on this distance does not work as well as the analysis on all three components as the calculated Euclidean distance becomes less informative. Therefore the HMMs are fitted to the north and east components only, and to the Euclidean distance calculated only using north and east components. As this issue will also affect the probability forecast using the Euclidean distance, another probability forecast is conducted only considering the north and east components.

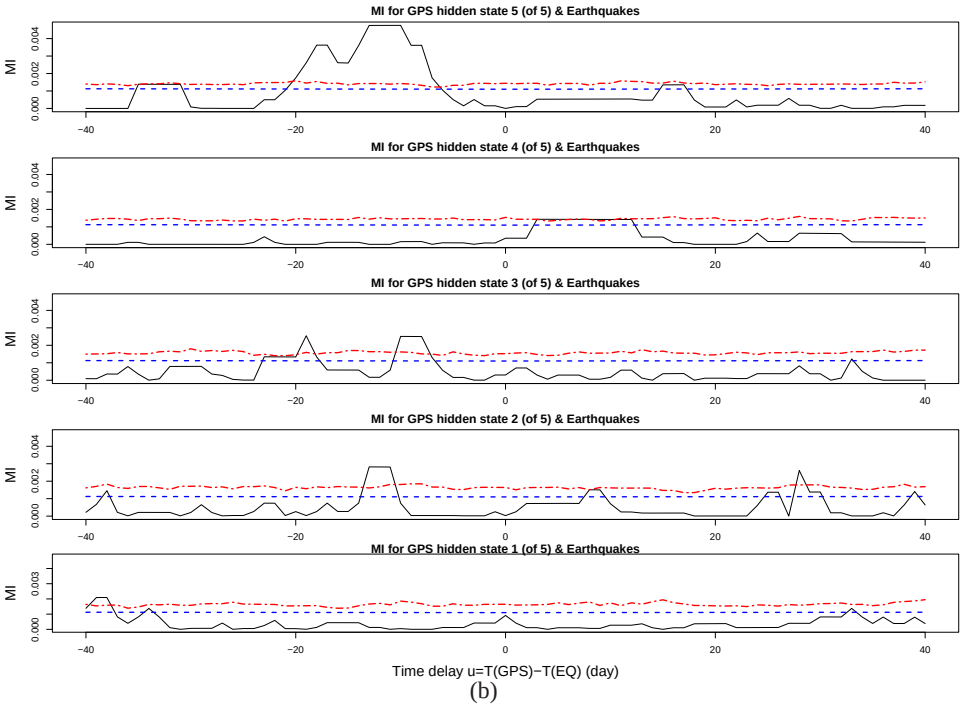
The HMMs are fitted to the deformation rate ranges from the north and east components with different numbers of hidden states. The state transitions for the fitted 3, 4, 5 and 6 state HMMs are shown in Figure 7.20. Note that the states are ordered according to the Euclidean distance of the estimated means in each state from the origin (0,0). The fitted results for the five-state model are

Table 7.7: The parameter estimates of the fitted 5 state HMM for the case when only the north and east components are considered (Taupo).

	State	1	2	3	4	5
Estimated means	north	0.175	0.345	0.227	0.227	0.525
	east	0.190	0.223	0.345	0.631	0.515
Estimated s.d.	north	0.045	0.061	0.063	0.062	0.133
	east	0.046	0.065	0.060	0.188	0.229
Estimated transition probability matrix	state 1	0.913	0.034	0.053	0.000	0.000
	state 2	0.039	0.887	0.041	0.000	0.033
	state 3	0.032	0.029	0.906	0.021	0.012
	state 4	0.024	0.000	0.043	0.923	0.010
	state 5	0.000	0.042	0.011	0.026	0.921

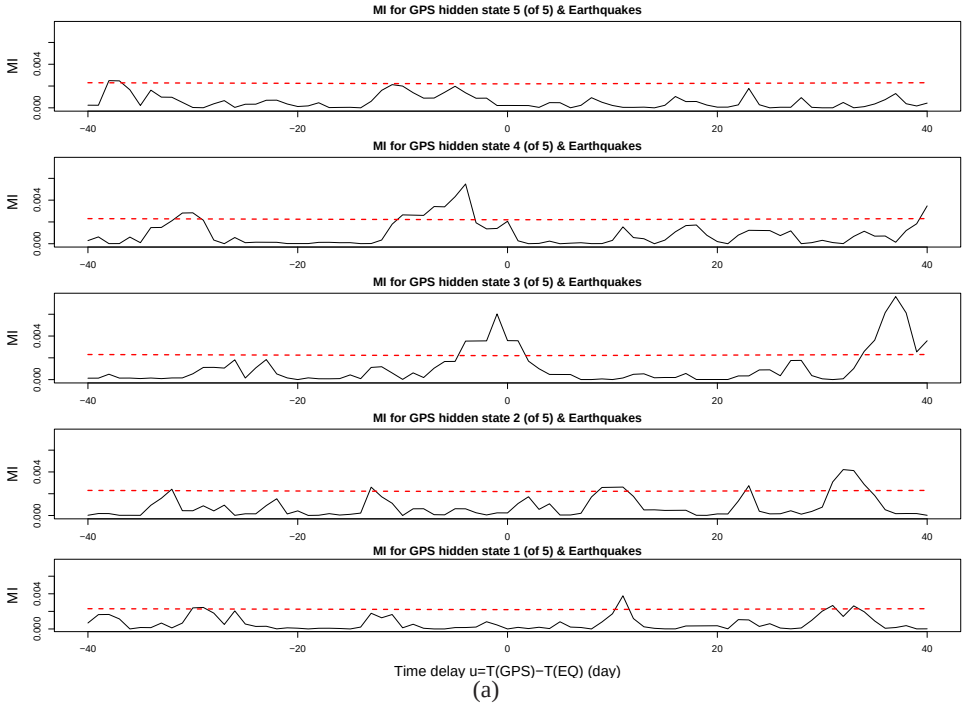


(a)

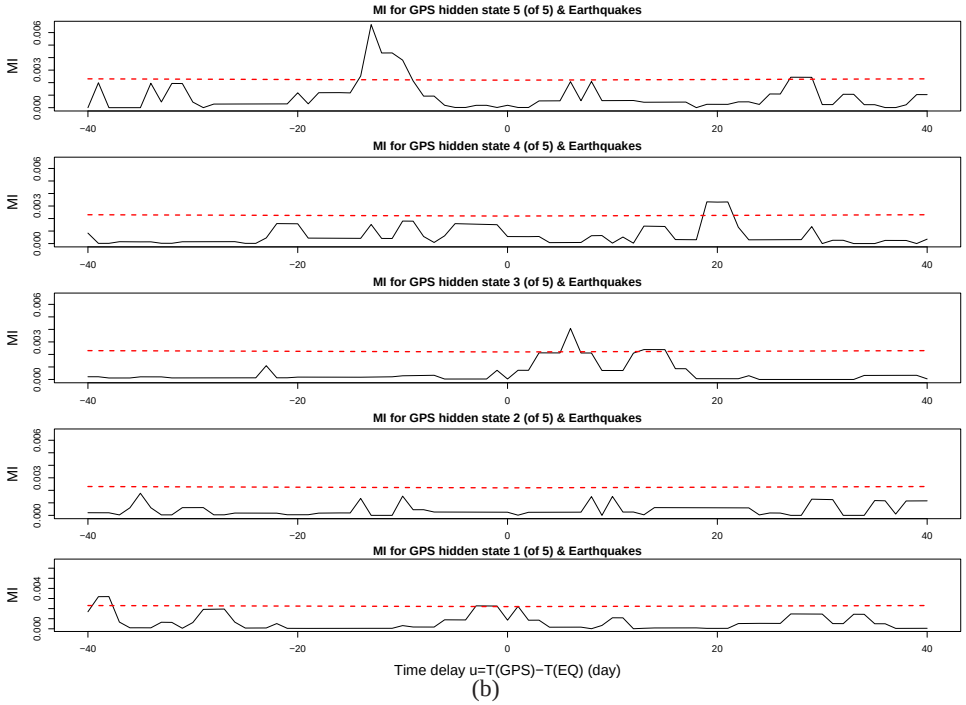


(b)

Figure 7.17: For data around Taupo. (a): Mutual information between U and V (for earthquake magnitude larger than or equal to 4.2) with a time lag u. (b): Mutual information between U and V (for earthquake magnitude larger than or equal to 5.1) with a time lag u. The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent. The dot-dash line in each plot is the calculated 95% confidence level from the 1000 simulated earthquake series and 1000 simulated Markov chains.

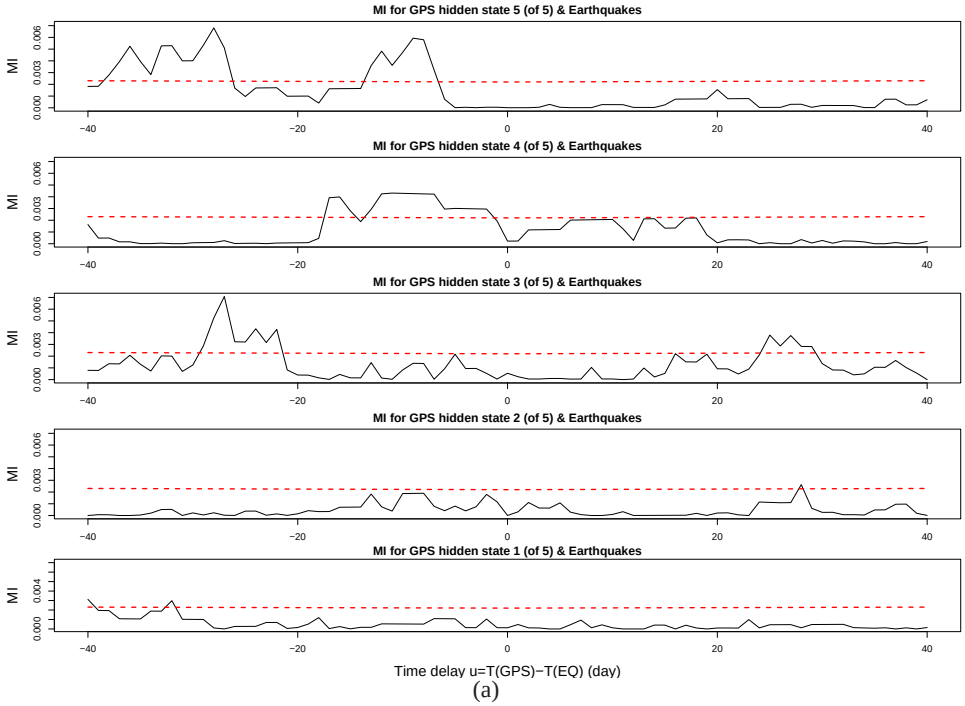


(a)

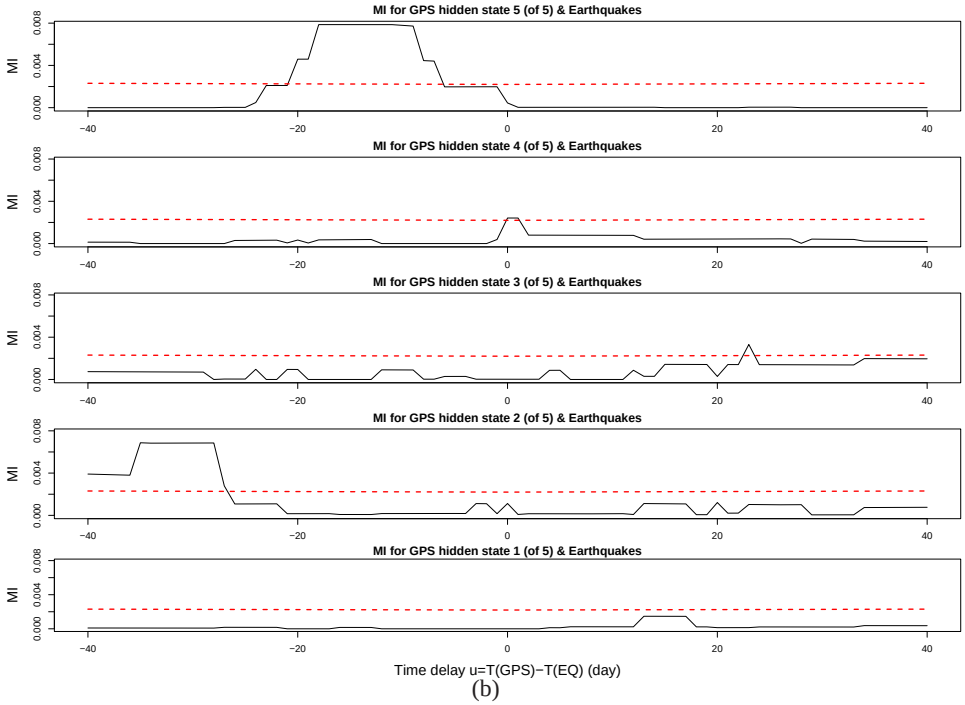


(b)

Figure 7.18: For data around Taupo. (a): Mutual information between U and V (for the second half with earthquakes greater than or equal to 4.2) with a time lag u. (b): Mutual information between U and V (for the second half with earthquakes greater than or equal to 5.1) with a time lag u. The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent.



(a)



(b)

Figure 7.19: For data around Taupo. (a): Mutual information between U and V (for the first half with earthquakes greater than or equal to 4.2) with a time lag u. (b): Mutual information between U and V (for the first half with earthquakes greater than or equal to 5.1) with a time lag u. The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent.

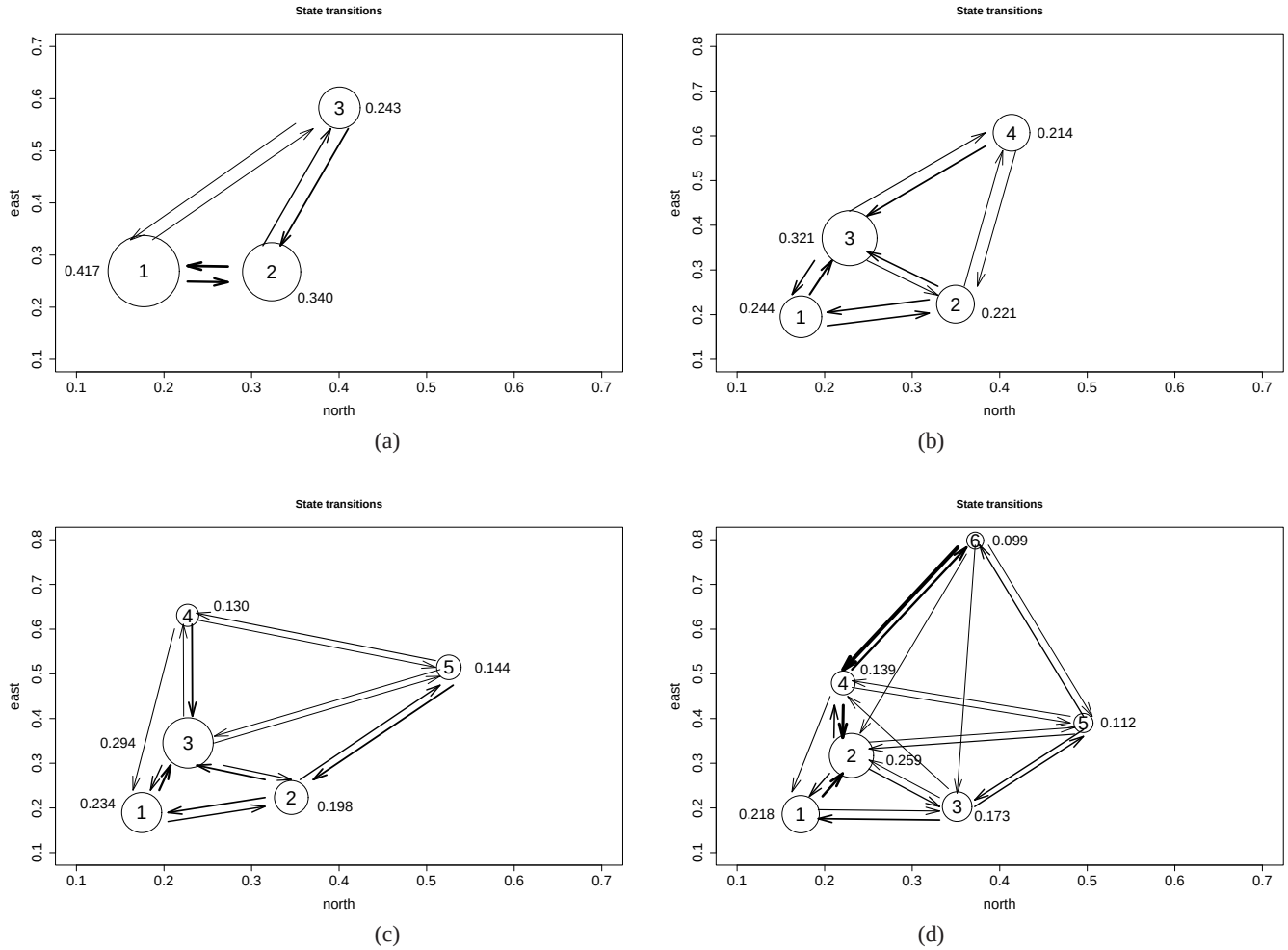


Figure 7.20: Illustration of state transitions for HMMs with 3 states (a), 4 states (b), 5 states (c) and 6 states (d) for data around Taupo, using the north and east components only. Circle sizes are proportional to the stationary distribution of the hidden Markov chain, i.e., the proportion of time in each state, which is shown as the number besides each circle. The thickness of the arrows is proportional to the transition probabilities. The numbers inside the circles indicate the states and are located at the means of the states estimated from the fitted model.

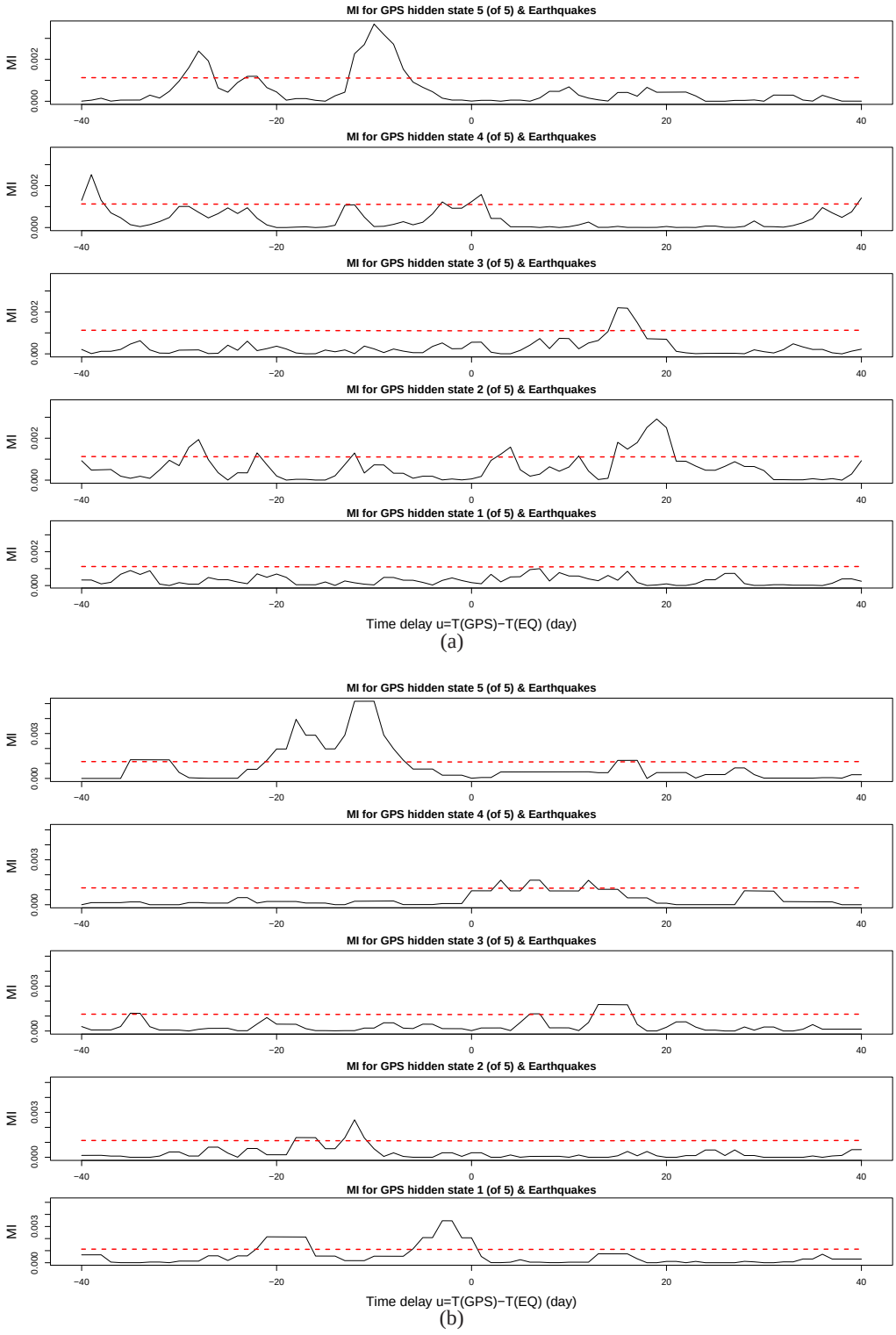


Figure 7.21: The mutual information for data around Taupo only using the north and east components. (a): Mutual information between U and V (for earthquakes with minimum magnitude 4.2) with a time lag u. (b): Mutual information between U and V (for earthquakes with minimum magnitude 5.1) with a time lag u. The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent.

listed in Table 7.7. The mutual information between the Viterbi path $\{U_{t+u}\}$ and the earthquakes $\{V_t\}$ for the five-state model is examined, where u is the time lag. The mutual information results are shown in Figure 7.21. The state furthest from the origin, State 5, which has the largest variances for both north and east components, and State 1, which is the closest state to the origin and has the smallest variances for both components, do not transit to each other. When the magnitude threshold is increased, the preseismic information in state 5 becomes stronger, whereas the mutual information for the other states becomes less significant.

The scatter plot of the deformation rate ranges from the east component versus that from the north component, with five symbols indicating the different states in which the points are tracked to be (using the Viterbi algorithm), is shown in Figure 7.22. It is obvious that all 5 states are

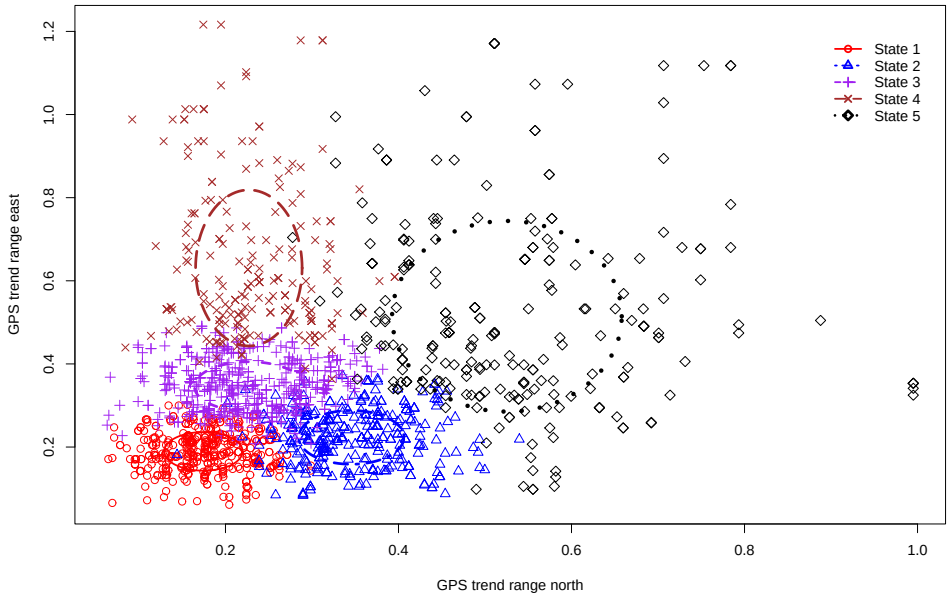


Figure 7.22: Scatter plots of the deformation rate ranges of the east component versus that of the north component (Taupo). The ellipses illustrate the projected standard deviations for the north and east directions centered at the estimated means from the HMM fitted to the north and east components only.

clearly separated from each other. The standard deviation contour projection of each state does not overlap with the others. The most likely state sequence tracked using the Viterbi algorithm and the deformation rate ranges of the GPS movements as calculated in Equation (7.1) are plotted in Figure 7.23, with the occurrence times of the earthquakes with minimum magnitude 5.1. We can clearly see spikes in the trend ranges before some of the earthquake occurrences. We also notice that

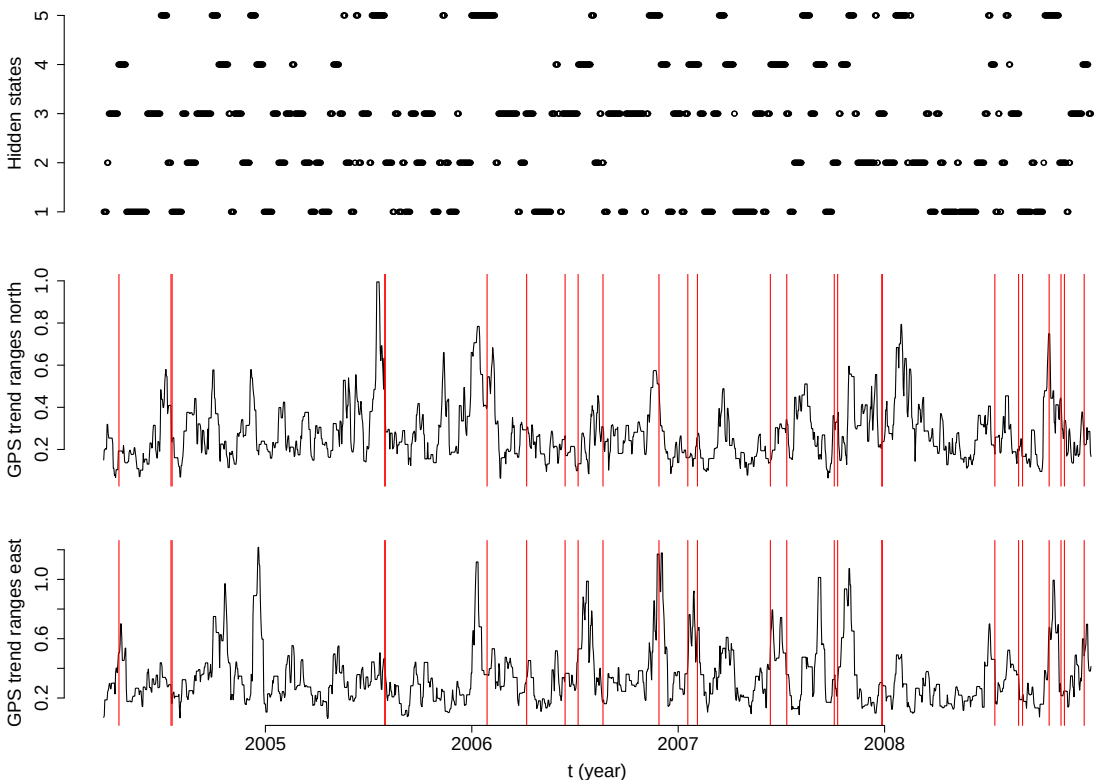


Figure 7.23: The Viterbi path and the trend changes of the GPS movements (north and east components) as calculated in Equation (7.1), with vertical lines indicating the earthquake occurrence times with magnitude larger than or equal to 5.1 (Taupo).

when the HMM visits State 5, then a large earthquake often follows. The sojourn time distribution for each state of the fitted 5 state model is shown in Figure 7.24. The P-values of the K-S tests for State 1 to State 5 are, respectively, 0.1154, 0.1195, 0.0832, 0.4084, and 0.3270, indicating that the distribution of the sojourn time in each state is approximately Geometric.

We consider the Euclidean distance of the ranges of the north and east components from the origin (0,0), $D_{1t} = \sqrt{R_{Nt}^2 + R_{Et}^2}$. The HMMs are fitted to the distance (with normal distribution for the observations) with different numbers of hidden states. We examine the mutual information between the Viterbi path $\{U_{t+u}\}$ and the earthquakes $\{V_t\}$ respectively for the fitted models with different states, where u is the time lag. The mutual information results (see, e.g., Figure 7.25 for the mutual information between the earthquakes and the Viterbi path of the fitted five-state model) suggest that for the purpose of extracting preseismic signals, using the Euclidean distance is not as informative as using the separate components.

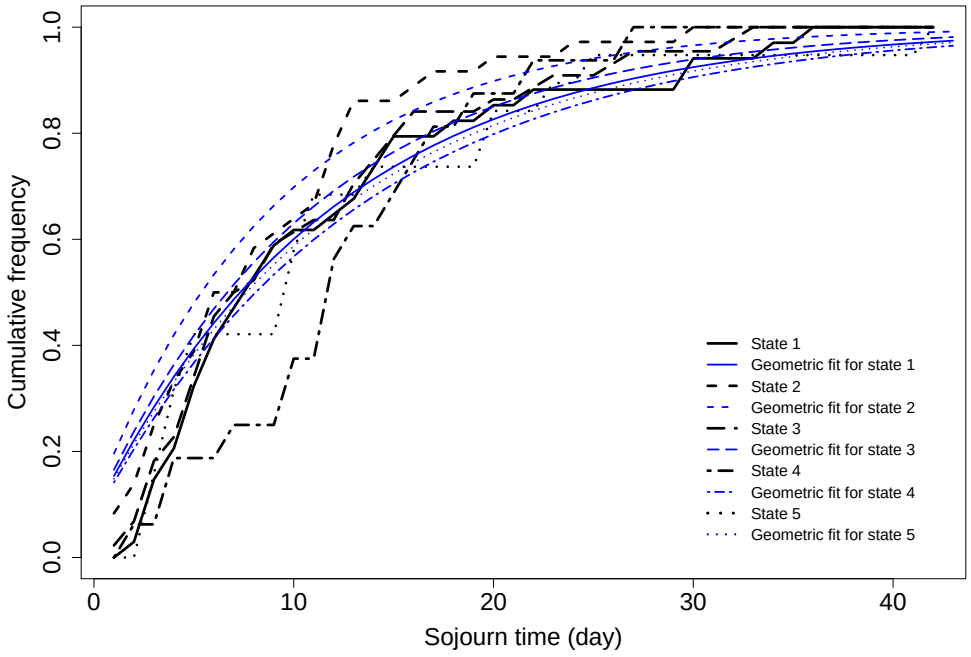


Figure 7.24: Sojourn time distribution for the fitted five-state HMM (thick lines) using the north and east components, with the thin lines (for comparison) indicating Geometric distributions, each with mean calculated from the sample mean of the sojourn time in each state (Taupo).

7.3.3 Probability Forecast Using Logistic Probability Model

Forecast Using All Three Components of GPS Measurements

According to the mutual information between State 5 and earthquake occurrences with minimum magnitude 5.1 for the entire data around Taupo (in Section 7.3.2), we define a TIP as follows. We divide the entire time period into non-overlapping intervals J_i with equal length of 10 days. For each interval J_i, if there is any day in which the HMM visits State 5, we consider the following interval, J_{i+1}, a TIP. A contingency table for testing whether the 5th state is related with the large earthquake occurrences can be conducted using the two categorical variables: for each interval, whether it is a TIP and whether there is any earthquake with minimum magnitude 5.1 occurring in this interval. The contingency table for this purpose is shown in Table 7.8. It indicates a sensitivity of 0.45 and a specificity of 0.77. The Chi-squared statistic of this contingency table with the null hypothesis that the two variables are statistically independent is 4.1865, with a P-value of 0.04. This shows some evidence (at significance level of 0.05) that, whether the HMM visits State 5 in the current interval is related to whether there is large earthquake occurrence in the following interval. This further confirms the mutual information conclusion for State 5.

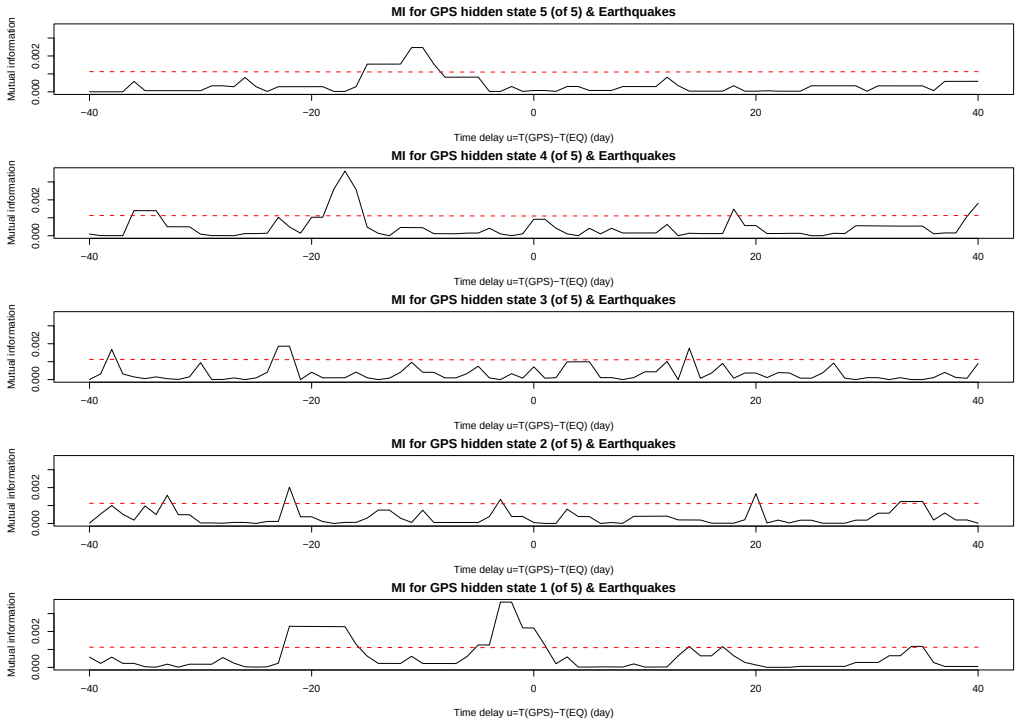


Figure 7.25: Mutual information between U (for the Euclidean distance calculated from the north and east components around Taupo, five-state HMM) and V (for earthquake magnitude larger than or equal to 5.1) with a time lag u. The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent.

Table 7.8: Contingency table, where 'Yes' indicates that there is an earthquake with minimum magnitude 5.1 occurring in the interval J_{i+1}, and 'No' indicates that there is no large earthquake occurrence in that time interval (Taupo).

	Yes	No	Total
J_{i+1} a TIP	10	34	44
J_{i+1} not a TIP	12	117	129
Total	22	151	173

Table 7.9 lists the details of the state transitions preceding and following the earthquakes with magnitude 5.1 and larger. Some 18 out of 26 large earthquakes (about 70 percent) are preceded by sojourning in State 5 in the previous three states and the state in which the earthquake occurs. The three state transitions are all within 70 days. Some 13 of them have the transition pattern 2–5, and 7 of them have the pattern 2–5–2. We then look at the three state transitions following the earthquake occurrences. After the earthquake occurrences, some 10 out of 26 correspond to that the process either stays in State 3 and then transits to State 1, or transits to State 3 and then to State 1. Some 8 of the 26 transition patterns transit to State 3 after visiting the state in which earthquakes occurred,

Table 7.9: State transition pattern for earthquakes with magnitude 5.1 or larger (Taupo). For each transition pattern, the first three states are the three consecutive states preceding the earthquake state, the fourth (in bold) is the state in which the earthquake occurred, and the last three are the three following the earthquake state.

Date	Mag	Latitude	Longitude	Depth (km)	state transition
2004.04.17	5.5	-37.99	176.69	133.06	2-3-1-**4**-1-3-1
2004.07.18	5.1	-38.00	176.51	5	1-2-5-**2**-1-3-2
2004.07.20	5.2	-38.98	174.76	578.47	2-5-2-**1**-3-2-3
2005.07.31	5.4	-38.38	175.99	192.67	3-1-2-**5**-2-3-2
2005.08.01	5.7	-37.89	176.36	217.11	1-2-5-**2**-3-2-1
2006.01.28	5.5	-38.57	175.82	164.36	1-2-5-**2**-5-3-1
2006.04.08	5.3	-38.36	176.03	155.55	1-3-2-**3**-4-3-2
2006.06.15	5.5	-38.50	175.86	167.08	1-3-4-**1**-3-1-4
2006.07.08	5.4	-39.17	176.82	32.78	1-3-1-**4**-5-2-3
2006.08.21	5.4	-38.55	175.83	158.67	4-5-2-**3**-1-2-1
2006.11.28	5.5	-38.15	176.86	84.17	2-1-2-**5**-4-1-4
2007.01.18	5.4	-39.54	175.85	50.85	2-5-4-**1**-4-1-3
2007.02.04	5.4	-39.09	176.29	56.33	5-4-1-**4**-1-3-5
2007.06.13	5.1	-38.20	176.34	157.76	1-3-1-**4**-3-1-2
2007.07.12	5.3	-38.73	176.22	93.69	3-1-4-**3**-1-2-5
2007.10.04	5.7	-37.84	176.15	282.24	3-4-1-**2**-1-4-5
2007.10.10	5.1	-38.29	176.42	155.19	3-4-1-**2**-1-4-5
2007.12.27	5.5	-38.95	175.67	117.48	2-5-2-**3**-2-5-2
2007.12.28	5.5	-38.77	176.29	79.76	2-5-2-**3**-2-5-2
2008.07.14	5.1	-38.28	175.84	179.36	1-3-5-**4**-3-1-3
2008.08.25	5.9	-39.71	176.85	31.83	3-5-4-**3**-1-3-1
2008.09.01	5.8	-39.10	175.89	85.09	5-4-3-**1**-3-1-2
2008.10.18	5.5	-38.00	176.24	203.61	3-1-2-**5**-3-1-2
2008.11.08	5.7	-38.15	176.06	200.83	3-1-2-**5**-3-1-2
2008.11.14	5.2	-38.28	176.24	156.89	1-2-5-**3**-1-2-5
2008.12.19	5.8	-38.24	176.13	186.88	1-2-5-**4**-

and 9 of the 26 transit to State 1 after visiting the state in which earthquakes occurred.

The frequency of the state transitions in Table 7.9 is provided in Table 7.10. The most frequent transitions are 3–1, 1–2 and 2–5. Moreover, some 13 of the 26 large earthquakes are preceded by the transition pattern 2–5. We therefore use this transition pattern to define a 'TIP' and see how it works. For each interval J_i, if there is any day in which the HMM visits State 5 and if state 2 is the state preceding State 5, we consider the following interval J_{i+1} a TIP. A contingency table for testing whether this transition pattern is related with the large earthquake occurrences can also be conducted using the two categorical variables: for each interval, whether it is a TIP and whether there is an earthquake with minimum magnitude 5.1 occurring in this interval (Table 7.11). The test

Table 7.10: Frequency of state transitions for Table 7.9.

	1	2	3	4	5	Total
1	0	18	14	10	0	42
2	7	0	9	0	18	34
3	21	8	0	4	3	36
4	9	0	6	0	4	19
5	0	11	4	7	0	22
Total	37	37	33	21	25	153

Table 7.11: Contingency table, where 'Yes' indicates that there is an earthquake with minimum magnitude 5.1 occurring in the interval J_{i+1}, and 'No' indicates that there is no large earthquake occurrence in that time interval (Taupo). A TIP is defined using the transition pattern 2–5.

	Yes	No	Total
J_{i+1} a TIP	8	22	30
J_{i+1} not a TIP	14	129	143
Total	22	151	173

has a sensitivity of 0.36 and a specificity of 0.85. The Chi-squared statistic of this contingency table with the null hypothesis that the two variables are statistically independent is 4.9334, with a P-value of 0.026. The P-value has a slight improvement (recall that the P-value for the case when only a visit to State 5 is used to define a TIP is 0.04). This suggests that a 2–5 transition pattern preceding large earthquakes may not be a coincidence. The average number of entries to each state between two consecutive events is listed in Table 7.12. In most cases, the HMM sojourns in State 1 between earthquakes.

Table 7.12: Average number of entries to each state between two consecutive earthquakes (Taupo).

State	1	2	3	4	5
Ave. number of entries	1.58	1.46	1.46	0.81	0.85

We now examine how the GPS deformation rate ranges influence the probability of declaring a TIP. We use a Logistic linear predictor for this purpose. Since the hidden State 5 is a combination of the three GPS components, we consider the Euclidean distance of the ranges from the origin (0,0,0),

$$D_t = \sqrt{R_{Nt}^2 + R_{Et}^2 + R_{Ut}^2}$$

Table 7.13: Logistic regression result for a target event in the time interval J_{i+1} with minimum magnitude 5.1 (Taupo).

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
β_0	-3.2273	0.7400	-4.361	1.29e-05
β_1	0.7886	0.4061	1.942	0.0522

where R_{Nt}, R_{Et} and R_{Ut} are the trend ranges of the GPS north, east and up components, respectively. In order to get a fairly robust linear predictor of a target event, instead of using some extreme values such as maximum or minimum, we use the mean of the distance D_t in interval J_i,

$$M_i = \frac{1}{10} \sum_{t \in J_i} D_t,$$

which is the average distance of the deformation rate ranges from the origin in 10 days, as a linear predictor of a target event in the time interval J_{i+1} using the regression $f(J_{i+1}) = \beta_0 + \beta_1 M_i$, where $f(J_{i+1})$ is the log-odds ratio. The fitted result is shown in Table 7.13. We can see that the regression coefficient for the predictor is not equal to 0 with a P-value of 0.0522 (or at a significance level of 0.1). The result is plotted in Figure 7.26. This shows that the larger the mean distance of the trend ranges from the origin in interval J_i, the higher the probability that there will be a large earthquake occurring in the interval J_{i+1}. We move the earthquake magnitude threshold lower to 5.0, i.e., the binary response variable takes on the value 1 when there is an earthquake with magnitude 5.0 or larger occurring in the interval J_{i+1}, and 0 otherwise. The fitting result is shown in Table 7.14. The

Table 7.14: Logistic regression result for a target event in the time interval J_{i+1} with minimum magnitude 5.0 (Taupo).

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
β_0	-2.9390	0.6727	-4.369	1.25e-05
β_1	0.7890	0.3737	2.111	0.0347

regression coefficient for the predictor M_i is now not equal to 0 with a P-value of 0.0347 (or at a significance level of 0.05). The result is plotted in Figure 7.27. The conclusion remains consistent with that of the magnitude threshold 5.1.

Figure 7.28 shows the scatter plot of the time from the moment of entry in State 5 to the next event versus the sojourn time of the HMM in State 5. It appears that a longer sojourn in State 5 tends

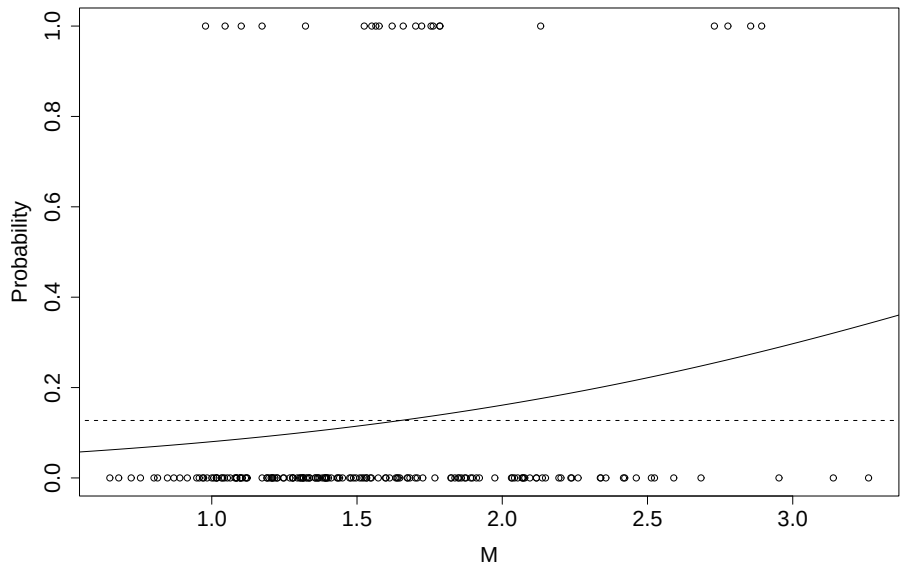


Figure 7.26: Earthquake occurrence (0-1) in the interval J_{i+1} versus M_i plot (points) with the solid line showing the probability of any magnitude 5.1 or larger earthquake occurrence in the interval J_{i+1} (Taupo). The dashed line indicates Total number of intervals in which there is earthquake occurrence/Total number of intervals.

to have a shorter time interval to the next event. Figure 7.29 is the scatter plot of the time to the next event versus the minimum distance of the deformation rate ranges from the origin when the HMM is sojourning in State 5, which does not suggest a strong correlation between the two variables.

Forecast Using Only North and East Components of GPS Measurements

According to the mutual information between State 5 and earthquake occurrences with minimum magnitude 5.1 when only considering the north and east components, we define a TIP as follows. We divide the entire time period into non-overlapping intervals J_i with equal length of 10 days. For each interval J_i, if there is any day in which the HMM visits State 5, we consider the following interval J_{i+1} a TIP. A contingency table for testing whether the fifth state is related with the large earthquake occurrences is shown in Table 7.15, which has a sensitivity of 0.45 and a specificity of 0.81. The Chi-squared statistic of this contingency table with the null hypothesis that the two variables are statistically independent is 6.1485, with a P-value of 0.0132. The P-value is smaller than when we use all three components (the P-value of which is 0.0408). This confirms that whether the HMM visits the furthest state from the origin (0,0) in the current interval, which has the largest

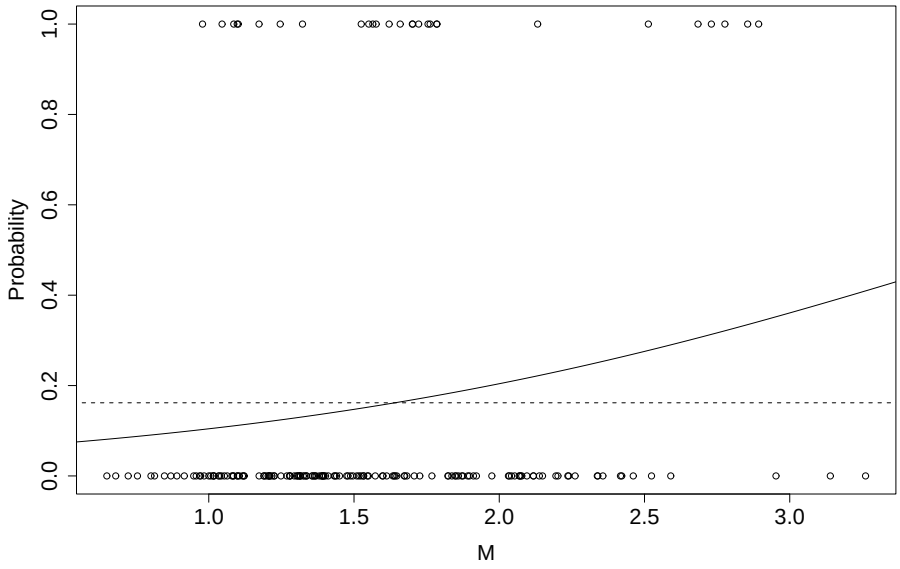


Figure 7.27: Earthquake occurrence (0-1) in the interval J_{i+1} versus M_i plot (points) with the solid line showing the probability of any magnitude 5.0 or larger earthquake occurrence in the interval J_{i+1} (Taupo). The dashed line indicates Total number of intervals in which there is earthquake occurrence/Total number of intervals.

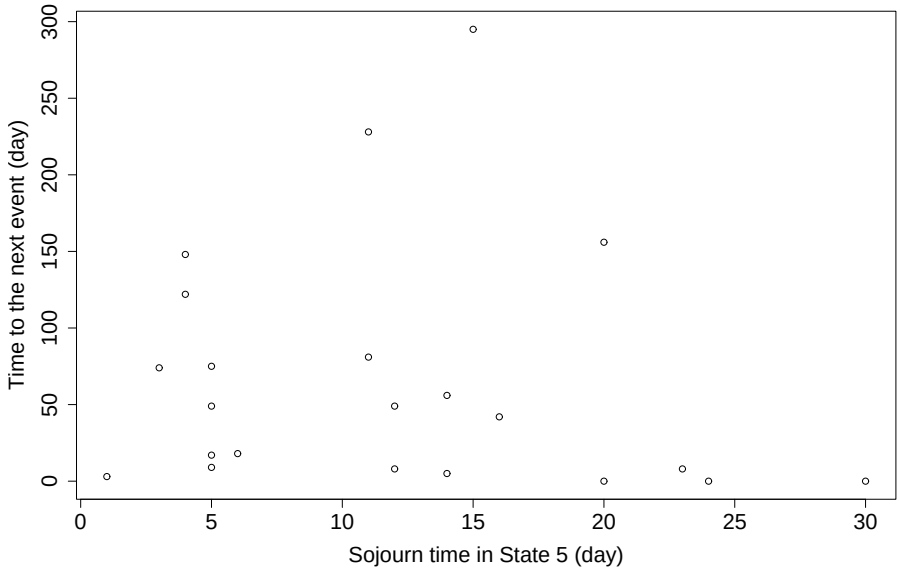


Figure 7.28: The scatter plot of the time to the next event versus the sojourn time of the HMM in State 5 (Taupo).

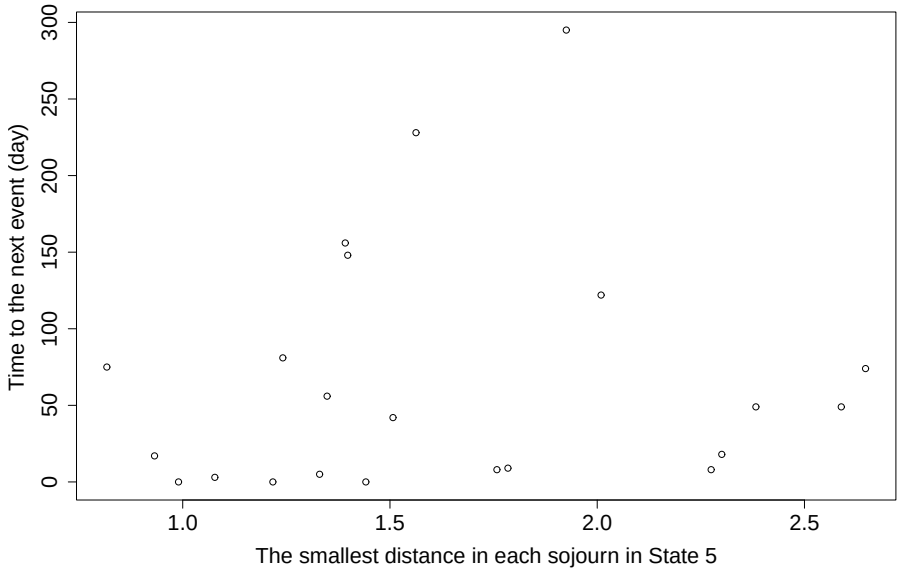


Figure 7.29: The scatter plot of the time to the next event versus the minimum distance of the deformation rate ranges to the origin when the HMM is sojourning in State 5 (Taupo).

Table 7.15: Contingency table for analysis using only north and east components (Taupo), where 'Yes' represents that there is an earthquake with minimum magnitude 5.1 occurring in the interval J_{i+1}, and 'No' indicates that there is no large earthquake occurrence in that time interval.

	Yes	No	Total
J_{i+1} a TIP	10	29	39
J_{i+1} not a TIP	12	122	134
Total	22	151	173

variances for both components, is related to whether there is large earthquake occurrence in the following interval.

Again, in order to get a relatively robust linear predictor of a target event, instead of using some extreme measures such as maximum or minimum, the mean of the distance D_{1t} in interval J_i

$$M_{1i} = \frac{1}{10} \sum_{t \in J_i} D_{1t},$$

is used as a linear predictor of a target event in the time interval J_{i+1} in a Logistic regression

$$f(J_{i+1}) = \beta_0 + \beta_1 M_{1i},$$

where $f(J_{i+1})$ is the log-odds ratio. The result is shown in Table 7.16. We can see that the regression

Table 7.16: Logistic regression result for a target event in the time interval J_{i+1} with minimum magnitude 5.1, using north and east components only (Taupo).

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
β_0	-3.1704	0.6146	-5.159	2.49e-07
β_1	2.5214	1.0823	2.330	0.0198

coefficient for the predictor is not equal to 0 with a P-value of 0.0198 (or at a significance level of 0.05). This result is more significant than when using all three components. Figure 7.30 shows that the larger the mean distance of the trend ranges from the origin in interval J_i, the higher the probability that there will be a large earthquake occurring in the interval J_{i+1}. We can see that there

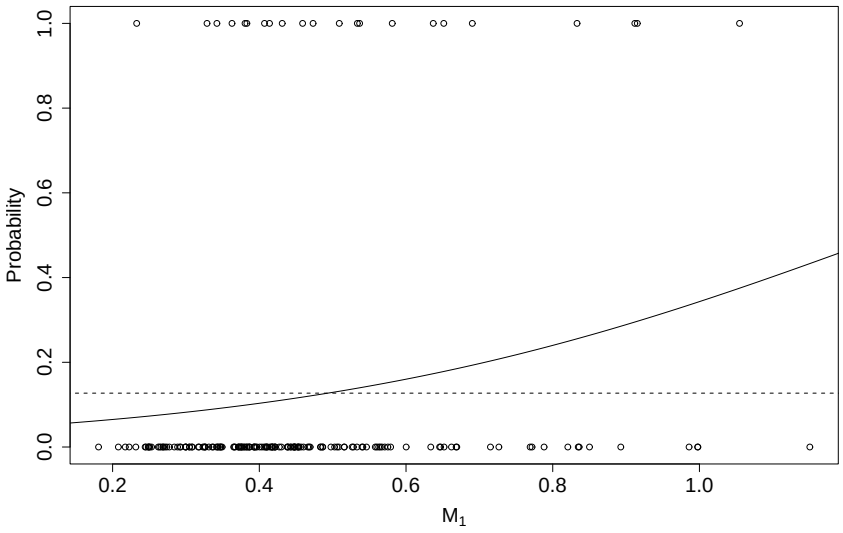


Figure 7.30: Earthquake occurrence (0-1) in the interval J_{i+1} versus M_{1i} plot (points) with the solid line showing the probability of any magnitude 5.1 or larger earthquake occurrence in the interval J_{i+1}, using the north and east components only (Taupo). The dashed line indicates Total number of intervals in which there is earthquake occurrence/Total number of intervals.

are less 0's for the earthquake occurrences at the further end of large distance values compared to the plot (Figure 7.26) for Euclidean distance using all three components.

7.4 Data from Southern California

Let us consider another data set from a different (strike-slip, rather than subduction-related rifting) tectonic environment, with longer sequences of observations, in Southern California. The southern part of the San Andreas fault (as shown in Figure 7.31) which forms the tectonic boundary between the Pacific Plate (on the west) and the North American Plate (on the east) runs through Southern California. The motion of the San Andreas fault is right-lateral strike-slip. The Pacific Plate moves

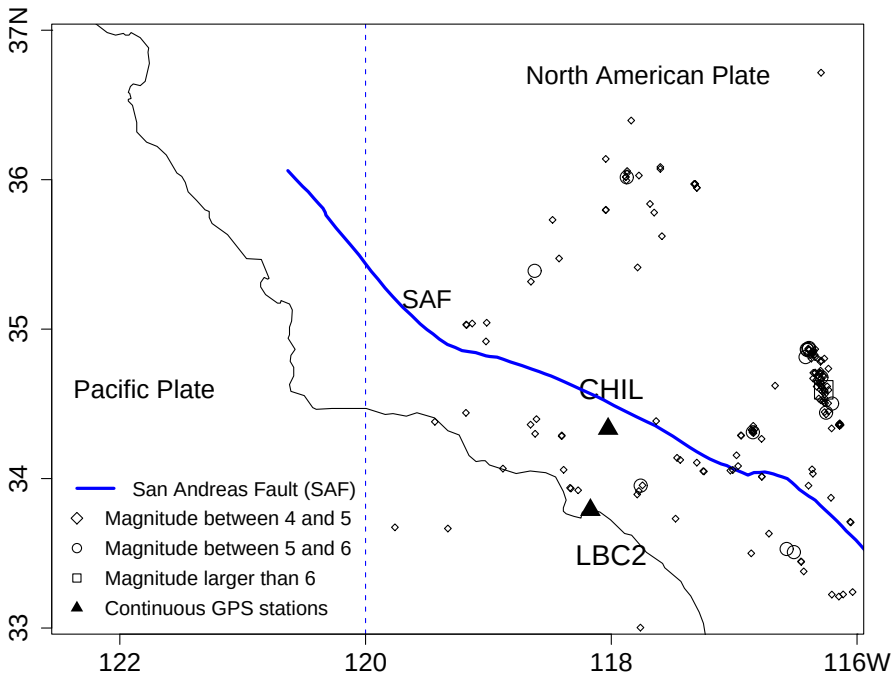


Figure 7.31: Location map for the GPS stations CHIL and LBC2 and earthquakes in Southern California. The symbol ▲ indicates the location of a GPS station. The earthquakes are selected from the rectangular area between latitude (33N,37N) and longitude (116W,120W). The small size of the symbols for earthquakes is for earthquakes with magnitude larger than 4.0 and smaller than 5.0; the medium size for earthquakes with magnitude larger than 5.0 and smaller than 6.0; and the large size for earthquakes with magnitude larger than 6.0. The maximum magnitude of the earthquakes in this area is 7.1.

approximately 48mm/yr to the northwest relative to the North American Plate (DeMets et al., 1987). The earthquakes occurring here are mainly shallow ones. We choose earthquakes, from the SCEC catalogue, in a rectangular area in Southern California with latitude between 33N and 37N and longitude between 116W and 120W, from 1999.01.01 to 2009.06.30. See Figure 7.31 for the location map of the earthquakes. The frequency–magnitude plot is shown in Figure 7.32, from which we

elect to consider the earthquakes with magnitude larger than or equal to 4.5 (50 events in total).

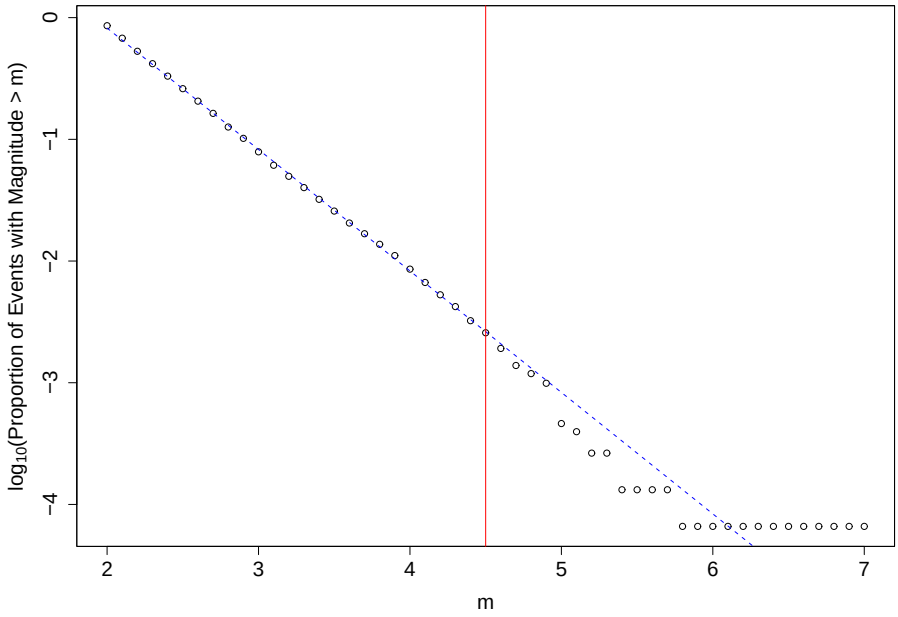


Figure 7.32: The frequency–magnitude plot for the earthquakes during the time period 1999.01.01 to 2009.06.30 in the selected area between latitude 33N and 37N and between longitude 116W and 120W.

The GPS data can be obtained from the SOPAC website (http://sopac.ucsd.edu/cgi-bin/refined-TimeSeriesListing.cgi, last accessed on August 3, 2009). The daily estimates of relative coordinates are created by a SOPAC refined model (Nikolaidis, 2002). We examine the resulting 'raw' daily GPS time series in Southern California. Unlike the tectonic environment around Taupo, where there is plate subduction, and active rifting and rotation of the TVZ, the most significant motion in Southern California is the right-lateral strike-slip of the San Andreas fault. The raw GPS time series shows an obvious long-term trend, for which a direct fit of HMM will not be useful for the purpose of probability forecast (c.f., Granat, 2003, 2006). Moreover, the GPS times series from many stations in this area display prominent heteroscedasticity. We therefore consider the baseline between two stations, one close to the San Andreas fault, CHIL, and one further away from this fault, LBC2 (as indicated in Figure 7.31), which have the longest records available online and the least amount of missing data. The missing data are again interpolated by setting each of them as the mean of the non-missing data within the 10 days ahead of and the 10 days following the missing point.

7.4.1 Hidden Markov Model and Mutual Information Analyses

Analysis of the Entire Data

The GPS measurements at LBC2 station are subtracted from the measurements at CHIL station. We will then get three series of baselines, north, east and up as shown in Figure 7.33. For the three

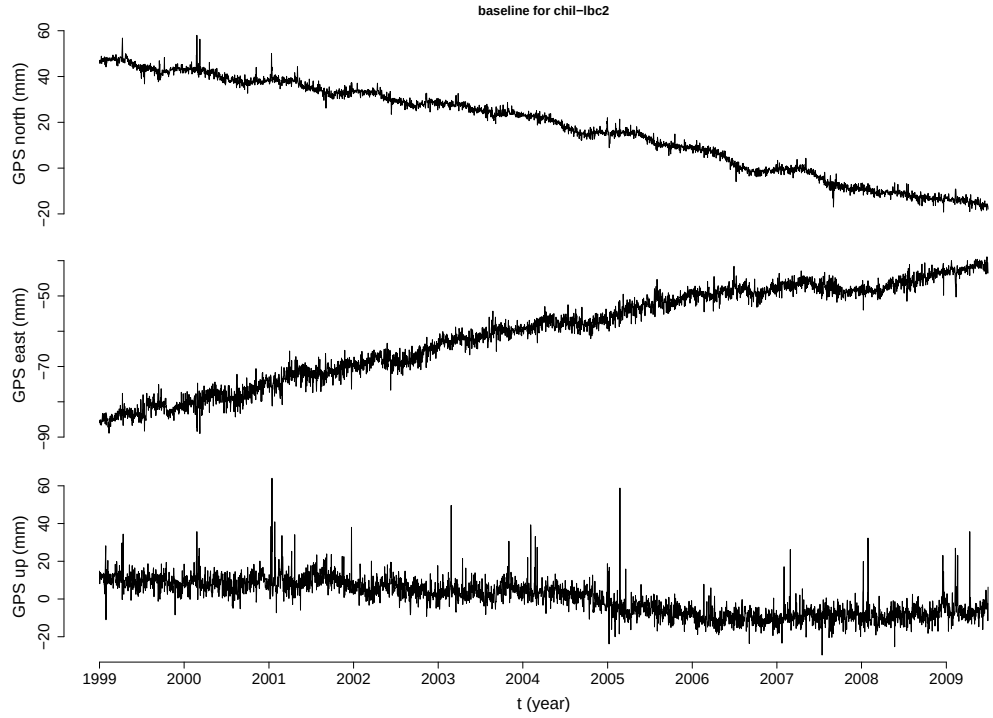


Figure 7.33: Baseline between the GPS stations CHIL and LBC2.

series, the analysis done in Section 7.3.2 is repeated. The log likelihood and BIC values for each of the model are listed in Table 7.2. Note that the increase of the log likelihood is similar to that of the data from Taupo. Again the five-state model is chosen by the criterion described in Section 7.2.1, which captures the most variable state consisting of about 6% of the entire data. The state transition diagrams for the fitted HMMs with 3, 4, 5 and 6 hidden states are shown in Figure 7.34.

The fitted five-state HMM results are shown in Table 7.17. State 5 has the largest deformation rate ranges and the largest variances for all three components. It does not transit to State 1 and vice versa. The stationary distribution for the Markov chain is shown in Figure 7.34(c) as the number besides each circle. The Markov chain spends the least amount of time in State 5 and mostly occupies State 1. State 1 has the least amount of changes in trend ranges, and hence is considered as a ground state. We normalize the five states by subtracting the mean vector of State 1 from the mean vector of each state and then dividing the result by the standard deviation vector of State 1.

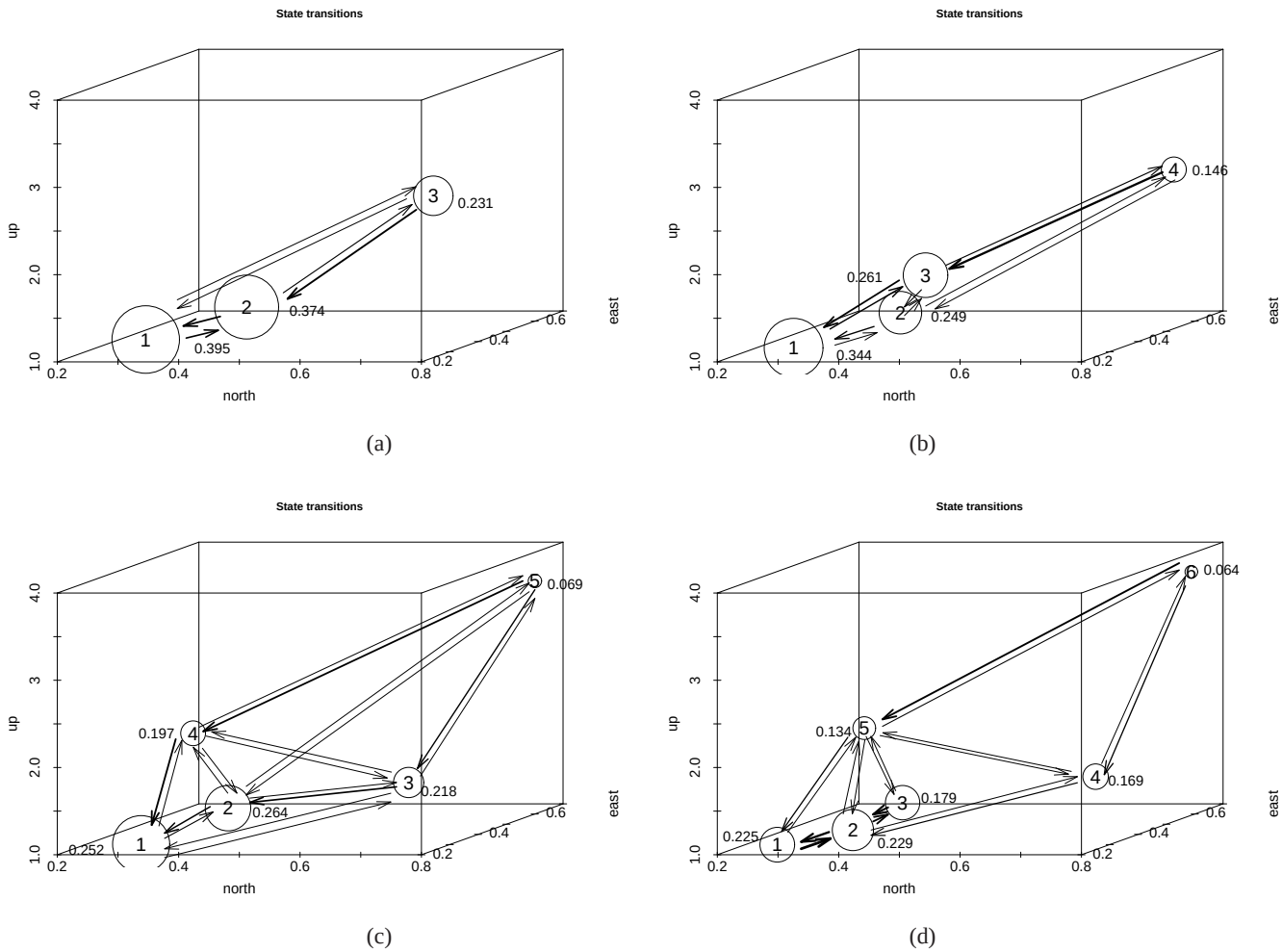


Figure 7.34: Illustration of state transitions for HMMs with 3 states (a), 4 states (b), 5 states (c) and 6 states (d) for data from Southern California. Circle sizes are proportional to the stationary distribution of the hidden Markov chain, i.e., the proportion of time in each state, which is shown as the number besides each circle. The thickness of the arrows is proportional to the transition probabilities. The numbers inside the circles indicate the states and they are located at the means of the states estimated from the fitted model.

Table 7.17: The parameter estimates of the fitted five-state HMM (Southern California). The normalized means are obtained by subtracting the mean vector of state 1 from the mean vector of each state and then dividing the result by the standard deviation vector of state 1.

	State	1	2	3	4	5
Estimated	north	0.296	0.307	0.588	0.336	0.771
means	east	0.289	0.572	0.608	0.387	0.661
	up	1.016	1.099	1.351	2.173	3.599
Estimated	north	0.102	0.085	0.138	0.093	0.472
standard	east	0.088	0.126	0.234	0.143	0.317
deviations	up	0.320	0.318	0.447	0.474	1.562
Estimated	state 1	0.943	0.026	0.013	0.018	0.000
transition	state 2	0.035	0.928	0.023	0.012	0.002
probability	state 3	0.020	0.039	0.910	0.019	0.012
matrix	state 4	0.043	0.023	0.020	0.891	0.023
	state 5	0.000	0.007	0.033	0.042	0.918
Normalized	north	0	0.1	2.9	0.4	4.7
means	east	0	3.2	3.6	1.1	4.2
	up	0	0.3	1.0	3.6	8.1

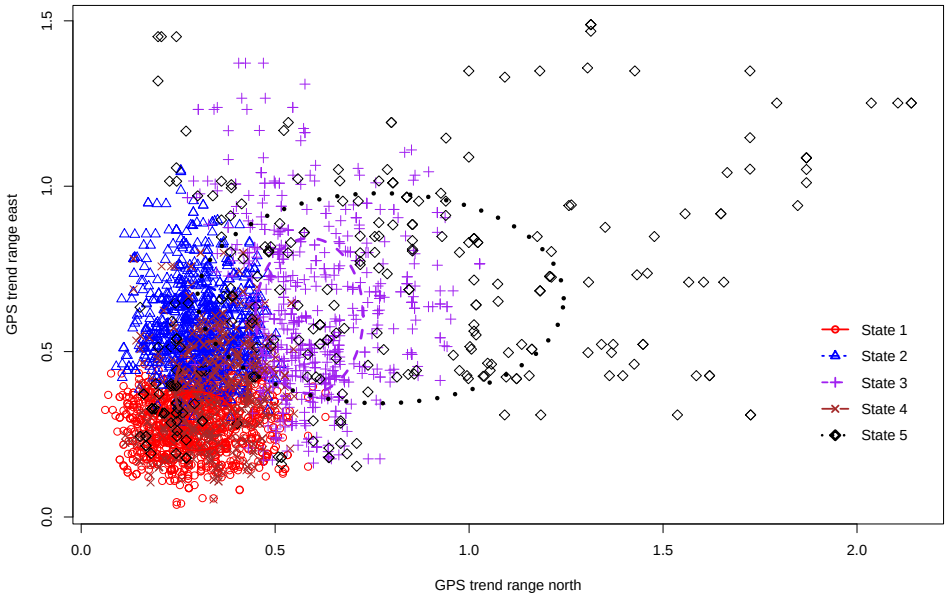


Figure 7.35: Scatter plots of the trend ranges of the east components versus that of the north component. The ellipses illustrates the projected standard deviations for the north and east directions centered at the estimated means (Southern California).

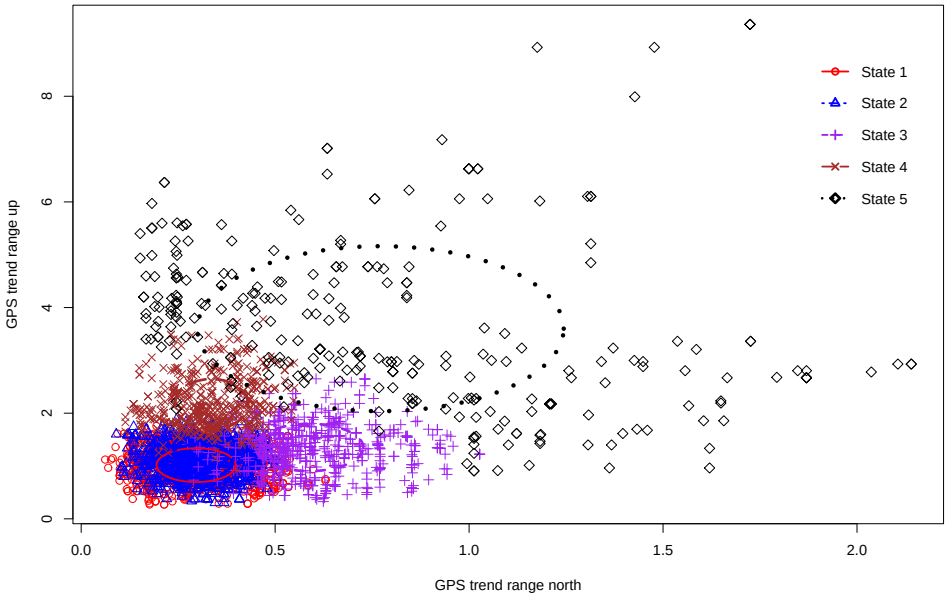


Figure 7.36: Scatter plots of the trend ranges of the up components versus that of the north component. The ellipses illustrates the projected standard deviations for the north and up directions centered at the estimated means (Southern California).

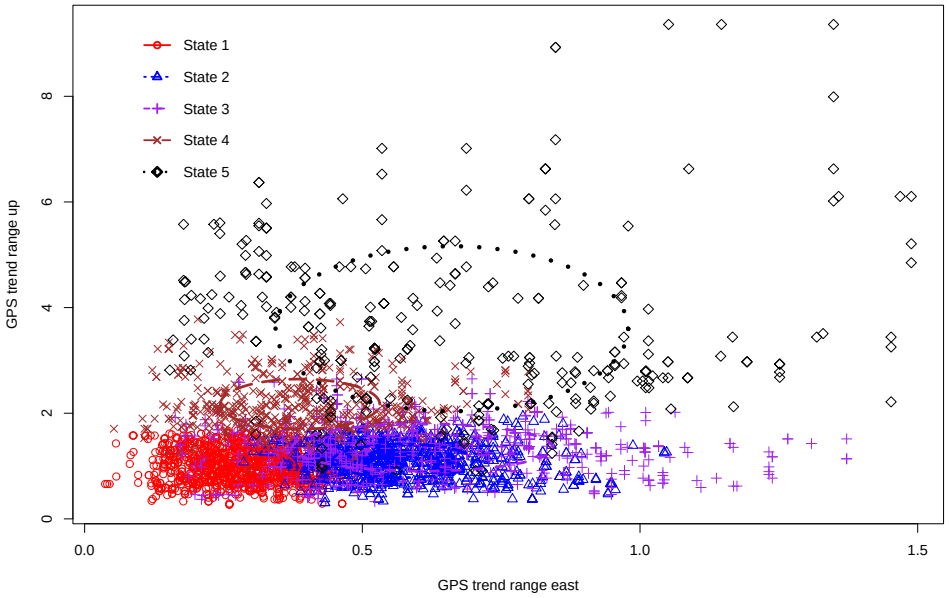


Figure 7.37: Scatter plots of the trend ranges of the up components versus that of the east component. The ellipses illustrates the projected standard deviations for the east and up directions centered at the estimated means (Southern California).

The resulting normalized means are shown in Table 7.17 (see the results of the data around Taupo in Table 7.4 for comparison). State 2 deviates from the ground state largely in the east component. The main deviation of State 4 from the ground state is in the up component. The 3rd state appears to deviate in the north and east components, whereas the 5th state has the largest deviation in all three components.

The scatter plots of the deformation rate ranges of the east component versus north, up versus north and up versus east, with five symbols indicating the different classes in which the points are tracked to be (using the Viterbi algorithm) are shown in Figures 7.35 to 7.37. Comparing to the scatter plots for the data around Taupo, on the north-east surface, unlike the result for Taupo (Figure 7.13) that State 5 is clearly separated from the other states, in Southern California, the data in State 5 are mixed with the other states on this surface (Figure 7.35). However, on the east-up surface, State 5 is separated from the other states in Southern California (Figure 7.37), whereas for the data around Taupo, State 5 overlaps with State 4 (Figure 7.15). The common feature of the two data sets is on the north-up surface, where the ellipse, which illustrates the projected standard deviations of State 5 centered at the estimated means, does not overlap with that of the other states (Figures 7.14 and 7.36). The most likely state sequence tracked using the Viterbi algorithm and the deformation rate ranges of the baselines as calculated in Equation (7.1) are shown in Figure 7.38, along with the occurrence times of earthquakes with minimum magnitude 4.5. Spikes in the deformation rate ranges occur before some of the earthquake occurrences as well. After the HMM visits State 5, a large earthquake often occurs. We calculate the mutual information between each state and the earthquake occurrences to examine whether this association is statistically significant.

The mutual information between the two series with a time lag u is shown in Figure 7.39, which suggests that State 5 shows some preseismic information, while State 3, which accounts for the movement in north and east directions, has a bump around 15 days following earthquake occurrences which may suggest some postseismic information.The other states show little association with the earthquakes. The dot-dash lines in Figure 7.39 indicate the simulated 95% confidence level for comparison. The preseismic information of State 5 is very clearly shown in the plot above the dot-dash line. This again verifies that the association is not arbitrary. We thus consider State 5 in the fitted five-state model as the precursory state. The sojourn time distribution for each state of the fitted five-state model is shown in Figure 7.40. The P-values of the K-S tests for State 1 to State 5 are 0.1693, 0.1605, 0.0439, 0.0425, and 0.3017, respectively. Although the P-values for States 3 and 4 are low, the model still appear reasonable, given the multiple comparison problem and small

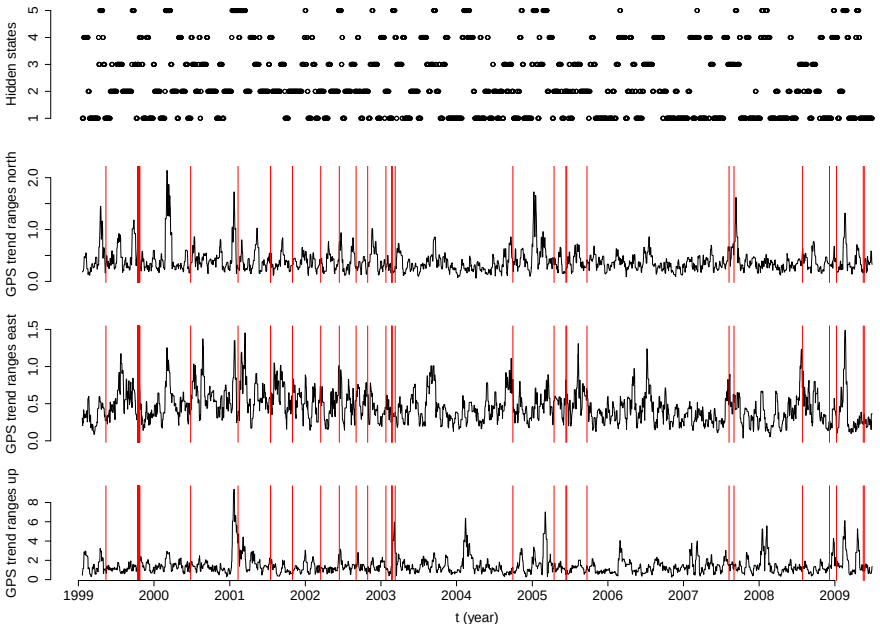


Figure 7.38: The Viterbi path and the trend changes of the GPS movements as calculated in Equation (7.1), with vertical lines indicating the earthquake occurrence times with magnitude larger than or equal to 4.5 (Southern California).

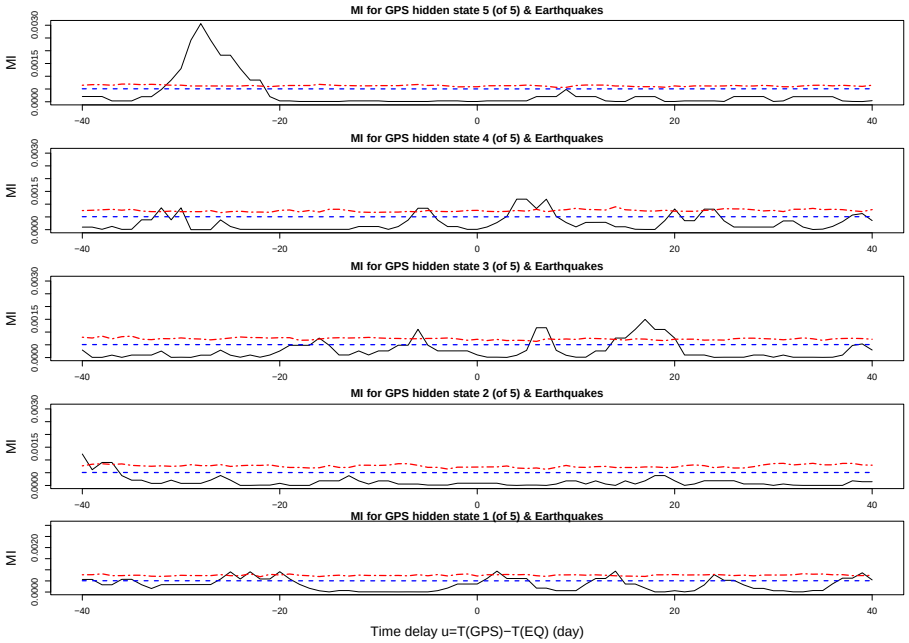


Figure 7.39: Mutual information between U and V (for earthquake magnitude larger than or equal to 4.5) with a time lag u (Southern California). The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent. The dot-dash line in each plot is the calculated 95% confidence level from the 1000 simulated earthquake series and 1000 simulated Markov chains.

sample sizes.

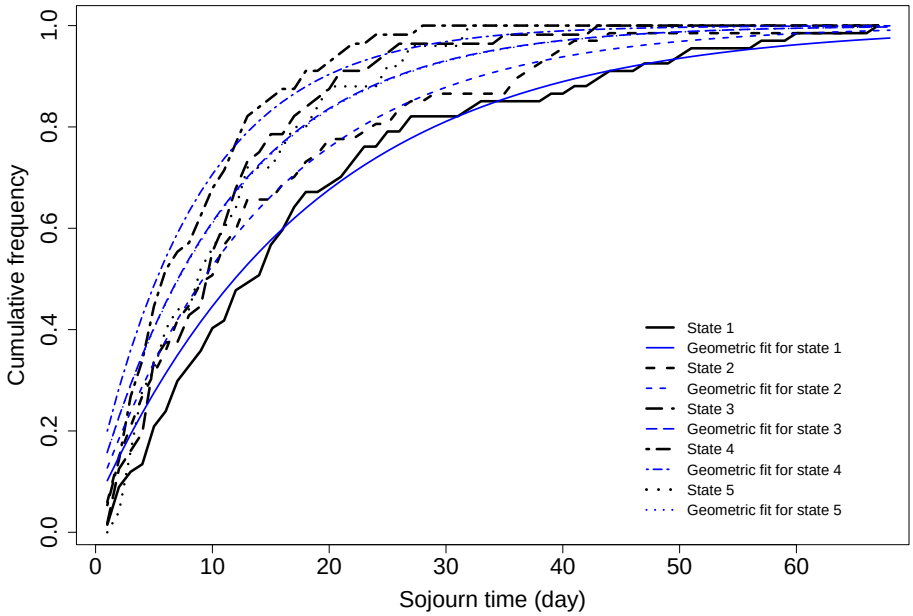


Figure 7.40: The sojourn time distribution for the fitted five-state HMM (thick lines), with the thin lines (for comparison) indicating Geometric distributions, each with mean calculated from the sample mean of the sojourn time in each state (Southern California).

Since there are 50 earthquakes with minimum magnitude 4.5 in a more than 10 year period in Southern California, including 15 earthquakes on the same day, October 16, 1999, we have about 30 occurrence times over the $10+$ years as the time unit we are considering is per day. A cross validation is thus not suitable for this data set due to the sparse data.

Analysis Using Only North and East Components

The Euclidean distance of the variable R_t to the origin $(0, 0, 0)$ is again dominated by the up component, given that the up component is approximately 4 times as large as the other two components and has over 3 times the standard deviation of the other two. It down-weights the north and east component effects. Therefore, the Euclidean distance using all three components is less informative. Thus the HMMs are fitted to the north and east components only, and to the Euclidean distance calculated only using north and east components. Subsequently, another probability forecast using the north and east components only is conducted.

The HMMs are fitted to the deformation rate ranges from the north and east components with different numbers of hidden states. The state transition diagrams of the fitted 3, 4, 5 and 6 state HMMs are shown in Figure 7.41. The mutual information between the Viterbi path $\{U_{t+u}\}$ and the

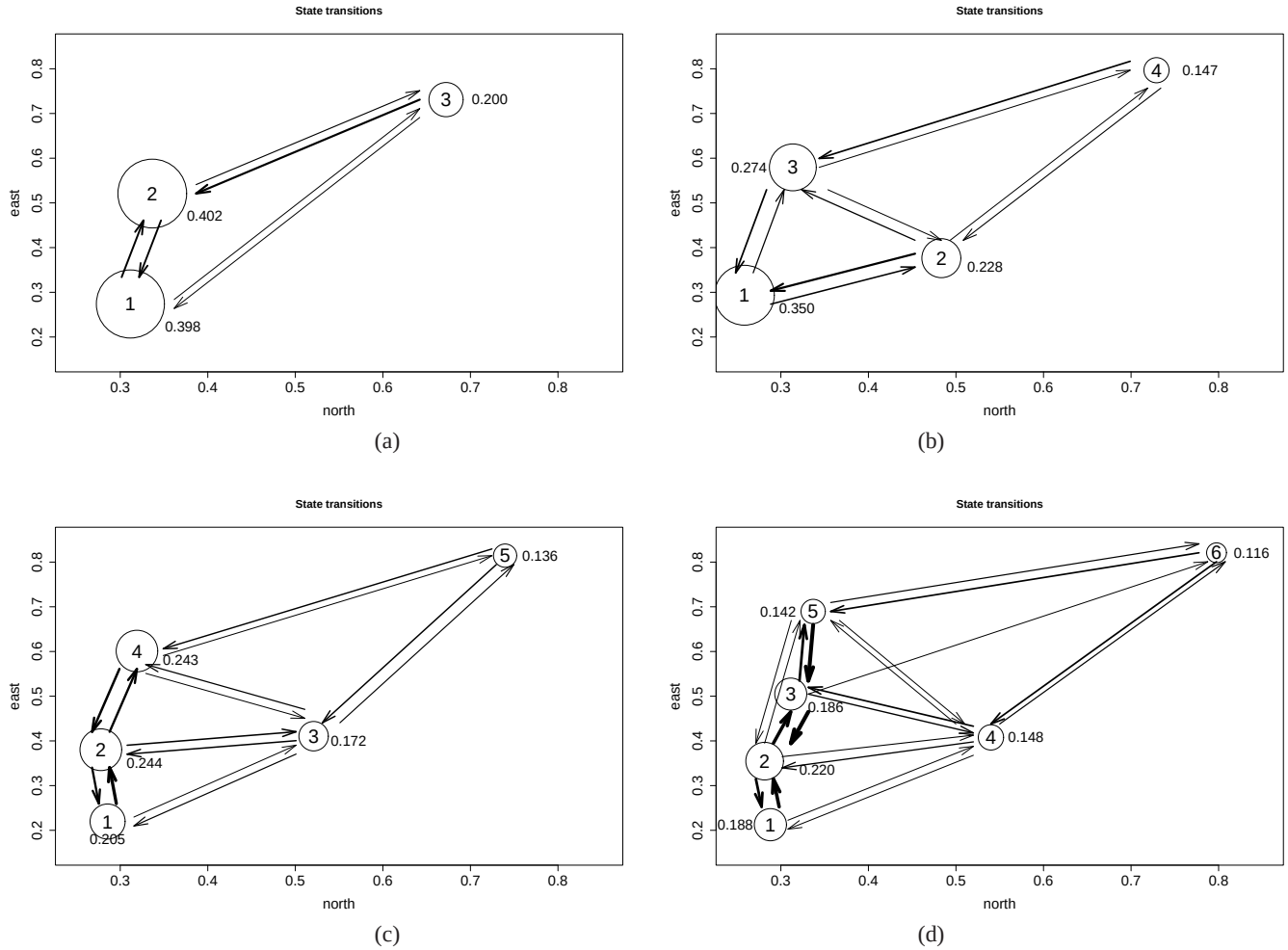


Figure 7.41: Illustration of state transitions for HMMs with 3 states (a), 4 states (b), 5 states (c) and 6 states (d) for data from Southern California, using the north and east components only. Circle sizes are proportional to the stationary distribution of the hidden Markov chain, i.e., the proportion of time in each state, which is shown as the number besides each circle. The thickness of the arrows is proportional to the transition probabilities. The numbers inside the circles indicate the states and are located at the means of the states estimated from the fitted model.

Table 7.18: The parameter estimates of the fitted five-state HMM for the case when only the north and east components are considered (Southern California).

	State	1	2	3	4	5
Estimated means	north	0.286	0.278	0.521	0.319	0.740
	east	0.220	0.380	0.411	0.601	0.814
Estimated s.d.	north	0.096	0.076	0.092	0.087	0.368
	east	0.055	0.058	0.107	0.096	0.226
Estimated transition probability matrix	state 1	0.910	0.077	0.013	0.000	0.000
	state 2	0.054	0.853	0.036	0.056	0.001
	state 3	0.028	0.035	0.881	0.029	0.027
	state 4	0.000	0.059	0.017	0.904	0.020
	state 5	0.003	0.000	0.035	0.033	0.929

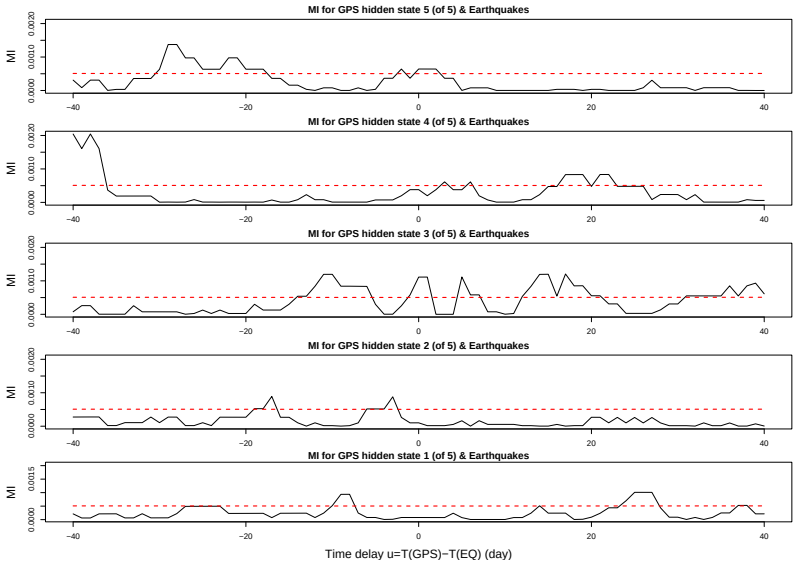


Figure 7.42: Mutual information between U (for trend ranges from north and east components, 5 state HMM) and V (for earthquake magnitude larger than or equal to 4.5) with a time lag u (Southern California). The dashed lines show the upper level of the approximate 95% confidence interval under the hypothesis that the two processes are independent.

earthquakes $\{V_t\}$ for the fitted five-state model is examined. The parameter estimates of the five-state model are shown in Table 7.18. The mutual information is shown in Figure 7.42. For the fitted five-state model, State 5 is the furthest from the origin (0,0) and has the largest variances for both components. State 1, which is the closest state to the origin, does not transit to State 5. Moreover, the other two pairs of states that have negligible probability of transiting to each other are States 1 (the ground state) and 4 (which accounts for the east movement), and States 2 (which deviates from

the ground State 1 in the east component but the mean of which is half of that of State 4) and 5. The mutual information suggests that State 5 may show some preseismic information, but not as strong as when we use all three components. It seems that there is a favored transition pattern to the precursory state, State 1–2–4–5. However, there are not enough (earthquake) data to test whether this pattern will provide additional precursory information as done in Section 7.3.3 for the data set around Taupo.

The scatter plot of the deformation rate ranges from the north component versus that from the east component, with five symbols indicating the different states in which the points are tracked to be is shown in Figure 7.43. The first four states cluster at their centers, while State 5 spread

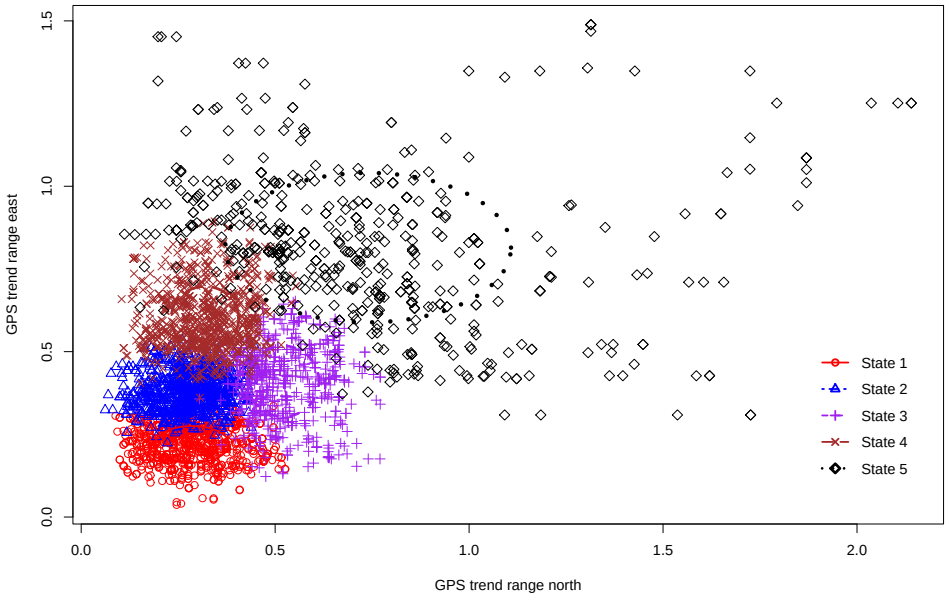


Figure 7.43: Scatter plot of the trend ranges of the east component versus that of the north component. The ellipses illustrates the projected standard deviations for the north and east directions centered at the estimated means (Southern California).

at large values for the two components. The standard deviation contour projection of each state is separated from the others. The most likely state sequence tracked using the Viterbi algorithm and the deformation rate ranges of the baselines as calculated in Equation (7.1) are plotted in Figure 7.44. Spikes in the deformation rate ranges show up before some of the occurrences of the earthquakes with minimum magnitude 4.5. When the HMM visits State 5, a large earthquake often follows. The sojourn time distribution for each state of the fitted 5 state model is as shown in Figure 7.45. The P-values for the K-S tests for State 1 to State 5 are 0.0292, 0.0025, 0.0216, 0.1235, and 0.2650,

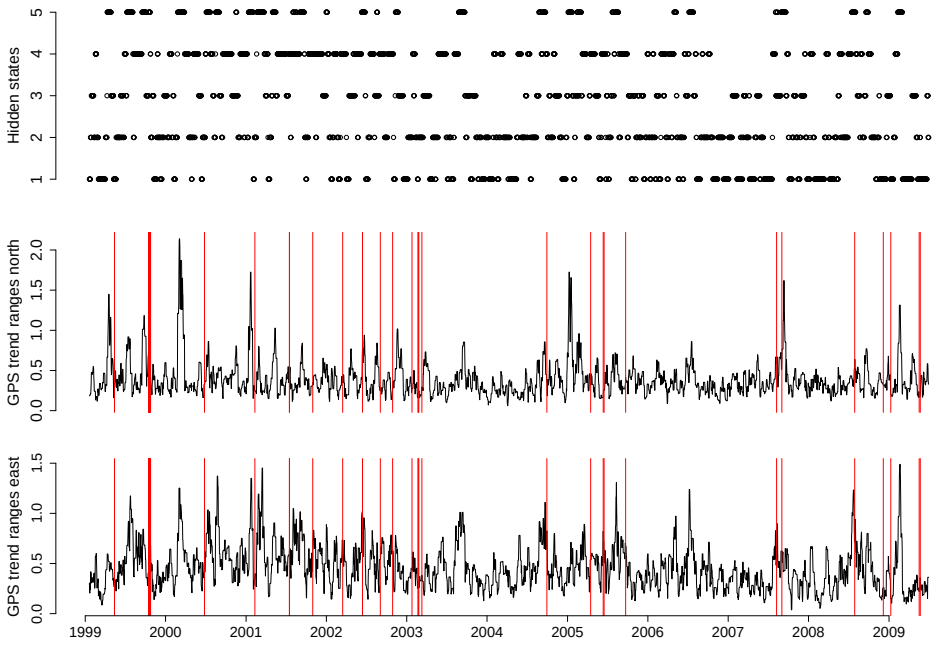


Figure 7.44: The Viterbi path and the trend changes of the GPS movements (north and east components) as calculated in Equation (7.1) (Southern California), with vertical lines indicating the earthquake occurrence times with magnitude larger than or equal to 4.5.

respectively.

7.4.2 Probability Forecast Using Logistic Probability Model

Forecast Using All Three Components of GPS Measurements

According to the mutual information between State 5 and earthquake occurrences with minimum magnitude 4.5 for the case when the entire data were used (see Section 7.4.1), we define a TIP with the interval length of 20 days. A contingency table for testing whether the fifth state is related with the large earthquake occurrences is conducted using the two categorical variables and is shown in Table 7.19, which has a sensitivity of 0.35 and a specificity of 0.83. The Chi-squared statistic of this contingency table with the null hypothesis that the two variables are statistically independent is 3.1232, with a P-value of 0.077. This shows weak evidence (at significance level of 0.1) that visits by the HMM to State 5 are not independent of large earthquake occurrences in the following interval.

Repeating the Logistic linear regression using the linear predictor M_i, the mean of the distance D_t in the interval J_i, with results shown in Table 7.20, we see that the regression coefficient for the predictor is not equal to 0 with a P-value of 0.0575 (or at a significance level of 0.1). The result is

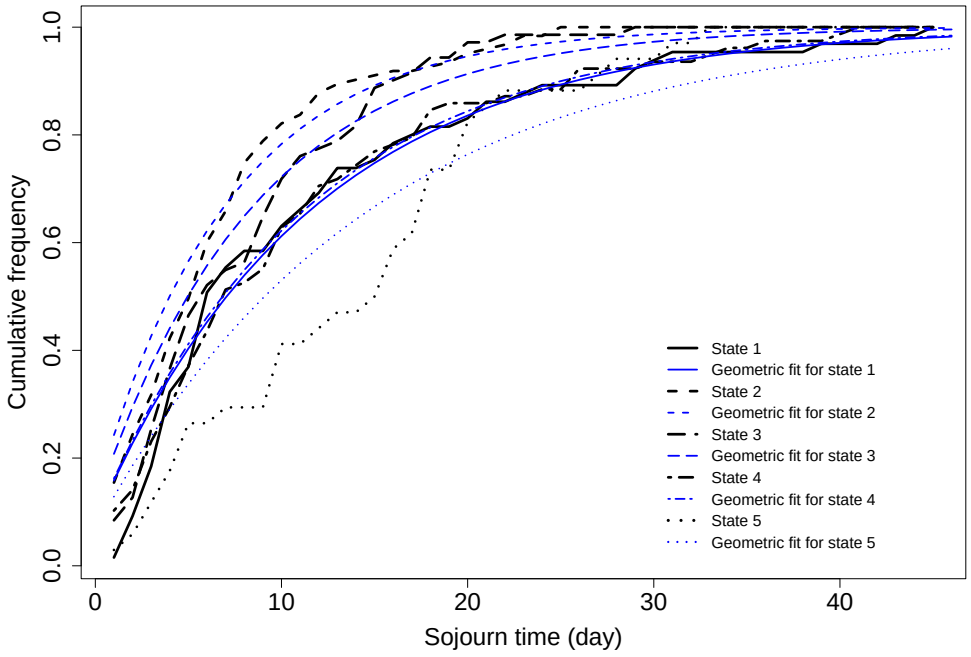


Figure 7.45: The sojourn time distribution for the fitted five-state HMM (thick lines), with the thin lines (for comparison) indicating Geometric distributions, each with mean calculated from the sample mean of the sojourn time in each state (Southern California).

Table 7.19: Contingency table for data from Southern California, where 'Yes' indicates that there is an earthquake with minimum magnitude 4.5 occurring in the interval J_{i+1}, and 'No' indicates that there is no large earthquake occurrence in that time interval.

	Yes	No	Total
J_{i+1} a TIP	8	28	36
J_{i+1} not a TIP	15	138	153
Total	23	166	189

plotted in Figure 7.46. It shows that the larger the mean distance of the deformation rate ranges from the origin in the interval J_i, the higher the probability that there will be a large earthquake occurring in the interval J_{i+1}.

Forecast Using Only North and East Components of GPS Measurements

We define a TIP using the five-state HMM results when we only consider north and east components with the interval length of 20 days. The contingency table for testing whether the fifth state is related with the large earthquake occurrences is shown in Table 7.21. The sensitivity of the test is 0.39, and the specificity is 0.73. The Chi-squared statistic of this contingency table with the null hypothesis

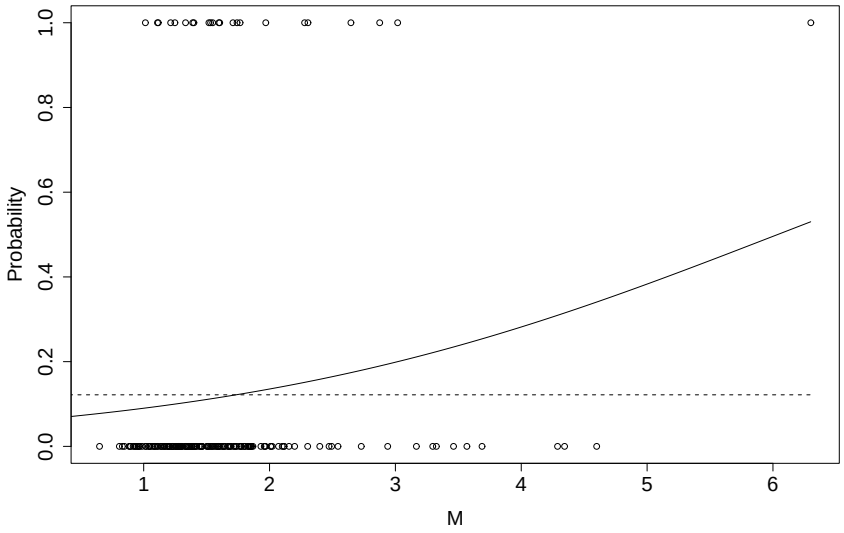


Figure 7.46: Earthquake occurrence in the interval J_{i+1} versus M_i plot (points) with the solid line showing the probability of any magnitude 4.5 or larger earthquake occurrence in the interval J_{i+1} (Southern California). The dashed line indicates Total number of intervals in which there is earthquake occurrence/Total number of intervals.

that the two variables are statistically independent is 0.9022, with a P-value of 0.3422. It becomes less significant when we use only the north and east components than when we use the entire data.

Table 7.20: Logistic regression result for a target event in the time interval J_{i+1} with minimum magnitude 4.5 (Southern California).

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
β_0	-2.7724	0.4955	-5.595	2.20e-08
β_1	0.4592	0.2418	1.899	0.0575

Table 7.21: Contingency table for analysis using only north and east components, where 'Yes' indicates that there is an earthquake with minimum magnitude 4.5 occurring in the interval J_{i+1}, and 'No' indicates that there is no large earthquake occurrence in that time interval (Southern California).

	Yes	No	Total
J_{i+1} a TIP	9	45	54
J_{i+1} not a TIP	14	121	135
Total	23	166	189

This suggests that ignoring the up component will remove some information from the baselines between the two stations. This is consistent with the mutual information results.

Similarly using a Logistic regression to the data using the mean of the distance D_{1t} in the interval J_i, M_{1i}, as a linear predictor of a target event in the time interval J_{i+1}, with result shown in Table 7.22, we see that the regression coefficient for the predictor is significantly not equal to 0 with a P-value of 0.0961 (or at a significance level of 0.1). This result is less significant than that of the Euclidean distance using all three components. The result is plotted in Figure 7.47. It

Table 7.22: Logistic regression result for a target event in the time interval J_{i+1} with minimum magnitude 4.5, using north and east components only (Southern California).

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
β_0	-2.9195	0.6316	-4.622	3.79e-06
β_1	1.4320	0.8606	1.664	0.0961

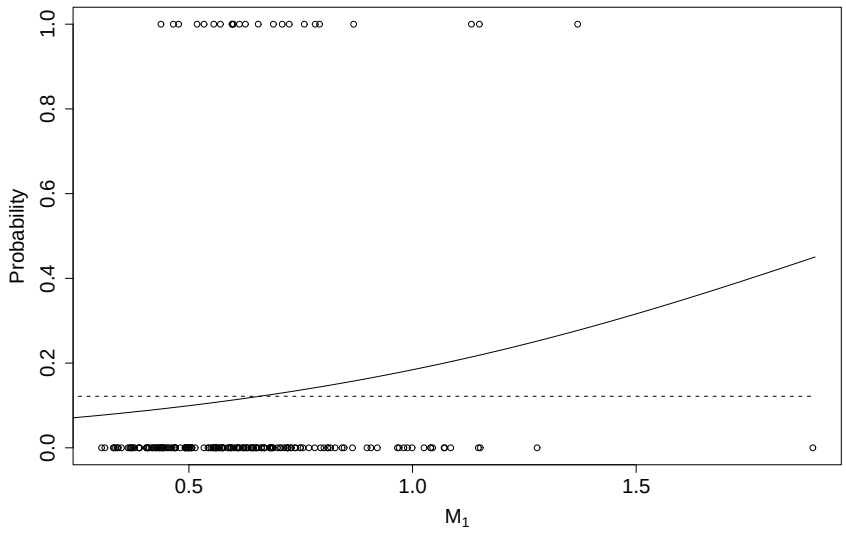


Figure 7.47: The earthquake occurrence in the interval J_{i+1} versus M_{1i} plot (points) with the solid line showing the probability of any magnitude 4.5 or larger earthquake occurrence in the interval J_{i+1} (Southern California). The dashed line indicates Total number of intervals in which there is earthquake occurrence/Total number of intervals.

again confirms that we may lose some information when we do not include the up component in our analysis on the baselines from the two stations.

7.5 Conclusion and Discussion

As discussed in the Section 7.1, the data from tectonically active areas may behave differently from that from a non-tectonic area. The trend of each of the GPS components from the tectonic areas varies on a large scale. In order to examine whether large variations in trend changes contain some precursory signals for large earthquakes, a non-linear filter is developed for the GPS process. It is a smoothing tool and is useful for extracting signals which are distinguished from the majority of the data. As defined in Section 7.2.1, this nonlinear filter essentially measures the maximum deformation rate changes in the previous 10 days. This filter and the underlying dynamics of earthquakes form the hidden Markov model framework. Therefore the HMM is used to investigate the filtered GPS data and then classify the data into different categories via the Viterbi algorithm. The association between each category and the earthquake sequence is then examined by mutual information.

For the data around Taupo, it seems that large variations in the deformation rate may provide precursory information for large earthquakes. This pre-seismic information becomes stronger after discarding smaller earthquakes, a lot of which are aftershocks. This confirms that State 5 may possess some precursory signals for large earthquakes. The Chi-squared test for the contingency table also shows that State 5 is related to subsequent earthquake occurrences. Later on, the regression analysis for the TIP confirms that the further the GPS deformation rate variations depart from the origin the more likely there will be an earthquake occurrence in the TIP. The largest state in each of the fitted HMMs with up to 10 hidden states also presents some preseismic information for large earthquakes. However, as mentioned in Section 7.3.2, the precursory state in the five-state model splits into two states in the fitted HMMs with 7, 8, 9 and 10 hidden states, both of which presented precursory information. Furthermore, the fitted 10 state model with the minimum BIC value has 159 parameters for 1747 days data, too many for stability. The data from Southern California may possess a similar precursory behavior. The chosen model also has 5 hidden states. The furthest state from the origin shows precursory information for large earthquakes, but more (earthquake) data is required for confirmation

For the data around Taupo, State 3 in the five-state model also shows some preseismic information, though when we increase the magnitude threshold to 5.1, the preseismic information becomes weaker than using magnitude threshold 4.2. This may correspond to some cluster effect either of small foreshocks or due to the previous large earthquakes. The post-seismic information (from the mutual information results) in State 2 disappeared after deleting the small earthquakes. The state transitions of the fitted hidden Markov model as illustrated in Figure 7.10(c) suggest that State 5 is

most likely to transit to State 2. State 2 is also likely to transit to State 5. This state may be the postseismic state which is more related with aftershocks. The mutual information shows little association of State 1 with the earthquake occurrences. State 1 has the smallest means and variations for both the north and up components and the process spends most of the time in this state. Note that States 1 and 5 do not transit to each other. Apart from each state being more likely to stay in its own, States 2, 3 and 4 are all most likely to transit to State 1. This may suggest that State 1 consists of the background noise for the GPS measurements of deformation and corresponds to a quiescence period of the underlying dynamics.

Note that State 3 in the five-state model for the data in Southern California shows postseismic information for the large earthquakes. This state deviates from the background State 1 mainly in the north and east components. We also notice that the main deviation of the postseismic State 2 from the background State 1 for the data around Taupo is in the north component. Both the postseismic states are not related to movements in the up component. Due to the nature of the GPS measurements, the up component is much larger than the north or east components and has larger standard errors than that of the other two components, hence the Euclidean distance calculated from all three components is dominated by the up component. Therefore, an HMM analysis on this distance does not perform as well as the analysis on the multivariate displacement data. We conducted the probability forecast only considering the north and east components for comparison. The result suggests that for the data around Taupo using only the two components improves the probability forecast, whereas for the data in Southern California, not considering the up component may remove some information. The Taupo Volcanic Zone is a subduction-related rift zone and hence deformation in the up component tends to follow crustal extension (and thinning) during an earthquake, i.e., the up component for the data around Taupo contains postseismic, not precursory information. However, the tectonic environment in Southern California is different, being predominantly strike-slip. If, prior to an event, the strike-slip fault is held fixed, the strain may cause the ground to buckle, which corresponds to up motion. Hence, ignoring the up component may lose some precursory information for large earthquakes.

Chapter 8

Conclusions and Future Research

8.1 Conclusions

This thesis has contributed to two topics in parallel. Firstly, assuming the existence of an earthquake cycle (for example, mainshock–aftershock–quiescence–precursory seismicity), a new HMM type model is proposed. This model can be used to capture a self-exciting process which switches among some different phases (or states) in each of which the process has a distinguishable characteristic. Secondly, two nonlinear filters are developed to extract signals from millions of data or data with subtle changes, which can not be easily detected by visual examination. A method to combine two very different methods, HMMs and mutual information, is introduced to investigate the link between two processes. For the former, unlike the traditional HMM and MMPP, the distinctive feature of this new HMM type model, MMHPSD, is the incorporation of a self-exciting point process into a continuous-time hidden Markov chain. The existing self-exciting models can only capture one or several fixed or pre-identified phases in one (seismic) cycle. For example, the ETAS model formulates aftershock sequences, while the two-node stress release/transfer model can capture main shocks and aftershocks. The new model, however, while characterizing the self-exciting feature of each phase, switches into a new regime automatically whenever the feature of the event occurrences evolves towards a different attractor. For the latter, the extracted signals from the Tangshan Well data using the nonlinear filter were demonstrated to be strongly associated with large global earthquakes. The combined method of an HMM and mutual information on the filtered GPS data around Taupo and in Southern California revealed some preseismic signals. With more testing and longer data sequences, GPS measurements may be able to provide some probability forecast for large earthquakes.

The parameter estimation for the new HMM type of model incorporating a time-varying conditional intensity function is nowhere near trivial. This issue is further discussed in Section 8.2.1. Due to the computational difficulty in solving the integration of a time-varying matrix exponential, we restricted the conditional intensity function to vary only when a new event occurs and remains constant between two consecutive events, which results in the MMHPSD. Under this constraint, a method for estimating the parameters via EM algorithm was developed, which involves a numerical optimization in the M-step for estimating the parameters in the Hawkes intensity function. The residual analysis for point processes is borrowed to evaluate the goodness-of-fit for this model. Simulation is employed to demonstrate the consistency of the parameter estimation.

Compared to the ETAS model, the Hawkes process always has lower intensity at very short and very long times after an event. This is due to the fact that the ETAS model assumes that the aftershocks decay in a power law fashion, whereas the Hawkes process assumes an exponential decay rate. As discussed in Section 4.3.1, the power law decay always tends to have larger values at extreme time intervals. Moreover, in the ETAS intensity function the decay rate is multiplied by the exponential term of the magnitude always larger than 1, which thus results in very large ETAS intensity. When we fit the MMHPSDs to the simulated ETAS sequence, on average, the estimated intensity of each of the fitted models is smaller than the true ETAS intensity. It appears that various hidden states of the MMHPSD correspond to different magnitude effects parameterized as $e^{\alpha(M_i-M_0)}$ in the ETAS model. In particular, the state with large (small) intensities captures large (small) earthquakes.

The exploratory data analysis of the MMHPSD to the earthquake sequence around Landers shows that this model is useful in modelling changes of the temporal patterns of seismicity. The states in the model can capture the behavior of main shocks, large aftershocks, secondary aftershocks and a period of quiescence with different background rates and decay rates. The state transitions can then explain the seismicity rate changes and hence indicate if there is any seismicity shadow or relative quiescence. The ETAS model is purely an immigration-branching process. The advantage of this model over the ETAS model is that this model uses seismic cycles rather than only immigration-birth framework, and when the seismicity rate changes, the model automatically switches into a different regime (or state) in a seismic cycle. This is especially useful for a long sequence with several state changes. One may use the change-point analysis for the ETAS model to account for the seismicity changes. Firstly, estimating the change point is not an easy procedure in the way that one has to try as many points as possible to determine one change point. Even if

the time points of the events are used as the candidates, a great number of models have to be fitted. Secondly, it is common that there may be more than one change point in the sequence. For the change point analysis, one has to determine one change point first, and then for the data preceding this change point, the analysis can be repeated to determine another change point, and so on. The procedure can be very complicated and time-consuming. The MMHPSDs, however, can handle a large data set with many state changes without manually dividing the data into different segments.

An R-package was developed for this model. Since this involves numerical optimization inside the EM loop, the parameter estimation program is somewhat time-consuming. In Chapter 4, the model was selected based on a residual analysis. Starting from a two state MMHPSD, one more state is added each time until the residual point process of the current model becomes a stationary Poisson process with unit rate, i.e. until the current model can capture the main features of the data. It is likely that there are still better models with more states which may have smaller AIC and BIC values than the ETAS model. But including more states means more parameters to be estimated via numerical optimization. In order to get estimates close to the true parameters, a wide range of initial values should be tried, resulting in a longer computation time.

For the investigation of a possible link between earthquakes and the ancillary data, we reviewed several statistics which can be used to quantify the association between series of events. Coherence describes the strength of linear association between two series in frequency domain. Mutual information measures the information that two random variables share. The advantage of using the mutual information is that it equals to 0 if and only if the two variables are statistically independent, whereas the coherence may be identically 0 when two series are actually related. Though the two statistics are not naturally set up for point process context, we can transform the processes of interest into 0-1 time-series. If we are further interested in whether one process causes another, the above two statistics are not adequate any more. In this case, we can use the Lin–Lin model to examine the linear causal relationship between two processes.

For the ancillary well data, we introduced a variance-based moving average method and showed how signals compatible with coseismic responses can be extracted using this method from approximately two million groundwater level data. We then adopted the aforementioned three statistics to investigate the relation between the identified well signals and 600 earthquakes of magnitude 6.0 or greater in the global catalogue during this same period. Identifiable coseismic responses are found for approximately 40 percent of the total number of such teleseismic earthquakes. The initial oscillations of the groundwater level appear to be strongly associated with the arrivals of the

earliest P phase, although the maximum amplitude usually follows the arrival of the later Love and Rayleigh waves. The detection probability, and well signal characteristics (delay, length, maximum amplitude) are quantified as functions of earthquake characteristics (magnitude, distance, depth and azimuth), showing that the response contains considerable variation, as yet not understood.

This idea is then extended to the analysis of possible link between GPS measurements of deformation and earthquakes. The difference is that the GPS measurements consist of multivariate data from three dimensions, north, east and up. Moreover, the well responses to earthquakes appear to be oscillations, however, the anomalous changes in the deformation data which may be related with earthquakes are either long-term apparent displacement (see, e.g., Ogata, 2007), or subtle changes for which it is necessary to have some techniques to detect or extract the anomalies. We developed an algorithm to first filter the raw GPS data and then used the HMMs on the filtered data to get signals which may be possible precursors.

For two case studies of a) deep earthquakes in central North Island, New Zealand, and b) shallow earthquakes in Southern California, an HMM fitted to the short-term deformation rate ranges of the GPS measurements can classify the deformation data into different patterns which form proxies for states of the earthquake cycle. Mutual information can be used to examine whether there is any relation between these patterns, in particular the Viterbi path, and subsequent (or previous) earthquakes. The class of GPS movements identified by the HMM as having the largest range of deformation rate changes, appear to have some precursory character for earthquakes with minimum magnitude 5.1 (central North Island, New Zealand, 26 earthquakes in 1747 days) and 4.5 (Southern California, 50 earthquakes in 3815 days). We defined a “Time of Increased Probability” (TIP) as being a 10-day interval (central North Island, New Zealand) or a 20-day interval (Southern California) following entry (as identified by the Viterbi algorithm) into the “precursory” hidden state, and examined the performance of this as a possible forecast of subsequent earthquakes. The purpose of this study was to examine whether there is any causal relationship between the GPS measurements of deformation and the earthquakes. The analysis shows that indeed there may be some weak causal relationship between the two.

8.2 Future Research

This research is working on the interface of statistics, geophysics and geodesy. Earthquake analysis is of great importance in terms of both exploring the nature of earthquakes and forecasting earth-

quakes. Statistical models play an important part not only in understanding the earthquake process itself, but also in the probability forecast of earthquakes and estimation of hazard from earthquakes. This research focuses on the two aspects in the way of developing suitable stochastic models to investigate the seismicity rate changes and incorporating ancillary data to examine the seismic responses and possible earthquake forecasting. The twofold nature of this research and the research findings in this thesis give rise to new challenges for the future research. Possible extensions and further analyses which this thesis does not cover are suggested in the following subsections.

8.2.1 Markov-modulated Hawkes Processes with Time-varying Decay

The proposed MMHPSD switches among some finite states according to a Markov transition rate matrix with a self-exciting occurrence rate of the events from a Hawkes process in which the intensity of this process changes after each event occurrence but remains a constant between each two consecutive events. The conditional intensity function of this process is

$$\lambda^*(t) = \lambda + \nu\eta \sum_{t_j < \max\{t_l : t_l < t\}} e^{-\eta(\max\{t_l : t_l < t\} - t_j)}, \tag{8.1}$$

not the original Hawkes process with conditional intensity function

$$\lambda^*(t) = \lambda + \nu\eta \sum_{t_j < t} e^{-\eta(t - t_j)}. \tag{8.2}$$

This is for the sake of parameter estimation. Note that when calculating the transition probability without arrival $H_{ij}^{(n)}(u)$, the ordinary differential equation (3.2) has to be solved. If the conditional intensity function takes the form of (8.2), the ordinary differential equation will become

$$\begin{cases} H^{(n)'}(u) = H^{(n)}(u)(Q - \Lambda^*(t_{n-1} + u)) \\ H^{(n)'}(0) = I \end{cases} \tag{8.3}$$

and the solution for this equation is $H^{(n)}(u) = \exp\{(Q - \Lambda)u + \Lambda^{I_n}(t_{n-1} + u)\}$ for $u \geq 0$, where

$$\Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \lambda_r \end{pmatrix}, \qquad \Lambda^{I_n}(t) = \begin{pmatrix} \lambda_1^{I_n}(t) & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \lambda_r^{I_n}(t) \end{pmatrix},$$

and $\lambda_i^{I_n}(t) = \nu_i \sum_{t_l<t_n}(e^{-\eta_i(t-t_l)} - e^{-\eta_i(t_{n-1}-t_l)})$. This will cause difficulty when carrying out the EM algorithm for the parameter estimation. The reason is that when the conditional intensity function (8.2) is used, as a consequence of the change of the ordinary differential equation, Equation (3.17) becomes

$$\begin{aligned}\hat{w}^T =& Q^T \odot \sum_{k=1}^{n} \frac{1}{c_k} \int_{t_{k-1}}^{t_k} \exp\{(Q-\Lambda)(t_k-t) + \Lambda^{J_k}(t)\}\Lambda^*(t_k)R(k+1)L(k-1) \\ & \times \exp\{(Q-\Lambda)(t-t_{k-1}) + \Lambda^{I_k}(t)\}\mathrm{d}t, \end{aligned} \tag{8.4}$$

where $\odot$ denotes term-by-term multiplication of the two matrices. Consequently, the integral

$$\begin{aligned}\mathcal{I}_k = & \int_{t_{k-1}}^{t_k} \exp\{(Q-\Lambda)(t_k-t) + \Lambda^{J_k}(t)\}\Lambda^*(t_k)R(k+1)L(k-1) \\ & \times \exp\{(Q-\Lambda)(t-t_{k-1}) + \Lambda^{I_k}(t)\}\mathrm{d}t, \end{aligned} \tag{8.5}$$

has to be calculated. To the best of our knowledge, there does not exist an easy way to do this. However, when the stepwise decay rate (8.1) is used, this problem can be easily solved. As an analytical solution for the integration (8.5) is difficult to solve, one may consider integrating it numerically. Again, a numerical solution is not easy as well given that there is a matrix exponential in the integrand with each element of the matrix being a function of t. If this integration problem is solved, then we can carry on analyzing the earthquake data using the HMM type model incorporating the original Hawkes process with exponential decay rate.

Another possible extension would be using the power law decay rate (i.e., Omori-Utsu formula), which is well known as the best empirical temporal-distribution of aftershocks, instead of the exponential decay rate. This would result in many new challenges to our mathematical calculations given that we will lose the Markovian property.

8.2.2 MMHPSD with Marks

When the MMHPSD is applied to the earthquake sequences from Landers, Big Bear, Hector Mine and Joshua Tree, this simple initial model accounts for different states of earthquake occurrence rates. It provides an exploratory analysis of earthquake occurrences. As discussed in Section 4.4.2, the model interprets the magnitude effect on the inter-event times although no magnitude term is included in the intensity function. The state with the biggest conditional intensity function captures the features of the largest earthquakes which have short inter-event times. The state with the smallest

intensities occupies the times when small earthquakes occur. This suggests that there may be a possible way to improve this model by including the magnitude effect in the intensity function. As pointed out in Chapter 4, the magnitude term in the ETAS model, parameterized as an exponential function, may be too strong along with the power law decay function. However, a power law, $(M_i - M_0)^\alpha$, may be suitable to be included in the intensity function of the MMHPSD. This may improve the fitting to the data if this magnitude term can capture the majority of the magnitude effect. Each state might account for a longer sequence of events which could be a period of an aftershock sequence following a large earthquake, a period of swarms, or a period of foreshocks indicating accelerating seismic activity preceding a large earthquake.

8.2.3 NHMM and MMGLM Analysis of Earthquakes with Ancillary GPS Data

It has been shown that indeed there may be some weak causal relationship between the GPS measurements of deformation and the earthquakes. The GPS measurements of deformation reflects some of the underlying dynamics for earthquake occurrences. For future research, the GPS measurements can be incorporated as ancillary data (or covariate) into the HMM framework to investigate earthquake occurrences. Whether/how the GPS measurements influence earthquake occurrence modelling can be examined using HMMs incorporating the GPS measurements via three different ways. For any day t, the observation of interest here can be any earthquake occurrence (with magnitude greater than or equal to a threshold M_0) in the time period $I_t = (t, t + l)$ days. It is a binary variable, $\mathbf{O}_t$, taking on the value 1 when there is any magnitude M_0 or greater earthquake occurring in the time interval I_t, and 0 otherwise. The covariate can be the current Euclidean distance of the GPS trend ranges to its origin (0,0,0), D_t. The following three models can be considered:

1. HMM with

$$\begin{cases} P(\mathbf{O}_t \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_{t-1}, S_1, \cdots, S_T) = P(\mathbf{O}_t = o_t \,|\, S_t = s) = p_s^{o_t}(1 - p_s)^{1-o_t} \\ P(S_t \,|\, S_1, \cdots, S_{t-1}) = P(S_t = j \,|\, S_{t-1} = i) = p_{ij} \end{cases} \tag{8.6}$$

   where p_s is the probability of earthquake occurrence in state s, and p_{ij} is the transition probability from state i to j;

2. NHMM with

$$\begin{cases} P(\mathbf{O}_t \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_{t-1}, S_1, \cdots, S_T, D_1, \cdots, D_T) = P(\mathbf{O}_t = o_t \,|\, S_t = s) = p_s^{o_t}(1-p_s)^{1-o_t} \\ P(S_t \,|\, S_1, \cdots, S_{t-1}, D_1, \cdots, D_T) = P(S_t \,|\, S_{t-1}, D_t) \end{cases} \tag{8.7}$$

where the parameter p_s remains the same as in the HMM, and the transition probabilities can be determined by

$$\begin{aligned} \text{logit}P(S_t = 2 \,|\, S_{t-1} = 1, D_t) &= a_1 + b_1 D_t, \\ \text{logit}P(S_t = 1 \,|\, S_{t-1} = 2, D_t) &= a_2 + b_2 D_t, \\ P(S_t = 1 \,|\, S_{t-1} = 1, D_t) &= 1 - P(S_t = 2 \,|\, S_{t-1} = 1, D_t) \\ P(S_t = 2 \,|\, S_{t-1} = 2, D_t) &= 1 - P(S_t = 1 \,|\, S_{t-1} = 2, D_t), \end{aligned}$$

with parameters a_1, b_1, a_2 and b_2;

3. MMGLM with

$$\begin{cases} P(\mathbf{O}_t \,|\, \mathbf{O}_1, \cdots, \mathbf{O}_{t-1}, S_1, \cdots, S_T, D_1, \cdots, D_T) = P(\mathbf{O}_t = o_t \,|\, S_t = s, D_t = d_t) \\ \qquad\qquad = \exp\left\{\left(o_t \log\left(\frac{\mu_{ts}}{1-\mu_{ts}}\right) - \log\left(\frac{1}{1-\mu_{ts}}\right)\right)\right\} \\ P(S_t \,|\, S_1, \cdots, S_{t-1}) = P(S_t = j \,|\, S_{t-1} = i) = p_{ij} \end{cases} \tag{8.8}$$

where the link function is $\log\left(\mu_{ts}/(1-\mu_{ts})\right) = \zeta_{0s} + \zeta_{1s} d_t$ with parameters ζ_{0s} and ζ_{1s} for state s at time t.

Model (8.6) does not include any covariate effect. It models the earthquakes in a simple way which has binary event occurrences. In Model 8.7, the covariate, GPS measurements, indirectly reflects the observed process (earthquakes) through the hidden process (underlying dynamics). The underlying dynamics classifies the GPS measurements into several classes, which most probably reflect certain earthquake occurrence patterns. Model 8.8 assumes that the earthquake occurrences are directly related to the GPS measurements and the GPS measurements do not influence the transitions of the hidden states. Another possibility would be considering the time to next event as the observation for the three models with an exponential distribution, and the Euclidean distance of the GPS trend ranges to its origin (0,0,0) as a covariate.

8.2.4 Testing of HMM Analysis of GPS Data for Earthquake Forecasting

A 10-day moving window was used in the nonlinear filter (7.1) introduced in Section 7.2.1 for analysis of the GPS measurements. A test of how the choice of the window length may influence the mutual information can be done to examine the window length effect. The window length here acts as a smoothing parameter which determines the degree of smoothness. In addition, analysis of the prediction success versus the earthquake characteristics, the depths, distances, focal mechanisms and sources of the earthquakes, can be carried out, along with a sensitivity analysis on the earthquake catalogue used (the choice of magnitude threshold, depth and location of the earthquakes). In Chapter 7, the TIP intervals for probability forecasts were chosen to be 10 days for Taupo and 20 days for Southern California. A further investigation of the selection of this TIP interval by balancing false alarms and missed events may improve the probability forecast.

In Chapter 7, the mutual information was calculated by transforming the data into bivariate binary series. According to the definition in Chapter 5, it may be worthwhile to directly consider the mutual information between the short-term deformation ranges of the GPS measurements (R_N, R_E, R_U) and the earthquake binary series. The definition of the mutual information would need to be extended to the multivariate case.

Also of interest would be examining whether the distance from the fault axis or station has any influence on the predictability of earthquakes using the procedure in Chapter 7. Moreover, GPS measurements at more locations from different tectonic settings, with various types of mechanisms, may provide further insights into probability forecasting for large earthquakes.

Appendix

A. Dispersion Test (Cox and Lewis, 1966)

Let $n_1, n_2, \cdots, n_k$ denote k observations for a discrete random variable N and let

$$\bar{n} = \frac{n_1 + \cdots + n_k}{k}. \tag{A-1}$$

A standard test for the hypothesis that the n_i's are observations of a Poisson variate is the dispersion test for homogeneity based on the statistic

$$d = \sum_{i=1}^{k} \frac{(n_i - \bar{n})^2}{\bar{n}}. \tag{A-2}$$

The statistic d divided by $k - 1$ is the ratio of the estimated variance of N to the estimated mean of N, and the test is roughly a comparison of this ratio with its true value of unity under the null hypothesis. Possible alternative hypotheses are that the n_i's are non-Poisson and serially correlated.

B. Logistic Regression Analysis

For the purpose of exposition, we will momentarily denote the magnitude M, log-scaled well–epicenter distance $\log_{10}(D)$, depth H, and the sine and cosine functions of the azimuth (given that the azimuth is an angle, we use the sine and cosine functions of the azimuth) S_ζ and C_ζ by $x_i, i = 1, 2, 3, 4, 5$. If the probability that an earthquake causes a coseismic response is p, then the probability of an observation y (where y is either a coseismic response or not, 1 or 0) has the Bernoulli distribution

$$P(y) = p^y (1 - p)^{1-y}.$$

The random variable y has a mean of p. Our objective is to determine how the explanatory variables (in this case, the magnitude, log-scaled well–epicenter distance, depth, sine and cosine functions of azimuth) influence the probability value p. These five continuous explanatory variables, which we will denote by x_1, x_2, x_3, x_4 and x_5, suggest that a multiple regression analysis may be suitable. The response y is binary. Instead of using a regression directly for the success probability, one usually transforms the probability scale from the range $(0, 1)$ to $(-\infty, \infty)$, and then use a regression for the transformed values. Logistic transformation is usually recommended because it is more convenient. Moreover, it provides a direct interpretation in terms of the logarithm of the odds of success, which is defined to be the ratio of the probability of a success to the probability of a failure, $p/(1-p)$. Hence we can do logistic regression with binomial errors (see Collett, 1991 for example).

In order to check the interactions between the earthquake statistics as well, we fit a logistic model with maximal interactions between explanatory variables,

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) =& b_0 + \sum_{i=1}^{5} b_i x_i + \sum_{i=1}^{4}\sum_{j=i+1}^{5} b_{ij} x_i x_j + \sum_{i=1}^{3}\sum_{j=i+1}^{4}\sum_{k=j+1}^{5} b_{ijk} x_i x_j x_k \\ &+ \sum_{i=1}^{2}\sum_{j=i+1}^{3}\sum_{k=j+1}^{4}\sum_{l=k+1}^{5} b_{ijkl} x_i x_j x_k x_l + b_6 x_1 \cdots x_5 + \varepsilon \end{aligned} \tag{A-3}$$

to the data, where ε is the residual. Then by using stepwise regression and comparing the Akaike information criteria (Akaike, 1974) of the resulting models, we can identify the best model. Here it turns out that none of the interactions had a significant effect, and that the AIC of the linear regression model without interactions was significantly smaller (difference in AIC is greater than 2) than that of the models with interactions. Therefore, we do not need to include the interactions between explanatory variables in our analysis. However, since non-parametric smoothers in generalized additive models suggest some nonlinearity of the variables, particularly the log-scaled well–epicenter distance and the depth, we did include squared terms for the explanatory variables in the logistic model.

Therefore, after eliminating the interaction terms $\sum_{i=1}^{4}\sum_{j=i+1}^{5} b_{ij}x_i x_j$, $\sum_{i=1}^{3}\sum_{j=i+1}^{4}\sum_{k=j+1}^{5} b_{ijk}x_i x_j x_k$, $\sum_{i=1}^{2}\sum_{j=i+1}^{3}\sum_{k=j+1}^{4}\sum_{l=k+1}^{5} b_{ijkl}x_i x_j x_k x_l$ and $b_6 x_1 \cdots x_5$ in model (A-3), and adding the quadratic terms $\sum_{i=1}^{4} b_{1i}x_i^2$, we fit the model

$$\log\left(\frac{p}{1-p}\right) = b_0 + \sum_{i=1}^{5} b_i x_i + \sum_{i=1}^{4} b_{0i} x_i^2 + \varepsilon. \tag{A-4}$$

Note that $x_4^2 + x_5^2 = \sin^2\zeta + \cos^2\zeta = 1$, and hence we only need to include one of the quadratic terms of x_4 and x_5.

C. Multiple Regression Analysis

We first use scatter plots and generalized additive models using non-parametric smoothers to examine whether there is nonlinearity in the relationship between Δ_P and the statistics x_i, between A_w and the earthquake statistics, and between L_w and the earthquake statistics. This will suggest whether we should include squared terms in the regression. Then the tree model will be used to indicate whether the interaction structure of the data is complex. If it is complex, we then need to include interaction terms in the analysis. Preliminary analysis also suggested that transformation of the response variable Δ_P, A_w and L_w improved the fitting results. Thus we fit the model

$$z = b_0 + \sum_{i=1}^{5} b_i x_i + \sum_{i=1}^{4}\sum_{j=i+1}^{5} b_{ij} x_i x_j + \sum_{i=1}^{4} b_{0i} x_i^2 + \varepsilon$$

(with squared effects and interactions between the earthquake statistics) to the data. Here z is $\sqrt{\Delta_P}$ in the case of delay analysis, $\log_{10}(A_w)$ and $\log_{10}(L_w)$ for amplitude and length analyses; ε is the residual and the b's are parameters to be estimated.

D. Stationary Distribution of a Markov Chain

Assume $\{S_n\}$ is a Markov chain in the state space $\{1, \cdots, r\}$ with $\mathbf{P} = (p_{ij})_{r\times r}$ and $\mathbf{P}_n = (p_{ij}^{(n)})_{r\times r}$, where

$$p_{ij} = P(S_n = j | S_{n-1} = i)$$

and

$$p_{ij}^{(n)} = P(S_n = j | S_0 = i).$$

According to the Chapman-Kolmogorov equation, we have

$$\mathbf{P}_n = \mathbf{P}_{n-k}\mathbf{P}_k.$$

Let $\mathbf{p}_n = (P(S_n = 1), \cdots, P(S_n = r))$ denote the probability distribution of S_n. Given that

$$\mathbf{p}_n = \mathbf{p}_{n-1}\mathbf{P}, \tag{A-5}$$

it follows that

$$\mathbf{p}_n = \mathbf{p}_0\mathbf{P}^n.$$

By letting $\mathbf{p}_n = \mathbf{p}_{n-1} = \pi$ in Equation (A-5), we have

$$\pi = \pi\mathbf{P}. \tag{A-6}$$

The solution of π to (A-6) is called the stationary distribution of the Markov chain.

Bibliography

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716-723.

[2] Antos, A. and Kontoyiannis, Y. (2001). Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, 19, 163-193.

[3] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37, 1554-1563.

[4] Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41, 164-171.

[5] Bebbington, M. S. (2007). Identifying volcanic regimes using hidden Markov models. *Geophysical Journal International,* 171, 921-942.

[6] Bebbington, M. S. (2008). Estimating rate- and state-fraction parameters using a two-node stochastic model for aftershocks. *Tectonophysics*, 457, 71-85.

[7] Bebbington, M. S. and Harte, D. S. (2001). On the statistics of the linked stress release model. *Journal of Applied Probability,* 38A, 176-187.

[8] Bebbington, M. S. and Harte, D. S. (2003). The linked stress release model for spatiotemporal seismicity: formulations, procedures and applications. *Geophysical Journal International*, 154, 925-946.

[9] Bebbington, M. S., Harte, D. S. and Jaumé, S. C. (2010). Repeated intermittent earthquake cycles in the San Francisco Bay Region. *Pure and Applied Geophysics*, to appear.

[10] Berman, M. (1983). Comment on "Likelihood analysis of point processes and its applications to seismological data," by Ogata, Y., *Bulletin of the International Statistical Institute,* 50, Book 3, 412-418.

[11] Blewitt, G. (1993). Advances in Global Positioning System technology for geodynamics investigations: 1978-1992. In *Contributions of Space Geodesy to Geodynamics: Technology,* (ed. by Smith, D. E. and Turcotte, D. L.), Geodynamics Series Vol. 25, Washington DC, pp. 195-213.

[12] Borovkov, K. and Bebbington, M. S. (2003). A stochastic two-node stress transfer model reproducing Omori's law. *Pure and Applied Geophysics,* 160, 1429-1445.

[13] Bower, D. R. and Heaton, K. C. (1978). Response of an aquifer near Ottawa to tidal forcing and the Alaskan earthquake of 1964. *Canadian Journal of Earth Sciences,* 15(3), 331-340.

[14] Bowman, D. D., Ouillon, O., Sammis, C. G., Sornette, A. and Sornette, D. (1998). An observational test of the critical earthquake concept. *Journal of Geophysical Research,* 103, 24359-24372.

[15] Bowsher, C. G. (2007). Modelling security market events in continuous time: intensity based, multivariate point process models. *Journal of Econometrics,* 141, 876-912.

[16] Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *Annals of Probability,* 24, 1563-1588.

[17] Brillinger, D. R. (1975). The Identification of Point Process Systems. *The Annals of Probability,* 3(6), 909-929.

[18] Brillinger, D. R. (1994). Examples of scientific problems and data analyses in demography, neurophysiology and seismology. *Journal of Computational and Graphical Statistics,* 3, 239-249.

[19] Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory,* SIAM, Philadelphia, USA.

[20] Brillinger, D. R. (2002). Second-order Moments and Mutual Information in the Analysis of Time Series. In *Recent Advances in Statistical Methods,* Imperial College Press, London, pp. 64-76.

[21] Brillinger, D. R. (2003). Modelling and analysis of some random process data from neurophysiology. *Investigacion Operacional,* 24, 206-218.

[22] Brillinger, D. R. (2004). Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics,* 18, 163-183.

[23] Brillinger, D. R., Bryant, H. L. and Segundo, J. P. (1976). Identification of synaptic interactions. *Biological Cybernetics,* 22, 213-228.

[24] Brillinger D. R. and Guha A. (2007). Mutual information in the frequency domain. *Journal of statistical planning and inference,* 137, 1076-1084.

[25] Brillinger, D. R. and Villa, A. E. P. (1997). Assessing connections in networks of biological neurons. In *The Practice of Data Analysis: Essays in Honor of John W. Tukey,* ed. D. R. Brillinger, L. T. Fernholz and S. Morgenthaler, Princeton Univ. Press, Princeton, pp. 77-92.

[26] Brodsky, E. E., Roeloffs, E. A., Woodcock, D. and Gall, I. (2003). A mechanism for sustained groundwater pressure changes induced by distance earthquakes. *Journal of Geophysical Research,* 108, 2390, doi:1029/2002JB002321.

[27] Bufe, C. G. and Varnes, D. J. (1993). Predictive modeling of the seismic cycle of the greater San Francisco Bay region. *Journal of Geophysical Research,* 98, 9871-9883.

[28] Bufe, C. G., Nishenko, S. P. and Varnes, D. J. (1994). Seismicity trends and potential for large earthquakes in the Alaska-Aleutian region. *Pure and Applied Geophysics,* 142, 83-99.

[29] Chia, Y., Chiu, J. J., Chiang, Y-H, Lee, T-P, Wu, Y-M and Horng, M-J (2008). Implications of coseismic groundwater level changes observed at multiple-well monitoring stations. *Geophysical Journal International,* 172, 293-301.

[30] Christensen, R. (1997). *Log-linear Models for Logistic Regression,* Springer, New York.

[31] Coble, R. (1965). The effects of the Alaskan earthquake of March 27, 1964, on ground water in Iowa. *Proceedings of Iowa Academy of Sciences,* 72, 323-332.

[32] Collett, D. (1991). *Modelling Binary Data,* Chapman & Hall, London.

[33] Cooper, Jr. H. H., Bredehoeft, J. D., Papadopulos, I. S. and Bennett, R. R. (1965). The response of well-aquifer systems to seismic waves. *Journal of Geophysical Research,* 70, 3915-3926.

[34] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory,* New York: Wiley.

[35] Cox, D. R. and Lewis P. A. W. (1966). *The Statistical Analysis of Series of Events,* Methuen & Co. Ltd, London.

[36] Daley, D. J. and Vere-Jones, D. (2003). *Introduction to the Theory of Point Processes*, 2nd ed. Springer, New York.

[37] DeMets, C., Gordon, R. G., Argus, D. F. and Stein, S. (1994). Effect of recent revisions to the geomagnetic reversal time scale on estimates of current plate motions. *Geophysical Research Letters*, 21, 2191-2194.

[38] DeMets, C., Gordon, R. G., Stein, S. and Argus, D. F. (1987). A revised estimate of Pacific-North America motion and implications for western North America plate boundary zone tectonics. *Geophysical Research Letters*, 14, 911-914.

[39] Dempster, A. P., Laird, N. M. and Bubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussions). *Journal of Royal Statistical Society, Series B,* 39, 1-38.

[40] Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ.

[41] Doan, M. L. and Cornet, F. H. (2007). Small pressure drop triggered near a fault by small teleseismic waves. *Earth and Planetary Science Letters,* 258, 207-218.

[42] Eaton, J. P and Takasaki, K. J. (1959). Seismological interpretation of earthquake-induced water-level fluctuations in wells. *Bulletin of the Seismological Society of America,* 49, 227-245.

[43] EERI Committee on the Anticipated Tokai Earthquake, Scawthorn C, ed. (1984). *Anticipated Tokai Earthquake: Japanese Prediction and Preparedness Activities.* Oakland, CA: EERI. 89 pp.

[44] Elkhoury, J. E., Brodsky, E. E. and Agnew, D. C. (2006). Seismic waves increase permeability. *Nature,* 441, 1135-1138.

[45] Fedotov, S. A. (1968). The seismic cycle, quantitative seismic zoning, and long-term seismic forecasting. In: *Seismic Zoning in the USSR,* (ed. Medvedev, S. V.) (Izdatel'stvo Nauka, Moscow 1968) pp. 133-166.

[46] Felzer, K. R., Becker, T. W., Abercrombie, R. E., Ekström, G. and Rice, J. R. (2002). Triggering of the 1999 M_w 7.1 Hector Mine earthquake by aftershocks of the 1992 M_w 7.3 Landers earthquake. *Journal of Geophysical Research,* 107, 2190, doi:10.1029/2001JB000911.

[47] Felzer, K. R., Abercrombie, R. E. and Ekström, G. (2003). Secondary aftershocks and their importance for aftershock forecasting. *Bulletin of the Seismological Society of America,* 93, 1433-1448.

[48] Fischer, W. and Meier-Hellstern, K. S. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation,* 18(2), 149-171.

[49] Fletcher, R. and Powell, M. J. D. (1963). A rapidly convergent method for minimization. *The Computer Journal,* 6, 163-168.

[50] Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE,* 61, 268-278.

[51] Gomberg, J., Reasenberg, P., Bodin, P. and Harris, R. (2001). Earthquake triggering by transient seismic waves following the Landers and Hector Mine, California earthquakes. *Nature,* 411, 462-466.

[52] Granat, R. A. (2003). A method of hidden Markov model optimization for use with geophysical data sets. *Computational Science - ICCS 2003, PT III, Proceedings,* 2659, 892-901.

[53] Granat, R. A. (2006). Detecting regional events via statistical analysis of geodetic networks. *Pure and Applied Geophysics,* 163, 2497-2512.

[54] Granat, R. A. and Donnellan, A. (2002). A hidden Markov model based tool for geophysical data exploration. *Pure and Applied Geophysics,* 159(10), 2271-2283.

[55] Granger, C. W. J. and Lin, J. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *Journal of Time Series Analysis,* 15, 371-384.

[56] Gutenberg, B. and Richter, C. F. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological Society of America,* 34, 185-188.

[57] Hall, R. W. (1991). *Queueing Methods: For Services and Manufacturing*, Prentice-Hall, Englewood Cliffs, New Jersey.

[58] Halliday, D. M., Rosenberg, J. R., Amjad, A. M., Breeze, P., Conway, B. A. and Farmer, S. F. (1995). A framework for the analysis of mixed time series/point process data - Theory and application to the study of physiological tremor, single motor unit discharges and electromyograms. *Progress in Biophysics and molecular Biology*, 64(2/3), 237-278.

[59] Hardebeck, J. L., Felzer, K. R. and Michael, A. J. (2008). Improved tests reveal that the accelerating moment release hypothesis is statistically insignificant. *Journal of Geophysical Research*, 113, B08310, doi:10.1029/2007JB005410.

[60] Harte, D. S. (2005). *Package "HiddenMarkov": Discrete Time Hidden Markov Models*. R statistical program routines. Statistics Research Associates, Wellington. URL: http://cran.at.r-project.org/web/packages/HiddenMarkov.

[61] Harte, D. S., Li, D.-F., Vreede, M. and Vere-Jones, D. (2003). Quantifying the M8 prediction algorithm: reduction to a single critical variable and stability results. *New Zealand Journal of Geology and Geophysics*, 46(1), 141-152.

[62] Harte, D. S., Li, D.-F., Vere-Jones, D., Vreede, M. and Wang, Q. (2007). Quantifying the M8 algorithm: Model, forecast and evaluation. *New Zealand Journal of Geology and Geophysics*, 50(2), 117-130.

[63] Hashimoto, C., Noda, A., Sagiya, T. and Matsu'ura, M. (2009). Interplate seismogenic zones along the Kuril-Japan trench inferred from GPS data inversion. *Nature Geoscience*, 2, 141-144.

[64] Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58, 83-90.

[65] Hawkes, A. G. and Adamopoulos, L. (1973). Cluster models for earthquakes-regional comparisons. *Bulletin of the International Statistical Institute*, 45, 454-461.

[66] Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3), 493-503.

[67] Heffes, H. and Lucantoni, D. (1986). A Markov modulated characterization of packetized voice and data traffic related statistical performance. *IEEE Journal on Selected Areas in Communications,* 4, 856-868.

[68] Helmstetter, A. and Sornette, D. (2002). Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *Journal of Geophysical Research,* 107, doi:10.1029/2001JB001580.

[69] Hill, D. P., Reasenberg, P. A., Michael, A., Arabaz, W., Beroza, G. C., Brune, J. N., Brumbaugh, D., Davis, S., DePolo, D., Ellsworth, W. L., Gomberg, J., Harmsen, S., House, L., Jackson, S. M., Johnston, M., Jones, L., Keller, R., Malone, S., Nava, S., Pechmann, J. C., Sanford, A., Simpson, R. W., Smith, R. S., Stark, M., Stickney, M., Walter, S. and Zollweg, J. (1993). Seismicity in the western United States remotely triggered by the $M7.4$ Landers, California, earthquake of June 28, 1992. *Science*, 260, 1617-1623.

[70] Hill, D. P., Johnston, M. J. S., Langbein, J. O. and Bilham, R. (1995). Response of Long Valley caldera to the $M_w = 7.3$ Landers, California, earthquake. *Journal of Geophysical Research*, 100, 12985-13005.

[71] Hughes, J. P. (1993). *A class of stochastic models for relating synoptic atmospheric patterns to local hydrologic phenomena*. PhD thesis, University of Washington, Seattle.

[72] Hughes, J. P. and Guttorp, P. (1994a). A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resources Research*, 30, 1535-1546.

[73] Hughes, J. P. and Guttorp, P. (1994b). Incorporating spatial dependence and atmospheric data in a model of precipitation. *Journal of Applied Meteorology,* 33, 1503-1515.

[74] Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika,* 76, 297-307.

[75] Igarashi, G. and Wakita, H. (1991). Tidal responses and earthquake-related changes in the water level of deep wells. *Journal of Geophysical Research,* 96, 4269-4278.

[76] Igarashi, G., Wakita, H. and Sato, T. (1992). Precursory and coseismic anomalies in well water levels observed for the February 2, 1992 Tokyo Bay Earthquake. *Geophysical Research Letters*, 19, 1583-1586.

[77] Ishimoto, M. and Iida, K. (1939). Observations of earthquakes registered with the microseismograph constructed recently. *Bulletin of the Earthquake Research Institute,* 17, 443-478.

[78] Iyengar, S. (2001). The analysis of multiple neural spike trains. In *Advances in Methodological and Applied Aspects of Probability and Statistics,* ed. N. Balakrishnan, Gordon and Breach, pp. 507-524.

[79] Jaumé, S. C. and Bebbington, M. S. (2004). Accelerating seismic release from a self-correcting stochastic model. *Journal of Geophysical Research,* 109, B12301.1-B12301.12.

[80] Jaumé, S. C. and Sykes, L. R. (1999). Evolving towards a critical point; a review of accelerating moment/energy release prior to large great earthquakes. *Pure and Applied Geophysics,* 155, 279-306.

[81] Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Statistical Mathematics,* 41, 683-697.

[82] Jowett, J. and Vere-Jones, D. (1972). The prediction of stationary point processes. In *Stochastic Point Processes: Statistical Analysis, Theory, and Applications,* ed. P. A. W. Lewis, Wiley, New York, pp. 405-435.

[83] Kagan, Y. Y. (1997). Are earthquakes predictable? *Geophysical Journal International,* 131, 505-525.

[84] Keilis-Borok, V. I. and Kossobokov, V. G. (1990). Premonitory activation of earthquake flow: algorithm M8. *Physics of the Earth and Planetary Interiors,* 61, 73-83.

[85] Kennett, B. L. N. and Engdahl E. R. (1991). Travel times for global earthquake location and phase identification. *Geophysical Journal International,* 105, 429-465.

[86] King, C.-Y., Azuma, S., Igarashi, G., Ohno, M., Saito, H. and Wakita, H. (1999). Earthquake-related water-level changes at 16 closely clustered wells in Tono, central Japan. *Journal of Geophysical Research,* 104, 13,073-13,082.

[87] King, G. C. P., Stein, R. S. and Lin, J. (1994). Static Stress Changes and the Triggering of Earthquakes. *Bulletin of the Seismological Society of America,* 84, 935-953.

[88] Kitagawa, G. and Gersch, W. (1984). A smoothness priors-state space approach to the modeling of time series with trend and seasonality. *Journal of the American Statistical Association,* 79(386), 378-389.

[89] Kitagawa, G. and Matsumoto, N. (1996). Detection of coseismic changes of underground water level. *Journal of the American Statistical Association,* 91(434), 521-528.

[90] Kitagawa, Y., Koizumi, N., Takahashi, M., Matsumoto, N. and Sato, T. (2006). Changes in groundwater levels or pressures associated with the 2004 earthquake off the west coast of northern Sumatra (M9.0). *Earth Planets Space,* 58, 173-179.

[91] Kleijnen, J. P. C. (1979). Regression metamodels for generalizing simulation results. *IEEE Transactions on System, Man, and Cybernetics,* SMC-9, 93-96.

[92] Kunugi, T., Fukao, Y. and Ohno, M. (2000). Underdamped responses of a well to nearby swarm earthquakes off the coast of Ito City, central Japan, 1995. *Journal of Geophysical Research,* 105, 7805-7818.

[93] Liu, J., Vere-Jones, D., Ma, L., Shi, Y. and Zhuang, J. (1998). The principle of coupled stress release model and its applications. *Acta Seismologica Sinica,* 11, 273-281.

[94] Liu, L.-B., Roeloffs, E. and Zheng, X.-Y.. (1989). Seismically induced water level fluctuations in the Wali Well, Beijing, China. *Journal of Geophysical Research,* 94, 9453-9462.

[95] Lombardi, A. M., Marzocchi, W. and Selva J. (2006). Exploring the evolution of a volcanic seismic swarm: The case of the 2000 Izu Islands swarm. *Geophysical Research Letters,* 33, L07310, doi:10.1029/2005GL025157.

[96] Lu, C., Harte, D. S. and Bebbington, M. S. (1999). A linked stress release model for historical Japanese earthquakes: coupling among major seismic regions. *Earth, Planets and Space,* 51(9), 907-916.

[97] Lv, S. (2009). *Extensions of Markov modulated Poisson processes and their applications to deep earthquakes.* Unpublished Ph.D. thesis, Victoria University of Wellington, New Zealand.

[98] Ma, L. and Vere-Jones, D. (1997). Application of M8 and Lin-Lin algorithms to New Zealand earthquake data. *New Zealand Journal of Geology and Geophysics,* 40, 77-89.

[99] MacDonald, I. and Zucchini, W. (1997). *Hidden-Markov and Other Models for Discrete-valued Time Series*, Chapman & Hall, New York.

[100] Mackay, R. J. (2002). Estimating the order of a hidden Markov model. *Canadian Journal of Statistics*, 30, 573-589.

[101] Mars, N. J. I. and van Aragon, G. W. (1982). Time delay estimation in non-linear systems using average amount of mutual information analysis. *Signal Process*, 4, 139-153.

[102] Marsan, D. (2003). Triggering of Seismicity at Short Timescales Following Californian Earthquakes. *Journal of Geophysical Research*, 108, 2266, doi:10.1029/2002JB001946.

[103] Marsan, D. and Nalbant, S. S. (2005). Methods for measuring seismicity rate changes: A review and a study of how the M_w 7.3 Landers earthquake affected the aftershock sequence of the M_w 6.1 Joshua Tree earthquake. *Pure and Applied Geophysics*, 162, 1151-1185.

[104] Matsumoto, N. (1992). Regression-analysis for anomalous changes of ground-water level due to earthquakes. *Geophysical Research Letters*, 19, 1193-1196.

[105] Matsumoto, N., Kitagawa, G. and Roeloffs, E. A. (2003). Hydraulic response to earthquakes in the Haibara well, central Japan. I. Groundwater level changes revealed using state space decomposition of atmospheric pressure, rainfall and tidal responses. *Geophysical Journal International*, 155, 885-898.

[106] Matsu'ura, M., Noda, A. and Fukahata, Y. (2007). Geodetic data inversion based on Bayesian formulation with direct and indirect prior information. *Geophysical Journal International*, 171, 1342-1351.

[107] McCloskey, J., Nalbant, S. S., Steacy, S., Nostro, C., Scotti, O. and Baumont, D. (2003). Structural constraints on the spatial distribution of aftershocks. *Geophysical Research Letters*, 30, 1610, doi:10.1029/2003GL017225.

[108] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Second Edition. Chapman & Hall/CRC.

[109] McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York, Wiley.

[110] Moddemeijer, R. (1989). On estimation of entropy and mutual information of continuous distributions. *Signal Processing*, 16, 233-248.

[111] Mogi, K. (1968). Source locations of elastic shocks in the fracturing process in rocks (1). *Bulletin of Earthquake Research Institute, University of Tokyo,* 46, 1103-1125.

[112] Montgomery, D. R. and Manga, M. (2003). Streamflow and water well responses to earthquakes. *Science,* 300, 2047-2049.

[113] Moon, Y. I., Rajagopalan, B. and Lall, U. (1995). Estimation of mutual information using kernel density estimators. *Physical Review E,* 52, 2318-2321.

[114] Nikolaidis, R. (2002) *Observation of geodetic and seismic deformation with the Global Positioning System.* PhD thesis, University of California, San Diego.

[115] Oakes, D. (1975). The Markovian self-exciting process. *Journal of Applied Probability,* 12(1), 69-77.

[116] Ogata, Y. (1983). Likelihood analysis of point processes and its applications to seismological data. *Bulletin of the International Statistical Institute,* 50, Book 2, 943-961.

[117] Ogata, Y. (1987). Long term dependence of earthquake occurrences and statistical models for standard seismic activity (in Japanese). In *Suri Zisin Gaku (Mathematical Seismology) II* (ed. Saito, M.), ISM Cooperative Research Report 3, pp. 115-124.

[118] Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association,* 83(401), 9-27.

[119] Ogata, Y. (1992). Detection of precursory relative quiescence before great earthquakes through a statistical model. *Journal of Geophysical Research,* 97, 19845-19871.

[120] Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics,* 50, 379-402.

[121] Ogata, Y. (1999). Seismicity analysis through point-process modeling: A review. *Pure and Applied Geophysics,* 155, 471-507.

[122] Ogata, Y. (2007). Seismicity and geodetic anomalies in a wide area preceding the Niigata-Ken-Chuetsu earthquake of 23 October 2004, central Japan. *Journal of Geophysical Research,* 112, B10301, doi:10.1029/2006JB004697.

[123] Ogata, Y., Akaike, H. and Katsura, K. (1982). The application of linear intensity models to the investigation of causal relations between a point process and another stochastic process. *Annals of the Institute of Statistical Mathematics*, 34(B), 373-387.

[124] Ogata, Y., Jones, L. M. and Toda, S. (2003). When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California. *Journal of Geophysical Research,*108(B6), 2318, doi:10.1029/2002JB002009 (ESE1-12).

[125] Ohno, M., Wakita, H. and Kanjo, K. (1997). A water well sensitive to seismic waves. *Geophysical Research Letters*, 24, 691694.

[126] Pievatolo A. and Rotondi, R. (2008). Statistical identification of seismic phases. *Geophysical Journal International,* 173, 942-957.

[127] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 257-286.

[128] Ramana, D. V., Chadha, R. K., Singh, C. and Shekar, M. (2007). Water level fluctuations due to earthquakes in Koyna-Warna region, India. *Natural hazards,* 40, 585-592.

[129] Reasenberg, P. A. and Jones, L. M. (1989). Earthquake hazard after a mainshock in California. *Science,* 243, 1173-1176.

[130] Reyners, M. (2009). Stress and strain from earthquakes at the southern termination of the Taupo Volcanic Zone, New Zealand. *Journal of Volcanology and Geothermal Research*, in press.

[131] Reyners, M., Eberhart-Phillips, D., Stuart, G. and Nishimura, Y. (2006). Imaging subduction from the trench to 300 km depth beneath the central North Island, New Zealand, with Vp and Vp/Vs. *Geophysical Journal International*, 165, 565-583.

[132] Richter, C. F. (1958). *Elementary Seismology,* W. H. Freeman, New York.

[133] Roberts, W. J. J., Ephraim, Y. and Dieguez, E. (2006). On Rydén's EM Algorithm for Estimating MMPPs. *IEEE Signal Processing Letters*, 13(6), 373-376.

[134] Roeloffs, E. (1998). Persistent water level changes in a well near Parkfield, California, due to local and distant earthquakes. *Journal of Geophysical Research,* 103, 869-889.

[135] Roeloffs, E. A. (2006). Evidence for aseismic deformation rate changes prior to earthquakes. *Annual Review of Earth and Planetary Sciences,* 34, 591-627.

[136] Rojstaczer, S., Wolf, S. and Michel, R. (1995). Permeability enhancement in the shallow crust as a cause of earthquake-induced hydrological changes. *Nature,* 373, 237-239.

[137] Rydén, T. (1994). Parameter estimation for Markov modulated Poisson processes. *Communications in Statistics–Stochastic Models,* 10(4), 795-829.

[138] Rydén, T. (1996). An EM algorithm for estimation in Markov-modulated Poisson processes. *Computational Statistics & Data Analysis*, 21, 431-447.

[139] Sammis, C. G. and Smith, S. W. (1999). Seismic cycles and the evolution of stress correlation in cellular automaton models of finite fault networks. *Pure and Applied Geophysics,* 155, 307-334.

[140] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6, 461-464.

[141] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, pp. 379-423 (July 1948) and pp. 623-656 (Oct 1948). URL: http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html

[142] Stein, S. and Wysession, M. (2003). *An Introduction to Seismology, earthquakes, and earth structure,* Blackwell Publishing Ltd., UK.

[143] Strong, S. P., Koberle, R., Steveninck, R. R. de Ruyter van and Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters,* 80, 197-200.

[144] Utsu, T. (1961). A statistical study on the occurrence of aftershocks. *Geophysical Magazine,* 30, 521-605.

[145] Van Loan, C. F. (1978). Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control*, AC-23(3), 395-404.

[146] Vere-Jones, D. (1970). Stochastic models for earthquake occurrence. *Journal of Royal Statistical Society, Series B,* 32, 1-62.

[147] Vere-Jones, D. (1978). Earthquake prediction – a statistician's view. *Journal of Physics of the Earth*, 25, 129-146.

[148] Vere-Jones, D. and Davies, R. B. (1966). A statistical survey of earthquakes in the main seismic region of New Zealand: Part 2 — time series analysis. *New Zealand Journal of Geology and Geophysics*, 9, 251-284.

[149] Vere-Jones, D. and Ozaki, T. (1982). Some examples of statistical estimation applied to earthquake data. *Annals of the Institute of Statistical Mathematics*, 34(B), 189-207.

[150] Vere-Jones, D., Robinson, R. and Yang, W. (2001). Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation. *Geophysical Journal International*, 144, 517-531.

[151] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13, 260-269.

[152] Wallace, L. M., Beavan, J., McCaffrey, R. and Darby, D. (2004). Subduction zone coupling and tectonic block rotations in the North Island, New Zealand. *Journal of Geophysical Research*, 109, B12406. doi:10.1029/2004JB003241.

[153] Wang, T., Bebbington, M. S. and Harte, D. S. (2010). A comparative study of coherence, mutual information and cross-intensity models. *International Journal of Information & Systems Sciences*, 6(1), 49-60.

[154] Wang, C.-Y., Wang, C.-H. and Kuo, C.-H. (2004). Temporal change in groundwater level following the 1999 (M_W = 7.5) Chi-Chi earthquake, Taiwan. *Geofluids*, 4, 210-220.

[155] Wilson, C. J. N., Houghton, B. F., McWilliams, M. O., Lanphere, M. A., Weaver, S. D. and Briggs, R. M. (1995). Volcanic and structural evolution of Taupo Volcanic Zone, New Zealand: a review. *Journal of Volcanology and Geothermal Research*, 68, 1-28.

[156] Wyss, M., and Wiemer, S. (2000). Change in the probabilities for earthquakes in Southern California due to Landers M7.3 earthquake. *Science*, 290, 1334-1338.

[157] Zhang, Z., Zhang, S., Zheng, Y. and Wang, H. (2002). Analysis on the ability of groundwater level for recording seismic wave in Shanxi Water 2 and Yue 42 wells, Tangshan, China, (in Chinese). *Northwestern Seismological Journal*, 24(3), 262-266.

[158] Zheng, X. and Vere-Jones, D. (1994). Further applications of stress release models to historical earthquake data. *Tectonophysics*, 229, 101-121.

[159] Zhuang, J. (2000). Statistical modeling of seismicity patterns before and after the 1990 Oct 5 Cape Palliser earthquake, New Zealand. *New Zealand Journal of Geology and Geophysics*, 43, 447-460.

[160] Zucchini, W. and Guttorp, P. (1991). A Hidden Markov Model for Space-Time Precipitation. *Water Resources Research*, 27(8), 1917-1923.