THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy, School of Geosciences

Modelling seismicity as a spatio-temporal point process using inlabru

by

Francesco Serafini

2023

Declaration

This project report is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, School of Geosciences. I declare that this thesis was composed by myself, that the work contained therein is my own, except where explicitly stated otherwise in the text, and that it has not been submitted, in whole or in part, for any other degree or professional qualification.

Francesco Serafini

Word Count: 85000 circa

This thesis was conducted under the supervision of Dr Mark Naylor, Prof Finn Lindgren, Prof Ian Main.

Abstract

Reliable deterministic prediction of earthquake occurrence is not possible at present, and may never be. In the absence of a reliable deterministic model, we need alternate strategies to manage the seismic hazard or the risk. This involves making statements of the likelihood or earthquake occurrence in space and time, including a fair and accurate description of the uncertainty around statements used in operational decision-making. *Probabilistic Seismic Hazard Analysis* (PSHA) and *Operational Earthquake Forecasting* (OEF) have the role of providing probabilistic statements on the hazard associated with earthquakes on long-term (decades to centuries) and short-term (days to decades) time frames respectively. Both PSHA and OEF rely on a source model able to describe the occurrence of earthquakes.

PSHA models are commonly modelled using a spatially-variable Poisson process to describe earthquake occurrence. Therefore, they are calibrated on declustered catalogues which retains only the largest earthquakes in a sequence. OEF models, on the other hand, are commonly time-dependent models which describes the occurrence of all the events above a certain magnitude threshold including dependent events such as afetrshocks or swarms. They are calibrated on the full earthquake catalogue and provide accurate descriptions of the clustering process and the time-evolution of earthquake sequences. The Epidemic-Type Aftershock Sequence (ETAS) model is the most commonly used model as time-dependent seismicity model and belongs to the general class of Hawkes (or *self-exciting*) processes. Under the ETAS model, any earthquake in the sequence has the ability of inducing (or triggering) its own subsequence of earthquakes in a cascade of events, as commonly observed in nature. The earthquake catalogue is then the union of a set of events occurring independently from each other (background events) and a set of events which have been induced or triggered by another (aftershocks).

The reliability of PSHA or OEF strategies depends upon the reliability of the source model used to describe earthquake occurrence. In order to improve the source model, we need the ability to (a) incorporate hypotheses on earthquake occurrence in a model, and (b) validate the model against observed data. Both tasks are problematic. Indeed, the complex mathematical form of the ETAS model requires ad-hoc methodologies to perform inference on the model parameters. These methodologies then need further modification if the classical ETAS model is adjusted to introduce new hypotheses. Comparing forecasts produced by models incorporating different hypotheses which are and calibrated with different methods is problematic because it is difficult (if not impossible) to determine where the differences in the forecasts are coming from. Therefore, a unique framework capable of supporting ETAS models incorporating different hypotheses would be beneficial. Similarly, the validation step has to be done on models calibrated on the same data and producing forecasts for the same spatio-temporal region. Moreover the validation must ultimately be done against future data, unknown in the moment in which the forecasts are produced, to ensure that no information about the data used to validate the models is incorporated in the models themselves. Hence, the Collaboratory for the Study of Earthquake Predictability (CSEP) has been founded with the role of gathering forecasting models and running fully-prospective forecasting experiments in an open environment. CSEP ensures that the models are validated fairly and using a set of community-agreed metrics which measure the agreement between forecasts and data on the outcomes.

In this thesis, I present and apply a new Bayesian approximation technique for Hawkes process models (including ETAS). I also demonstrate the importance of one of the statistical properties that scores used to rank competing forecasts need to have in order to provide trustworthy results. The Bayesian framework allows an accurate description of the uncertainty around model parameters which can then be propagated to any quantity of interest. In the context of Bayesian statistics, the most commonly used techniques to perform inference are Markov Chain Monte Carlo (MCMC) techniques which are sampling-based methods. Instead, I use the Integrated Nested Laplace Approximation (INLA) to provide a deterministic approximation of the parameter posterior distribution instead of the random sampling. INLA is faster than MCMC for problems involving a large number of correlated parameters and offers an alternative way to implement complex statistical models which are infeasible (from a computational point of view) with MCMC. This provides researchers and practitioners with a statistical framework to formulate ETAS models incorporating different hypotheses, produce forecasts that accounts for uncertainty, and test them using CSEP procedures. I build on the work done to implement time-independent models for seismicity with INLA which provided a framework to study the effect of covariates such as depth, GPS displacement, heatflow, strain rate, and distance to the nearest fault but lacked the ability to describe the clustering process of earthquakes. I show that this work can be extended to include time-dependent Hawkes process models and run in a reasonable computational time using INLA. In this framework, the information from covariates can be incorporated both in modelling the rate of background events, and in modelling the number aftershocks. This resembles how information on covariates is incorporated in Generalized Linear Models (GLMs) which are widely used to study the effect of covariates on a range of phenomena. Indeed, this work offers a way to borrow ideas and techniques used with GLMs and apply them to seismicity analyses. To make the proposed technique widely accessible, I have developed a new R-package called `ETAS.inlabru` which offers user-friendly access to the proposed methodology. The `ETAS.inlabru` package is based on the `inlabru` R-package which offers access to the INLA methodology. In this thesis, I compared our approach with the MCMC technique implemented through the `bayesianETAS` package and shows that `ETAS.inlabru` provides similar results to `bayesianETAS`, but it is faster, scales more efficiently increasing the amount of data, and can support a wider range of ETAS models, specifically those involving multiple covariates. I believe that this work provides users with a reliable Bayesian framework for the ETAS model alleviating the burden of modifying/coding their own optimization routines and allowing more flexibility in the range of hypotheses that can be incorporated and validated. In this thesis, I have analysed the 2009 L'Aquila and 2016 Amatrice seismic sequences occurred in central Italy and found that the depth of the events have a negative effect on the aftershock productivity, and that models involving covariates show a better fit to the data than the classical ETAS model.

On the statistical properties that scores needs to posses to provide trustworthy rankings of competing forecasts, I focus on the notion of *proper* scores. I show that the Parimutuel Gambling (PG) score, used to rank forecasts in previous CSEP experiments, has been used in situations in which is not proper. Indeed, I demonstrate that the PG score is proper only in a specific situation and improper in general. I compare its performances with two proper alternatives: the Brier and the Logarithmic (Log) scores. The simulation procedure employed for this part of the thesis can be easily adapted to study the properties of other validation procedures as the ones used in CSEP or to determine important quantities for the experimental design such as the amount of data with which the comparison should be

performed. This contributes to the wider discussion on the statistical properties of CSEP tests, and is an additional step in determining *sanity-checks* that scoring rules have to pass before being used to validate earthquake forecasts in CSEP experiments.

Lay Summary

Neither national institutions nor any scientists have been able to provide a method to reliably predict major earthquakes. As scientists, we can only provide the probability that an earthquake above a certain magnitude will happen in a specified spatio-temporal window. This is the role of Probabilistic seismic hazard analysis (PSHA) and Operational Earthquake Forecasting (OEF) activities that have the scope to provide probabilistic statements on future seismicity to inform institutions and civil populations on the risk associated with earthquake occurrence on a long (decades to centuries) and short term (days to decades) time frames respectively. Our inability to predict major earthquakes is mainly due to the limited amount of information that we have on earthquakes, the large scale of the geological processes involved, and the inability to reproduce them in laboratory experiments. Indeed, earthquakes are generated by the interaction between tectonic plates. As they move, stress is accumulated along the faults for periods of time ranging from decades to centuries, and when the stress level exceeds a certain (unknown) critical value we have an earthquake. The amount of energy released determines the size of the earthquake and, possibly, triggers more earthquakes. Therefore, relevant quantities to predict earthquakes may be the stress accumulated along the faults, the direction in which the plates are moving, the material of which the plates are made at their boundaries, the heat flow from Earth's interior to the surface, and potentially many more. Most of these quantities are not monitored on a continuous basis and only in a few places around the globe. Indeed, many models do not use this kind of information, and questions on which aspects of the earthquake generation process are explained by these quantities are still open. Furthermore, a robust statistical framework to study the effect of available information on earthquake occurrence, as it is done in other fields ranging from ecology to psychology, is still missing.

In absence of a reliable physical model able to describe observed seismicity, scientists have had to resort to statistical models describing the frequency of earthquakes occurrence in time, space, and magnitude. Their findings are summarised by well-established empirical laws describing these frequencies such as Omori's law, the Utsu-Seki law, and the Gutenberg–Richter law. The Epidemic-Type Aftershocks Sequence (ETAS) model is the most used point process model used in statistical seismology. The main characteristic that made ETAS successful over the years is its ability to incorporate the aforementioned empirical laws in a unified framework. Another advantage is that ETAS allows us to directly model the short-term clustering of seismicity by modeling the ability of each earthquake to induce additional ones. Specifically, it models the number and spatio-temporal distribution of the aftershocks induced by each event. In the case of the classical ETAS model, the only information used to predict the number of induced events and how this number decays with time and space is the magnitude and location of the event. However, as time passes and technology develops, a welth of new information is available with great potential of improving our ability to predict earthquakes. For example, faults maps are more detailed, GPS measures allow us to calculate the displacements associated with seismic events, updated heatflow and stress rate maps could be incorporated as well as information on the earth's focal mechanism. In this

context, it is crucial to have a reliable statistical framework to formulate and test hypotheses on how this additional information can be incorporated in the model. This involves being able to formulate different hypotheses, build models incorporating them, and fairly compare the models against observed data to determine which hypotheses are most useful to predict future seismicity.

My Ph.D. project can be divided into two parts: a modelling part and a testing part. Regarding the modelling part, I develop a framework to perform Bayesian inference on the ETAS model. Being Bayesian is foundamental for us because it is the only statistical framework that allows us to explicitly quantify the uncertainty around the model parameters. This is crucial, especially in an OEF context, because it allows measuring the level of trust we should have around the information provided by the model on future seismicity. Not considering the uncertainty explicitly may lead to overconfident forecasts and potentially underestimate the seismic risk, which, in turn, may lead to catastrophic consequences. Our approximation method is based on the Integrated Nested Laplace Approximation (INLA) and constructed on the R-package `inlabru`. INLA is a method to perform Bayesian inference on complex models which are usually not practicable with alternative methods due to the huge computational time required. The `inlabru` R-package facilitates the use of INLA for spatial models and generalizes the INLA method to even more complicated models. Both of them have been largely used to study ecological processes, especially in presence of multiple covariates. My project aims to bring the experience maturated by this community in studying how available information helps in understanding and forecasting complex ecological processes to the seismological community. To make our approach as accessible as possible, we made an R-package `ETAS.inlabru` to allow users to use our ETAS model implementation with minimum coding efforts. This provides an accessible framework to researchers to develop spatio-temporal seismicity models incorporating available information with the potential of improving our ability to predict earthquakes. We provide applications of our methodology on the temporal and spatio-temporal ETAS models to seismic sequences in Italy.

Regarding the testing part of my project, I focus on one statistical property that scoring rules used to rank competing forecasts must have. I consider only positively oriented scoring rules that are functions of the forecast and the observed data returning a single value, the highest the value the better the forecasts. In this way, having a set of competing forecasts and the observed data we can rank them and determine which forecast is *better*. Different scoring rules penalize the competing forecasts differently and the obtained rankings may vary. However, they should always advantage, on average, of the data-generating model. This property is known as *propriety* or *properness* and a score that has this property is said *proper*. If a scoring rule is not proper means that it can be biased and favours models that under/overpredict the target quantity. I prove that the Parimutuel Gambling (PG) score, which has been used to compare forecasts in recent papers, is indeed proper only in a very specific situation and not in general. I explore the consequences of using an improper scoring rule comparing the rankings obtained with the PG score with two proper alternatives. Finally, I show that we can easily find situations in which the PG score assigns the highest score to forecasts underestimating the probability of observing earthquakes above a certain magnitude threshold, and how this does not happen using proper scoring rules. This work shows how scores that seem proper at first glance may be not if analysed more deeply and shows different techniques to check for propriety using simulated data.

Dedication

First and foremost, I want to thank my supervisory team composed of Mark Naylor, Finn Lindgren, and Ian Main who made this work possible.

Thank you Mark for your unwavering support, guidance, and enthusiasm throughout this journey. I am thankful for the way you have taken the time to get to know me on a personal level before establishing a working relationship. Your kindness, empathy, and humor have made these years not only productive but also enjoyable.

Thank you Finn, your wealth of knowledge and expertise in the field has been an invaluable resource to me, and I am grateful for your willingness to share your insights and expertise with me. Despite your busy schedule, you have always made time to provide me with guidance, feedback, and suggestions, which have been instrumental in shaping the direction of this project.

Thank you Ian for all the advice and suggestions that brought a broader perspective and vision to the project. Thank you also for the efforts in building a friendly seismological community and (of course) the music.

A special thanks to Kirsty Bayliss who guided me in my first steps in seismology, your willingness to share your knowledge and expertise, and your unwavering support and encouragement, have been incredibly helpful to me. You have been a constant source of motivation and positivity.

Thanks to the whole CSEP community who warmly welcomed me and for all the efforts you made daily in advancing this field of science.

Last but not least, I dedicate this thesis to my wife, Susanna. You put up with my long hours, and weekend work, always providing me with the support and encouragement that I needed to keep going. Your love and support have given me the strength to overcome the challenges and obstacles that I faced along the way.

Contents

List of Figures

List of Tables

Abbreviations

AIC Akaike Information Criterion.

CSEP Collaboratory for the Study for Earthquake Predictability.

ETAS Epidemic-Type Aftershock Sequence.

GMRF Gaussian Markov Random Field.

GR-law Gutenberg-Richter law.

INLA Integrated Nested Laplace Approximation.

LGCP Log-Gaussian Cox Process.

MCMC Markov Chain Monte Carlo.

OEF Operational Earthquake Forecasting.

PG score Parimutuel Gambling score.

PSHA Probabilistic Seismic Hazard Analysis.

SPDE Stochastic Partial Differential Equantion.

Note: Author abbreviations are shown in their corresponding reference entry.

Nomenclature

Term	Description	Units
$\Lambda(\cdot)$	Integrated intensity	—
$\bar{f}(\cdot)$	Linear approximation of the function $f(\cdot)$	—
$\boldsymbol{\theta}$	Model parameters	—
$\eta(\cdot)$	Linear predictor	—
$\lambda(\cdot\|\mathcal{H})$	Conditional intensity	—
$\mathbb{E}(\cdot)$	Expected value	—
$\mathcal{H}$	History of the process	—
$\mathcal{L}$	Log-likelihood function	—
$\pi(\cdot)$	Probability distribution	—

Chapter 1

Introduction

1.1 The importance of the source model

Large earthquakes have a devastating impact on our society. During the last century, it has been estimated that more than 8.5 million people have died and more than 2 trillion dollars in damage are due to earthquakes (Daniell et al., 2011). Despite strong efforts through the last century, the aspiration of a model capable of predicting earthquakes, within a narrow time, space magnitude window, has proven elusive, and may well be impossible [1]. Nevertheless, we need short and long-term strategies to mitigate this risk. We need long-term strategies (decades to centuries) to strategically determine where sensitive infrastructures should be placed and design the building codes based on the expected maximum level of ground shaking. At the same time, we need short-term strategies (days to decades) to inform people and organizations and guide the recovery efforts *while* a sequence is happening. Given that deterministic predictions that try to provide the exact space-time location of future large events are unreliable, the only other choice is to study at which probability earthquakes occur, and how these probabilities evolve in time and space. In both kinds of analysis, the main interest lies in forecasting how much the ground will shake which, in turn, depends on forecasting the occurrence of earthquakes. The two main branches of earthquake forecasting are Probabilistic Seismic Hazard Analysis (PSHA, Cornell (1968)) which focuses on long-term (decades to centuries) aspects of seismicity, and Operational Earthquake Forecasting (OEF, Jordan and Jones (2010)) which focuses on the short-term (days to decades) aspects of seismicity. Other techniques such as Deterministic hazard assessment exists (Connor et al., 2009) but in this thesis, I will focus on probabilistic methods. Both PSHA and OEF have the goal of providing authoritative probabilistic statements on future earthquake occurrence to inform people, institutions, and organizations involved in operations to mitigate this risk. PSHA has been widely used for decades and is now a fundamental ingredient in the process of designing reliable building codes (Solomos et al., 2008; Hanks et al., 2009) and hazard maps are routinely produced for many countries such as the United States (Frankel et al., 2002), New Zealand (Stirling and Wilson, 2002), Italy (Gruppo di Lavoro, 2004a), for the entire globe (Silva et al., 2020; Pagani et al., 2020). These maps are usually freely available and can be navigated interactively. Examples are the map provided by the European Facilities for Earthquake Hazard and Risk (EFEHR) [2] or the one provided by the Global Earthquake

[1] The United States Geological Survey (USGS) states *"Neither the USGS nor any other scientists have ever predicted a major earthquake. We do not know how, and we do not expect to know how any time in the foreseeable future."* https://www.usgs.gov/faqs/can-you-predict-earthquakes#:~:text=No.,time%20in%20the%20foreseeable%20future.

[2] http://www.efehr.org/earthquake-hazard/hazard-map/

Model (GEM) foundation [3]. On the other hand, OEF activities are still at the beginning (Jordan et al., 2014; Marzocchi et al., 2014).

The standard technique to calculate the probability that the maximum ground-shaking will exceed a certain threshold over a certain period of time was developed by Cornell (1968) in the context of PSHA and is composed of 4 steps (Figure 1.1). The first step is to identify the earthquake sources, essentially where and when the earthquakes are more likely to occur. This is done using a source model able to represent the properties of seismicity in space and time. The second step is to determine the magnitude distribution providing information on how likely it is to observe large earthquakes at a particular site. Information from step 1 and step 2 are combined in step 3 to calculate the expected peak ground acceleration distribution which states how probable it is to observe a certain maximum level of shaking. This is done using ground-motion equations (Douglas, 2003). Step 4 just calculates the probability that the peak ground acceleration will exceed a certain threshold. Varying the threshold yields exceedance probability functions can then be used in what is called Probabilistic Seismic Risk Analysis (PSRA) to calculate expected damages such as structural failures, fatalities, and economic losses (Baker et al., 2021). The main difference between PSHA and OEF is in step 1: the source model used to describe earthquake occurrence. Indeed, in PSHA the interest is on the spatial distribution of earthquakes over extended periods of time (decades to centuries), and the model employed as source models are time-independent models. The focus is on estimating the rate of background events, where the latter indicates events that are assumed to occur spontaneously and not be *triggered* by other events known as *aftershocks* (Chapter 2 gives a more formal definition of background and aftershock events). The time of the events is used to label the events as background or aftershocks using a technique called declustering methods (Gardner and Knopoff, 1974; Reasenberg, 1985; Zhuang et al., 2002), and only the background events are then retained for the analysis. On the other hand, in OEF, the greatest interest is in describing the temporal evolution of earthquake sequences and the spatio-temporal distribution of aftershocks to be able to promptly respond to destructive events. Therefore, OEF analyses make use of the whole earthquake catalogue. For both of them, the reliability of the final results strictly depends on the reliability of the source model which needs to provide results as consistent as possible with observed seismicity patterns. Having an unreliable source model leads to underestimating/overestimating the risk associated with seismic activity. In this thesis, we propose a new framework to build complex source models describing earthquake occurrence and we explore methods to validate them against observed data.

1.2 Evolution of the source model

The source model has the role of describing the evolution of seismic patterns in space and time and needs to incorporate all the knowledge gained over centuries of observations. Ancient Greek philosophers, such as Thales and Aristotle, already noticed that earthquakes do not occur randomly in space and time, and that an earthquake can affect places very far from where it starts (Oeser, 1992). They explained this by supposing the existence of a complex system of underground caves where earthquakes were generated by water or wind and that allowed the effect of an earthquake *to travel* from one place to another. Similar theories were also proposed by Kant (1756) who added that these caverns should run parallel to mountain ranges and large rivers because it was there that the majority of earthquakes happen. Nowadays, we know that the majority of earthquakes are generated by plate tectonic movement near plates boundaries and the resulting seismic release of energy is through slip on

[3] https://www.globalquakemodel.org/gem-maps/global-earthquake-hazard-map

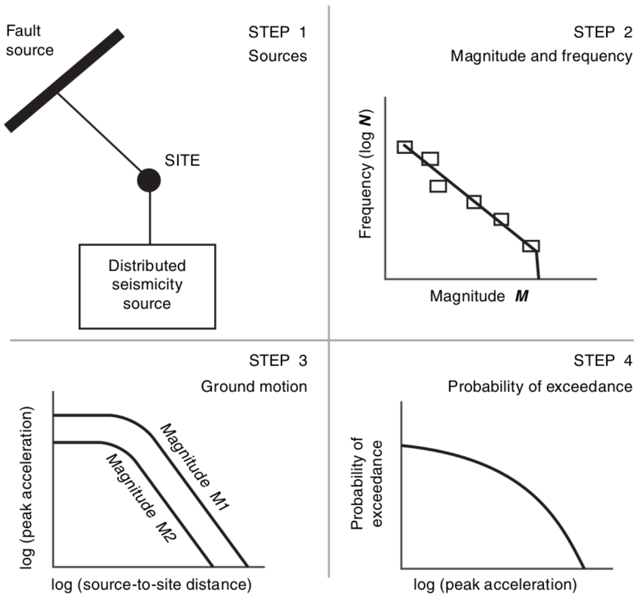


Figure 1.1: The four steps of probabilistic seismic hazard analysis (PSHA). Figure taken from Chapter 10 of Connor et al. (2009).

faults. It is widely accepted that the earth's crust is composed of tectonic plates (Wegener, 1912) and that they move relatively to each other (Morgan, 1968; McKenzie and Parker, 1967). Friction prevents the plates from slipping smoothly, and causes the plates to lock, stress builds up until there is enough to generate a rupture. The rupture creates seismic waves that travel through the earth's crust and may trigger additional earthquakes. Scientists have invented instruments to record these waves (seismometers) and their output (seismograms) is analysed to determine the spatial location, the time, and the energy released (usually quantified by a logarithmic measure such as the magnitude) by an earthquake. Seismic recordings are even used to determine the composition of the Earth's crust and can be applied to study the structure of other planets (Shearer, 2019; Stähler et al., 2021). In this thesis, we used as data the inferred epicenter location, the time, the magnitude, and the depth of the events composing an earthquake catalogue.

Early models describing seismicity were based on the hypotheses that earthquakes occurrence has some sort of periodicity and that large earthquakes occurring on the same fault should have similar characteristics. These are respectively known as the seismic gap hypotheses (Fedotov, 1965; Sykes, 1971) and the characteristics earthquake hypotheses (Schwartz and Coppersmith, 1984) and have been the basis of some PSHA models for many years. These models are not interested in describing the clustering behavior of seismicity but only the occurrence of large earthquakes (mainshocks). For these reason, PSHA still relies on declustering methods which removes from the catalogue all the earthquakes happened before (foreshocks) and after (aftershocks) the mainshocks. Then, they assume that the mainshocks occur independently in time and follow a spatially variable but temporally stationary Poisson process. This approach resents of various problems, first of all the identification of mainshocks is non-trivial, there are many alternative declustering algorithms (Gardner and Knopoff, 1974; Reasenberg, 1985; Zhuang et al., 2002), and this introduces a degree of subjectivity in the analyses. Second, acquiring more data on earthquake occurrence and rigorous testing showed that observed seismicity does not verify the seismic gap hypothesis nor the characteristic earthquake one (Geller et al., 2015; Kagan et al., 2012). This led to the diffusion (and use) of hazard maps that failed to describe observed seismicity and damaging earthquakes that happened in areas considered as low-risk by PSHA maps (Muir-Wood, 1993; Geller, 2011; Stein et al., 2012; Mulargia et al., 2017). This induced part of the community to deem PSHA methods as unreliable and state that should be abandoned (Geller et al., 2015; Stark, 2022). A counterargument is that this models tries to capture departure from the Gutenberg-Richter law (Gutenberg and Richter, 1944) for the magnitude frequency distribution on data regarding individual faults in cases where the largest earthquakes are over-represented (Field et al., 2014). Also, a longer observation period suggests that earthquakes are happening in high and low risk areas in correct proportion (Hanks et al., 2012). From a different point of view, these failures are empirical proof of how important is to properly validate models for seismicity against future data before using them operationally (Marzocchi et al., 2014; Strader et al., 2018).

A more realistic model to describe seismicity patterns is the Epidemic Type Aftershocks Sequence (ETAS) model (Ogata, 1988). The ETAS model assumes that the events can be divided into background events happening spontaneously and independently from each other, and aftershocks that are triggered by a parent event, and in turn can trigger others. In this way, events do not need to be discarded and the ETAS model provides a description of the entire earthquake catalogue. It is considered the state-of-the-art in modern earthquakes forecasting due to its ability to incorporate the most widely accepted (and empirically validated) laws of seismicity such as the Omori's law for aftershock decay (Omori, 1894) and the Gutenberg-Richter law describing the magnitude frequency distribution (Gutenberg and Richter, 1944). The ETAS model is based on the idea that any earthquake is capable of creating its own aftershock sequence, this allows for cascades of events as they are observed in nature, as opposed to the models described above where only the mainshocks have this capability. This makes ETAS particularly suited to quantify the risk due to aftershock activity (Iervolino et al., 2018) and is used or considered in the majority of OEF applications (Marzocchi et al., 2014; Rhoades et al., 2016; van der Elst et al., 2022). Moreover, the supremacy of the ETAS model against competitors has been proved in many prospective forecasting experiments in different tectonic settings (Woessner et al., 2011; Strader et al., 2017; Taroni et al., 2018; Nandan et al., 2019; Savran et al., 2020). For these reasons, ETAS is now considered the state-of-the-art in modern earthquake forecasting. In this thesis, I offer a new way to estimate the parameters of the ETAS model with the ability to quantify the uncertainty relative to these estimates and to extend the classic ETAS formulation to include possible covariates. A detailed description of the ETAS model and the empirical laws of seismicity incorporated in it is provided in Chapter 2.

1.3 Uncertainty

Accounting for uncertainty is fundamental to have a fair representation of our knowledge around future earthquakes; not including uncertainty may lead to under/overestimating the risk associated with earthquakes and, in turn, to make poor decisions (Crowley et al., 2005). In the same way, the act of rejecting or accepting a hypothesis based on observed data is rooted in the concept of uncertainty, and decisions are taken based on how the uncertainty of the result is quantified. This is the case for example of classical hypotheses testing where a hypothesis is rejected based on the probability distribution of appropriate statistics (function of the data) under the hypothesis under analysis. The common ontological framework for probabilistic forecasting models considers two kinds of uncertainty: aleatory and epistemic (Budnitz et al., 1997; Marzocchi and Jordan, 2014; Marzocchi et al., 2015). The aleatory uncertainty stems from the intrinsic randomness of the system itself, for example, the result of tossing a coin is an aleatory event because every time we toss the coin a potentially different result may occur. Indeed, *aleator* is the Latin word for *dice thrower*. Aleatory uncertainty is well represented by a single model with fixed parameters, for instance, statistical models like ETAS are capable of generating many different realistic synthetic catalogues from the same set of parameters. Instead, the epistemic uncertainty stems from our lack of knowledge of the system and our ignorance of the laws governing the phenomenon under study. This includes uncertainty around the value of the parameters of the model and, on a higher level, on the model formulation itself. Aleatory variability is independent of our knowledge of the system and is often characterized as *known unknowns*, while epistemic uncertainty comes from our state of knowledge and is characterized as *unknown unknowns*. Epistemic uncertainty can be reduced by acquiring more information on the process, while aleatory uncertainty can not be reduced within one representation or model (Goldstein, 2013), but could be reduced by more adequate representations.

The Bayesian approach to statistical analysis provides a theoretical framework to quantify the epistemic uncertainty around a model. Indeed, in Bayesian statistics, every parameter is considered as a random variable with its own probability distribution, as opposed to the frequentist approach where is supposed to exist a *true* value of the parameters and uncertainty about the estimates only comes from the aleatory variability of the data used to calibrate the model (Jeffreys, 1998). Moreover, the Bayesian approach offers a way to quantify the reduction in epistemic uncertainty due to the information provided by an experiment. Any Bayesian analysis is composed of three main ingredients: the prior distribution, the likelihood, and the posterior distribution. The prior distribution synthesises the knowledge on the parameters *before* the experiment is run, or before looking at the data. Any knowledge coming from previous experiments and expert opinions can be used to determine the prior distribution. The likelihood describes the aleatory variability of the data, and is a way to formalise the information provided by the observations on the parameters of interest. Frequentist analyses rely on analysing only the likelihood and parameters are estimated as the set of parameters under which the probability of observing the data is maximized. The posterior distribution is a normalised combination of the prior and the likelihood and synthesises the updated level of knowledge on the parameters *after* the experiment. In other words, the prior distribution describes the epistemic uncertainty before the experiment while the posterior distribution after. Comparing the two gives us a measure of how much information we gained running the experiment. In the case of a new experiment, the posterior distribution can be used as prior and the whole process repeated. This offers the possibility of tracking the evolution of our knowledge through multiple experiments.

The epistemic uncertainty, however, does not include only the parameter values but also the model formulation. This includes which hypotheses are incorporated in the model, examples of unresolved questions are, is the aftershock triggering isotropic? If not, which shape is more appropriate? Do the parameters vary with time or space or both? Is there any external factor influencing the earthquake generation process? If yes, on which aspects? and which factors? There is no clear answer to these questions, and the answer may change with the data. This uncertainty led to the formulation of many different alternative models for seismicity. The collection of all the possible outputs from all the viable models gives a representation of the level of epistemic uncertainty around the earthquake-generating process. A standard approach based on this idea is the logic tree which offers a hierarchical framework to handle the problem (Kulkarni et al., 1984; Henley and Kumamoto, 1996; Musson, 2012; Marzocchi et al., 2015). The nodes of the tree represent sources of uncertainty, like hypothesis A versus hypothesis B where the alternatives are mutually exclusive and each one creates its own branch. This is repeated until all the sources of uncertainty are considered. The collection of the end nodes represents the complete epistemic uncertainty and nodes are combined using a probabilistic structure. The logic tree falls under the more general class of ensemble models (Lorenz, 1965; Parker, 2013), and offers a way to combine possibly contradictory hypotheses combining the output of different models. The impact of each model on the final result is expressed by the weight associated with that branch of the tree. The weights are usually inferred from expert judgement or based on a measure of how well the models fit the data. However, also in this case, the Bayesian approach can be used to quantify the uncertainty around the logic tree weights (Kwag and Gupta, 2017).

The *only* difficulty of applying the Bayesian approach is actually retrieving the posterior distribution which often involves integrals with no closed form solution. This task is particularly challenging when coming to the ETAS model which has a mathematically complex likelihood function and parameters are correlated with each other (meaning that different combinations of parameters explain equally well the observed data). The technique that has become standard in the last decades to perform bayesian inference is called Markov Chain Monte Carlo (MCMC, Robert et al. (1999)) and relies on the ability to sample from the posterior distribution of the parameters without having explicitly calculated it. Then, the posterior distribution is empirically estimated using the obtained samples. This means that to have a trustworthy representation of the posterior distribution over the whole domain we need to sample extensively from the posterior and we do not have to specify the uncertainty structure a priori. This makes MCMC methods notoriously slow compared to frequentist alternatives, and the problem is exacerbated by the number of parameters and the correlation between them. This is because the interaction between parameters may lead to part of the domain being undersampled which, in turn, increases the number of samples needed for a decent representation of all parameters. But MCMC methods are not the only option. In this work, we develop a new approximation method for the Bayesian ETAS model based on the Integrated Nested Laplace Approximation (INLA, Rue et al. (2009, 2017)) and implemented through the package `inlabru` (Bachl et al., 2019). INLA is an alternative method to MCMC for bayesian inference of Latent Gaussian models, and `inlabru` is an R-package providing user-friendly access to this methodology. Chapter 3 provides an introduction to bayesian inference, latent Gaussian models, in there I explain how the INLA methodology works, and it is extended by `inlabru`.

1.4 Validation of the source model

The key to advance our knowledge on the earthquake-generating process, in order to make *better* forecasts, is to be able to falsify hypotheses based on the evidence provided by observed data. Validating models (and hypotheses there-in) against observed data is a formidably challenging task when it regards natural phenomena like earthquakes. Indeed, in fields such as seismology or astronomy, data can not be acquired from experiments but it must come from monitoring naturally occurring events over long periods of time. This makes verifying hypotheses against observed data a long-term enterprise that could last decades. A typical example is the time required to validate the hypothesis that earthquakes are periodic or quasi-period events. The idea was first proposed by Gilbert (1884) more than 100 years ago. Schwartz and Coppersmith (1984) introduced the characteristic earthquake hypothesis that gives the theoretical foundation to study the quasi-periodicity of earthquakes by assuming that there is a class of recognizably similar events happening on the same fault branch from which statistics about recurrence intervals can be calculated. The notion of a characteristic earthquake is the base of any model based on the seismic gap (or seismic cycle) hypothesis. The debate over these assumptions went over in the 90s, but it was not until large earthquakes contradicting forecasts based on these hypotheses occurred that it was reconsidered (Jackson and Kagan, 2006; Kagan et al., 2012; Geller et al., 2015), even though the debate is still open today in some areas (Husker et al., 2023). Forecasts based on the characteristic earthquake hypothesis failed to describe seismicity in several occasions, the most noticeable being the 2004 Parkfield earthquake (Bakun and Lindh, 1985; Bakun et al., 2005), the 2004 Sumatra earthquake (McCaffrey, 2007; Okal and Stein, 2009) and, the 2011 Tohoku earthquake (Geller, 2011; Onishi, 2011). The fact that took 30 years to accumulate enough data to discredit a hypothesis gives the idea of how difficult hypothesis testing could be when dealing with natural phenomena of this scale.

The previous example shows the importance of prospective testing, i.e. validating forecasts against future data, which is the only tool we have to validate hypotheses on the earthquake-generating process through the comparison between forecasts produced accordingly to a hypothesis and observed data. This process has to be done as rigorously as possible, meaning that the validation of a model should be done independently from the calibration of the model, competing models should be compared against the same data and using the same metrics, and the results of the experiments should be fully-reproducible and accessible. To smooth the difficulties of the forecast validation process and create community standards on the way in which models are evaluated against future seismicity the Southern California Earthquake Centre (SCEC) launched the Regional Earthquake Likelihood Models experiment (RELM, Field (2007)). Researchers joining the RELM experiment were asked to submit their forecasts (or code) to a testing center (Schorlemmer and Gerstenberger, 2007b) in which they have been tested prospectively between January, 1, 2006 and December, 31, 2010. The submitted models were evaluated using a set of community-agreed likelihood based tests (Schorlemmer et al., 2007b; Zechar et al., 2010b). The RELM experiment is the first case where researchers agreed to submit their models in a standardized format to a common, community-agreed testing center to be validated independently against future observations.

The RELM experiment was a success (Zechar et al., 2013) and was extended to the period from 2011 to 2015 (Strader et al., 2017). As the idea of rigorous independent testing gained support from the community, the Southern California Earthquake Center (SCEC) founded the Collaborative study of earthquake predictability (CSEP[4], Jordan (2006); Schorlemmer et al. (2010); Zechar et al. (2010b)) with the aim of expanding internationally the methods and philosophy applied in the RELM experiment. CSEP is now composed of three testing centers besides the SCEC one: the GNZ Science in New Zealand Gerstenberger and Rhoades (2010), the Earthquake Research Institute (ERI) of the University of Tokyo in Japan Tsuruoka et al. (2012), and the ETH Zurich in Switzerland Marzocchi et al. (2010) in Europe (Figure 1.2). This gave the possibility of validating more than 400 hundred models in different tectonic regimes (Taroni et al., 2014, 2016, 2018; Eberhard et al., 2012; Bayona et al., 2021; Savran et al., 2020; Rhoades et al., 2018).

The experiments organized by CSEP, however, highlighted the problems of having physical testing centers in terms of reproducibility, accessibility, and flexibility in accommodating new types of forecasting experiments (Schorlemmer et al., 2018). Specifically, the main problem was that the testing center software was strongly entangled with the testing center system architectures, and, despite the code always being open source, reproducing the experiment results outside of the testing center was impracticable. Furthermore, maintaining a physical testing center is expensive in terms of economic costs for extended periods of time. To increase the reproducibility and availability of the experimental results, in accordance with the modern open-science principle (Wilkinson et al., 2016), CSEP decided to decouple the testing center from the evaluation routine. This led to the development of an open-source Python library, called pyCSEP (Savran et al., 2022), which provides a beginner-friendly, extendible interface for practitioners to validate earthquake forecasts. `pyCSEP` aims to be the

[4] website: https://cseptesting.org/

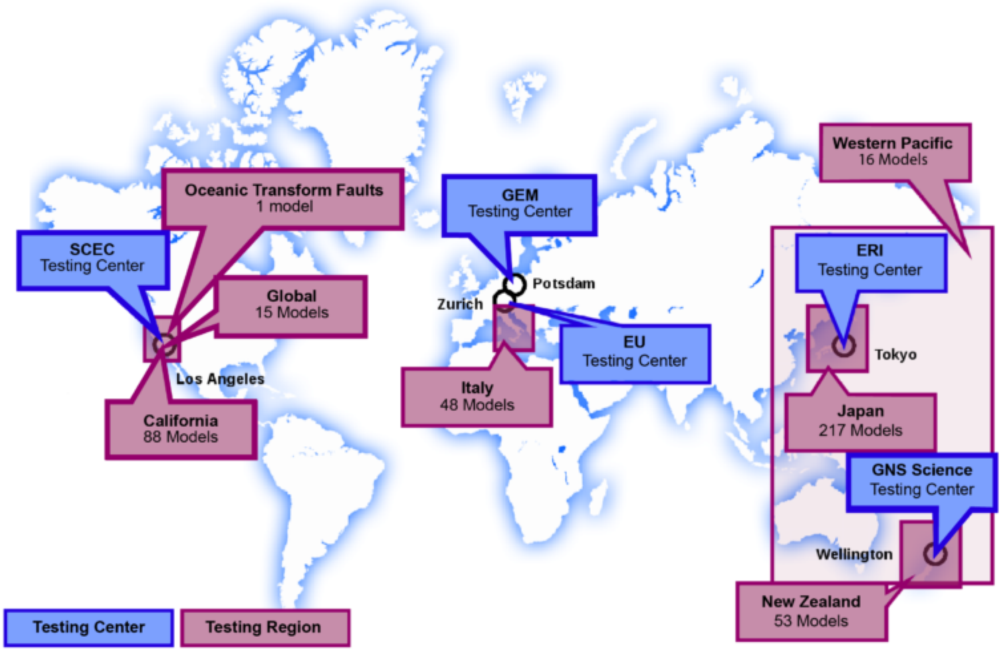


Figure 1.2: Global map of CSEP testing centers (Blue) and testing regions (Red)

first step towards making research in earthquake forecasting more sustainable (Anzt et al., 2020) providing a bridge between software developers and scientists. Indeed, the open-source nature of the pyCSEP toolkit allows researchers to contribute to the refinement and development of testing procedures by identifying potential issues and creating new features by themselves, as it has been successfully done in other contexts (Hunter, 2007; McKinney et al., 2010; Team, 2020). As part of CSEP, I have been involved in discussions around the statistical properties of the metrics used within CSEP to validate earthquake forecasts. In Chapter 7, I explore the notion of *proper* scoring rules used to rank competing earthquake forecasts based on their consistency with observed data. We show that being *proper* is an essential requirement and provide ways to check if a score is proper or not. Chapter 2.5 reviews the main metrics provided by pyCSEP to validate earthquake forecasts.

1.5 My project

In this thesis, I develop a new bayesian approximation technique to perform inference on the ETAS model. The novelty resides in a new log-likelihood approximation and differs from alternative bayesian techniques (Rasmussen, 2013; Ross, 2021). The main difference resides in the technique used to retrieve the posterior distribution. The methodology proposed in this thesis is based on the Integrated Nested Laplace Approximation (INLA) which, unlike MCMC techniques which are based on sampling the posterior distribution and approximating the posterior density using the samples, relies on a deterministic approximation of the joint posterior distribution. This has two major implications. The first one is that our methodology is faster than MCMC alternatives and scales more efficiently increasing the amount of data provided. I show in Chapter 4 that it can be 10 times faster than MCMC alternatives for the temporal ETAS model on catalogues with 2000-3000 observations, and we can expect the computational gain to be even larger for the spatio-temporal case. Second, the same data and initial settings produce exactly the same results, increasing the level of reproducibility of any result obtained with this technique. Moreover, the approximation technique I developed is not limited to the ETAS model but could be applied, in principle, to any Hawkes (or *self-exciting*) process model.

This new methodology is implemented through the R-package inlabru which provides a user-friendly interface to use the INLA method. On top of inlabru, I have developed a new R-package called ETAS.inlabru (Naylor and Serafini, 2023) with the aim of providing a user-friendly, extendible framework to work with our implementation of the ETAS model. Indeed, we want to provide a methodology able to accommodate different modifications of the ETAS model, including the possibility of accounting for the effect of external covariates (e.g. strain rate maps, fault geometry, fault displacement, heat flow), with minimal variations so that differences in the forecasts can be imputed only to differences in the model formulation. This build on the work presented by Bayliss et al. (2020) in which the inlabru methodology is used for spatially variable time-independent models of seismicity and the effect of including different combinations of available covariates is studied. My aim was to extend that approach to time-dependent models. Having the ability to accommodate a variety of models within the same framework is useful because when models are implemented using different methodologies is difficult to distinguish the effect of different hypotheses from the effect of different methodologies. In this way, researchers could run their own forecasting experiments producing forecasts with the ETAS.inlabru package and testing them with pyCSEP being sure that the parameters of each model are estimated with the same technique. All the results showed in Chapter 4, 5, and 6 are obtained with the ETAS.inlabru package. The last part of the thesis is dedicated to my work on the statistical properties of scoring rules employed to rank earthquake forecasts. I focus on the notion of *proper* scoring rule and illustrate why this property is fundamental for a score to be effective and to provide trustworthy results. I show that scores may be proper only in specific situations and when used in a certain way and not in general, and therefore, a rigorous check before using them is always needed. The Parimutuel Gambling (PG) score (Zhuang, 2010; Zechar and Zhuang, 2014) was shown to be an example of a score that is proper only in a specific situation and its performance when improper compared poorly with two proper alternatives: the Brier (Brier, 1950) and the logarithmic (Good, 1952) scores. This work provides an additional step towards having a suite of community-agreed safety checks that testing procedures and scores need to pass prior to being employed by CSEP or other organizations/users to rank earthquake forecasts.

1.6 Thesis overview

Chapter 2 provides an introduction to the most widely accepted empirical laws of seismicity (i.e. the Omori's law (Omori, 1894) and the Gutenberg-Richter law (Gutenberg and Richter, 1944)) as well as a detailed illustration of the ETAS model. The chapter also shows the difficulties in estimating the parameters and some of the many ETAS modifications that have been proposed during the last 20 years. For space-time reasons, the list will not be exhaustive but gives an idea of the possible generalizations. The last part of Chapter 2 is dedicated to model validation. The section is divided into consistency tests to measure the agreement between a forecast and the observations, and comparison scores to rank competing forecasts in the light of the observations. We describe only the main validation techniques used within CSEP, and again many others have been proposed in the literature but their analysis is beyond the scope of this chapter.

Chapter 3 describes the Integrated Nested Laplace Approximation (INLA) and the extensions provided by the inlabru R-package. The INLA methodology is designed for the general class of Latent Gaussian models, and therefore the chapter starts introducing this class. Then, I describe the Laplace approximation and how this is applied to various parts of the bayesian inferential problem to obtain the final INLA output. The second part of the chapter is dedicated to `inlabru` and especially to Log-Gaussian Cox Processes (LGCP), which are a class of point processes widely used for time-independent applications. We describe the method with which LGCP models are approximated by `inlabru`. This is relevant because we use a modification of this technique for the ETAS model and illustrating it is useful to fix the ideas used in Chapter 4 on a simpler case. The last section is dedicated to the `inlabru` iterative method which is used to extend the INLA methodology to Latent Gaussian models with non-linear predictors which are also needed to approximate Hawkes process models.

Chapter 4 is an article recently submitted to the Enivonmetrics journal and currently under review. The article is also available in pre-print at (Serafini et al., 2022a). The article describes the approximation method for general Hawkes process models and provides an application to the temporal ETAS model. The article compares my new implementation with the one provided by the `bayesianETAS` R-package (Ross, 2021) using data regarding the 2016 Amatrice seismic sequence in central Italy. In the article, I compare, retrospectively, the two implementations in terms of the number of expected events, branching ratio, and a generic measure of goodness-of-fit. Results on simulated sequences show the advantages of our approach in terms of computational time when increasing the number of events used to estimate the parameters.

Chapter 5 is an article submitted to the Frontiers in Earth Science journal and accepted for publication. The article is also available in pre-print at (Naylor et al., 2022). The article shows the capabilities of the `ETAS.inlabru` package on synthetic catalogues representing a set of situations in which standard methods struggle to correctly retrieve the ETAS parameters. We limit ourselves to the temporal case and we demonstrate that reliable estimates of the model parameters require that the catalogue data contains periods of relative quiescence as well as triggered sequences. We explore the robustness under stochastic uncertainty in the training data and show that the method is robust to a wide range of starting conditions. We show how the inclusion of historic earthquakes prior to the modelled domain affects the quality of the inversion. Finally, we show that incompleteness after large earthquakes has a significant and detrimental effect on the ETAS posteriors.

Chapter 6 generalizes the method described in Chapter 4 and used in Chapter 5 to the spatio-temporal case and shows how to include available covariates to model the number of aftershocks generated by an event. The chapter considers three modifications of the classical ETAS model: one accounting for the depth of the event, one for the mean strike of the nearest fault, and one with both covariates. The models are compared retrospectively using data on the 2009 L'Aquila sequence and the 2016 Amatrice sequence in central Italy, and the models are ranked using the Akaike Information Criterion (AIC). The aim of this chapter is not to evaluate how well the models perform in forecasting future seismicity but rather if introducing additional information from the covariates yields models offering a more precise description of seismicity. The AIC shows that all the modified ETAS model provides a better (in terms of likelihood) description of the observed data in a retrospective analysis. Forecasts produced by the models considered in this chapter would be part of the next prospective Italian forecasting experiments, and they are being used in a study on the feasibility of real-time rapid loss assessment.

Chapter 7 is an article published by the Geophysics Journal International (GJI) and can be found at Serafini et al. (2022b). The article explores the notion of proper scoring rules and applies the concept to ranking earthquake forecasts. Specifically, the paper investigates the consequences of using an improper scoring rule using as an example the Parimutuel Gambling (PG) score (Zhuang, 2010; Zechar and Zhuang, 2014), and proves analytically that the PG score is proper only in a specific situation and improper in all the others. Its performance using simulated data is compared with two proper alternatives. The simulation example can be used to test if a score is proper if it can not be proved analytically. The article provides an additional step in defining sanity checks for a scoring rule that needs to be passed before a scoring rule is used to validate earthquake forecasts.

Chapter 8 contains a discussion of the proposed methodology and obtained results and outlines future research directions.

Chapter 2

Earthquake modeling and validation

2.1 Introduction

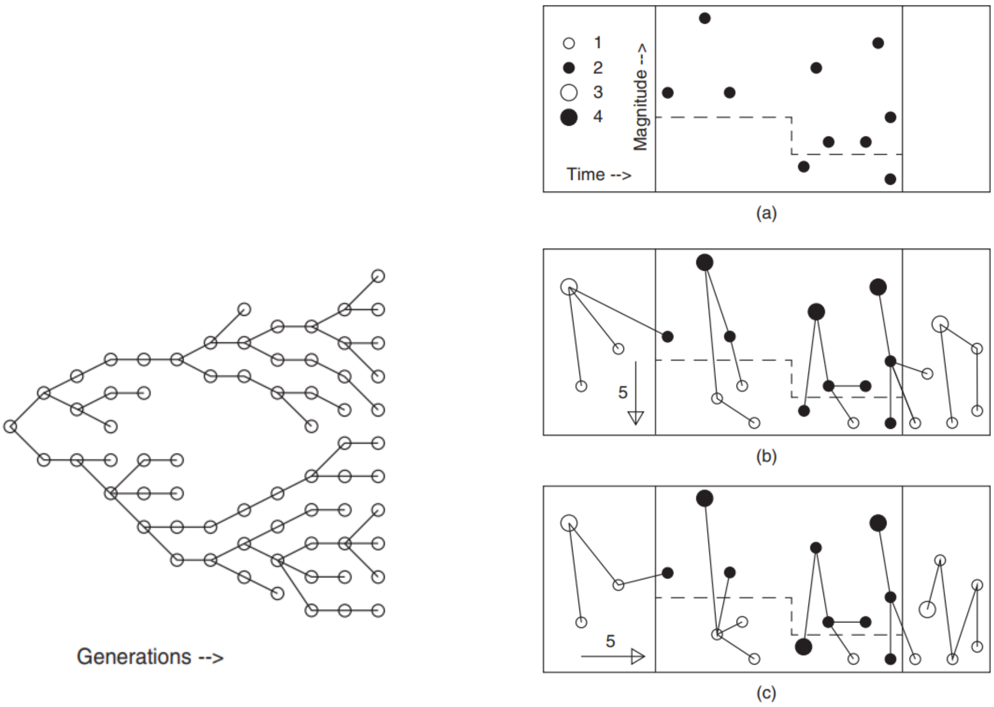
Modeling seismicity is a challenging task (Kagan and Vere-Jones, 1996). Indeed, earthquakes are usually conceptualized as multidimensional entities characterized by an origin time, a space location (2-D or 3-D), a measure of magnitude, and a focal mechanism. This means that, excluding the focal mechanism, we already have at least 4 dimensions. In addition, earthquakes tend to cluster around large events which occur relatively rarely and for which we have observations on a small interval of time compared to the recurrence of these events. Also, having more accurate or larger catalogs does not reduce the uncertainty around their occurrence (Marzocchi et al., 2015). Moreover, earthquake occurrence exhibits a fractal behavior and, therefore, models aiming to describe seismicity have to be scale-invariant over a certain range of scales (Sornette, 2006; Mandelbrot and Mandelbrot, 1982). Furthermore, earthquake occurrence is influenced by fault geometry, the presence of volcanos, the level of heat flow, the material composing the lithosphere, the deformation rate, and many other possible covariates. These complexities have so far prohibited the development of a reliable forecasting model describing future earthquakes and, therefore, the analysis of earthquake occurrence focused on describing the statistical properties of the earthquake patterns. This led to the formulation of empirical laws describing the frequencies of earthquakes in different domains (magnitude, time, space). More specifically, they are the Gutenberg-Richter (GR) law (Gutenberg and Richter, 1944, 1956), the Omori's law (Omori, 1894) and the Utsu-Seki law (Utsu, 1955; Utsu and Okada, 1969). These laws were founded to describe seismicity well in a variety of tectonic settings. Nowadays all models used in practice have to be consistent with these empirical laws, and models which do not explicitly include these laws, like machine learning models or non-parametric models, are thought to be satisfactory if they are able to *learn* them from the data van der Elst and Page (2018); Zhu et al. (2021).

Statistical analysis of earthquake occurrence based on stochastic point process theory (Daley and Vere-Jones, 2008) provides the groundwork to analyse the multidimensional structure of seismicity and to provide long and short-term forecasts. The first application of a stochastic process to the earthquake process was made by Vere-Jones (1970) who used a Neymann-Scott (Neyman and Scott, 1958) cluster process. The Neymann-Scott process was developed to study the clustering properties of galaxies in cosmology, specifically, how stars cluster to form a galaxy and how galaxies are distributed. The model is based on the idea that there are a number of unobserved cluster centers forming a homogeneous Poisson process, each center generates its own offsprings following an inhomogeneous Poisson process, the collection of all offsprings of all centers (but not the centers themselves) forms a realization of the Neymann-Scott process. Earthquakes can be described by this model thinking of stars as earthquakes and galaxies as seismic sequences. The underlying hypothesis is that there exist a number of unknown cluster centers which generate their own sequences and we observe the superposition of all offspring sequences. This is in line with the idea that there exists a number of mainshocks capable of triggering their own seismic sequence (aftershocks) and a catalogue contains a mixture of random, independent parent events and the collection of all aftershocks. The limitations of this model are that only the mainshocks are capable of generating aftershocks which means that the model does not explain secondary triggering (aftershocks of aftershocks) and the analysis is influenced by how mainshocks are determined.

The idea in Vere-Jones (1970) was generalized by Kagan (1973) who instead takes inspiration from population modeling and uses an "immigration and birth" process. In this process, a number of immigrants appear independently and spontaneously, and each immigrant generates their own offsprings who, in turn, generate their own offsprings, and so on. The cluster (or family) is composed of the immigrant, its offsprings, the second generation offsprings, and so forth. In contrast to the previous model, each point now has the ability to generate its own sequence. The analogy with the earthquake process is natural, the immigrants are the mainshocks and the offsprings are the aftershocks. These types of processes belong to the class of branching processes (Harris et al., 1963; Athreya et al., 2004) where the observations can be grouped hierarchically following a tree structure (Fig 2.1 left). Kagan (1973) considered a branching process along the magnitude axis (Fig 2.1 right (b)) where each parent can only trigger events with a lower magnitude. A branching process on the magnitude axis has the advantage that irrespective of the magnitude of completeness chosen for the analysis, the parent nodes are always retained and only offspring are removed. On the other hand, this model does not allow an earthquake to trigger a larger one, therefore, smaller foreshocks cannot be used as precursory signals of large earthquakes (Jones, 1985).

Nowadays, due to its intuitive appeal, statistical analysis of earthquake patterns is carried out almost exclusively with a branching process on the time axis (Fig 2.1 right (c)) in which each event can generate only future events irrespective of the magnitude. Branching processes in time are also known as Hawkes processes (Hawkes, 1971a,b), or *self-exciting* processes due to the fact that every observation increases the probability of additional observations in its surroundings. The clear advantage is now that any earthquake can be induced by a previous one paving the way to study precursory signals. On the other hand, excluding the events below a certain magnitude threshold may lead to breaking the link between parent and offspring (*broken linkages*), which, in turn, may bias the parameter estimates (Harte, 2016). The first application of Hawkes processes to seismicity can be dated back to Kagan and Knopoff (1987) who call it the Critical Branching Model (CBM, Kagan (2013) Chapter 9 Section 4). A different version of this model was proposed by Ogata (1988), namely, the Epidemic-Type-Aftershock-Sequence (ETAS, Ogata and Zhuang (2006); Ogata (2011)) model. Due to its ability to incorporate all the most established empirical laws of seismicity, the ETAS model is now the most commonly used model to describe the earthquake process and has been used to analyze seismicity in various countries such as Japan (Zhuang, 2011; Ogata, 2011), California (Field et al., 2017, 2021; Schneider and Guttorp), Italy (Lombardi and Marzocchi, 2010b; Marzocchi and Zhuang, 2011; Lombardi, 2017), and New Zealand (Cattania et al., 2018) to name a few.

Despite being widely used, unbiasedly estimating the ETAS model parameters is a challenging task. The nature of the inversion process means it is far too easy to invert for a set of ETAS parameters and then uncritically progress assuming they are correct. The problem of broken linkage is just the tip of the iceberg, indeed the ETAS likelihood is afflicted by multimodality and flatness around the optima which makes the application of standard estimation methods, such as maximum likelihood, unstable (Kagan, 1991; Veen and Schoen-



(a)

(b)

(c)

Figure 2.1: Left: Illustration of branching process hierarchical structure. Right: (a) Illustrative time vs magnitude scatter plot with magnitude of completeness varying over time. (b) Example of a branching process along the magnitude axis. (c) Example of a branching process along the time axis. The figures are taken from Kagan (2013)

berg, 2008; Lombardi, 2015; Harte, 2013). This is true for the most basic version of the ETAS model which makes extending it to include additional information or random effects a highly challenging task. Another issue is model validation which comprehends ranking models and evaluating how *good* the probabilistic forecasts are (Bray and Schoenberg, 2013; Savran et al., 2020). Indeed, validation is typically problematic for point process models with almost no widely accepted method to rank competing forecasts or measure the agreement between forecasts and observed data (Daley and Vere-Jones, 2004; Brehmer et al., 2021; Heinrich-Mertsching et al., 2021). As Kagan said *"the major challenge facing earthquake seismology is that new methods for hypothesis verification need to be developed. These methods should yield reproducible, objective results, and be as effective, for instance, as double-blind, placebo-controlled randomized testing in medical research"* (Kagan, 2013).

Developing such methods, along with running prospective forecasting experiments, is the aim of a global initiative such as the Collaboratory Study for Earthquakes Predictability (CSEP, Zechar et al. (2010b); Schorlemmer et al. (2018); Savran et al. (2022)) which builds on the progress made during the Regional Earthquake Likelihood Models (RELM, Schorlemmer and Gerstenberger (2007b); Zechar et al. (2013)) experiment. The goal of CSEP is indeed to organize global forecasting experiments, gather forecasts from participants, collect data and test the models against future observations in a fully prospective fashion. This allows to test hypothesis on the earthquake generation process in a scientific fashion (Popper, 2015; Jordan et al., 2011). However, part of the seismological community believe that probabilistic forecasts for earthquake occurrence are not falsifiable (Stark, 1997; Luen et al., 2008; Freedman and Stark, 2003) and therefore, these kind of experiments are meaningless. At the same time, it is not clear what the alternative would be and the debate is still open.

This Chapter is structured as follows: Section 2.2 describes earthquake catalogues and illustrates some of the problems relative to working with this data. Section 2.3 describes the most established empirical laws incorporated in almost any statistical model for earthquake occurrence. Section 2.4 describes the classical ETAS model, illustrates the methods currently used in research to estimate the parameters (Section 2.4.1), possible ETAS extensions and limitations are described in Section 2.4.2. Section 2.5 provides more detail on the validation problem and describes the statistical tests employed by CSEP.

2.2 Earthquake catalogues and missing data

Statistical analysis of real data is always challenging due to the fact that statistical models offer only a simplified (but nevertheless useful) description of reality (see Section 2.4.2), and the data collection procedure may introduce additional biases. The latter is particularly relevant in seismology where earthquake-specific information such as location, time, and magnitude is estimated from the signals provided by the seismograms network. Indeed, an earthquake's location and magnitude are not directly recorded but calculated from the signal captured by seismograms. Different methods can be applied to retrieve earthquakes summary information and therefore alternative earthquake catalogues exist. Regarding Italy, for example, there is the new Italian Homogenized instrumental seismic catalogue (HORUS, Lolli et al., 2020), the 'Catalogo Parametrico dei Terremoti Italiani' (CPTI15, Rovida et al., 2020), and the Italian Seismological Instrumental and Parametric Database (ISIDE, Group, 2007), all providing slightly different information for the same events. New algorithms to detect earthquakes are proposed every year, for example, Wimez and Frank (2022) develop a deep-learning method to identify earthquakes which found 97% more events than (Lough et al., 2013) in Antarctica known as template matched events. However, catalogues based on template matching may lead to numerical problems when analysed given that many events are reported as having the same location and many models are based on the distance between events (which would be zero in this case). Besides the problems, the choice of the catalogue is subjective and there is no standard method to choose one or the other. This is again problematic when comparing models incorporating different hypotheses and calibrated on different data because it is impossible to determine whether the difference in the results is due to the difference in the hypotheses or the data. For the analyses presented in Chapter 4 and Chapter 6 I have used data from the Horus catalogue.

As we said, the ability of an algorithm to detect earthquakes strictly depends on our ability to extract the event information from seismograms. This ability is limited and is commonly accepted that earthquake catalogues are incomplete (Ogata, 1983; Utsu et al., 1995; Kagan, 2013). The most typical case is when an event is not detected because its signal is covered by the signal of a stronger one. This may happen if i) the earthquake is so small that the noise covers its signal; ii) a large earthquake has just happened and its signal dominates the seismogram. This means that missing events are likely to be preferentially clustered in time and space around large events. For example, Kagan (2004) calculated that 1/4 of all events above magnitude 2 following the 1992 Landers earthquake were missing from the California catalogue of Hauksson et al. (2012). Therefore, catalogues are usually characterized by a magnitude of completeness above which the catalogue is thought to be complete. This magnitude of the completeness is a *property* of the catalogue that varies smoothly over time and space, due to changes in the seismographic network, and, sharply after large earthquakes, due to interference between the seismograms of the large and smaller event. Also, different estimation techniques can be used providing different values (Kagan, 2004; Hainzl, 2016a; Helmstetter et al., 2006a).

The usual approach to handle missing data is to define a magnitude threshold M_0, called cutoff magnitude, for which all events with magnitude $m > M_0$ are assumed to be detected (Mignan and Woessner, 2012). The cutoff magnitude is a subjective choice of the researcher doing the analysis and is usually different (larger) than the magnitude of completeness to account for possible biases in the estimation of the latter. However, in this way, we are wasting potentially useful information from small events. Indeed, including small events highlights important aspects of seismicity such as the cluster spatial distribution, which may help in identifying the extent of the rupture plane, or anomalies in the earthquake's occurrence before large events, that could in principle be used as precursors (Ebel, 2008; Mignan, 2012; Schurr et al., 2014). Moreover, including small events provides more accurate parameter estimates due to increased sample size (Wang et al., 2010; Schoenberg et al., 2010; Harte, 2016).

Many studies investigate how the choice of M_0 influences the estimates of the ETAS model (Wang et al., 2010; Schoenberg et al., 2010; Hainzl et al., 2013; Harte, 2016; Seif et al., 2017). The main problem is that estimation techniques such as maximum likelihood assume complete data and using incomplete data may lead to biased estimates. This is due to the problem of *broken linkage*. The problem is that the parent of an event may be missing, which leads to associating the event with another parent, which leads to biased estimates of the clusters, which determines the bias in the parameter estimates. The problem is exacerbated by the strong correlation between parameters. Indeed, if we underestimate one parameter, we are likely to underestimate all the parameters positively correlated with this one, and overestimate all the negatively correlated. The opposite happens if one parameter is overestimated.

Focusing on data incompleteness, some approaches have been proposed to circumvent the problem. Hainzl et al. (2008) discard all events recorded within a certain time from a large event. Helmstetter et al. (2006a) and Werner et al. (2011) describe incompleteness in terms of mainshock characteristics but this approach relies on the identification of mainshocks and an ad-hoc estimation procedure. A more general approach is the one proposed by Omi et al. (2014) which extends the idea of Ogata and Katsura (1993). Specifically, they assume that each earthquake has a detection probability, which is a function of the time elapsed from the last large events. In this way, they can calculate the expected number of *missed* events and adjust the estimates based on this value. However, the parameters of the detection function have also to be estimated from incomplete data. Alternatively, Zhuang et al. (2017) propose an algorithm to add synthetic observations to the real catalogue, however, the space-time evolution of M_0 must be defined. Lastly, Hainzl (2022) developed a method based on the simple assumption that an earthquake cannot be detected if it occurs within a certain time after a larger event. All of these methods use maximum likelihood and there is no bayesian counterpart nor commonly accepted method to handle the problem.

Regarding the temporal ETAS model, how characteristics of the data influence parameter estimates are described in Chapter 5. Specifically, I investigate the effect of estimating parameters from sequences with zero, one, or more large events, the effect of data incompleteness, and the effect of considering events before the time interval chosen for the study.

2.3 Fundamental Empirical Laws

Statistical models for seismicity are based on three well-established empirical laws. The Gutenberg-Richter (GR) law (Gutenberg and Richter, 1944), Omori's law ((Omori, 1895), and the Utsu-Seki law (Utsu, 1955; Utsu and Okada, 1969). The first describes the magnitude distribution, the second describes the time distribution of aftershocks, and the third relates the rupture area to the magnitude of the event. These laws have been successfully applied to a variety of tectonic settings and they now constitute the backbone of any statistical model for seismicity. Indeed, the magnitude of completeness is determined by checking *how well* the data abide by the GR-law (Woessner and Wiemer, 2005) and the same is done to measure the quality of the data provided by high-resolution catalogs (Herrmann and Marzocchi, 2021). In the same way, models which do not explicitly assume Omori's law are considered successful if they retrieve a similar relationship.

2.3.1 The Gutenberg-Richter law

The GR law provides a relationship between a magnitude value m and the logarithm of the number of earthquakes $N(m)$ with a magnitude greater or equal to m:

$$\log_{10} N(m) = a - b(m - M_0), \tag{2.1}$$

where M_0 is the cutoff magnitude which should be always higher than the magnitude of completeness of the catalogue. The quantity a is the intercept of the model and it is usually estimated as the logarithm of the total number of events with magnitude above m_0. The term b is called b-value and characterizes how the number of events scales with the magnitude. Despite the b-value being thought to be a universal constant (Kagan, 1999), many studies have investigated its variation in space and time (Herrmann et al., 2022; El-Isa and Eaton, 2014; Lombardi, 2022).

Regarding the classical GR law, Aki (1965) showed that equation 2.1 can be rewritten as

$$N(m) = \exp\{a \log(10)\} \exp\{-\log(10) b(m - m_0)\}, \tag{2.2}$$

where $\exp\{a \log(10)\}$ can be seen as the total number of events and $\exp\{-\log(10)b(m-m_0)\}$ as the probability of having an event with magnitude greater than m. Framed this way, the GR law assumes an exponential distribution for the quantity $m - m_0$ with parameter $\beta = \log(10)b$. This implies that the maximum likelihood estimator for β is

$$\hat{\beta} = \frac{1}{\bar{m} - m_0}, \tag{2.3}$$

where $\bar{m}$ is the average magnitude in the catalogue.

Assuming an unbounded exponential distribution, however, is not realistic. Specifically, the exponential distribution has no upper bound, implying the possibility of an infinite seismic energy, which is unphysical. Furthermore, this is problematic when considering catalogue-based forecasts, in which a forecast is a collection of synthetic catalogues (usually 10000 or 100000) simulated according to the forecasting model. In this case, it is highly probable to generate an event with magnitudes over 9 even if the catalogue used to estimate the GR law parameters has a maximum magnitude equal to 7. This leads to overpredicting the expected number of events in an area.

Two approaches have been proposed to address this problem leading to two modifications of the classical GR law to model the tails of the magnitude distribution. They are known as the *Truncated* GR law (Cosentino et al., 1977) and the *Tapered* GR law (Utsu, 1999; Kagan, 2002). The former requires the specification of a maximum magnitude and applies a hard bound to the tail of the distribution (the probability of observing a magnitude greater than the maximum is zero). The latter, instead, requires the specification of a corner magnitude and applies a softbound to the tail meaning the probability density function rapidly goes to zero after the corner magnitude. Here we describe only the Tapered GR law (Tap-GR) because estimating there is no reliable method to estimate the maximum magnitude needed to define the Truncated GR law (Zöller and Holschneider, 2016). Instead, we have reliable methods to estimate the corner magnitude of the Tapered GR law (Kagan and Schoenberg, 2001).

The idea behind the Tapered and Truncated GR law is that the GR law on the magnitude (m) domain is equivalent to the Pareto distribution on the seismic moment (M) domain (Kagan, 2002). The Tapered and Truncated GR laws are then obtained considering a tapered or truncated Pareto distribution on the seismic moments. The density of a tapered Pareto distribution with cutoff seismic moment M_0 and corner moment M_c is:

$$f_{tap}(M) = \left(\frac{\beta}{M} + \frac{1}{M_c}\right)\left(\frac{M_0}{M}\right)^{\beta} \exp\left\{\frac{M_0 - M}{M_c}\right\}, \tag{2.4}$$

for $M > M_0$. The parameter β regulates the slope of the GR law and is linked to the b-value by $\beta = b2/3$.

A source of bias in the b estimates obtained with 2.3 comes from the binning of the magnitude as shown in (Marzocchi and Sandri, 2003). In fact, assuming an exponential distribution implicitly assumes that the magnitude is a continuous random variable while the observed values are binned through the rounding to 1 decimal place, this is true also for the Tapered or Truncated GR laws. This brings two sources of bias in the estimator $\hat{b}$ i) the distribution in each bin is not uniform and, therefore, the average of the binned variable is different from the average of the continuous one; ii) the minimum magnitude is not m_0. The average $\bar{m}$ using binned observations systematically overestimates the true mean. However, this effect is negligible when the bin size is $\Delta_m = 0.1$ (Bender, 1983). The second and more important source of bias is that by fixing a threshold on the binned magnitude, observations with $m = m_0$ are ranging from $m_0 - \Delta_m/2$ to $m_0 + \Delta_m/2$ and, perhaps, m_0 is not the minimum. Utsu (1966) proposed a simple but effective modification of $\hat{\beta}$ in which m_0 is replaced with $m_0 - \Delta_m/2$.

Typical observations report *b*-values in the range $0.5 - 2$. Variations of *b*-value in time and space have been largely investigated in past years and have proven to be a rich source of information about the seismotectonic of a region (Christensen and Olami, 1992; Wiemer and Wyss, 2002; Boettcher et al., 2009). For example, *b*-value variations have been linked to type of faulting (Schorlemmer et al., 2005), surface creep rate (Tormann et al., 2013), property of the materials composing the earth crust (Goebel et al., 2017), more examples can be found in Herrmann et al. (2022). In fact, even if the debate on the source of b-value variations is still open (Marzocchi et al., 2020), it is widely accepted that they reflect the heterogeneity in the earth's crust. In general, variations in the *b*-value need to be evaluated carefully, they can depend on varying magnitude of completeness during seismic sequences (Herrmann and Marzocchi, 2021), sample size and magnitude range (Nava et al., 2017; Geffers et al., 2022), the used magnitude scale, magnitude binning, windowing, and maximum likelihood estimator (Marzocchi et al., 2020). Some progress has been made to develop methods that do not require estimates of the magnitude of completeness or data windowing (van der Elst, 2021b), however, most analyses heavily rely on subjective choices (Herrmann et al., 2022). In Chapter 6, we use a Tapered GR law to produce the forecasts.

2.3.2 Omori's Law

This law (Omori, 1895) was the first discovered empirical law that holds for different earthquake sequences taking place in different tectonic settings. A sequence can be identified as a cluster of events in time and space, where the one having the highest magnitude is retrospectively identified as the mainshock, the events before this are called foreshocks, and the ones after are referred to as aftershocks. This classification can be made only once the entire sequence has been observed and, thus, it can not be done in real-time when forecasting.

The Omori law describes the rate R of occurrence of aftershocks as a function of time t from the mainshock. The first version of Omori's law stated that:

$$R(t) = \frac{K}{t+c}, \tag{2.5}$$

where $K, c > 0$, and $t \geq 0$. This version was generalized by Utsu (1957, 1966) to

$$R(t) = \frac{K}{(t+c)^p}, \tag{2.6}$$

with $p > 1$. The parameter p has to be greater than one otherwise an earthquake can generate an aftershock sequence with an infinite number of events in infinite time which is unphysical. This is known as the modified Omori law. A complete review of the evolution of Omori's law can be found in Utsu et al. (1995).

The parameter p regulates how fast the number of aftershocks decays with time (the larger the faster). It needs to be greater than 1 otherwise the total number of aftershocks generated by a mainshock in infinite time is infinite, which is unphysical. Mogi (1962) studied more than 30 sequences in Japan and observed that the spatial distribution of p values was similar to the surface heat-flow distribution. In particular, higher p values (faster aftershocks decay) were found in regions with higher temperatures, such as volcanic regions, where the stress is supposed to relax faster, with the opposite being true for regions with low temperatures. Kisslinger and Jones (1991) found the same relationship while studying California, and tried to link the p values and the surface temperature using a linear regression model (Kisslinger, 1993). However, data about surface temperature is not always available, which prevents proper testing of this hypothesis. The relationship between p and the magnitude of the mainshock has been studied in different articles, early studies stated no relationship was found between the two quantities Utsu et al. (1995), Hainzl and Marsan (2008); Ouillon and Sornette (2005) found a significant increase of p with m. However, this is usually not taken into account in forecasting models.

The parameter K regulates the number of aftershocks and, therefore, it is strictly connected to the magnitude of completeness applied to the data used to estimate it, while parameter c prevents $R(0) = \infty$. The parameter c acts as a minimum time interval and describes the aftershock decay close to the mainshock. The value of c is determined both by temporal incompleteness near the mainshock and foundamental physics. Being able to detect a greater number of aftershocks a lower magnitudes leads to smaller value of c (Narteau et al., 2002; Peng and Zhao, 2009). However, using standard catalogues, the intervals of time just after the mainshock are the ones most affected by missing data, especially for lower magnitudes, Section 5.2.4 shows the consequences of using incomplete data to estimate the parameters of the temporal ETAS model. Therefore, lowering the magnitude of completeness may lead to biased estimates of c.

Omori's law has been applied in a plethora of studies. However, as formulated above, it presents some problems in application. For example, if we have two mainshocks close enough in time, their aftershock sequences will be superposed, and parameter estimates may be biased. One approach has been to study superposed sequences assuming that all of them should obey the same Omori's law (Davis and Frohlich, 1991; Utsu, 1992). However approached, those analyses relied on an underlying declustering algorithm in order to distinguish aftershocks generated by mainshocks from background seismicity. Most estimation methods for the ETAS model are affected by the same problem.

2.3.3 The Utsu-Seki Law

The magnitude is a measure of the energy released by an event and is therefore linked to the extent of the rupture generated by an earthquake. Such rupture generates the aftershock sequence that will be concentrated near the rupture, resulting in clustering in space. The area S containing the aftershocks well approximates the rupture area (Marsan and Lengline, 2008; Grimm et al., 2022a), and is natural to search for a mathematical relationship between the magnitude of the event m and the rupture area S. Utsu (1955) formulated that this relationship has the same log-linear form as the GR law

$$\log S = a_u + b_u m, \tag{2.7}$$

with $a_u \in \mathbb{R}$, and $b_u \in \mathbb{R}^+$ to reflect the fact that stronger earthquakes generate larger ruptures. Equation 2.7 is referred as the Utsu-Seki law. The other empirical finding that is usually reported with the Utsu-Seki law is the fact that aftershocks are usually contained in an ellipsoid around the mainshock or, at least, containing also the mainshock (Utsu and Okada, 1969) and that aftershocks diffuse anomalously slowly in space (Huc and Main, 2003).

2.4 Epidemic-type Aftershock Sequence model

The Epidemic-Type Aftershocks Sequence model (ETAS, Ogata (1988, 1998) can be considered the state-of-the-art model for describing seismicity having outperformed competitors in several experiments (Taroni et al., 2018; Cattania et al., 2018; Nanjo et al., 2012; Schorlemmer et al., 2018). The ETAS model belongs to the family of Hawkes process models (Hawkes, 1971a,b), or *self-exciting* processes, which are point process models designed to model phenomena exhibiting clustering in time and space-time. This is done by considering every earthquake as a point in time (or space-time) equipped with a magnitude and allowing every point to generate its own sequence of aftershocks (or offspring). The branching is done with respect to time, meaning that events can only trigger events in the future. This gives Hawkes process models the ability to model phenomena with cascades of events strongly clustered in time (space-time). For a more general introduction and examples of Hawkes process applications, I refer to Chapter 3 or Laub et al. (2021).

Point process models are completely specified by defining a conditional intensity function Daley et al. (2003). Indeed, given a point $\mathbf{x}$ in a space $\mathcal{X} \subset \mathbb{R}^n$, a set of N observations constituting the history of the process $\mathcal{H} = \{\mathbf{x}_h : \mathbf{x}_h \in \mathcal{X}\, \forall h = 1, ..., N\}$, and a point process with conditional intensity $\lambda(\cdot|\mathcal{H}) : \mathcal{X} \to (0, \infty)$, the probability of observing a point in ball $b(\mathbf{x})$ around $\mathbf{x}$ is given by $\lambda(\mathbf{x})|b(\mathbf{x})|$ where $|b(\mathbf{x})|$ is the volume of the ball. Section 4.2.3 gives a more formal conditional intensity definition. Hawkes process models share the same functional form of the conditional intensity, considering $\mathbf{x} = t$ consisting of only time, the Hawkes process conditional intensity is

$$\lambda(t|\mathcal{H}_t) = \mu + \sum_{h:t_h<t} g(t - t_h), \tag{2.8}$$

where $\mathcal{H}_t$ is the history of the process, namely, the collection of all recorded events that happened strictly before t, $\mathcal{H}_t = \{t_h \in \mathcal{H} : t_h < t\}$. The quantity $\mu > 0$ is referred to as the background rate and is the rate at which events arise spontaneously, in other words, μ is the rate of the Poisson process regulating the immigrants' arrival. The summation is over all the events recorded before t at which we wish to evaluate the conditional intensity and represent the aftershocks rate a time t and is given by the sum of the rate of the aftershocks sequence initiated by events in the past. The function $g(\cdot)$ is referred to as *triggering* or *excitation* function and it is the conditional intensity of the Poisson process regulating the offspring's arrival. The model can be easily extended to account for space, magnitude, or other variables changing the form of the triggering function.

The basic temporal ETAS model as formulated by (Ogata, 1988) uses the following conditional intensity

$$\lambda(t|\mathcal{H}_t) = \mu + \sum_{h:t_h<t} K e^{\alpha(m_h - M_0)}(t - t_h + c)^{-p}, \tag{2.9}$$

where M_0 is the magnitude of completeness, and the history of the process is composed by time-magnitude couples $\mathcal{H}_t = \{(t_h, m_h) \in \mathcal{H} : t_h < t, m_h > M_0\}$ and the parameters of the model are $\mu, K, \alpha, c > 0$ and $p > 1$. The parameter K regulates the number of events generated by an event with magnitude $m_h = M_0$, while the parameter α regulates how the number of aftershocks scales with the magnitude, it is expected to be positive because it has to reflect the fact that stronger earthquakes generate more aftershocks. The parameters c and p are the parameters of the Omori law and regulate how the aftershocks' number decays with time.

This model describes only the occurrence times of earthquakes. The magnitude in a given seismic zone is usually considered to be independent of time and space and follows a GR law described in Section 2.3.1. Ogata (1988) extends the ETAS model to include also space. Indeed, it is sufficient to multiply the triggering function for a function describing the aftershock decay in space $g_s(\mathbf{s}, \mathbf{s}_h)$. This function usually accounts also for the magnitude of the triggering event and is parametrized as:

$$g_s(\mathbf{s}, \mathbf{s}_h, m_h) = \frac{1}{\pi\sigma(m_h)} h\left(\frac{(\mathbf{s} - \mathbf{s}_h)A(\mathbf{s} - \mathbf{s}_h)^t}{\sigma(m_h)}\right), \tag{2.10}$$

where $\sigma(m_h)$ is an increasing function of the magnitude there to represent the fact that the aftershock region increases with the magnitude of the triggering event as a consequence of the Utsu-Seki law, and the usual choice is $\sigma(m_h) = e^{\gamma(m_h - M_0)}$. The quantity A is a 2×2 matrix (3×3 if $\mathbf{s}$ is 3-D). The quadratic form is in place because the aftershock occurrences are supposed to be contained in an ellipsoid whose shape is regulated by A and its size by $\sigma(m_h)$. Considering A equal to the identity matrix we have the function g_s is a function of the Euclidean distance between $\mathbf{s}$ and $\mathbf{s}_h$, which means that the aftershock process is isotropic.

Ogata (1998) presented different choices of $h()$ and compared them using data from Japan. The most natural choices are the exponential for which $g(\mathbf{s}, \mathbf{s}_h, m_h)$ is a bivariate Gaussian density with mean $\mathbf{s}_h$ and variance $\sigma(m_h)$, or a power-law function such as

$$g_s(\mathbf{s}, \mathbf{s}_h, m_h) = \left(\frac{(\mathbf{s} - \mathbf{s}_h)A(\mathbf{s} - \mathbf{s}_h)^t}{\sigma(m_h)} + d \right)^{-q}. \tag{2.11}$$

Ogata (1998) and Zhuang et al. (2004) analyzing sequences in Japan found that equation 2.11 provides better results in terms of Akaike information criterion (AIC, Akaike, 1974) than the bivariate Gaussian density. This is due to the fact that equation 2.11 is more flexible and can model a spatial decay slower than exponential. However, the Gaussian case should not be ruled out given the mathematical advantages and the fact that while AIC is most effective for regular processes (Daley and Vere-Jones, 2004), its applicability to Hawkes processes is questionable (Kagan, 2013).

The complete conditional intensity for a point $\mathbf{x} = (t, \mathbf{s}, m)$ is then given by:

$$\lambda(\mathbf{x} = (t, \mathbf{s}, m)|\mathcal{H}_t) = \left(\mu + \sum_{h:t_h<t} K e^{\alpha(m_h - M_0)}(t - t_h + c)^{-p} g_s(\mathbf{s}, \mathbf{s}_h, m_h) \right) \pi(m), \tag{2.12}$$

where $\pi(m)$ is a magnitude distribution derived from the GR-law.

The main advantage of ETAS is that provides a theoretical framework to incorporate in the model all the most established empirical laws regarding earthquake occurrence. The GR-law is taken into account by the term $\pi(m)$, while Omori's law determines how the aftershocks sequences distribute in time, and the Utsu-Seki law is incorporated considering a space triggering function $g_s()$ that scales with the magnitude of the event.

2.4.1 Parameter Estimation for simple ETAS

One of the main goals of any statistical analysis based on the ETAS model is estimating the parameters based on observed data. The most important quantity in doing this is the log-likelihood of the model. It is convenient to group the parameters to be estimated in one parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, for the temporal ETAS model $\boldsymbol{\theta} = (\mu, K, \alpha, c, p)$, while for the spatio-temporal model with space-triggering function given by equation 2.11 and $\sigma(m) = \exp\{\gamma(m - M_0)\}$ is $\boldsymbol{\theta} = (\mu, K, \alpha, c, p, d, q, \gamma)$. The information on the parameters carried by the data is usually synthesized by the likelihood of the model. The likelihood of a set of parameters, given an observed point pattern, is the probability of observing that pattern with those parameter values. For point process models, calling $\mathcal{H}$ the set of all observed points in a region of interest $\mathcal{X}$, and $\boldsymbol{\theta}$ the parameters of the model, the log-likelihood of the model is given by:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{H}) = -\Lambda(\mathcal{X}) + \sum_{x_h \in \mathcal{H}} \log \lambda(\mathbf{x}_h|\mathcal{H}_{t_h}), \tag{2.13}$$

where $\Lambda(\mathcal{X})$ is the integral of the conditional intensity over the domain $\mathcal{X}$. In the case where $\mathbf{x} = (t, \mathbf{s})$ is composed by a time t, a 2D location $\mathbf{s}$, the domain is $\mathcal{X} = (T1, T2] \times W \times [M_0, M_c)$, with $0 < T_1 < T_2$, and $W \subset \mathbb{R}^2$ is the spatial domain. Given that the frequency-magnitude distribution $\pi(m)$ is independent of the space-time location and that the magnitude of the event does not appear in other parts of the intensity, the integral over the magnitude domain of the intensity is equal to the intensity multiplied by the integral of $\pi(m)$. The latter is 1 by definition if the magnitude distribution is a proper probability distribution and therefore it can be omitted from the calculations. Then, the integral of the conditional intensity is given by

$$\begin{aligned}\Lambda(\mathcal{X}) &= \int_W \int_{T_1}^{T_2} \lambda(\mathbf{x} = (t, \mathbf{s})|\mathcal{H}_t) dt d\mathbf{s} \\ &= (T_2 - T_1)|W|\mu + \sum_{h:x_h \in \mathcal{H}} K e^{\alpha(m_h - M_0)} \int_W \int_{T_1}^{T_2} g(t - t_h, \mathbf{s} - \mathbf{s}_h) dt d\mathbf{s},\end{aligned} \tag{2.14}$$

where

$$g(t - t_h, \mathbf{s} - \mathbf{s}_h) = (t - t_h + c)^{-p} g_s(\mathbf{s}, \mathbf{s}_h, m_h). \tag{2.15}$$

Equation 2.14 refers to the simple case in which the background rate is assumed to be homogeneous in space and time. This assumption can be relaxed, in that case, it is sufficient to substitute $(T_2 - T_1)|W|\mu$ with the integral of the inhomogeneous background rate on the domain.

The ETAS log-likelihood just described is problematic under multiple aspects when inference needs to be performed. The first thing is the evaluation of the integral $\Lambda(\mathcal{X})$ that often has no closed-form solution and researchers had to resort to numerical integration (Ogata, 1998; Schneider and Guttorp) or to the use of simplifying assumptions (Schoenberg, 2013). Also, the sum of the logarithm of the conditional intensity presents its own issues, indeed just calculating it for n points scales as $\mathcal{O}(n^2)$ which slows down any numerical iterative method employed to optimize the log-likelihood. Regarding the likelihood optimization, this is probably the most problematic part, indeed, the maximum likelihood estimator has been reported multiple times to be unstable (strongly dependent from the starting point of the optimization algorithm) and potentially biased, especially for short catalogues (Harte, 2013; Seif et al., 2017). This is due to the fact that the ETAS likelihood is multimodal and flat near the optima (Veen and Schoenberg, 2008; Lombardi, 2015), which, in turn, is due to the fact that the parameters are strongly correlated with each other (Guo and Ogata, 1997). The correlation stems from the fact that different combinations of parameters may assign the same probability to a given catalogue making them indistinguishable from a likelihood point of view.

To mitigate the problems arising from the ETAS likelihood function, the most commonly used methods for ETAS parameters' estimation are based on a different likelihood function leveraging on the unknown branching structure of the process. Indeed, both the EM algorithm developed by Veen and Schoenberg (2008) (frequentist) and the Gibbs sampler developed by Ross (2021) (bayesian) are based on the same conditional likelihood. For each event i, a variable B_i is defined assuming value

$$B_i = \begin{cases} 0 & \text{if event } i \text{ is a background event} \\ j & \text{if event } i \text{ is an aftershock of event } j \end{cases}. \tag{2.16}$$

Then, the likelihood conditional on the knowledge of B_i is

$$\begin{aligned}\mathcal{L}_{br} =& |A_0| \log \mu + \sum_{h=1}^{n} \left(|A_h| (\log(K) + \alpha(m_j - M_0)) + \sum_{j:\mathrm{x}_j \in A_h} \log g(t_j - t_h, \mathbf{s}_j - \mathbf{s}_h) \right) + \\ & - \Lambda(\mathcal{X}), \end{aligned} \tag{2.17}$$

where A_0 is the set of background events composed by $|A_0|$ elements, while A_h is the set of aftershocks of the event h (set of events with $B_i = h$), and $|A_h|$ the number of such aftershocks.

Both (Veen and Schoenberg, 2008) and (Ross, 2021) report that inference based on the log-likelihood expressed in equation 2.17 is more robust than using the classical one. Both methods are based on an iterative procedure. For the EM algorithm, starting from an initial guess of the branching structure B_i, the conditional likelihood is optimized with respect to the ETAS parameters $\boldsymbol{\theta}$, then, a new branching structure is estimated from $\boldsymbol{\theta}$, and a new optimization step is performed. This is repeated until convergence is reached. The MCMC method developed by Ross (2021) is similar, they start from an initial guess of the branching structure B_i given by the prior, then sample from the conditional (on B_i) distribution of $\boldsymbol{\theta}$, the sample is used to estimate a new branching structure, which, in turn, is used to update the conditional distribution of the ETAS parameters $\boldsymbol{\theta}$. This is repeated until a sufficiently large number of posterior samples is obtained. In both cases, the branching structure given the value of the ETAS parameters is given by

$$\Pr[B_i = j | \boldsymbol{\theta}] = \begin{cases} \frac{\mu}{\lambda(\mathrm{x}_i)} & \text{if } j = 0 \\ & \\ \frac{g(t_i - t_j, \mathrm{s}_i - \mathrm{s}_j)}{\lambda(\mathrm{x}_i)} & \text{if } j = 1, 2, ..., i-1 \end{cases} . \tag{2.18}$$

Once the parameters have been estimated, another challenging task is to estimate the uncertainty around the parameter values. This is not a problem for Bayesian methods, for which each parameter is a random variable and has its own distribution, but it is a problem for frequentist methods for which the uncertainty around the parameter comes only from the uncertainty in the observations. The classical way to estimate the uncertainty around maximum likelihood estimators is based on the use of the second-order derivatives (Hessian matrix) of the likelihood (Wilks, 1964). One of the assumptions of this method is that the likelihood isoline should be elliptic, but this assumption is rarely satisfied by earthquake catalogues due to their limited time-extension and uncertainty estimates based on this method are unstable and may be unreliable (Harte, 2013). Moreover, they require the likelihood function to be differentiable twice, however, parameters are subject to constraints that introduce points in which the likelihood is not differentiable (Jackson and Matsu'Ura, 1985; Kagan, 2013). For this reason, the bayesian approach seems most appropriate.

Other techniques have been proposed based on maximizing the likelihood Lombardi (2015); Chiodi and Adelfio (2011); Kasahara et al. (2016) but do not address the problems above nor biases due to data incompleteness. For example, Seif et al. (2017) reports that productivity parameter K is usually overestimated by maximum likelihood methods which is correlated with the underestimation of p. In general, p is thought to decrease if M_0 is increasing due to reduced sample size (Harte, 2016), however, Schoenberg et al. (2010) and Seif et al. (2017) observed that it increases in simulated sequences. This apparent contradiction can be explained with the *broken linkage* problem. In fact, Harte (2016) studies events that may have been associated with the wrong parent and they will be likely associated with the tails of the nearest large events widening the distribution of its aftershocks and inducing smaller values of p. All the reported studies also remark that the bias depends on the poor performance of the maximum likelihood for sequences with a branching ratio close to 1 (critical regime). The bias is smaller considering sequences with a smaller branching ratio (Seif et al., 2017).

2.4.2 ETAS limitations and extensions

The ETAS model as presented in 2.12 provides a simplified description of reality and, if applied to real earthquake catalogues, parameter estimates can be biased by incorrect assumptions. For example, various studies observe a bias in the estimation of parameter α (Seif et al., 2017; Hainzl et al., 2008, 2013). They argue that the bias may be due to the assumption of isotropic spatial triggering function or constant background rate which have been reported to influence the estimates of α. In fact, also the background rate estimates are influenced by the choice of M_0. Specifically, have been observed that higher M_0 values lead to smaller background rates (Sornette, 2006). This is expected, and the background rate varies similarly to what the GR law predicts (Seif et al., 2017). Other model assumptions which may lead to biased estimates are the infinite duration of triggering, temporally and spatially varying parameters, or 3-D earthquake locations, we show in this section how these assumptions can be relaxed.

In the classical ETAS formulation, the background rate is supposed to be constant over space, meaning that background points are assumed to be homogeneous over space and time. This assumption is usually relaxed considering a substituting $\mu(\mathbf{s})$ to μ in expression 2.12 and considering:

$$\mu(\mathbf{s}) = \mu\nu(\mathbf{s}), \tag{2.19}$$

for $\mathbf{s} \in W \subset \mathbb{R}^2$ and assuming $\int_W \nu(\mathbf{s})d\mathbf{s} = 1$.

Different choices of $\nu(\mathbf{s})$ can be employed. Ogata (1998) uses bi-cubic B-splines for $\log \nu(\mathbf{s})$ and optimizes a penalized log-likelihood assuming a non-homogeneous Poisson model for the observations and a smoothness penalty (Goodd and Gaskins, 1971) using a declustered catalogue as observed data. Many researchers have employed similar procedures with different kernel estimators for $\nu(\mathbf{s})$ (Lombardi and Marzocchi, 2010b; Zhuang, 2011; Ogata, 2011; Nandan et al., 2021b). The usual choice is to use a Gaussian kernel with adaptive bandwidth, however, the bandwidth is usually not estimated from the data and estimate the other ETAS parameters using Zhuang et al. (2002) or Veen and Schoenberg (2008) methods. The results of these types of procedures depend on the choice of declustering algorithm. Alternatively, non-parametric approaches have been used to estimate a spatially varying background rate (Adelfio and Chiodi, 2015; Zhuang et al., 2002). On the Bayesian side, Molkenthin et al. (2022) modified the MCMC algorithm proposed by Ross (2021) and models the background rate as a spatially varying Gaussian process.

In the same way, researchers have investigated the hypothesis that the other ETAS parameters are also spatially varying. The most intuitive method is to divide the region of interest into sub-regions and estimate the parameters in each region separately (Veen and Schoenberg, 2008) but this approach is not completely satisfactory and the values of the correlation between parameters values in adjacent regions is not accounted for. Another approach is to define spatially varying parameters using the same functional form used for the background rate in equation 2.19 (Ogata, 2011). Specifically, (Ogata, 2011) defines the spatially varying parameters using a triangulation of the space and estimating the value of the parameter only at the nodes. Then, a piecewise linear basis function assuming value 1 at the node and 0 at adjacent nodes is used to calculate the value of the parameter at different locations. This is a similar approach used in Section 3.3.2 used to approximate Gaussian fields. Despite the generality of the approach, in practice, only the background rate and the productivity parameters (α or K) are considered as spatially varying. As before, gaussian kernels with adaptive bandwidth are also considered (Mizrahi et al., 2021; Nandan et al., 2021a) but the bandwidth is not estimated from the data.

A different approach is the one used by Adelfio and Chiodi (2021) which provides a link between Generalized Linear Models (GLM) and the ETAS model. Their approach is based on the observation that the number of aftershocks generated by an event $\mathbf{x}_i = (t_i, \mathbf{s}_i, m_i)$ is given by:

$$n(\mathbf{x}_i) = \pi(m_i) \int_W \int_{\max(T_1, t_h)}^{T_2} K \exp\{\alpha(m_i - M_0)\} g(t - t_i, \mathbf{s} - \mathbf{s}_i) dt d\mathbf{s}. \tag{2.20}$$

Excluding boundary effects, the expected number of aftershocks depends only on the magnitude. The idea of Adelfio and Chiodi (2021) is to replace the term $\alpha(m_i - M_0)$ with a linear predictor

$$\eta_i = \beta_0 + \boldsymbol{\beta}\mathbf{z}_i^t, \tag{2.21}$$

where $\mathbf{z}_i$ is the vector of covariates relative to the i-th point, $\boldsymbol{\beta}$ is the vector of coefficient, and β_0 is an intercept term. In this way, two points with the same magnitude can still generate a different number of aftershocks if, for example, they are at different depths, at different distances from mapped faults, or in regions with different levels of heatflow. In Chiodi et al. (2021) they found out that the depth and a measure of the displacement rate improve the model in terms of AIC. A further extension of their approach would be to include a structured random effect to account for any spatial (or temporal) correlation not explained by the covariates. We describe this kind of random effect in Section 3.3.2.

Many studies have shown that considering an isotropic kernel for the spatial triggering function is not appropriate (Ogata, 2011; Hainzl et al., 2013; Seif et al., 2017; Zhang et al., 2018). Indeed, the isotropic assumption may be acceptable only for moderate and small earthquakes, while large earthquakes usually produce a rupture with an elongated shape and, therefore, aftershocks usually form an elliptic shape (Utsu, 1955). The usual method is to define a number of mainshocks and consider different triggering functions for each of them Ogata (2011); Bach and Hainzl (2012); Grimm et al. (2022a). For example, Ogata (2011) uses a valid covariance matrix as S in expression 2.10 with different parameters for each mainshock. Bach and Hainzl (2012) uses information provided by ShakeMaps, ground motion maps, and static Coulomb stress changes to add anisotropy. Grimm et al. (2022b) uses the distance from an estimated rupture segment.

The list of extensions we mentioned above is not meant to be exhaustive but just gives an idea of the efforts made by the seismological community to improve the ETAS model. Other noticeable examples are the *Renewal* ETAS model (Stindl and Chen, 2021) that uses a renewal process instead of a homogeneous Poisson process as background. The *Restricted* ETAS model (Gospodinov and Rotondi, 2006) allows only aftershocks above a certain threshold to generate secondary earthquakes. (Mizrahi et al., 2021) and (Nandan et al., 2021a) use an Exponentially Tapered Omori Kernel instead of the classical Omori's law to describe aftershocks decay in time. The ETAS *Incomplete* model Hainzl (2016a, 2022) assumes a blind time of detection after the occurrence of a large event.

The flexibility of the ETAS model is what makes it a powerful tool to describe earthquake occurrence. On the other hand, every extension uses a different optimization algorithm. If two of these models had to be compared on the basis of the produced forecasts, understanding the impact of the different optimization procedures on the difference in the results would be impossible. Having a unified framework able to accommodate different models using the same optimization routine would certainly increase the robustness of the comparison. The methodology proposed in this thesis to approximate the Bayesian ETAS model have the potential to accommodate a large number of different models.

2.5 Validation

Earthquake model validation is the process of evaluating the accuracy and reliability of seismicity forecasts produced by a model against observed data. This process helps to ensure that the models are accurately representing real earthquake occurrences or to identify aspects where there could be potential improvements. Furthermore, they provide a measure of how well a model represents the earthquake-generating process, and therefore, in presence of multiple competing models, these techniques can be used to rank models in terms of agreement between forecasts and observations. The statistical tests used to verify the agreement between a forecast and the observations are called consistency tests and the output is usually binary (passed or failed). On the other hand, ranking competing models is usually done using positively (or negatively) oriented scoring rules. A scoring rule is a function of the data and the forecast, and the higher (or lower) the score the *better* the forecast. Both consistency tests and scores can be designed to target specific aspects of seismicity such as the number of events, the magnitude distribution, or the spatial distribution. Therefore earthquake model validation usually comprises the application of multiple tests/scores in order to validate earthquake models on different aspects.

In order to avoid potential biases, model validation has to be done in a prospective, or, at least, pseudo-prospective fashion. An earthquake forecasting experiment is said to be prospective if the forecast is tested against data that have yet to be recorded at the moment in which the forecast is produced. On the other hand, the experiment is said pseudo-prospective the data is divided into a training and testing catalog, the former is used to calibrate the model while the latter is used for validation. Pseudo-prospective experiments are similar to how models are evaluated by the machine learning community but they often overestimate the forecast skills because knowledge about the test data may still be incorporated into the model (even unconsciously). Therefore, the fairest way to validate the model is through fully prospective experiments, which require considerable effort to run. Indeed, forecasts need to be collected before the testing period, then a data collection period needs to pass (usually 5 or 10 years), and only afterward forecasts can be validated and compared against each other. Additionally, prospective (or pseudo-prospective) model validation can not be done just by collecting results from the literature. Indeed, differences in the data, region of interest, time scale, and reported metrics, make it impossible to have reliable and reproducible results.

In order to test the models in a rigorous, standardized, and reproducible way the Collaboratory Study for Earthquake Predictability was founded (Jordan, 2006). Having such infrastructure ignited the discussion about earthquake forecast validation and forecast formatting. It drove the development and refinement of statistical procedures for this forecast validation, and around which format should be used to represent adequately the uncertainty around the forecasted values. This section revises the metrics used provided by the pyCSEP library to validate earthquake forecasts (Savran et al., 2022). We start by defining the currently preferred forecast format within CSEP in Section 2.5.1 and introducing the relevant notation. After, the next two Sections are dedicated, respectively, to consistency tests and scores for the comparison of competing forecasts.

2.5.1 Earthquake forecasts

In the early CSEP experiments, modelers were provided with a regular space-time-magnitude grid covering a certain region and were asked to provide their estimate of the expected number of events above a certain magnitude threshold M_0 for each bin (Rhoades et al., 2011; Zechar et al., 2010b; Schorlemmer and Gerstenberger, 2007a; Werner et al., 2011). Then, models were evaluated under the assumption that the bin counts are independent, Poisson distributed, and that points in each bin are homogeneously distributed in space. This kind of forecast is known as grid-based forecast. With this format, models without a likelihood can also be evaluated with likelihood-based techniques. On the other hand, observed, as well as simulated with *self-exciting* models, bin counts are not independent (in space and time), not Poisson distributed, and not homogeneously distributed. This led to models verifying the false assumptions to be unfairly advantaged as pointed out by many authors (Harte, 2015; Werner and Sornette, 2008; Nandan et al., 2019; Lombardi and Marzocchi, 2010a).

Most recently, catalog-based earthquake forecasts are considered (Savran et al., 2020; Brehmer et al., 2021) in which each forecast is a collection of simulated catalogues ($\sim 10,000$ or $\sim 100,000$). Considering catalog-based forecasts allows for a fair evaluation of the competing forecasts because using the sampled point process directly rather than smoothing into bins removes many unnecessary assumptions and allows the application of Monte Carlo methods to estimate the distribution of quantities of interest under the model. To better describe the advantages of having catalog-based forecasts is useful to introduce some notation first.

Any experiment has its own testing region $\mathcal{R} = [0, T] \times W \times [M_0, M_c]$ where $[0, T]$ is the time domain, W is the space domain, and $[M_0, M_c]$ is the magnitude domain composed by a magnitude of completeness M_0 and a corner magnitude M_c. Therefore the generic event in $\mathcal{R}$ is defined by a time t a location $\mathbf{s}$ and a magnitude m, namely $\mathbf{e} = (t, \mathbf{s}, m) \in \mathcal{R}$. The observations against which the models will be evaluated are defined as $\Omega = \{e_i : i = 1, ..., N_{obs}, e_i \in \mathcal{R}\}$. In the same way, a synthetic catalogue is $\Lambda = \{\tilde{e}_i : i = 1, ..., N, \tilde{e}_i \in \mathcal{R}\}$. A forecast is a collection of $J \in \mathbb{N}$ synthetic catalogues and can be written as $\mathcal{F} = \{\Lambda_1, ..., \Lambda_J\}$ with $\Lambda_j = \{\tilde{e}_{ij} : i = 1, ..., N_j, \tilde{e}_{ij} \in \mathcal{R}\}$.

Any characteristic of the earthquake generation process can be seen as a measurable mapping $\mathcal{S} : \mathcal{P} \to \mathcal{B}$ from the space of possible point patterns $\mathcal{P}$ to a simpler space. The synthetics catalogues composing a forecast $\Lambda_1, ..., \Lambda_J$, and the observed catalogue Ω, all belong to the space of possible point patterns $\mathcal{P}$, and a mapping $S : \mathcal{P} \to \mathcal{B}$ is just a function of the catalogue returning a number or another function. For example, the number of earthquakes can be seen as a mapping from $\mathcal{P}$ to the set of natural numbers ($\mathcal{B} = \mathbb{N}$), so that $S(\Lambda_j) = N_j$ and $S(\Omega) = N_{obs}$. In the same way, Ripley's K-functions (Ripley, 1976, 1977) is a mapping from $\mathcal{P}$ to the space of univariate functions on the positive real line, so that $S(\Lambda_j) = K(t), t \geq 0$. Modern model testing is based on the intuitive idea of comparing the mapping calculated on the observed catalogue with the mapping calculated on the synthetics (Savran et al., 2020; Brehmer et al., 2021; Heinrich-Mertsching et al., 2021). Indeed, the values of the mapping on the synthetics can be used to empirically estimate the distribution of the mapping (the characteristic) under the model. This is used as null hypothesis, and if the mapping calculated on the observations falls in the tail of the distribution, we can reject the hypothesis that the model and the forecasts come from the same distribution (on the specific aspect under study) and consider the forecast as inconsistent with the observations at some level of confidence. Therefore, the mapping is used as test statistic, and the critical values are determined by the distribution of the test statistics provided by the model. The suite of consistency tests applied in CSEP applies the above principle to different test statistics (mappings) to study the consistency between observations and forecasts on different aspects.

I will now show how this can be applied also when two competing models need to be compared against observed data.

2.5.2 Consistency tests

Below, we describe one by one the consistency tests implemented in the pyCSEP package, namely the N-test, M-test, S-test (where N refers to number, M to magnitude, and S to space), and the Pseudolikelihood test. Besides the scores presented here many more techniques were developed to test the consistency between forecasts and observations. Remarkable examples are: error diagrams or receiver operating characteristic (ROC) curves (Swets, 1973), also known as Molchan diagrams (Molchan, 1991, 2010; Kagan, 2009) are diagnostic tools testing the performance of a model in casting alarms; residual analysis techniques for point process models such as thinning (Lewis and Shedler, 1979; Schoenberg, 2003), superposition (Brémaud, 1981), super-thinning (Clements et al., 2012), rescaling (Meyer, 1971; Schoenberg, 2004). We refer the reader to Clements et al. (2012) and Bray and Schoenberg (2013) for a review.

N-test

The N-test assesses if the forecasted number of events is consistent with the observed one. It was introduced in Kagan and Jackson (1995) and refined in Schorlemmer and Gerstenberger (2007a); Zechar et al. (2010b); Savran et al. (2020). The test statistics is the number of events per catalogue in a given window in time and space. The distribution of the number of events provided by the model is estimated empirically from $N_j, j = 1, ..., J$, and the following two quantities are calculated

$$\delta_1 = 1 - F_N(N_{obs} - 1) = \Pr(N_j \geq N_{obs}) \tag{2.22}$$

$$\delta_2 = \gamma_N = F_N(N_{obs}) = \Pr(N_j \leq N_{obs}), \tag{2.23}$$

where $F_N(\cdot)$ is the empirical predictive cumulative distribution of the number of events. The two quantities are the probability of predicting at least (δ_1) or at most (δ_2) the observed number of events N_{obs}.

The test is based on the fact that if the observed number of points is distributed accordingly to F_N, if we repeat this process over multiple independent periods of time obtaining a sequence of values of γ_N, then, the γ_N values are uniformly distributed between 0 and 1. In practice, if γ_N values calculated over multiple testing periods (like daily or monthly forecasts) are not uniformly distributed in $[0, 1]$ the model fails the test.

M-test

The M-test assesses if the forecasted magnitude frequency distribution is consistent with the observed one (Savran et al., 2020), therefore only the magnitude component is used. The test statistic relies on the magnitude domain to be divided into bins and the logarithm of bin counts is used for comparison. Using the logarithm places more weight on the bins with smaller counts which will be the ones at higher magnitudes for which it is desirable to have more weight.

The forecasted magnitude distribution is determined by merging all the magnitudes in the synthetics catalogues composing the forecast, namely

$$\Lambda_U = \{\tilde{e}_{ij}, i = 1, ..., N_j, j = 1, ..., J\}, \tag{2.24}$$

then, given a partition of magnitude domain composed by k bins, the counts per bin are calculated, $\Lambda_U^{(m)}(k)$ is the number of events in the union catalogue with magnitude falling inside the k-th bin. We refer to the total number of simulated events as $N_U = \sum_k \Lambda_U^{(m)}(k)$. The counts per magnitude bin are calculated also for each synthetic catalogue and for the observed catalogue, respectively $\Lambda_j^{(m)}(k)$ and $\Omega^{(m)}(k)$.

The idea is to compare the counts from the synthetic catalogues $\Lambda_j^{(m)}(k)$ against the union catalogue counts $\Lambda_U^{(m)}(k)$ using test statistic. This allows to estimate empirically the distribution of the test statistic according to the model, and see in which part of the distribution the test statistics calculated between the observed bin counts $\Omega^{(m)}(k)$ and the union catalogue counts falls. All the bin counts are normalized such that the sum of the bin counts is equal to the number of observations. The test statistics is the sum of the square differences between the logarithm of the normalized bin counts, for the observed catalogue is

$$d_{obs} = \sum_k \log\left(\frac{N_{obs}}{N_U}\Lambda_U^{(m)}(k) + 1\right) - \log\left(\Omega^{(m)}(k) + 1\right). \tag{2.25}$$

The test statistic for the j-th synthetic catalogue is

$$D_j = \sum_k \log\left(\frac{N_{obs}}{N_U}\Lambda_U^{(m)}(k) + 1\right) - \log\left(\frac{N_{obs}}{N_j}\Lambda_j^{(m)}(k) + 1\right). \tag{2.26}$$

Unity is added to each bin to avoid zero-count bins, for which the test statistic is $\mp\infty$. We can use the sequence $D_j, j = 1, ..., J$ to estimate the distribution of the test statistic under the model $F_D(\cdot)$, and calculate the quantile score

$$\gamma_M = F_D(d_{obs}) = \Pr(D_j \leq d_{obs}). \tag{2.27}$$

As for the N-test, the values of γ_M for different independent observed catalogues are uniformly distributed if the observations and the forecasts are consistent with each other.

S-test

The S-test assesses if the spatial distribution of the forecasts is consistent with the observed one. Similarly to the magnitude test is based on the discretization of the domain (space in this case) in bins. The test statistic is based on the spatial distribution of the expected rates. For each bin b_k, the average number of events per bin per catalogue is calculated, $\tilde{\lambda}(b_k)$ which can be seen as the approximate mean rate per bin provided by the forecast. Then, it is normalized to

$$\tilde{\lambda}^*(b_k) = \frac{\tilde{\lambda}(b_k)}{\sum_k \tilde{\lambda}(b_k)}. \tag{2.28}$$

The quantity $\tilde{\lambda}^*(b_k)$ approximates the probability of having an event in bin b_k under the model. Therefore, it can be used as a likelihood to calculate a statistic. For the observed catalogue Ω composed by events $\mathbf{e}_1, ..., \mathbf{e}_{N_{obs}}$ we call the bin in which each observation falls $b_1, ..., b_{N_obs}$. Then, the test statistic calculated on the observed catalogue is

$$S_{obs} = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} \log \tilde{\lambda}^*(b_i), \tag{2.29}$$

where $\tilde{\lambda}^*(b_i)$ is the normalized approximated forecasted rate in bin b_i where the i-th observation has fallen. Essentially S_{obs} can be interpreted as the average probability at which observed events would have occurred under the forecasting model. The same quantity is calculated for each synthetic catalogue

$$S_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \log \tilde{\lambda}^*(b_{ij}), \tag{2.30}$$

where b_{ij} is the bin containing the synthetic event $\tilde{e}_{ij}$.

As before, the distribution of the test statistic $F_S(\cdot)$ can be approximated empirically from $S_1, ..., S_J$, and the following quantile score calculated

$$\gamma_S = F_S(S_{obs}) = \Pr(S_j \leq S_{obs}), \tag{2.31}$$

and, when γ_S is computed for multiple independent observed catalogues, the consistency of the forecast and the observations can be assessed by testing the uniformity of the quantile scores γ_S for the different catalogues.

Pseudolikelihood-test

The Pseudolikelihood-test assesses the consistency between forecast and observations in general, without focusing on a specific aspect. Being a likelihood test, it accounts for all aspects of the distribution of the events, i.e number, spatial, and magnitude. The first version of this test was the L-test as presented in Schorlemmer and Gerstenberger (2007a); Rhoades et al. (2011). The L-test however, was relying on the Poisson log-likelihood and penalized models with non-Poissonian event distribution. The present test relies on the use of the more general point process log-likelihood (Daley and Vere-Jones, 2004). The log-likelihood of a point process model with conditional intensity $\lambda(\mathbf{e}|\mathcal{H}_t)$ having observed $\Omega = \{\mathbf{e}_i \in \mathcal{R}, i = 1, ..., N_{obs}\}$ is given by

$$\mathcal{L} = -\int_{\mathcal{R}} \lambda(\mathbf{x}|\mathcal{H}_t) d\mathbf{x} + \sum_{i=1}^{N_{obs}} \log \lambda(e_i|\mathcal{H}_{t_i}). \tag{2.32}$$

The idea is to consider an approximate conditional intensity function that can be calculated from the set of synthetic catalogues. This plays the role of the forecasted conditional intensity. Then, the log-likelihood can be calculated for the observed catalogue and for all the synthetics, and, as before, compared.

More formally, to calculate the log-likelihood we need to approximate the two log-likelihood components: the integral and the summation. The integral is the expected number of points by the model and can be easily approximated by the average number of synthetic events per catalogue $\bar{N} = \sum_j N_j/J$. The elements of the summation represent the expected rate at which points are supposed to occur at the observed location. They can be approximated by dividing the entire space-time-magnitude region in bins and calculating the average number of synthetic events per catalogue per bin. This gives us a discrete representation of $\lambda(\S|\mathcal{H})$ that can be calculated also for models which do not present an explicit likelihood.

As before, we indicate with $b_1, ..., b_{Nobs}$ the space-time-magnitude bin in which observation e_i, then the test statistic for the observed catalogue is

$$\tilde{\mathcal{L}}_{obs} = -\bar{N} + \sum_{i=1}^{N_{obs}} \log \tilde{\lambda}(b_i), \tag{2.33}$$

where $\tilde{\lambda}(b_i)$ is the average number of synthetic events in the i-th space-time-magnitude bin where observation $\mathbf{e}_i$ have fallen.

Similarly, the test statistics can be calculated for each synthetic catalogue as

$$\tilde{\mathcal{L}}_j = -\bar{N} + \sum_{i=1}^{N_j} \log \tilde{\lambda}(b_{ij}), \tag{2.34}$$

where b_{ij} is the bin containing the synthetic event $\tilde{\mathbf{e}}_{ij}$.

As usual, the sequence of values of $\tilde{L}_j, j = 1, ..., J$ is used to estimate empirically the distribution of the test statistics under the model $F_L(\cdot)$ and the quantile is calculated

$$\gamma_L = F_L(\tilde{L}_{obs}) = \Pr[\tilde{L}_j < \tilde{L}_{obs}]. \tag{2.35}$$

If the model and the observations are coherent, if tested over multiple independent catalogues, the values of γ_L will be uniformly distributed. Any departure from uniformity can be seen as an inconsistency between the model and the observations. Thus all four tests result in a uniformly distributed metric if the model is consistent with the observations. This rises the question - how do we measure the departure of the metrics from a uniform distribution for different models, and decide which is best? This is done by a scoring rule, as described in the next section.

2.5.3 Comparison scores

The comparison between models is usually carried out with the use of a scoring function. A scoring function is just a function of the data and forecasts can be ranked according to its values. Scoring functions, as the test statistics seen in the previous section, target a specific property of the forecast and assign to each competing model a number based on the similarity between forecasts and observations. To be effective the scoring rule has to possess some statistical properties, the most important being consistency and properness (Gneiting and Raftery, 2007). In Chapter 7 we define formally when a score is proper and explore the consequences of using an improper scoring rule. Here, it is sufficient to say that, for positively oriented scoring rules (the higher, the better) a scoring function is consistent for a property if the expected score with respect to the data-generating model is maximized (minimized if negatively oriented) by the value of the property calculated for the data-generating model. A scoring rule is said to be proper if the expected score with respect to the data-generating model is maximized (minimized) by the data-generating model. In essence, they assure that a model will get on average the highest (lowest) score if tested against data simulated with the model. Scoring rules that are consistent and proper can be used for a variety of tasks other than forecasts comparison, indeed they can be employed in regression and M-estimation (Gneiting, 2011; Fissler and Ziegel, 2016), as loss functions to estimate the value of tuning parameters (Steinwart et al., 2014; Frongillo and Kash, 2021) or to calculate the weights of an ensemble model (Marzocchi et al., 2012). We refer to Brehmer et al. (2021) for an extensive review of the topic and on the link between the consistency tests introduced above and the theory of scoring rules.

A common approach is to use the log-likelihood of the model to build scoring functions. This led to the application of information criteria such as the Akaike Information Criterion (AIC, Akaike (1974) to compare earthquake models (Ogata, 1998, 1999; Bayliss et al., 2020). Information criteria such as AIC are usually composed of a goodness-of-fit measure (the likelihood) and a penalty (in the case of AIC the number of free parameters). They are thought for retrospective testing, in which models are tested against the same data used to estimate the parameters. The penalty component is there to prevent overfitting models to get the highest score. However, this makes the scoring rule not consistent and therefore not ideal for prospective testing (Brehmer et al., 2021). In the bayesian framework, the Bayes factor is usually employed to compare models. The Bayes factor indicates which model is more likely after the data has been observed. It also heavily relies on the log-likelihood of the model, see Marzocchi et al. (2012) and Bray and Schoenberg (2013) for applications. Also, residual methods can be used for models comparison (Baddeley et al., 2005; Clements et al., 2012). The approach currently used within CSEP is to compare the models in terms of Information Gain (Vere-Jones, 1998; Rhoades et al., 2011), and therefore, we are going to describe only this one.

Information Gain

The Information Gain (IG) approach was first proposed by Vere-Jones (1998) for temporal models and extended by Zechar et al. (2010b) and finalized by Rhoades et al. (2011). The IG approach compares two models in terms of log-likelihood given the same observed data. Intuitively, the log-likelihood measures how likely is the observed point pattern under the model, and therefore, the model with the highest log-likelihood should be preferred. As before, given a partition of the domain $\mathcal{R}$ in non-overlapping bins, a testing catalogue $\Omega = \{\mathbf{e}_i \in \mathcal{R}, i = 1, ..., N_{obs}\}$, and two competing forecasts $\mathcal{F}_A = \{\Lambda_{A1}, ..., \Lambda_{AJ}\}$ and $\mathcal{F}_B = \{\Lambda_{B1}, ..., \Lambda_{BJ}\}$, we can estimate the average rate per bin provided by the two forecasts $\tilde{\lambda}_A(b_i)$ and $\tilde{\lambda}_B(b_i)$ for $i = 1, ..., k$. Then, the pseudolikelihood in equation 2.33 can be used to approximate the log-likelihood assigned by each forecast to the observations and the difference can be considered, namely

$$R_{obs} = -(\bar{N}_A - \bar{N}_B) + \sum_{i=1}^{N_{obs}} \log \tilde{\lambda}_A(b_i) - \log \tilde{\lambda}_B(b_i), \tag{2.36}$$

where $\bar{N}_A$ and $\bar{N}_B$ are the average numbers of events per catalogue provided by the forecasts and b_i, $i = 1, ..., N_{obs}$ are the bins where the observed events $\mathbf{e}_i$ falls.

Then, Rhoades et al. (2011) propose the R-test, in which we have to calculate the test statistics R_{Aj} and R_{Bj}, with $j = 1, ..., J$, using the synthetic catalogues composing the two forecasts. Doing this provides two sequences of test statistics from which we can empirically estimate the distribution of the test statistics under the two models $F_{R_A}(\cdot)$ and $F_{R_B}(\cdot)$. If R_{obs} lies in the tail of the distribution $F_{R_A}(\cdot)$ then, model A is considered worse than model B, and vice-versa. As noted in Rhoades et al. (2011) the R-test is more a consistency test than a way to compare models because does not provide information on which model has the highest likelihood or if the difference is significantly different from zero. Also, it can lead to contradictory results, i.e model A is preferred when using $F_{R_A}(\cdot)$ and model B is preferred when using $F_{R_B}(\cdot)$ (Bray and Schoenberg, 2013; Brehmer et al., 2021), as well as models mutually rejecting each other (Gerstenberger et al., 2009).

Consequently, Rhoades et al. (2011) provides two modifications to circumvent the above problems. They proposed the T-test and W-test, both based on the Information Gain per earthquake (IGPE) (Harte and Vere-Jones, 2005) which is given by

$$I_{N_{obs}}(A, B) = -\frac{\bar{N}_A - \bar{N}_B}{N_{obs}} + \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} \log \tilde{\lambda}_A(b_i) - \log \tilde{\lambda}_B(b_i) = \frac{R_{obs}}{N_{obs}}. \tag{2.37}$$

The idea is to test whether $I_{N_{obs}}(A, B)$ is significantly different from zero or not. This approach is rooted in the view that $I_{N_{obs}}(A, B)$ is an unbiased estimate of the true average IGPE $I(A, B)$ and that $I_{N_obs}(A, B) \to I(A, B)$ as $N_{obs} \to \infty$. Assuming a distribution for

$I(A,B)$ it is possible to test whether, based on the sequence $\log\tilde{\lambda}_A(b_i) - \log\tilde{\lambda}_B(b_i)$, the IGPE is significantly different from zero or not. The T-test and the W-test differ in the distribution assumed for $I(A,B)$

More formally, the T-test assumes that the differences $\log\tilde{\lambda}_A(b_i) - \log\tilde{\lambda}_B(b_i)$ are independent samples from a Gaussian distribution with mean μ and variance σ^2. This implies that $I_{N_{obs}}$ has a Gaussian distribution with mean $\mu - (\bar{N}_A - \bar{N}_B)/N_{obs}$ and variance σ^2/N_{obs}. Then, a classic paired t-test (Student, 1908) can be used to test the null hypothesis $\mu - (\bar{N}_A - \bar{N}_B)/N_{obs} = 0$ versus the alternative $\mu - (\bar{N}_A - \bar{N}_B)/N_{obs} \neq 0$. This requires the variance σ^2 to be estimated from the data, (Rhoades et al., 2011) proposes to use the following estimator

$$\tilde{\sigma}^2 = \frac{1}{N_{obs}-1}\sum_{i=1}^{N_{obs}}\left(\log\tilde{\lambda}_A(b_i) - \log\tilde{\lambda}_B(b_i)\right)^2 + \\ + \frac{1}{N_{obs}^2 - N_{obs}}\left(\sum_{i=1}^{N_{obs}}\log\tilde{\lambda}_A(b_i) - \log\tilde{\lambda}_B(b_i)\right)^2 . \tag{2.38}$$

Under the hypothesis that $I(A,B) = 0$ the quantity $T = I_{N_{obs}}(A,B)/(\tilde{\sigma}/\sqrt{N_{obs}})$ follows a t-student distribution with $N_{obs} - 1$ degrees of freedom, and the null hypothesis can be rejected if $|T|$ exceed a certain quantile of the t-student distribution, t_q. In the same way, this allows to construct confidence intervals for the score difference as $I_{N_{obs}}(A,B) \mp t_q \tilde{s}igma/\sqrt{N}$ on which a decision rule can be constructed following the method illustrated in Chapter 7.

The W-test was developed to relax the assumption that the differences $\log\tilde{\lambda}_A(b_i) - \log\tilde{\lambda}_B(b_i)$ are normally distributed. Indeed, it works as the T-test but applies a Wilcoxon signed-rank test (Wilcoxon, 1992) instead of a paired t-test. The W-test verifies, in a non-parametric fashion, the hypothesis that the median of the IGPE is significantly different from zero and does not assume symmetry in the distribution of $\log\tilde{\lambda}_A(b_i) - \log\tilde{\lambda}_B(b_i)$. Both the T-test and W-test can be expressed as modifications of the Diebold-Mariano test (Diebold and Mariano, 2002; Diebold, 2015), we refer to Brehmer et al. (2021) for more details on the connection with scoring theory. This terminates the review of earthquake forecast validation techniques.

2.6 Conclusion

In this chapter, we have described earthquake catalogues, the difficulties that they bring, how to model earthquake occurrence using point process models such as the ETAS model, how catalogue-based forecasts are defined, and how to validate them in the light of observed data.

In this thesis, I propose a novel Bayesian technique to approximate the ETAS model. This technique is not limited to the ETAS model but could be applied (in principle) to any Hawkes process model. The technique is described in Chapter 4 where it is compared with a MCMC alternative. There, I show that it can be up to 10 times faster than the latter on catalogues with more than 2000 events, and that scales more efficiently increasing the amount of data. Chapter 5 shows in detail the difficulties linked to the quality of the data used to estimate the parameters. In there, I investigate how the parameters' posterior distributions are influenced by the number of large events in the sequence if including quiet periods yields more robust parameter estimates. In the same way, I investigate if including a period of pre-conditioning (events that are accounted in the history but not in the set of events on which the likelihood is calculated) helps. Finally, I investigate the effect of missing data on the posterior distributions of parameters. This is done using the temporal ETAS model.

Chapter 6 generalizes the proposed approach to the spatio-temporal case with a spatially varying background rate and covariates described in Section 2.4.2. In there, we compare retrospectively models with different combinations of covariates and rank them using the Akaike Information Criterion. Retrospective forecasts produced with our approach for the 2009 L'Aquila sequence and the 2016 Amatrice sequence that happened in central Italy will be used in a novel study on the feasibility of real-time rapid loss forecasting. Moreover, forecasts produced with this approach will be submitted to the next Italian CSEP experiments where they will be evaluated prospectively with the validation metrics described in this chapter. For both forecasts, I consider a Tapered GR law to model the magnitude distribution.

The validation metrics used within CSEP needs to be validated themselves, at least to check if they are biased or to estimate the power of statistical test as the one described in this chapter based on these quantities. For example, we have found that the M-test is problematic when applied to forecasts overestimating the number of events. More details on the issue can be found at https://github.com/SCECcode/pycsep/issues/196. I have proposed a solution and also designed a modified multinomial likelihood score test that does not suffer from this problem and is more powerful than the M-test. We plan to write an article assessing the performance of the M-test similarly to what is done in Khawaja et al. (2023) for the S-test. We plan to use a simulation technique similar to the one described in Chapter 7 for the Parimutuel Gambling score.

The next chapter introduces the Bayesian methodology, namely the *Integrated Nested Laplace Approximation* (INLA, Rue et al., 2009) which is the base of our method.

Chapter 3

Methodology

3.1 Introduction

Bayesian methods have seen increasing popularity in the last decades, especially due to the development of efficient Markov Chain Monte Carlo (MCMC) schemes (Robert et al., 1999) to obtain samples from the posterior distribution without explicitly calculating it and to software packages such as `WinBUGS` (Lunn et al., 2000), `JAGS` (Plummer et al., 2003), and `Stan` (Carpenter et al., 2017) which gives the possibility to use such schemes without having to explicitly code them. Bayesian methods are appealing in many fields of science, they provide a complete description of the uncertainty about the parameters, specifically, they provide the posterior distribution of the parameters combining the likelihood of the observed data and a prior distribution that elicits the state of our knowledge about the parameters before the data is observed. Thus, they provide a clean and clear way to merge what we know about the phenomenon under study before an experiment and the information provided by the experiment itself. This mechanism of updating our knowledge, which resembles how humans learn things, is particularly appealing for forecasting purposes.

As such, Bayesian methods have seen increasing popularity also in seismology (Holschneider et al., 2012; Shcherbakov, 2014; Omi et al., 2015), however, an efficient, general, extendible Bayesian framework is still missing. This is mainly due to the fact that, the ETAS model has all the characteristics to make MCMC methods inefficient, i.e. complex likelihood preventing analytical results, highly correlated parameters, non-Markovianity. Indeed, early studies had to resort to frequentist-style estimation techniques (Ebrahimian et al., 2014; Omi et al., 2015). First attempts to develop a fully Bayesian framework are Vargas and Gneiting (2012) and Ebrahimian and Jalayer (2017) which used the estimation technique introduced by Rasmussen (2013), however, their methods provide biased estimates and did not scale well when the number of events were increased. A new framework mitigating these problems was developed by Ross (2021), which uses a latent variable formulation, and has been extended to have a spatially varying background rate (Molkenthin et al., 2022; Ross and Kolev, 2022). However, the proposed method still does not scale efficiently with the number of events, especially in the most complicated cases. In Chapter 4, we show that our method is 10 times faster than the MCMC method developed by Ross (2021) for a simple temporal ETAS model using catalogs with more than 2000 events.

Alternative methods to MCMC exist, and we should use them whenever MCMC methods are inappropriate. The Integrated Nested Laplace Approximation (INLA, Rue et al., 2009; Bakka et al., 2018) is one of these. INLA was designed to handle efficiently large Latent Gaussian models with strongly correlated parameters and has been used in many applied fields such as air pollution (Forlani et al., 2020), disease mapping (Riebler et al., 2016; Santermans et al., 2016; Schrödle and Held, 2011a,b), genetics (Opitz et al., 2016), public health (Halonen et al., 2015), ecology (Roos et al., 2015; Teng et al., 2022), more examples can be found in Bakka et al. (2018); Blangiardo et al. (2013); Gómez-Rubio (2020). INLA is based on a deterministic approximation of the posterior distribution which makes it substantially faster than MCMC methods. Also, being deterministic, any result is easier to reproduce on different machines. The `inlabru` R-package (Bachl et al., 2019) facilitates the use of INLA and extends its capability to more complex models, such as LGCP which has already been applied to seismic data for time-independent models (Bayliss et al., 2020, 2022). Furthermore, the `inlabru` package uses an iterative method to handle more complex models. This iterative method will be heavily used in the approximation of Hawkes process models presented in Chapter 4.

This chapter describes the Integrated Nested Laplace Approximation (INLA) and the `inlabru` iterative method. The first section describes the INLA algorithm starting from the definition of a Latent Gaussian model, the Laplace approximation, and how this is used to retrieve the posterior of the parameters. The second section is about `inlabru` describing how spatial LGCP models are implemented when including a particularly relevant type of random effect and then, the iterative method. Having a clear idea of INLA and `inlabru`, and how they work together is essential to understand better the Hawkes process approximation technique described in the next chapter.

3.2 Integrated Nested Laplace Approximation (INLA)

3.2.1 Latent Gaussian models

The class of Latent Gaussian models (LGMs) is a flexible and powerful class of statistical models, particularly useful when the aim is to describe the dependency between observations and covariates and between observations themselves. Rue et al. (2017) reports a long list of successful applications of Latent Gaussian models implemented with R-INLA. The components of such models form a three-stage hierarchy composed by: the observations ($\mathbf{y} \in \mathbb{R}^n$), the latent field ($\mathbf{x} \in \mathbb{R}^m$), and the hyper-parameters ($\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^p$), with $n, m, p \in \mathbb{N}$. In this context, $\boldsymbol{\theta}_1$ are parameters determining the distribution of the data, while $\boldsymbol{\theta}_2$ are hyperparameters determining the distribution of the latent field. From now on $\pi(\cdot)$ will be used to generically represent distributions and $\pi(\cdot|\cdot)$ for conditional distributions. The hierarchy is then,

$$\mathbf{y}|\eta(\mathbf{x}), \boldsymbol{\theta}_1 \sim \prod_i \pi(y_i|\eta(\mathbf{x}), \boldsymbol{\theta}_1) \tag{3.1}$$

$$\mathbf{x}|\boldsymbol{\theta}_2 \sim N(\mu(\boldsymbol{\theta}_2), Q^{-1}(\boldsymbol{\theta}_2)) \tag{3.2}$$

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}). \tag{3.3}$$

The vector $\mathbf{x}$ contains all the unobservable quantities also called the *latent field*, which is supposed to have a multivariate Gaussian distribution with mean function $\mu(\boldsymbol{\theta}_2)$ and covariance function $Q^{-1}(\boldsymbol{\theta}_2)$, so that $Q(\boldsymbol{\theta}_2)$ is the precision matrix. The latent field includes the effect of the covariates and possible random effects which influences the observations. The observations depends on the latent field only through the function $\eta(\mathbf{x})$ which usually represents the expected value of $\mathbf{y}$ given the latent field $\mathbf{x}$. The hyper-parameters of the model are $\boldsymbol{\theta}$ which is divided into hyper-parameters of the likelihood $\boldsymbol{\theta}_1$ and hyper-parameters of the latent field $\boldsymbol{\theta}_2$. Assuming to have observed $\mathbf{y} = y_1, \ldots, y_n$, with $n \in \mathbb{N}$, in Bayesian statistics we are interested in the conditional distribution of the latent field and the hyper-parameter given the observations $\mathbf{y}$. This is called the posterior distribution and it is given by

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}_2)\prod_i \pi(y_i|\mathbf{x}, \boldsymbol{\theta}_1), \tag{3.4}$$

where $\pi(\boldsymbol{\theta})$ and $\pi(\mathbf{x}|\boldsymbol{\theta}_2)$ are, respectively, the prior of the hyper-parameters and the prior of the latent field. The product of $\pi(y_i|\mathbf{x}, \boldsymbol{\theta}_1)$ is the likelihood of the observed sample.

The likelihood synthesizes the information coming from the data and reflects our hypotheses on the observed variable, e.g. is the distribution symmetric around the mean? Is the mean comparable with the variance? Is the distribution heavy-tailed or not? The above formulation assumes that observations are independent of each other conditionally on the latent field and the hyper-parameters. In fact, the role of the latent field is to explain the dependence between observations which typically depends on the observations having similar levels of the covariates or similar spatio-temporal locations. For example, consider a region divided into small non-overlapping bins, $\mathbf{y}$ represents the number of earthquakes in each bin during a defined spatio-temporal window. For spatially varying time-independent models, given the clustering (in space) nature of the earthquake generation process, if the bin i and j are close enough, one would expect y_i and y_j to be correlated. However, the observed counts are generated by the same *latent* mechanism, let's say the movements of the tectonic plates, which will be described by the latent field $\mathbf{x}$. The likelihood of each observation y_i depends only on the corresponding x_i, and the dependence between y_i and y_j is then described by the dependence between x_i and x_j. In other words, knowing the underlying process makes the observations independent. For time-dependent models using real earthquakes data the situation is more complicated because the presence of large events could make events in their neighbours undetectable, and at the same time increases the number of events in its surroundings.

The priors, instead, should reflect our knowledge of the process that we have before running any experiment. This usually comes from previous studies or expert knowledge. It also reflects our hypotheses on the processes generating the observations such as: What is the domain of the parameters? Are they correlated? Do they vary over space or time? For LGMs the latent field has a Gaussian distribution and, from now on, we will always assume it is a Gaussian Markov Random Field (GMRF, Rue et al., 2009).

A GMRF can be defined through two functions $\mu(\boldsymbol{\theta}_2)$ (mean function) and $Q^{-1}(\boldsymbol{\theta}_2)$ (covariance function). A GMRF is completely specified by these two functions, in fact, if

$$\mathbf{x}|\boldsymbol{\theta}_2 \sim \mathit{GMRF}(\mu(\mathbf{x}, \boldsymbol{\theta}_2), Q(\mathbf{x}, \boldsymbol{\theta}_2)),$$

the vector $\mathbf{x}$ has a multivariate Gaussian distribution with mean vector $\mu(\mathbf{x}, \boldsymbol{\theta}_2)$ and precision matrix $Q(\mathbf{x}, \boldsymbol{\theta}_2)$. We use the precision matrix instead of the covariance matrix because it comes with the following useful property: if the elements i and j are conditionally independent, then the component ij of the precision matrix is zero ($Q_{ij} = 0$). Therefore, the dependency structure between elements of the latent field is entirely specified by the non-zero elements of the precision matrix.

We use additive models to give an example of the components usually included in the latent field. For additive models, the observations y_i depend on the latent field $\mathbf{x}$ only through the linear predictor η_i. The linear predictor is usually the linear combination of covariates and random effects, formally,

$$\eta_i = \mu + \sum_{j \in \mathcal{J}} \beta_j z_{ij} + \sum_{k \in \mathcal{K}} f_{k, j_k(i)} + \epsilon_i, \tag{3.5}$$

where $\mu \in \mathbb{R}$ is the overall intercept, $z_{ij} \in \mathbb{R}$ is the value of the j-th covariate for observation i with coefficient β_j to be estimated. We refer to those as *fixed effects* because depends on known covariates and are the same as the effects founded in classical linear regression problems. The terms $\mathbf{f}_k, k \in \mathcal{K}$ represent, instead, *random effects*, and they are assumed to be Gaussian processes. The values $f_{k,j_k(i)}$ are the components of $\mathbf{f}_k$ influencing the i-th observations. Being a Gaussian process, $\mathbf{f}_k = f_{k,1}, ..., f_{k,n}$ have a multivariate Gaussian distribution with given mean and covariance function. Examples of model components $\mathbf{f}_k$ include auto-regressive models, stochastic spline models, models for smoothing, and random effects models with different types of correlations.

The variety of possible alternatives for this model component reflects the flexibility of this class of models. The quantity ϵ_i is a small error component. The latent field in the general is

$$\mathbf{x} = (\boldsymbol{\eta}, \mu, \boldsymbol{\beta}, \mathbf{f}_1, \mathbf{f}_2, ...). \tag{3.6}$$

The components $\mathbf{f}_k$ are usually represented by a vector of length m_k at which the component is evaluated. In the case of spatial random effect, the component $\mathbf{f}_k$ may be the spatial field evaluated at a set of fixed locations m_k. The dimension of the latent field is given by the number of observations plus the number of the fixed effects and intercept, plus the number of the m_k's. This number (especially for spatial models) is usually between 10^3 and 10^5 in real data applications. The large dimension of the latent field and the correlation structure in each component usually prevent MCMC methods to be practical. I'll present another example in section 3.3.1.

3.2.2 Laplace Approximation

The Laplace approximation was introduced by Pierre-Simon Laplace (1774) and has been used for centuries to approximate integrals. In modern times, the Laplace approximation was one of the main tools to evaluate high-dimensional integrals in pre-MCMC times, however, it was quickly replaced once computers became fast enough to mitigate the high computational cost of MCMC. The Laplace approximation is meant to evaluate integrals of the form

$$I_n = \int_{\mathcal{X}} \exp\{nf(x)\}dx, \tag{3.7}$$

as $n \to \infty$. The function $f(\cdot)$ is assumed to be smooth in the sense that the first and second derivative exists and to have a unique global maximum. We assume that $f(\cdot)$ is a function of a scalar x, but what is said can be extended to the case in which x is a vector. Further, assuming $x_0 = \text{argmax}_x f(x)$, the second order Taylor expansion of $f(x)$ around x_0 is

$$\begin{aligned} I_n &\approx \int_{\mathcal{X}} \exp\{n(f(x_0) + \frac{1}{2}(x-x_0)^2 f''(x_0)\}dx \\ &= \exp\{nf(x_0)\}\sqrt{\frac{2\pi}{-nf''(x_0)}}, \text{ as } n \to \infty \end{aligned} \tag{3.8}$$

where x_0 is the maximum of $f(x)$ for $x \in \mathcal{X}$, $f''(\cdot)$ is the second derivative of $f(\cdot)$; the first derivative is zero when evaluated at x_0. This shows that the integral of $\exp\{nf(x)\}$ can be approximated by the integral of a Gaussian density matching the value and curvature of the original function at the mode. The more the function is *close* to be Gaussian the more the Laplace approximation is accurate. For our purposes, the function $nf(x)$ will be interpreted as a sum of log-likelihoods. In this case, if the central limit theorem holds, the Gaussian approximation matches perfectly the objective function as n goes to infinity. Following this interpretation, we will refer to the quantity n as the number of observations, this allows us to reformulate the sentence above as "the accuracy of the approximation increases as the number of observations increases". Keeping the parallelism with the central limit theorem, fixing the number of observations n the approximation will be more accurate if the curve $\exp\{nf(x)\}$ satisfies typical Gaussian properties such as uni-modality, symmetry, and tail behavior.

The efficiency of the INLA method is based on the fact that the Laplace approximation is deterministic and fast to compute once we have the mode. Hence, the most computationally expensive steps are i) find the mode, and ii) computing the log-determinant of the Hessian (in the multivariate case). For the first problem, INLA uses a gradient base method, which requires the Hessian to be calculated, and this is usually retrieved computing the Cholesky factorization of the Hessian which decompose the matrix in the product of a lower triangular matrix and its transpose. This speeds up the calculations because the Cholesky factor is now a sparse matrix, also, this reduces also significantly the time needed to compute the log-determinant of the Hessian which can also be expressed in terms of the same Cholesky factor which needs to be calculated only one time. When the function of interest is multidimensional or multi-modal (or both), determining the mode could be problematic, however, the original integration problem is now a maximization problem that is more much manageable and faster to solve.

For our purposes, the Laplace approximation will be used to compute marginal distributions from the joint distribution. Assuming to know the joint distribution $\pi(\boldsymbol{\theta})$ and to be interested in the marginal distribution of the first component $\pi(\theta_1)$, formally

$$\pi(\theta_1) = \frac{\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}_{-1}|\theta_1)}, \tag{3.9}$$

where $\pi(\boldsymbol{\theta}_{-1}|\theta_1)$ is the joint distribution of the $\boldsymbol{\theta}$ components except the first one conditional on the first one. Tierney and Kadane (1986) showed that approximating the denominator of Equation 3.9 with a Gaussian distribution is equivalent to the Laplace approximation of the marginal, namely

$$\pi(\theta_1) \approx \left. \frac{\pi(\boldsymbol{\theta}}{\pi_G(\boldsymbol{\theta}_{-1}; \mu(\theta_1), Q(\theta_1)} \right|_{\boldsymbol{\theta}_{-1}=\mu(\theta_1))}, \tag{3.10}$$

where $\pi_G(\cdot, \mu, Q)$ is a multivariate Gaussian density with mean μ and precision matrix Q. This approximation, however, gives the approximated value for only one value of θ_1; if are interested in a large number of θ_1 values, we have to solve as many maximization problems. Before going there, however, we give more details on the actual approximation used by INLA.

3.2.3 Integrated Nested Laplace Approximation

Here, I describe the classic INLA model formulation as originally proposed by Rue et al. (2009), in which the predictors are part of the latent field. Recent developments (van Niekerk and Rue, 2021; Van Niekerk et al., 2023) have shown that the INLA framework can be rewritten without the nested component, by using a Variational Bayes correction to the Gaussian approximation, and that this provides faster inference, improved numerical stability, and scalability. Since November 2022 version 22.11.06 the modern INLA model formulation has become the default used by the `R-INLA` package on which the implementation proposed in this thesis relies. The method proposed in this thesis works with both formulations and the advantages of the new formulation for our problem have yet to be quantified properly, therefore I describe only the classical INLA formulation.

Any Bayesian analysis has the purpose of retrieving the marginal posterior distributions of the latent field and the hyper-parameters, namely $\pi(\theta_j|\mathbf{y})$ for all $\theta_j \in \boldsymbol{\theta}$ and $\pi(x_j|\mathbf{y})$ for all $x_j \in \mathbf{x}$, where $\mathbf{y}$ represents the observed data. Once we have the marginal posterior distributions we can compute all the quantities of interest such as mean, median, and quantiles. The INLA approach is tailored to Latent Gaussian models in which $\boldsymbol{\theta}$ is low dimensional (usually between 3 and 5, never greater than 20), and $\mathbf{x}$ is a GMRF, the observations $\mathbf{y}$ are conditionally independent given the field $\mathbf{x}$ and hyper-parameters $\boldsymbol{\theta}$ and each y_i depends only on the one element of the field x_i.

The analytical expression of the marginal posterior distributions that we want to retrieve is

$$\pi(\theta_j|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j} \tag{3.11}$$

$$\pi(x_j|\mathbf{y}) = \int \pi(x_j|\boldsymbol{\theta},\mathbf{y})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}. \tag{3.12}$$

The idea behind the INLA methodology is to approximate $\pi(\theta_j|\mathbf{y})$ and $\pi(x_j|\mathbf{y})$ by approximating $\pi(x_j|\boldsymbol{\theta},\mathbf{y})$ and $\pi(\boldsymbol{\theta}|\mathbf{y})$ in equation 3.11 and 3.12. We use the notation $\tilde{f}(\cdot)$ to indicate an approximation of the function $f(\cdot)$.

The first step is to approximate $\pi(\boldsymbol{\theta}|\mathbf{y})$ which appears in both equations 3.11 and 3.12. This is given by

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\theta)\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}{\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}, \tag{3.13}$$

which can be approximated using the Laplace approximation as done in equation 3.10. This leads to the following approximate joint $\boldsymbol{\theta}$ posterior

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) = \left.\frac{\pi(\theta)\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}{\pi_G(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})}\right|_{x=x_0(\boldsymbol{\theta})}, \tag{3.14}$$

where $\pi_G(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ is the Gaussian approximation of $\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ with mean $\mathbf{x}_0(\boldsymbol{\theta})$ and variance depending on the second derivative of $\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ calculated at $\mathbf{x}_0(\boldsymbol{\theta})$. The function in 3.14 is then normalized to be a valid density distribution. The value $\mathbf{x}_0(\boldsymbol{\theta})$ is the mode of the full conditional $\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ for a given value of the hyper-parameters $\boldsymbol{\theta}$. The above approximation is likely to be accurate because $\mathbf{x}|\boldsymbol{\theta}$ is a GMRF. Therefore, it has a Gaussian distribution, and conditioning on the data will not dramatically change the shape of the distribution and we can expect $\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ to be close to a Gaussian. The only drawback here is that if we are interested in evaluating the approximation for k different sets of hyper-parameters $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ this would require solving k maximization problems to find the corresponding $\mathbf{x}_0(\boldsymbol{\theta}_i)$.

The next step is to approximate $\pi(x_j|\boldsymbol{\theta},\mathbf{y})$. Three methodologies were available in `R-INLA` R-pacakge to perform this step, now the default is the methodology proposed by Van Niekerk et al. (2023). The fastest between the three is to marginalize the Gaussian approximation $\pi_G(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ with respect to x_j to retrieve the marginal distribution. The only extra cost of this operation is to compute the marginal variance, which is frequently not very expensive. Despite this method working fine in many cases, there are situations in which it may be not accurate, for example in the cases of highly skewed or highly asymmetric posterior distributions. In such cases, two other approximations have been implemented, details about these methods can be found in Martins et al. (2013). Once a method has been chosen, $\pi(x_j|\mathbf{y})$ is approximated by numerically integrating an approximate version of the function in equation 3.12 with respect to $\boldsymbol{\theta}$. This is done considering k values $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ with weights $w_1, ..., w_k$ and

$$\tilde{\pi}(x_j|\mathbf{y}) = \sum_i \tilde{\pi}(x_j|\boldsymbol{\theta}_i, \mathbf{y})\pi(\boldsymbol{\theta}_i|\mathbf{y})w_i. \tag{3.15}$$

The choice of $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k$ is crucial for the efficiency and reliability of the INLA methodology. The trade-off is that for each value $\boldsymbol{\theta}_i$ we need to solve a maximization problem to obtain $\tilde{\pi}(x_j|\boldsymbol{\theta}_i, \mathbf{y})$, therefore we want k to be as small as possible. On the other hand, the smaller k is the worst the numerical approximation in equation 3.15 will be. This is also the reason why $\boldsymbol{\theta}$ has to be low dimensional, otherwise, we would need a large number of values for the approximation to have sufficient accuracy losing too much in efficiency. Bearing this in mind we arrive at the last step of INLA: approximation of the posterior marginal distribution of the hyper-parameters θ_j.

The approximation of $\pi(\theta_j|\mathbf{y})$ could be obtained by integrating numerically expression 3.11 considering $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ instead of $\pi(\boldsymbol{\theta}|\mathbf{y})$. However, such integration scheme would require evaluating the function $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ for many points and we already said that retrieving these values could take a large amount of time. The idea proposed by Rue et al. (2009) was to use the values $\tilde{\pi}(\boldsymbol{\theta}_i|\mathbf{y})$ that have already been calculated for equation 3.15. Specifically, they interpolate the values $\tilde{\pi}(\boldsymbol{\theta}_1|\mathbf{y}), ..., \tilde{\pi}(\boldsymbol{\theta}_k|\mathbf{y})$ obtaining the function $I(\boldsymbol{\theta}|\mathbf{y})$ which approximate the true $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$, and the integral in equation 3.11 is performed with respect to $I(\boldsymbol{\theta}|\mathbf{y})$. The obtained approximated marginal posterior distribution is given by

$$\tilde{\pi}(\theta_j|\mathbf{y}) = \int I(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-j}. \tag{3.16}$$

This concludes the review of the INLA methodology. The approach is not flawless and can give unrealistic results when one of the many approximations involved in the process is biased. However, it allows applied researchers to fit Bayesian models with complicated correlation structures that are usually impractical using MCMC methods. Indeed, many of the features introduced in the R-INLA package come from the needs manifested by the users themselves. For this thesis, I implemented for the first time a self-exciting process with the `inlabru` R package showing that the INLA methodology can be extended to this class of processes. These processes are the most commonly used model to describe earthquake occurrence, and the INLA methodology provides a robust framework to perform Bayesian inference on complex models. In this way, researchers and practitioners in statistical seismology have access to the complex models supported by INLA and successfully applied in disease mapping, ecology, and environmental statistics, to name a few.

3.3 inlabru

The `inlabru` R-package (Bachl et al., 2019) facilitates fitting spatial models to non-specialist users, it simplifies the syntax and extends the class of models that can be fitted to include models with non-linear predictors. Indeed, the `R-INLA` R-package requires the user to implement their own likelihood approximation scheme, while `inlabru` does that automatically using an iterative algorithm. Also, `inlabru` provides an automated way to implement Log-Gaussian Cox processes (LGCP, Cox, 1955; Møller et al., 1998) which are an important class of models widely used in spatial statistics. An important feature of `inlabru` is the possibility of including a Gaussian random field (GRF) in the predictor, a continuous random process in which values at different locations are normally distributed and correlated depending on the distance between each other. Specifically, the `inlabru` package considers a GRF with a Matérn covariance function approximated using the Stochastic Partial Differential Equation (SPDE) approach (Lindgren et al., 2011, 2022). The SPDE approach is fundamental for the efficiency of the approach which otherwise would be infeasible. The GRF has the role of capturing the correlation not explained by the covariates and can be interpreted as the combined effect of the non-observed processes influencing the observations. LGCP models including covariates and a GRF have already been used to model seismicity in a time-independent framework (Bayliss et al., 2020). They offer a robust statistical framework to compare models including different combinations of covariates and determine the best-performing combination.

3.3.1 Log-Gaussian Cox Process models

Cox processes (Cox, 1955; Møller et al., 1998) are a fruitful class of models particularly suitable for problems involving prediction of a partially observed spatio-temporal process. Thus, it seems to us appropriate to model earthquake data in which observations below a certain magnitude threshold are discarded. The generic Log-Gaussian Cox process in d dimensions has to respect the following two postulates:

- The log-intensity is given by $\lambda(\mathbf{x}) = \exp\{S(\mathbf{x})\}$ for $\mathbf{x} \in \mathbb{R}^d$, where $S(\mathbf{x})$ is a Gaussian process.
- Conditionally on a realization of $S(\mathbf{x})$ the process is an inhomogeneous Poisson process with intensity function $\lambda(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$

Specifically, the first point of the definition highlights why they are called Log-Gaussian Cox processes, indeed, the logarithm of the intensity has to be a Gaussian process. To establish a link with the INLA methodology, the Gaussian field $S(\mathbf{x})$ is the latent field. The second point implies that knowing the latent field, the observations follow a Poisson process, and therefore, they are independent of each other. This means, that the correlation between points is supposed to be completely explained by the latent field and thus, particular attention has to be devoted to its construction.

The Gaussianity of the latent field makes the moments of an LGCP model analytically tractable, which is not the case when we depart from Gaussianity. Furthermore, this formulation provides an elegant way to model the observations hierarchically, using the same framework of generalized linear models (GLM). In fact, assuming to have observed N points at locations $\mathbf{s}_1, ..., \mathbf{s}_N \in W \subset \mathbb{R}^d$, let $\mathbf{S} = (S(\mathbf{s}_1), ..., S(\mathbf{s}_N))$ be the vector containing the latent field value at the observed locations. Given that, $S(\mathbf{x})$ is a Gaussian field, the vector $\mathbf{S}$ has an N-dimensional Gaussian distribution with mean vector $\mu(\mathbf{S}) = \mu(\mathbf{s}_1), ..., \mu(\mathbf{s}_N)$ and $N \times N$ covariance matrix Q^{-1} such that Q^{-1}_{ij} is equal to the covariance between $S(\mathbf{s}_i)$ and $S(\mathbf{s}_j)$ for $i \neq j$ and, to the variance of $S(\mathbf{s}_i)$ when $i = j$. The corresponding LGCP model can be formulated in a way that resembles the Latent Gaussian models' formulation:

$$\begin{aligned}
\lambda(\mathbf{s}_i)|\mathbf{x} &\sim \exp\{\eta(\mathbf{s}_i)\} \\
\eta(\mathbf{s})|\mathbf{x} &= \beta_0 + \sum_{j=1}^{p} \beta_j z_j(\mathbf{s}) + u(\mathbf{s}) \\
\mathbf{u}|\boldsymbol{\theta} &\sim N(0, Q^{-1}(\mathbf{u})|\boldsymbol{\theta}) \\
\boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}),
\end{aligned} \tag{3.17}$$

where $z_j(\mathbf{s}) \in \mathbb{R}$ is the value of the $j-$th available covariate at location $\mathbf{s}$, $\mathbf{u} = \{u(\mathbf{s}_1), ..., u(\mathbf{s}_N)\}$, and $u(\mathbf{s})$ is a Gaussian field with 0 mean and precision matrix $Q(\mathbf{S})$ determined by specifying a covariance function $r(\mathbf{s}_i, \mathbf{s}_j)$ which returns the covariance between location $\mathbf{s}_i$ and $\mathbf{s}_j$. The quantities $\beta_0, ..., \beta_N \in \mathbb{R}^{p+1}$ are coefficients; they are assumed to be independent and to have a normal distribution with 0 mean and standard deviation σ_j. The vector of hyper-parameters $\boldsymbol{\theta}$ is composed of the variances of the coefficients β_js and the parameters determining the covariance function $r(\mathbf{s}, \mathbf{s}_0)$. The linear predictor $\eta(\mathbf{s})$, being a linear function of $u(\mathbf{s})$, is in turn a Gaussian field with mean $\mu(\mathbf{s}) = \mu + \sum_j \beta_j z_j(\mathbf{s})$ and same covariance matrix as $u(\mathbf{s})$. To continue the parallelism with the Latent Gaussian models, the latent field here is composed by $\mathbf{x} = \{\eta(\mathbf{s}_1), ..., \eta(\mathbf{s}_N), \beta_0, ..., \beta_N, u(\mathbf{s}_1), ..., u(\mathbf{s}_N)\}$ which is again a Gaussian field.

From the above formulation, it is possible to understand the centrality of the role played by the latent field in LGCP models. The role of the latent field is to explain the correlation between observations due to an unobserved underlying phenomenon. Examples include unobserved soil characteristics in studying plant locations; the presence of competing animal species in studying animals' behaviour; underlying factors in studying the spread of disease. The value β_0 has the same interpretation of the intercept in a simple Linear model, it represents the mean value of the linear predictor when the effect of the covariates is equal zero. The value β_0, the vector $\boldsymbol{\beta}$, the value of the random field $\mathbf{u}$, and the hyper-parameters $\boldsymbol{\theta}$ have to be estimated from the data.

The choice of the covariance function $r(\mathbf{s}, \mathbf{s}_0)$ is more critical because it has to reflect the hypothesis on the correlation between points. In my project, we are going to use the Matérn covariance function (Matérn, 1960) which defines a flexible and widely used class of Gaussian fields. The Matérn covariance function is specified as follows:

$$r(\mathbf{s}, \mathbf{s}_0) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}(k\|\mathbf{s} - \mathbf{s}_0\|)^\nu K_\nu(k\|\mathbf{s} - \mathbf{s}_0\|), \tag{3.18}$$

where K_ν is the modified Bessel function of second kind of order $\nu > 0$, $\sigma > 0$ is the marginal standard deviation, $k > 0$ is a scaling parameter, and $\|\mathbf{s} - \mathbf{s}_0\|$ is the Euclidean distance between $\mathbf{s}$ and $\mathbf{s}_0$. The parameter ν determines the differentiability of the process and, due to identifiability issues, it is usually fixed. The scaling parameter k does not have an intuitive interpretation, however, it is strictly linked to the quantity $\rho = \sqrt{8\nu/k}$ which is the Euclidean distance at which the correlation between $\mathbf{s}$ and $\mathbf{s}_0$ is around 0.13. The `inlabru` R-package provides an efficient way to estimate σ and ρ which are part of the hyper-parameters of the model.

The Matérn covariance function accounts only for the relative position of the points, specifically, it is a function only of the distance between points. Thus, processes with such a covariance function and with a constant mean are invariant to translation (stationary) and rotation (isotropic). In other words, the closer two points are to each other the more they are correlated. This is coherent with the interpretation that the field represents an unknown latent phenomenon that influences the point-generating process. This phenomenon is supposed to be continuous and thus, closer points will tend to have more similar values. Due to its nice property and clear interpretation of the parameters, the Matérn covariance function has become one of the most commonly used covariance functions in spatial statistics applications (Stein, 1999) and machine learning (Williams and Rasmussen, 2006).

3.3.2 The SPDE approach

Despite their popularity, models involving a continuously indexed Gaussian field with Matérn covariance function are not very manageable. Having observed N points, the latent field values have a multivariate normal distribution $\mathbf{S} = (S(\mathbf{s}_1), ..., S(\mathbf{s}_N)) \sim N(\mu(\mathbf{S}), \Sigma_N)$. The matrix Σ_N is an $N \times N$ dense matrix which makes the cost to evaluate the likelihood $\mathcal{O}(N^3)$

(Heaton et al., 2019). This means that any analysis involving more than hundreds of observations become infeasible. A common approach is to approximate the value of the Gaussian field using a number $n, N \gg n$ of basis functions such that:

$$S(\mathbf{s}) = \sum_{i=1}^{n} \omega_i \psi_i(\mathbf{s}), \tag{3.19}$$

where, $\omega_1, ..., \omega_n \sim N(0, \Sigma_n)$ are Gaussian weights and $\psi_1(\cdot), ..., \psi_n(\cdot)$ are basis functions. With this method evaluating the likelihood costs now n^3, which is a noticeable gaining when $N \gg n$. However, this approach prevents the model to capture fine-scale variations if n is too low. On the other hand, the computational time is cubic in the number of basis functions. As before, this is due to the fact that Σ_n, as well as Σ_n^{-1}, is a dense matrix that prevents the application of fast sparse matrix methods to perform calculations.

The SPDE approach (Lindgren et al., 2011, 2022) consists in considering $n \approx N$ and to approximate the precision matrix Σ_n^{-1} with a sparse matrix Q_n^{-1}. In this way, no local variation is discarded and the Gaussian weights are a Gaussian Markov Random Field (GMRF, Rue and Held, 2005). The fact that now the weights are from a GMRF brings two advantages: the computational time required to evaluate the likelihood is now $\mathcal{O}(n^{3/2})$ which is feasible for n in the order of thousands, and GMRF models are supported by INLA. Lindgren et al. (2011) shows that this approximation is valid and draw an explicit link between continuously indexed Gaussian fields (which in general are not tractable) and GMRF which are tractable.

Lindgren et al. (2011) exploited the fact that a Gaussian field $S(\mathbf{x}), \mathbf{x} \in \mathbb{R}^d$ with Matérn covariance function is a solution of the following Stochastic Partial Differential Equation (SPDE, Whittle, 1954):

$$(k^2 - \Delta)^{\alpha/2} S(\mathbf{x}) = \mathcal{W}(\mathbf{x}), \tag{3.20}$$

where $\Delta = \sum_{i=1}^{d} \partial^2/\partial x_i^2$ is the Laplacian operator, and $\mathcal{W}(\mathbf{x})$ is Gaussian white noise. There exists a one-a-one correspondence between the parameters of the above SPDE and the parameters of the Matérn covariance function of the solution as expressed above.

The link between GF and GMRF is, then, that it is possible to find a finite element representation of the SPDE solution with the form given by equation 3.20. Such representation is based on a discretization of the space. In particular, considering a triangularization of the space composed of a set of n nodes, the basis functions $\psi_i(\cdot)$ are piecewise linear functions assuming value 1 at node i and 0 at the other nodes $j \neq i$. Using such basis functions, the value ω_i can be interpreted as the value of the Gaussian field at the node and the discrete approximation as piecewise linear in each triangle.

The advantages of this approach which made us believe it may be effective in modelling earthquake data are:

1. The Matérn covariance parameters, being the same as in the SPDE formulation, has a clear physical meaning and can be used as quantities of interest in analysing earthquake data.

2. The number n of mesh points is independent of the number of observations N. It is related, perhaps, to the accuracy of the approximation, to the ability to capture local variations of the field, and to the computational time required to evaluate the likelihood. The trade-off, as usual, is between accuracy and computational time.

3. The sparsity of the precision matrix of the weights, allows us to exploit the advantages of the INLA approach. Furthermore, `inlabru` and `R-INLA` offer a general framework to perform Bayesian analysis using such models.

4. The SPDE representation allows to obtain *different* Matérn covariance functions changing the differential operator. In particular, this approach can be generalized to include: i) models with physical barriers (Bakka et al., 2019) ii) non-stationary fields (Fuglstad et al., 2015), iii) fields on general spaces (e.g. spheres, Lindgren et al., 2011).

A vast number of examples of applications of the SPDE approach can be found in Bakka et al. (2018), while Bolin and Kirchner (2020) extends the approach further allowing to work with precision matrices for the Gaussian weight which are not sparse.

3.3.3 LGCP model approximation

This section shows how LGCP models are approximated by `inlabru`. A modified version of this method will be used in the proposed Hawkes process approximation in Chapter 4, therefore, going through the LGCP approximation detail now will fix the fundamental ideas for later. The approximation relies on the fact that the INLA method can deal with Poisson Counts (PC) models but not with Poisson Process (PP) models. We start showing how point patterns can be represented as PP models and PC models.

Given a set of observations $\{\mathbf{y}_i : \mathbf{y}_i \in W, i = 1, ..., n\}$ of a point process in a region W we can model this data in two ways: as a PP model or as a PC model. The point process log-likelihood for this data is given by:

$$\mathcal{L}_{\text{PP}} = -\int_W \lambda(\mathbf{s})d\mathbf{s} + \sum_{i=1}^{n} \log \lambda(\mathbf{y}_i), \tag{3.21}$$

where $\lambda(\cdot)$ is the intensity function of the point process. In this case, the data is composed of the actual observations $\mathbf{y}_1, ..., \mathbf{y}_n$.

The second way (PC model) relies on a discretization of the region W. Suppose that the region W is divided in K non-overlapping bins $b_1, ..., b_K$ such that $\cup_k b_k = W$ and $b_k \cap b_j = \emptyset$ for any $k \neq j$. For each bin b_k, N_k represents the number of observations $\mathbf{y}_1, ..., \mathbf{y}_n$ in b_k. Also, for each bin b_k, λ_k represents the expected number of points by the model in b_k. The log-likelihood for the PC model is given by:

$$\mathcal{L}_{\text{PC}} = -\sum_{k=1}^{K} \lambda_k + \sum_{k=1} N_k \log \lambda_k. \tag{3.22}$$

In the equation above, we have omitted the term $N_k!$ because it is a known quantity. Here, the data is composed of the Poisson counts per bin $N_1, ..., N_K$.

The goal is to find a way to represent the PP log-likelihood using the PC log-likelihood to ensure $\mathcal{L}_{\text{PP}} \approx \mathcal{L}_{\text{PC}}$. To achieve this task we need to approximate the integral and the sum of log intensities in a way that

$$\int_W \lambda(\mathbf{s})d\mathbf{s} \approx \sum_{k=1}^{K} \lambda_k \qquad \sum_{i=1}^{n} \log \lambda(\mathbf{y}_i) \approx \sum_{k=1}^{K} N_k \log \lambda_k. \tag{3.23}$$

The first bit regards the total number of points expected in the region W. In fact, the integral of the intensity represents the expected value of the number of points in W and the sum of the expected number of points in each bin represents exactly the same thing. The sum of the intensity calculated at the observed points is a measure of *how likely* is to observe the present point patterns while the second summation is a measure of *how likely* is to observe the present counts.

We can reformulate the PC model in a way that is more convenient to approximate the PP model. Suppose that the intensity, in each bin b_k, is constant and equal to $\lambda(\mathbf{p}_k)$, where $\mathbf{p}_k$ is the centroid of the bin b_k. Suppose, also, that the bin b_k has volume (in 3D, area in 2D, length in 1D) E_k. E_k is also known as the exposure of the bin b_k. In this case, the expected number of points in the bin b_k is given by the product between $\lambda(\mathbf{p}_k)$ and E_k, namely $\lambda_k = \lambda(\mathbf{p}_k)E_k$. The log-likelihood of the PC model becomes:

$$\mathcal{L}_{\text{PC}} = -\sum_{k=1}^{K} \lambda(\mathbf{p}_k)E_k + \sum_{k=1} N_k \log \lambda(\mathbf{p}_k), \tag{3.24}$$

we have ignored the term $\log(E_k)N_k$ in the second summation because it is a known quantity.

The problem now is to find $\mathbf{p}_k$, N_k, and E_k, such that

$$\int_W \lambda(\mathbf{s})d\mathbf{s} \approx \sum_{k=1}^{K} \lambda(\mathbf{p}_k)E_k \qquad \sum_{i=1}^{n} \log \lambda(\mathbf{y}_i) \approx \sum_{k=1}^{K} N_k \log \lambda(\mathbf{p}_k). \tag{3.25}$$

The idea is to use two different sets of $\mathbf{p}_k$, N_k, and E_k, one to approximate the integral and one to approximate the sum of log intensities.

Approximation of the integral

In order to provide a better approximation of the integral of the intensity, it is convenient to base the approximation on a triangulation (or mesh) of the region W. Calling $\mathbf{s}_1, ..., \mathbf{s}_J$ the mesh points with weights $\omega_1, ..., \omega_J$, the integral is approximated by:

$$\int_W \lambda(\mathbf{s})d\mathbf{s} = \sum_{j=1}^{J} \lambda(\mathbf{s}_j)\omega_j. \tag{3.26}$$

Essentially is like considering a PC model with J bins defined by the triangulation. Here, $\mathbf{s}_j$ is the centroid of the bin b_j and ω_j is its exposure. The likelihood of this PC model is

$$\mathcal{L}_{\text{int}} = -\sum_{j=1}^{J} \lambda(\mathbf{s}_j)\omega_j + \sum_{j=1}^{J} N_j \log \lambda(\mathbf{s}_j). \tag{3.27}$$

Considering $N_j = 0, \forall j$ we have that

$$\mathcal{L}_{\text{int}} = -\sum_{j=1}^{J} \lambda(\mathbf{s}_j)\omega_j \approx -\int_W \lambda(\mathbf{s})d\mathbf{s}, \tag{3.28}$$

which approximates the target integral.

Approximation of the summation

To approximate the summation, it is convenient to consider as centroids the observed points $\mathbf{y}_1, ..., \mathbf{y}_n$. In this way, the summation is approximated by:

$$\sum_{i=1}^{n} \log \lambda(\mathbf{y}_i) \approx \sum_{i=1}^{n} N_i \log \lambda(\mathbf{y}_i). \tag{3.29}$$

The associated PC model log-likelihood is given by:

$$\mathcal{L}_{\text{sum}} = -\sum_{i=1}^{n} \lambda(\mathbf{y}_i)E_i + \sum_{i=1}^{n} N_i \log \lambda(\mathbf{y}_i). \tag{3.30}$$

Considering $N_i = 1$, and ω_i for all $i = 1, ..., n$ we have

$$\mathcal{L}_{\text{sum}} = \sum_{i=1}^{n} \log \lambda(\mathbf{y}_i). \tag{3.31}$$

Putting it all together

In order to provide a reliable approximation we need to combine the approximation of the integral with the approximation of the summation in a single PC model. Following the terminology defined above, we need to specify a set of centroids $\mathbf{p}_1, ..., \mathbf{p}_P$ representing the bins, a vector of exposures $E_1, ..., E_P$ representing the "size" of the bins and a vector of counts $N_1, ..., N_P$ representing the number of observed events in each bin.

The dimension P is given by the the number of mesh points J and the number observations n, such that $P = J + n$. Binding the specifications of centroids, exposures, and counts used previously, the three vectors are specified as follows:

$$\text{centroids} = \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_J \\ \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}, \quad \text{exposures} = \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_J \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \text{counts} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

The resulting PC model has log-likelihood given by

$$\begin{aligned} \mathcal{L}_{\text{PC}} &= \mathcal{L}_{\text{int}} + \mathcal{L}_{\text{sum}} \\ &= -\sum_{j=1}^{J} \lambda(\mathbf{s}_j)\omega_j + \sum_{i=1}^{n} \log \lambda(\mathbf{y}_i) \\ &\approx -\int_W \lambda(\mathbf{s})d\mathbf{s} + \sum_{i=1}^{n} \log \lambda(\mathbf{y}_i) = \mathcal{L}_{\text{PP}}. \end{aligned}$$

With the above approximation, the error depends only on how well we approximate the integral, given that the sum of log intensities is exact. The accuracy of the approximation of the integral depends on the mesh used to discretize the space and can be decided by the user. The trade-off is between accuracy and computational time, indeed, considering a finer mesh means considering more mesh points which, in turn, means considering more points in the surrogate PC model.

3.3.4 `inlabru` iterative method

The approximation method used for LGCP models requires the log intensity to be linear in the parameters, however, we will need to relax this hypothesis to approximate Hawkes process models. The `inlabru` package offers a way to approximate LGCP models in the case of a non-linear log intensity function. The framework was developed to allow the users to include in the linear predictor non-linear effects of the covariates and therefore, we are going to refer to this framework to explain how the approximation can be extended.

Consider the following model,

$$\log \lambda(\mathbf{s}) = \eta(\mathbf{s}, \boldsymbol{\theta}) \tag{3.32}$$

$$\eta(\mathbf{s}) = \beta_0 + \sum_{i=1}^{p} f_i(z_i(\mathbf{s}, \boldsymbol{\theta}_i)) + u(\mathbf{s}), \tag{3.33}$$

where $\beta_0 \in \mathbb{R}$ is a scalar and $u(\mathbf{s})$ is a GMRF, and $f_i(\cdot)$ is a deterministic function of the covariate $z_i(\cdot)$ depending on a set of parameters $\boldsymbol{\theta}_i$. The function $f_i(\cdot)$ is supposed to be smooth, meaning that first and second derivatives exist. The linear predictor depends on the set of all parameters $\boldsymbol{\theta} = \cup_i \boldsymbol{\theta}_i$ given by the union of the parameters needed by each function of the covariate. The parameters $\boldsymbol{\theta}$ have to be estimated from the data. For example, in the linear case $f_i(z_i(\mathbf{s})) = \beta_i z_i(\mathbf{s})$, $\boldsymbol{\theta}_i = \beta_i$.

The idea is to work with a linearised predictor with respect to $\boldsymbol{\theta}_0$ given by the first order Taylor expansion of $\eta(\mathbf{s}, \boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$.

$$\overline{\eta}(\mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\theta}_0) = \eta(\mathbf{s}, \boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}} \eta(\mathbf{s}) \bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0} . \tag{3.34}$$

The approximation is exact at $\boldsymbol{\theta}_0$ and degrades for values of $\boldsymbol{\theta}$ far away from that point. It is natural to choose $\boldsymbol{\theta}_0$ to be the value of the parameters in which we are most interested. In a Bayesian setting, this point is the posterior mode of $\boldsymbol{\theta}$, in this way, the approximation is exact at the posterior mode (where there is the greatest amount of probability) and degrades in the tails. However, knowing the posterior mode implies knowing the posterior distribution which is the final goal of any Bayesian analysis. The `inlabru` package uses an iterative method to determine the posterior mode around which the model is approximated.

The iterative method works as follows:

1. Let $\boldsymbol{\theta}_0$ be an initial linearisation point.

2. Compute the linearised predictor at $\boldsymbol{\theta}_0$.

3. Run INLA on the linearised model and obtain posterior mode $\boldsymbol{\theta}_1$

4. Let $\boldsymbol{\theta}_\alpha = \alpha \boldsymbol{\theta}_0 + (1 - \alpha) \boldsymbol{\theta}_1$ and find the value of α that minimises $\|\eta(\boldsymbol{\theta}_\alpha) - \overline{\eta}(\boldsymbol{\theta}_1)\|$

5. Set $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_\alpha$ as a new linearisation point and repeat from step 2.

The procedure ends when the maximum component-wise difference between $\boldsymbol{\theta}_\alpha$ and $\boldsymbol{\theta}_0$ is less than 1% of the component posterior standard deviation. The default value 1% can be changed by the user. A potential improvement to step 4 is to also take into account the prior distribution for $\boldsymbol{\theta}$ as a minimisation penalty, to avoid moving further than would be indicated by a full likelihood optimisation. A schematic of the iterative method can be found in Figure 5.3 in Chapter 5.

The properties of this method of providing reliable posterior distributions of the parameters are part of an ongoing study which still needs to be finalized, preliminary details can be found at https://inlabru-org.github.io/inlabru/articles/method.html.

3.4 Time-independent LGCP models for seismicity

Here, we show how the `inlabru` approximation method for LGCP models described in this Chapter can be used to build spatially-varying time-independent models for seismicity. This is the approach used in Bayliss et al. (2020) and Bayliss et al. (2022) to model seismicity in Southern California. In this section, we show the results obtained in Bayliss et al. (2022). The articles consider different LGCP models with intensity for a generic location $\mathbf{s} \in W$, where $W \subset \mathbb{R}^2$ is the spatial domain represented by the polygon in Figure REF, given by:

$$\begin{aligned} \lambda(\mathbf{s}) &= \exp\{\beta_0 + \sum_{j=1}^{J} \beta_j z_j(\mathbf{s}) + u(\mathbf{s})\} \\ u(\mathbf{s}) &\sim GMRF(\mathbf{0}, Q^{-1}(\mathbf{s}, \mathbf{s}')), \end{aligned} \tag{3.35}$$

where $\beta_j \in \mathbb{R}$ for $j = 0, ..., J$ and $J \in \mathbb{N} : J \geq 1$, and $u(\mathbf{s})$ is a Gaussian random field with Matérn covariance function and parameters estimated using the SPDE approach. The model is identical to the one described by Equation REF (3.17).

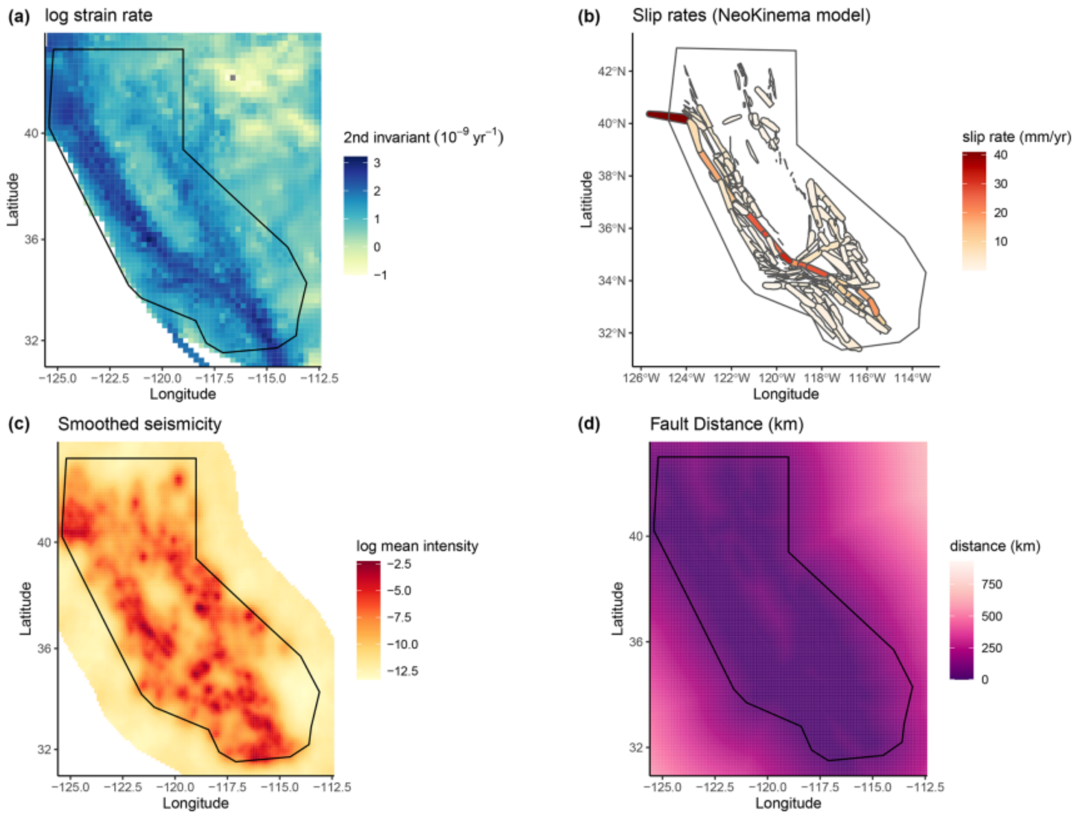


Figure 3.1: Input model covariates: (a–d) strain rate (SR), NeoKinema slip rates from UCERF3 (NK), smoothed seismicity from a Gaussian random field for events before 1984 (MS), distance to nearest (UCERF3, dip and uniformly buffered) fault in km (FD). This Figure is the same as Figure 2 of Bayliss et al. (2022).

The quantities $z_j(\mathbf{s})$ for $j = 1, ..., J$ represent the value of the covariates included in the model at location $\mathbf{s}$. Bayliss et al. (2022) considers three models based on different combinations of four spatially-varying covariates which were found to perform well in terms of Deviance Information Criterion (DIC, REF) in Bayliss et al. (2020). The covariates considered are shown in Figure REF and include the strain rate (Kreemer et al., 2014) (SR) map, NeoKinema model slip rates (NK) attached to mapped faults in the UCERF3 model (Field et al., 2014), a past seismicity model (MS) and a fault distance map (FD) constructed using the UCERF3 fault geometry, with fault polygons buffered by their recorded dip. The past seismicity model is a smoothed seismicity map derived from events in the UCERF3 catalogue that occurred prior to 1984 and not included in the data used to fit the models. The smoothed seismicity map is the posterior mean of an LGCP model composed of an intercept and a random field as $u(\mathbf{s})$ defined above.

The models considered are

1. SRMS model uses the strain rate and the past seismicity.

2. SRMSNK model uses the strain rate, the past seismicity, and the NeoKinema slip rates.

3. FDSRMS model uses the fault distance, the strain rate, and the past seismicity.

The LGCP models considered in Bayliss et al. (2022) are time-independent meaning that although the rates are spatially varying and modeled as an inhomogenous Poisson process with intensity given by 3.35, they are assumed to be homogeneous in time. However, this assumption is not verified by observed data which is not Poissonian due to spatiotemporal clustering (Vere-Jones, 1970; Gardner and Knopoff, 1974). The aim of the article is to investigate the effect of clustering on the models and does that by fitting the above three models on two different catalogues. The first one is composed by all the $M4.95+$ events from 1985 to 2005 from the UCERF3 dataset (Field et al., 2014), while the second one is obtained by considering only the mainshocks of the first one determined using the Gardner and Knopoff (1974) declustering algorithm (UCERF3 Appendix K). This results in 6 different spatial models. The models using the declustered catalogue are named SRMSDC, SRMSNKDC, and FDSRMSDC which include the same combinations of covariates described above.

Figures 3.2 and 3.3 show the logarithm of the posterior median of the intensity (diagonal), the differences in the log median (top-right side), and the differences in model variances (bottom-left side) of the models using, respectively, the full catalogue and the declustered one to estimate the parameters of the model. The models are then used to produce grid-based and catalogue-based forecasts and tested pseudo-prospectively against data in the periods 2006-2011 and 2011-2016 using the CSEP consistency tests described in Chapter 2. The models passed the CSEP tests in the period 2006-2011 while performing poorly in the period 2011-2016. The declustered catalogue models performed better than the full catalogue models in the latter period.

The approach presented in this section is coherent with current practice in time-independent forecasting and PSHA where the catalogues are declustered to conform to the Poisson assumption. It also provides a general framework to test the importance of different covariates in the model and a fully Bayesian method for forecast generation.

The main limitation of this approach is how aftershocks are handled and the dependence of the results upon a declustering algorithm. In fact, many different declustering algorithms exist and there is no formal way to identify which one is best. Also, the effect of declustering in the results presented here is attenuated by the (relatively) high magnitude threshold. With a smaller magnitude threshold, the number of discarded events by the declustering procedure would have been higher increasing the differences between models. These problems, however, are not exclusive of the `inlabru` approach but affect most time-independent models for seismicity. The real solution to this is to formally model the clustering process as it is done by time-dependent models such as the Epidemic-Type Aftershock Sequence (ETAS) model.

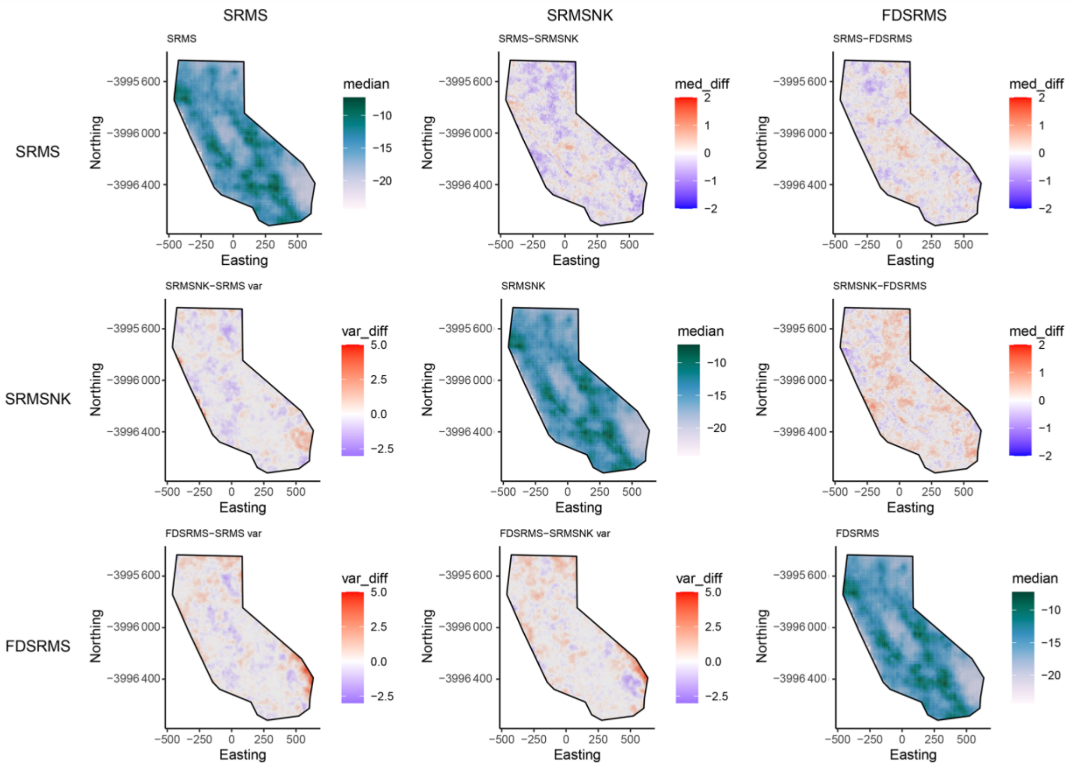


Figure 3.2: Pairwise comparison of models for full catalogue models. The top-right side of the plot shows differences in log median intensity and the lower left section shows the differences in model variances between the different models. The median log intensities for each model are shown on the diagonal. Models include combinations of smoothed past seismicity (MS), strain rate (SR), fault distance (FD) and fault slip rates (NK). This Figure is the same as Figure 3 of Bayliss et al. (2022).

3.5 Chapter Summary

In this chapter, I have described the INLA methodology to perform inference on the parameters of a Latent Gaussian model. The methodology relies on Gaussianly approximating parts of the posterior distribution while it does not alter the likelihood of the model. This means that the approximation is less accurate for models for which the posterior of the parameters departs significantly from normality (e.g. highly skewed distributions). However, the R-INLA R-package is able to detect these cases and apply a correction to the posterior to reduce the approximation bias.

I have shown how this methodology is extended in the inlabru R-package to support LGCP models and the iterative method employed to handle models with non-linear predictors. The ability of the algorithm to converge depends on the degree of non-linearity of the predictor and by possible numerical problems. However, it is usually fine for common functions such as exponentials, logarithms, and power laws. I have shown how the LGCP approximation method and the SPDE approach have been combined to produce time-independent models of seismicity (Bayliss et al., 2020, 2022). The advantage of this approach is to provide a Bayesian framework to study the effect of including different covariates combinations and whether this produces improved forecasts with respect to other approaches. This approach also supports time-dependent models if time-varying covariates are included, however, the model does not account for the interaction between points and all events occur indepen-

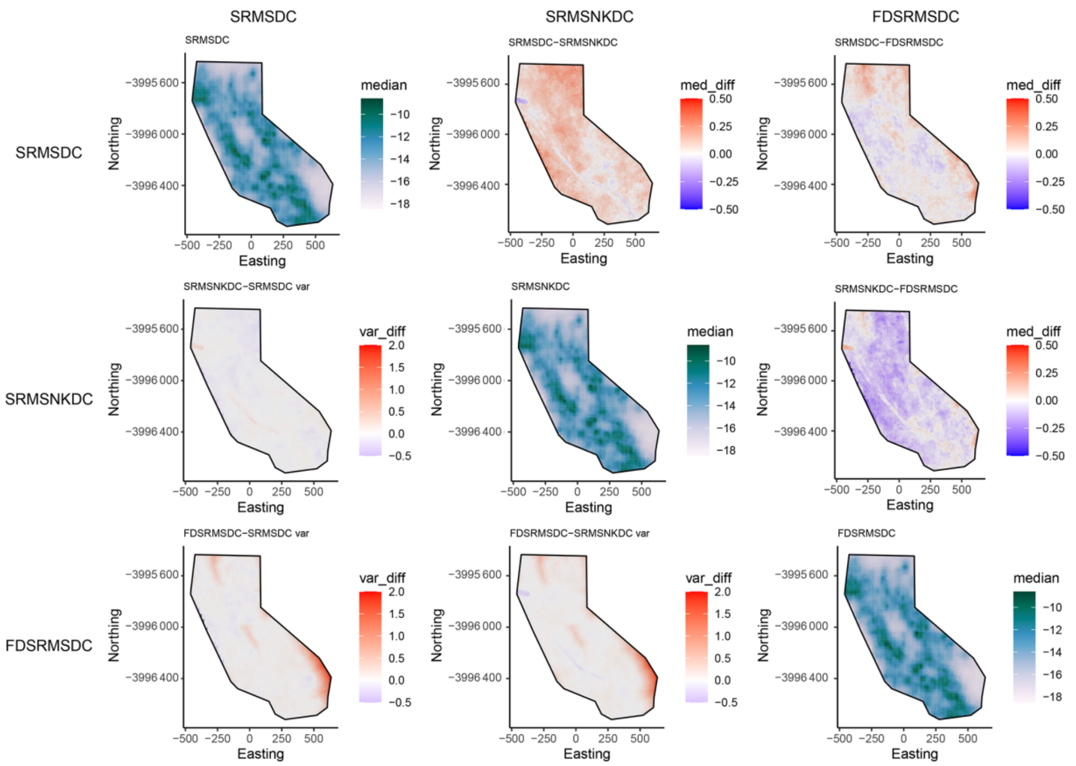


Figure 3.3: Pairwise comparison of models for declustered catalogue models. The top-right side of the plot shows differences in log median intensity and the lower left section shows the differences in model variances between the different models. The median log intensities for each model are shown on the diagonal. Models include combinations of smoothed past seismicity (MS), strain rate (SR), fault distance (FD) and fault slip rates (NK). This Figure is the same as Figure 4 of Bayliss et al. (2022).

dently from each other. In other words, this approach does not allow to explicitly model the clustering process and there is no distinction between background events and aftershocks. This can only be done using a Hawkes process (or *self-exciting* model as the ETAS model.

The main contribution of this thesis is to build upon the approximation method presented here for the LGCP model to approximate Hawkes process models. In Chapter 4 I generalize the approximation method presented in Section 3.3.3 to support Hawkes process models. The proposed method works for many different Hawkes process models and we use it, in the context of seismicity, to approximate the ETAS model. The main advantage is that it extends the `inlabru` approach to explicitly model the clustering of earthquakes in time and space. I start focusing on time only and Chapter 4 describes the methodology in general but applies it only to the temporal ETAS model. To make our approach accessible, I have developed an R-package called `ETAS.inlabru` which provides a user-friendly implementation of the proposed method. In Chapter 4 we compare the results obtained with our approach with the ones obtained using the `bayesianETAS` R-package which is based on an MCMC technique.

Chapter 5 makes use of the `ETAS.inlabru` R-package on simulated data to explore potential biases in the parameters' posterior distribution arising from choices regarding the data used as input (e.g. catalogue length, inclusion/exclusion of quiescence periods, number of large earthquakes, and catalogue incompleteness). Chapter 6 extends the proposed methodology to the spatio-temporal ETAS model and describes a way to include covariates in modelling the number of aftershocks relative to each event. This provides a general Bayesian framework to study the effect of available covariates on the produced forecasts while explicitly describing the clustering process. The approach described in this thesis generalizes to time-dependent models the approach described in this chapter for time-independent models.

Chapter 4

Approximation of Hawkes process models with application to temporal ETAS model

4.1 Introduction

This chapter includes an article accepted for publication by the Environmetrics[1] journal and available in preprint (Serafini et al., 2022a). The authors of the paper are Francesco Serafini (me), Mark Naylor, and Finn Lindgren. As the first author, I contributed by writing the article, gathering comments from the other authors and reviewers, and leading the review process until the final version was reached. I have also developed the methodology described in this article and provided the code to implement it. The methodology presented here is implemented in the `ETAS.inlabru` (Naylor and Serafini, 2023) R-package which was used to produce the results. The `ETAS.inlabru` is available on Git-Hub and will be soon submitted to CRAN.

Hawkes process models (or *self-exciting* processes Hawkes, 1971a,b) are a flexible class of point process models particularly suited to describe phenomena having a self-exciting behavior in which cascades of events are observed. Typical examples are infectious diseases, crimes, wildfires, droughts, neuronal activity, viral social media contents, and earthquakes. Indeed, the Epidemic Type Aftershock Sequence (ETAS) model belongs to this class. Despite being widely used, Hawkes process models have characteristics that makes it challenging to perform Bayesian inference on them. First of all, the likelihood is usually complex in non-trivial cases, and consequently, there is no close-form analytical solution for the posterior distribution. Second, the parameters are strongly correlated with each other, which poses difficulties in applying standard Markov Chain Monte Carlo (MCMC) techniques to retrieve the posterior. Moreover, these difficulties are exacerbated when considering more complex formulations (e.g. anisotropic spatial kernel, non-stationary processes), and the inclusion of covariates or structured random effects requires ad-hoc modifications of the procedure used to perform inference on the parameters. This, in turn, complicates hypotheses testing and makes it difficult to discriminate between behaviours that emerge from model formulation versus differences emerging from the methodologies to perform inference.

Here, I develop a new methodology to perform approximate Bayesian inference on the Hawkes process parameters. The technique is based on the Integrated Nested Laplace Approximation (INLA) and is implemented through the `inlabru` R-package. INLA is an alter-

[1] website: https://onlinelibrary.wiley.com/journal/1099095x

native to MCMC, while `inlabru` provides user-friendly access to INLA extending the classic methodology to point process models. In contrast to MCMC, INLA is based on a deterministic approximation which makes it faster than MCMC competitors. Further, it yields fully reproducible results. These advantages are particularly evident in models considering large structured random effects. Although the proposed technique is general, potentially applicable to any Hawkes process model, in this chapter we focus on the temporal ETAS model. The novelty of the approach resides in a new log-likelihood approximation. We compare our methodology with the MCMC technique proposed by Ross (2021) and implemented through the `bayesianETAS` R-package using the 2016 Amatrice earthquake seismic sequence (Michele et al., 2016).

4.2 The paper

4.2.1 Abstract

Hawkes process are very popular mathematical tools for modelling phenomena exhibiting a self-exciting or self-correcting behaviour. Typical examples are earthquakes occurrence, wild-fires, drought, capture-recapture, crime violence, trade exchange, and social network activity. The widespread use of Hawkes process in different fields calls for fast, reproducible, reliable, easy-to-code techniques to implement such models. We offer a technique to perform approximate Bayesian inference of Hawkes process parameters based on the use of the R-package `inlabru`. The `inlabru` R-package, in turn, relies on the INLA methodology to approximate the posterior of the parameters. Our Hawkes process approximation is based on a decomposition of the log-likelihood in three parts, which are linearly approximated separately. The linear approximation is performed with respect to the mode of the parameters' posterior distribution, which is determined with an iterative gradient-based method. The approximation of the posterior parameters is therefore deterministic, ensuring full reproducibility of the results. The proposed technique only requires the user to provide the functions to calculate the different parts of the decomposed likelihood, which are internally linearly approximated by the R-package `inlabru`. We provide a comparison with the `bayesianETAS` R-package which is based on an MCMC method. The two techniques provide similar results but our approach requires two to ten times less computational time to converge, depending on the amount of data.

4.2.2 Introduction

Hawkes processes or *self-exciting* processes, first introduced by Hawkes (1971a,b), are counting processes often used to model the "arrivals" of some events over time, when each arrival increases the probability of subsequent arrivals in its proximity. Typical applications can be found in seismology (Ogata, 1988; Ogata and Zhuang, 2006; Ogata, 2011; Paik Schoenberg, 2022), capture-recapture (Altieri et al., 2022; Weller et al., 2018), invasive species (Balderama et al., 2012), droughts (Li et al., 2021), crime (Mohler et al., 2011; Mohler, 2013; Mohler et al., 2018), finance (Azizpour et al., 2018; Filimonov and Sornette, 2012; Hawkes, 2018), disease mapping (Chiang et al., 2022; Garetto et al., 2021), wildfires (Peng et al., 2005), and social network analysis (Kobayashi and Lambiotte, 2016; Zhou et al., 2013).

Hawkes process, and more in general point processes, are counting processes assuming a value equal to the cumulative number of points recorded in a bounded spatio-temporal region. The main characteristic of a Hawkes process is its ability to model the effect of a point on the probability of observing additional points in its surroundings. For example, in seismology, it is often assumed that each earthquake has the ability to *induce* other earthquakes, and therefore observing an earthquake at a space-time location increases the probability of observing additional earthquakes in its proximity. Therefore, each observed point can be classified as *induced*, if it was induced by another point in the history of the process, or as *background* if it arose spontaneously. In this framework, a Hawkes process can be seen as the superposition of a background process, describing the occurrence of background events, and a sub-process for each observation in the history, describing the occurrence of events induced by that observation. This implies that the rate at which points occur at each space-time location is potentially influenced by the whole history of the process. This makes Hawkes process models non-Markovian. More formal definitions of the Hawkes process, its history, and its conditional intensity are given in Section 4.2.3.

The application of the Bayesian approach has become increasingly popular also in the Hawkes process field (Rasmussen, 2013; Donnet et al., 2020; Holbrook et al., 2021). In fact, Hawkes process models are often used in hazard or risk analyses, in which the ability to quantify the uncertainty around quantities of interest (e.g. number of events, probability of events of a certain class, inter-event time distribution) is of paramount importance (Marzocchi et al., 2015; Smit et al., 2019). However, applying the Bayesian framework, in these cases, is difficult, given the complex form of the posterior distribution and the high degree of correlation between Hawkes process parameters, and researchers had to resort to frequentist-like estimation techniques (Ebrahimian et al., 2014; Omi et al., 2015). Also, an easy-to-use, extendible, Bayesian technique to handle Hawkes process models is still missing, one of the few examples to the authors' knowledge is represented by Ross (2021). Furthermore, the techniques habitually used in the literature are based on the Markov-Chain Monte Carlo ((MCMC, Robert et al., 1999) method which limits the reproducibility of the results and resent from the presence of highly correlated parameters.

In this paper, we propose a novel approximation technique for Hawkes process models based on the use of the Integrated Nested Laplace Approximation (INLA, Rue et al., 2017) method. The INLA method is a well-known alternative to MCMC methods to perform Bayesian inference. It has been successfully applied in a variety of fields such as seismology (Bayliss et al., 2020), air pollution (Forlani et al., 2020), disease mapping (Riebler et al., 2016; Santermans et al., 2016; Schrödle and Held, 2011a,b), genetics (Opitz et al., 2016), public health (Halonen et al., 2015), ecology (Roos et al., 2015; Teng et al., 2022), more examples can be found in Bakka et al. (2018); Blangiardo et al. (2013); Gómez-Rubio (2020). Our approach aims to bring the INLA's advantages to the Hawkes process community and is implemented through the R-package `inlabru`. Specifically, the novelty of our approach resides in the likelihood approximation, indeed, the log-likelihood is decomposed in the sum of many small pieces, and each piece is linearly approximated with respect to the posterior mode. This means that the log-likelihood is exact at the posterior mode and the accuracy of the approximation decreases as we move away from that point. Furthermore, the linear approximation and the optimization routine to determine the posterior mode are internally performed by the `inlabru` package. The user only has to provide the functions to be approximated, the data, and the priors. The advantages of our approach are both in terms of computational time and simplicity to be extended to include covariates and/or to introduce structure in the parameters (e.g. considering one of them as temporally, or spatially, varying).

The article is structured as follows: Section 4.2.3 introduces the basic definition of a counting process, a Hawkes process, and defines its history and conditional intensity; Section 4.2.4 describes how Hawkes processes are used in practice and provides some examples on possible choices of the conditional intensity; Section 4.2.5 describes our novel approximation method for the log-likelihood; Section 4.2.6 provides a real data example on the Amatrice seismic sequence and compares the results obtained with our approach with the ones from the `bayesianETAS` R-package. For the Amatrice seismic sequence, we also provide a retrospective forecasting experiment in which we predict the daily number of earthquakes; Section 4.2.7 shows the results of a simulation experiment in which we simulate the data from a known model and compare the `inlabru` and `bayesianETAS` implementations. This is done to illustrate how the computational time scales increasing the amount of data. The three appendices at the end of the article (4.4.1,4.4.2, 4.4.3) provide the posterior distributions of the parameters for the two implementations considered and perform a sensitivity analysis of the `inlabru` results with respect to the binning strategy and the prior choice.

4.2.3 Notation and definitions

In this section, we give the basic definitions of a counting process, its history, and conditional intensity. Some definitions are only given with respect to time, but they can be easily extended to include space and marking variables. We start with the definition of a counting process. A counting process is a stochastic process assuming integer values changing over time. The value of a counting process at time $t \geq 0$ is equal to the number of observations with time less or equal than t. More formally,

Definition 4.2.0.1. *A counting process $\{N(t), t \geq 0\}$ is a stochastic process assuming values in the set of non-negative integers $\mathbb{N} \cup \{0\}$, such that: i) $N(0) = 0$; ii) $N(t)$ is a right-continuous step function with unit increments; iii) $N(T) < \infty$ almost surely if $T < \infty$. Also, given a time interval $[0, T)$ with $T < \infty$, we define the complete set of observations up to time T as $\mathcal{H}_T = \{t_h : t_h \in [0, T) \forall h = 1, ..., N(T^-)\}$. Given a random $t \in [0, T)$ we define the* history *of the process up to time t as the subset of elements of $\mathcal{H}_T$ recorded* strictly *before t and we call it $\mathcal{H}_t = \{t_h \in \mathcal{H}_T : t_h < t\}$.*

Definition 4.2.0.1 can be extended to the marked spatio-temporal case. In this case, a generic observed point is $\mathbf{x} = (t, \mathbf{s}, m)$ and is composed of a time t, a spatial location $\mathbf{s}$, and a marking variable m. The domain is given by $\mathcal{X} = [0, T) \times W \times M$, where $T > 0$, $W \subset \mathbb{R}^2$ and $M \subseteq \mathbb{R}$. The value of the counting process at time t is the number of events recorded before t (included), with spatial location in W and marking variable in M. Assuming that the spatial region of interest (W) and the marking variable's domain (M) are constant over time, we can use the same notation for the complete set of observations and the history of the process. In this case, the complete set of observations is $\mathcal{H}_T = \{\mathbf{x}_h = (t_h, \mathbf{s}_h, m_h) : \mathbf{x}_h \in \mathcal{X} \forall h = 1, ..., N(T^-)\}$, and the history of the process becomes $\mathcal{H}_t = \{\mathbf{x}_h = (t_h, \mathbf{s}_h, m_h) \in \mathcal{H}_T : t_h < t\}$.

Any counting process can be defined by specifying its conditional intensity. The conditional intensity of a counting process at time t is the expected infinitesimal rate at which events occur around time t given the history of the process $\mathcal{H}_t$. More formally,

Definition 4.2.0.2. *For a counting process $\{N(t), t \geq 0\}$ with history $\mathcal{H}_t$, the conditional intensity function of the process $N(t)$ is:*

$$\lambda(t|\mathcal{H}_t) = \lim_{\Delta_t \downarrow 0} \frac{\mathbb{E}[N(t + \Delta_t) - N(t^-) \mid \mathcal{H}_t]}{\Delta_t}.$$

For $\Delta_t, t \geq 0$. Assuming that the limit exists, the conditional intensity is left-continuous and $\lambda(t|\mathcal{H}_t) \geq 0$, $\forall t \geq 0$.

Definition 4.2.0.2 can also be extended to include a space location and a marking variable. The conditional intensity $\lambda(\mathbf{x}|\mathcal{H}_t)$ is the expected infinitesimal rate at which points occur in $(t, t + \Delta_t), \Delta_t > 0$, around space location $\mathbf{s}$, with marking variable around m.

The first characteristic for a Hawkes process as defined in Hawkes (1971b) Equation (4) is that the probability of the number of events in $(t, t + \Delta_t)$ being equal to $n = 0, 1, \dots$ is given by:

$$\Pr(N(t + \Delta_t) - N(t) = n|\mathcal{H}_t) = \begin{cases} 1 - \lambda(t)\Delta_t - o(\Delta_t) & if\, n = 0 \\ \lambda(t)\Delta_t + o(\Delta_t) & if\, n = 1 \\ o(\Delta_t) & if\, n > 1 \end{cases}. \tag{4.1}$$

Equation 4.1 has two major implications. The first one is that the probability of having more than one event in an infinitesimal interval around t goes to zero faster than the length of the interval. This implies that the probability of observing two events at the same time is zero and that the number of events in $\mathcal{H}_T$ is equal to $N(T)$ with probability one. However, recorded data does not have to obey that (due to time discretisation). The second is that the probability of having an event in $(t, t + \Delta_t)$ conditional on the history $\mathcal{H}_t$, for small $\Delta_t > 0$, is completely specified by the conditional intensity.

Now, we can define a Hawkes process model through its conditional intensity:

Definition 4.2.0.3. *A Hawkes process is a counting process with conditional intensity given by:*

$$\lambda(\mathbf{x}|\mathcal{H}_t) = \mu(\mathbf{x}) + \sum_{\mathbf{x}_h \in \mathcal{H}_t} g(\mathbf{x}, \mathbf{x}_h), \tag{4.2}$$

where $\mu : \mathcal{X} \to [0, \infty)$*, and* $g : \mathcal{X} \times \mathcal{X} \to [0, \infty)$

The conditional intensity is composed of a part $\mu(\mathbf{x})$ usually called the background rate, which does not depend on the history; and a second part representing the contribution to the intensity from the points in the history. The function $g : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is known as *excitation* or *triggering* function and measures the influence of observation $\mathbf{x}_h$ on the point $\mathbf{x}$.

Definition 4.2.0.3 implies that the whole history of the process is important to determine the current level of intensity. In this view, Hawkes processes can be seen as a non-Markovian extension of inhomogeneous Poisson processes. Both the background rate and the triggering function depends on a set of parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ which determines the properties of the Hawkes process under study (e.g. number of events per time interval, probability of a certain type of events, average number of induced events, type of clustering). Our technique provides a way to have a fully-Bayesian analysis of the parameters $\boldsymbol{\theta}$.

4.2.4 Hawkes process modelling

The Hawkes process intensity in Equation 4.2 is composed by two part, a background rate $\mu(\mathbf{x})$ and an *excitation* or *triggering* function $g(\mathbf{x}, \mathbf{x}_h)$. The background rate and the triggering function depend upon a number of parameters $\boldsymbol{\theta}$. Our objective is to provide a technique to determine the posterior distribution of $\boldsymbol{\theta}$ having observed points in $\mathcal{X} = [0, T] \times W \times M$. Equation 4.2 also shows that a Hawkes process can be thought of as the sum of $n+1$ Poisson processes, where $n = N(T)$ is the number of observations in the history of the process up to time $T < \infty$. One Poisson process represents the background rate and has intensity $\mu(\mathbf{x})$, the others n Poisson processes are each one generated by an observation $\mathbf{x}_h$ and have intensity $g(\mathbf{x}, \mathbf{x}_h)$. Many algorithms for fitting Hawkes process models are based on this decomposition and make use of a latent variable assigning the points to one of those $n+1$ Poisson processes (Ross, 2021; Veen and Schoenberg, 2008). Our approach is different because there is no explicit or implicit classification of the points into background and induced events.

Regarding marked spatio-temporal Hawkes process models, we only report the case where the marking variable distribution is independent of space and time, we refer to this distribution with $\pi(m)$. For the case where this assumption does not hold, and we have $\pi(\mathbf{x} = (t, \mathbf{s}, m))$, we just need to substitute $\mu(\mathbf{x})$, and $g(\mathbf{x}, \mathbf{x}_h)$ with $\mu(\mathbf{x})\pi(\mathbf{x})$, and $g(\mathbf{x}, \mathbf{x}_h)\pi(\mathbf{x})$ in all the following expressions without loss of generality. This is valid for both discrete and continuous distribution of the marking variable. Assuming an independent marking variable distribution the Hawkes process conditional intensity is given by:

$$\lambda(\mathbf{x} = (t, \mathbf{s}, m)|\mathcal{H}_t) = \left(\mu(\mathbf{x}) + \sum_{\mathbf{x}_h \in \mathcal{H}_t} g(\mathbf{x}, \mathbf{x}_h)\right) \pi(m). \tag{4.3}$$

Given the assumption of independence between the process representing the space-time locations and the marking variable's distribution, we only focus on the distribution of the space-time locations. The parameters of the marking variable distribution will be estimated independently and based on the observed marks solely. This is the usual situation in seismology, where the marking variable is the magnitude of the event, and its distribution is usually assumed to be independent of the space-time location of the events. If the assumption does not hold, applying the substitution described above allows us to estimate the marking variable distribution's parameters along with the Hawkes process parameters.

In this paper, we consider a spatially varying background rate that remains constant over time. This is done mainly to limit the number of modes in the likelihood and the correlation between parameters. Furthermore, we are going to consider a background rate parameterized as

$$\mu(\mathbf{x}) = \mu u(\mathbf{s}), \tag{4.4}$$

with $\mu \geq 0$ representing the number of expected background events in the area for a unit time interval, and $u(\mathbf{s})$ represents the spatial variation of the background rate and we assume it is normalized to integrate to one over the spatial domain. Different techniques have been employed to estimate $u(\mathbf{s})$. For example, in seismology, it is common practice to estimate it independently from the parameters of the triggering function smoothing a declustered set of observations (Ogata, 2011).

The common approach to model the triggering function is to factorize it in different components representing the effect of the observations $\mathbf{x}_h$ on the evaluation point $\mathbf{x}$ on the different dimensions (i.e. time, space, marking variable). More formally,

$$g(\mathbf{x}, \mathbf{x}_h) = g_m(m_h) g_t(t - t_h) g_{\mathrm{s}}(\mathbf{s} - \mathbf{s}_h) \mathbb{I}(t > t_h), \tag{4.5}$$

where, $I(t > t_h)$ is an indicator function assuming value one when the condition holds, and zero otherwise. The function $g_m(m_h)$ is the marking variable triggering function representing the effect of different values of the marking variable (e.g. if m is the magnitude of an earthquake, large earthquakes have a stronger influence); $g_t(t - t_h)$ is the time triggering function determining the time decay of the observed point's effect, and it is usually a decreasing function of $t - t_h$; $g_{\mathrm{s}}(\mathbf{s} - \mathbf{s}_h)$ is the space triggering function which has the same role of the time triggering function but in space and is usually a function of the *distance* between points (different distances may be employed).

Following this decomposition, also the parameter vector $\boldsymbol{\theta}$ can be decomposed in $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(\mu)}, \boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(\mathrm{s})})$, where $\boldsymbol{\theta}^{(\mu)}$ represents the parameters of the background rate, and $\boldsymbol{\theta}^{(m)}$, $\boldsymbol{\theta}^{(t)}$, $\boldsymbol{\theta}^{(\mathrm{s})}$ represent, respectively, the parameters of the magnitude, time and space triggering functions. We call J_μ, J_m, J_t, J_{s} the set of indexes indicating, respectively, the position of the background rate, marking variable triggering function, time triggering function, and space

Table 4.1: Typical choices of time and space triggering functions

	Time triggering	
Name	function	parameters
Exponential	$\beta e^{-\alpha(t-t_h)}$	$\alpha, \beta \geq 0$
Power Law	$k\left(1+\frac{t-t_h}{c}\right)^{-p}$	$k \geq 0, c > 0, p > 1$
	Space triggering	
Gaussian	$\det(2\pi\Sigma)^{-1/2}e^{-\frac{1}{2}(\mathrm{s}-\mathrm{s}_h)^T\Sigma^{-1}(\mathrm{s}-\mathrm{s}_h)}$	Σ positive semi-definite
Power Law	$(1+\frac{d(\mathrm{s},\mathrm{s}_h)}{\gamma})^{-q}$	$\gamma > 0, q > 1$

triggering function parameters inside $\boldsymbol{\theta}$, so we can write $\boldsymbol{\theta}_\mu = \{\theta_j \in \boldsymbol{\theta} : j \in J_\mu\}$. This notation will be particularly useful in Section 4.2.5.

Table 4.1 reports some of the typical choices for the space-time triggering function. Many modifications of these functions are used in real-data applications. For example, we can imagine a different time or space effect for different values of the marking variable. In seismology, it is common to consider a magnitude-dependent space triggering function representing the fact that earthquakes with large magnitudes affect wider areas. Another modification usually found in applications is to consider the normalized version of the reported functions to ensure they integrate to one over the (respective) domain.

As explained in Laub et al. (2021), the choice of the triggering function is crucial to the reliability and stability of any estimation procedure for Hawkes process parameters. For example, many techniques use triggering functions normalized to integrate to 1 over an infinite domain. For the approximation illustrated in this paper, we recommend using functions as close to linearity as possible with respect to the parameters. The approximation works for the ones in Table 4.1 which are not linear but at least monotonic. For the author's experience, the unnormalized version works best. The motivations behind this requirement will be illustrated in the next section.

In the real data example provided in Section 4.2.6, we apply our technique to earthquake data. The data is supposed to come from a spatio-temporal marked Hawkes process model, where the marking variable is the magnitude, however, we will consider it as a temporal marked point process, ignoring the information on the spatial location. The effect of that is to replace the full space-time intensity with a spatially integrated intensity. Indeed, assuming that the region of interest is constant over time, any temporal model, with intensity λ' can be seen as a spatio-temporal model (with intensity λ) integrated over space,

$$\lambda'(t, m|\mathcal{H}_t) = \int_W \lambda(t, \mathbf{s}, m|\mathcal{H}_t)d\mathbf{s}, \tag{4.6}$$

where $W \subset \mathbb{R}^2$. For the spatio-temporal model, if the background rate is given by equation 4.4 and the triggering function by equation 4.5, the temporal background rate (μ') and triggering function (g'_t) are given by

$$\mu' = \mu \int_W u(\mathbf{s})d\mathbf{s} \tag{4.7}$$

$$g'_t(t - t_h) = g_t(t - t_h) \int_W g(\mathbf{s} - \mathbf{s}_\mathrm{h})d\mathbf{s}. \tag{4.8}$$

Regarding the background rate, if $u(\mathbf{s})$ is normalized to integrate to 1 over the domain, the background rate is the same as in the spatio-temporal. For the triggering function, if there were no boundary effects, the integral would be independent of $\mathbf{s}_h$, so it would just be a common amplitude scaling. This seems a reasonable simplification to be able to treat space-time data as temporal only.

4.2.5 Hawkes process log-likelihood approximation

In this section, we illustrate our Hawkes process log-likelihood approximation technique. This approximation technique is new and allows us to express the Hawkes process log-likelihood as a sum of linear functions of the parameters $\boldsymbol{\theta}$. Suppose to have observed n events $\mathcal{H}_{T_1,T_2} = \{\mathbf{x}_1, ..., \mathbf{x}_n : \mathbf{x}_i \in \mathcal{X}\, \forall i = 1, ..., n\}$, where $\mathcal{X} = [T_1, T_2] \times W \times M$, with $0 \leq T_1 < T_2 < \infty$, $W \subset \mathbb{R}^2$, and $M \subseteq \mathbb{R}$. To ease the notation in the next steps we are using $\mathcal{H} = \mathcal{H}_{T_1,T_2}$ to indicate the complete set of observations. The general point process model log-likelihood given the observations is:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{H}) = -\Lambda(\mathcal{X}|\mathcal{H}) + \sum_{h=1}^{n} \log \lambda(\mathbf{x}_h|\mathcal{H}_{t_h}), \tag{4.9}$$

where $\mathcal{H}_{t_h}$ is the subset of $\mathcal{H}_{T_1,T_2}$ of events recorded strictly before t_h and,

$$\Lambda(\mathcal{X}|\mathcal{H}) = \int_{\mathcal{X}} \lambda(\mathbf{x}|\mathcal{H})d\mathbf{x}, \tag{4.10}$$

is the integrated conditional intensity corresponding to the expected number of points in $\mathcal{X}$. The integrated conditional intensity can be decomposed using the branching structure of Hawkes processes, indeed, we can think of the expected number of points in an area as the expected number of background points plus the expected number of points induced by each observation in the history. Formally, having observed $n = |\mathcal{H}_{T_1,T_2}|$ events,

$$\Lambda(\mathcal{X}|\mathcal{H}) = \Lambda_0(\mathcal{X}) + \sum_{h=1}^{n} \Lambda_h(\mathcal{X}), \tag{4.11}$$

where,

$$\Lambda_0(\mathcal{X}) = \int_{\mathcal{X}} \mu(\mathbf{x})d\mathbf{x} = (T_2 - T_1)\mu \tag{4.12}$$

is the integrated background rate, and is interpreted as the number of expected background events. The last equation only holds if the background rate follows the definition in Equation 4.4. The other quantity is given by

$$\Lambda_h(\mathcal{X}) = \int_{\mathcal{X}} g(\mathbf{x}, \mathbf{x}_h)d\mathbf{x} = g_m(m_h) \int_{\max(T_1,t_h)}^{T_2} \int_W g_t(t - t_h)g_s(\mathbf{s} - \mathbf{s}_h)dtd\mathbf{s}, \tag{4.13}$$

and is interpreted as the number of expected points generated by the observation $\mathbf{x}_h$. The last equation only holds if we use Equation 4.5 to define the triggering function.

The log-likelihood can be decomposed into three main components:

$$\mathcal{L}(\boldsymbol{\theta}) = -\Lambda_0(\mathcal{X}) - \sum_{h=1}^{n} \Lambda_h(\mathcal{X}) + \mathrm{SL}(\mathcal{H}). \tag{4.14}$$

The expected number of background events $\Lambda_0(\mathcal{X})$, the expected number of induced events $\sum_h \Lambda_h(\mathcal{X})$, and the sum of the log-intensities $\mathrm{SL}(\mathcal{H}) = \sum_h \log \lambda(\mathbf{x}_h|\mathcal{H}_{t_h})$.

Our technique is based on approximating these three components separately. The approximation is such that the value of the log-likelihood is exact at the posterior mode $\boldsymbol{\theta}^*$, and the degree of accuracy decays as we move from there. The level of accuracy for values of the parameters far from the posterior mode strongly depends on the choice of the triggering functions. Specifically, we separately perform a linear approximation of $\log \Lambda_0(\mathcal{X})$, $\log \Lambda_h(\mathcal{X})$, and $\log \lambda(\mathbf{x}_h)$, for $h = 1, ..., n$, and therefore, these functions should be as close to being linear as possible.

The next subsections illustrate the approximation of the different log-likelihood components. The last subsection reports some details on the iterative algorithm used to determine the mode of the posterior distribution around which the approximation is performed. For all of them, we will make explicit the dependence of the log-likelihood components from $\boldsymbol{\theta}$ and omit dependence from the domain $\mathcal{X}$, formally, $\Lambda(\mathcal{X}) = \Lambda(\mathcal{X}, \boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta})$. Also, if a quantity is approximated we use the Tilde symbol, such that $\tilde{f}(x)$ is the approximation of $f(x)$, while over-lined quantities stand for linearised, such that $\overline{f}(x, x_0)$ is the linear version of $f(x)$ with respect to x_0.

Part I - Expected Number of background events

We approximate the integrated background rate using a linear approximation of its logarithm. Namely,

$$\tilde{\Lambda}_0(\boldsymbol{\theta}) = \exp\{\overline{\log \Lambda}_0(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\}, \tag{4.15}$$

where,

$$\overline{\log \Lambda}_0(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \log \Lambda_0(\boldsymbol{\theta}^*) + \frac{1}{\Lambda_0(\boldsymbol{\theta}^*)} \sum_{j=1}^{m} (\theta_j - \theta_j^*) \frac{\partial}{\partial \theta_j} \Lambda_0(\boldsymbol{\theta}) \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}. \tag{4.16}$$

This approach is particularly convenient if the background rate has the form reported by Equation 4.4. The only parameter to estimate using this approximation is $\mu \geq 0$. Changing parameter to $\theta_\mu = \log \mu$, we have two huge advantages. First, $\theta_\mu \in (-\infty, \infty)$ is a free-constraint parameter, and second, the logarithm of the expected number of background events is linear in θ_μ, which means that there will be no approximation at this step and this component will be exact for any value of θ_μ.

Part II - Expected Number of triggered events

We start the approximation of the expected number of triggered events by considering the expected number of events triggered by a single observation $\mathbf{x}_h$. This is given by Equation 4.13. Considering a partition of the space $\mathcal{X}$, namely $b_{1,h}, ..., b_{B_h,h}$ such that $\bigcup_i b_{i,h} = \mathcal{X}$ and $b_{j,h} \bigcap b_{i,h} = \emptyset$, $\forall i \neq j$, we can write:

$$\Lambda_h(\boldsymbol{\theta}) = \sum_{i=1}^{B_h} \int_{b_{i,h}} g(\mathbf{x}, \mathbf{x}_h) d\mathbf{x} = \sum_{i=1}^{B_h} \Lambda_h(b_{i,h}, \boldsymbol{\theta}). \tag{4.17}$$

We approximate the above quantity linearly approximating the logarithm of the elements of the summation. This increase the computational time and memory required by the algorithm but it provides a much better approximation than considering one bin only. More formally,

$$\tilde{\Lambda}_h(\boldsymbol{\theta}) = \sum_{i=1}^{B_h} \exp\{\overline{\log \Lambda}_h(b_{i,h}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\}, \tag{4.18}$$

where $\overline{\log \Lambda}_h(b_{i,h}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the linear approximation with respect to the posterior mode of the expected number of generated events by the observation $\mathbf{x}_h$ in the area $b_{i,h}$ and has the same form of Equation 4.16.

Assuming that we are dealing with a spatio-temporal marked Hawkes process model with triggering function given by Equation 4.5 and bins partitioning the time domain only, such that $b_{i,h} = [t_{i-1,h}, t_{i,h}) \times W$ for $i = 1, ..., B_h$ and $t_{i,h} < t_{j,h} \forall i < j$ and $t_0 = \max(T_1, t_h)$ and $t_B = T_2$, we have that:

$$\begin{aligned}\Lambda_h(b_{i,h}, \boldsymbol{\theta}) &= g_m(\mathbf{x}_{t_h}, \boldsymbol{\theta}^{(m)}) \left(\int_{t_{i-1,h}}^{t_{i,h}} g_t(t - t_h, \boldsymbol{\theta}^{(t)}) dt \right) \left(\int_W g_{\mathrm{s}}(\mathbf{s} - \mathbf{s}_h, \boldsymbol{\theta}^{(\mathrm{s})}) d\mathbf{s} \right) \\ &= g_m(m_h, \boldsymbol{\theta}^{(m)}) I_t(b_{i,h}, \boldsymbol{\theta}^{(t)}) I_{\mathrm{s}}(\boldsymbol{\theta}^{(\mathrm{s})}),\end{aligned} \tag{4.19}$$

where $I_t(b_{i,h}, \boldsymbol{\theta}^{(t)})$ and $I_{\mathrm{s}}(\boldsymbol{\theta}^{(\mathrm{s})})$ are, respectively, the integral of the time and space triggering function. The derivative of the logarithm of $\Lambda_h(b_{i,h}, \boldsymbol{\theta})$ with respect to $\theta_j \in \boldsymbol{\theta}$ is given by

$$\frac{\partial}{\partial \theta_j} \log \Lambda_h(b_{i,h}) = \begin{cases} \frac{\partial}{\partial \theta_j} \log g_m(m_h), & \text{if } j \in J_m \\ \frac{\partial}{\partial \theta_j} \log I_t, & \text{if } j \in J_t \\ \frac{\partial}{\partial \theta_j} \log I_{\mathrm{s}}, & \text{if } j \in J_{\mathrm{s}} \end{cases}, \tag{4.20}$$

where J_m, J_t, J_{s} are defined in Section 4.2.4.

Therefore, the accuracy of the approximation depends on *how close* to be linear the functions $\log g_m(\cdot), \log I_t(\cdot), \log I_{\mathrm{s}}(\cdot)$ are with respect the parameters $\boldsymbol{\theta}$. In the case of normalized triggering functions, we have $\Lambda_h(\mathcal{X}) = g_m(m_h)$. This means that, on one hand, we don't need to split the integral in different bins saving computational time and memory; on the other hand, the information on the parameters $\theta_j \in \boldsymbol{\theta}^{(t)} \bigcup \boldsymbol{\theta}^{(\mathrm{s})}$ provided by this likelihood component is lost. Also, normalized triggering functions tend to be *farther* from linearity than the corresponding unnormalized versions and this is crucial for the approximation of the sum of log-intensities.

We remark that the division in bins is essential for the accuracy of the approximation and the ability to converge of the algorithm. Different binning strategies can be employed, and their performance depends on the form of the triggering function. For example, in the case in which the time triggering function represents the time-decay of the influence of an observation on the intensity, we expect it to be a monotonic decreasing function of the time difference and, therefore, a convenient strategy would be to consider a denser partition around zero and larger bins far from it where the function flattens. In Appendix 4.4.2 we illustrate the binning strategy used in the real data and simulation examples which has the characteristics described above. In there, we perform a sensitivity analysis fitting the same Hawkes process model using different binning strategies, and Table 4.6 compares the different binning strategies in terms of computational time and ability to converge.

Part III - Sum of log-intensities

For the sum of log-intensities calculated at the observed points, we simply consider the linear approximation of the elements of the summation, namely

$$\widetilde{\mathrm{SL}}(\mathcal{H}) = \sum_{h=1}^{n} \overline{\log \lambda}(\mathbf{x}_h, \boldsymbol{\theta}, \boldsymbol{\theta}^*), \tag{4.21}$$

where, omitting the dependence from $\mathbf{x}_h$,

$$\overline{\log \lambda}(\mathbf{x}_h, \boldsymbol{\theta}, \boldsymbol{\theta}^*) = \log \lambda(\boldsymbol{\theta}^*) + \frac{1}{\lambda(\boldsymbol{\theta}^*)} \sum_{j=1}^{m} (\theta_j - \theta_j^*) \frac{\partial}{\partial \theta_j} \lambda(\boldsymbol{\theta}) \Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}, \tag{4.22}$$

which is the same as Equation 4.16.

Assuming to be interested in a spatio-temporal marked Hawkes process model, with background rate specified by Equation 4.4, considering $u(\mathbf{s})$ known for any $\mathbf{s} \in W$, and triggering function specified by Equation 4.5, the conditional intensity is given by:

$$\lambda(\mathbf{x}_h | \mathcal{H}_{t_h}) = \mu u(\mathbf{s}_h) + \sum_{k : x_k \in \mathcal{H}_{t_h}} g_m(m_k) g_t(t_h - t_k) g_{\mathrm{s}}(\mathbf{s}_h - \mathbf{s}_k), \tag{4.23}$$

with derivative with respect to $\boldsymbol{\theta}$ equal to

$$\frac{\partial}{\partial \theta_j} \lambda(\mathbf{x}_h) = \begin{cases} u(\mathbf{s}_h), & \text{if } \theta_j = \mu \\ \\ \sum_k g_t(t_h - t_k) g_{\mathrm{s}}(\mathbf{s}_h - \mathbf{s}_k) \frac{\partial}{\partial \theta_j} g_m(m_k), & \text{if } j \in J_m \\ \\ \sum_k g_m(m_k) g_{\mathrm{s}}(\mathbf{s}_h - \mathbf{s}_k) \frac{\partial}{\partial \theta_j} g_t(t_h - t_k), & \text{if } j \in J_t \\ \\ \sum_k g_m(m_k) g_t(t - t_k) \frac{\partial}{\partial \theta_j} g_{\mathrm{s}}(\mathbf{s}_h - \mathbf{s}_k), & \text{if } j \in J_{\mathrm{s}} \end{cases}. \tag{4.24}$$

The above expression indicates that the accuracy of the approximation depends on how close to linearity the different triggering function components are.

Full approximation and `inlabru` implementation

Putting all together, the Hawkes process log-likelihood approximation used by our technique is:

$$\begin{aligned} \tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) &= -\tilde{\Lambda}_0(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \sum_{h=1}^{n} \sum_{i=1}^{B_h} \tilde{\Lambda}_h(b_{i,h}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) + \tilde{SL}(\mathcal{H}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \\ &= - \exp\{\overline{\log \Lambda}_0(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} - \sum_{h=1}^{n} \sum_{i=1}^{B,h} \exp\{\overline{\log \Lambda}_h(b_{i,h}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\} + \sum_{h=1}^{n} \overline{\log \lambda}(\mathbf{x}_h, \boldsymbol{\theta}, \boldsymbol{\theta}^*). \end{aligned} \tag{4.25}$$

The approximation is performed with respect to the mode of the posterior distribution $\boldsymbol{\theta}^*$, which is determined by an iterative algorithm. The algorithm starts from a linearisation point $\boldsymbol{\theta}_0^*$ (provided by the user), finds the mode of the linearised (with respect to $\boldsymbol{\theta}_0^*$) posterior using the INLA method, namely $\overline{\boldsymbol{\theta}}_1^*$, the value of the linearisation point is updated to $\boldsymbol{\theta}_1^* =$

$\gamma\boldsymbol{\theta}_0^* + (1-\gamma)\overline{\boldsymbol{\theta}}_1^*$, where the scaling $\gamma \in \mathbb{R}$ is determined by the line search method described here https://inlabru-org.github.io/inlabru/articles/method.html. This process is repeated until, for each parameter, the difference between two consecutive linearization points is less than 1% of the marginal posterior standard deviation. The value 1% is the default value used by the R-package inlabru and can be changed by the user. Regarding $\boldsymbol{\theta}_0^*$ provided by the user, we suggest setting the parameters to a value which do not lead to extreme cases. In our experience, using $\boldsymbol{\theta}_0^*$ such that all the parameters are equal to 1 is a safe choice. Another option may be to set it equal to the maximum likelihood estimators. We recommend avoiding cases where parameters are equal, or very close, to zero (e.g. $< 10^{-10}$), as well as far from it (e.g. > 1000), which may prevent the algorithm from converging.

The proposed method is implemented in inlabru combining three Poisson models on different datasets. The reference to a Poisson model is merely artificial and used for computational purposes, it does not have any specific meaning. Specifically, we leverage the internal log-likelihood used for Poisson models by INLA (and inlabru) to obtain the approximate Hawkes process log-likelihood. This is the only reason why we chose to implement our Hawkes process approximation using different Poisson models.

More formally, INLA has the special feature of allowing the user to work with Poisson counts models with exposures equal to zero (which should be improper). A generic Poisson model for counts $c_i, i = 1, ..., n$ observed at locations $\mathbf{x}_i, i = 1, ..., n$ with exposure $E_1, ..., E_n$ with log-intensity $\log \lambda_P(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta})$, in inlabru has log-likelihood given by:

$$\mathcal{L}_P(\boldsymbol{\theta}) \propto -\sum_{i=1}^{n} \exp\{\overline{f}(\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\} * E_i + \sum_{i=1}^{n} \overline{f}(\mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\theta}^*) * c_i. \tag{4.26}$$

Each Hawkes process log-likelihood component is approximated using one surrogate Poisson model with log-likelihood given by Equation 4.26 and appropriate choice of counts and exposures data. Table 4.2 reports the approximation for each log-likelihood component with details on the surrogate Poisson model used to represent it. For example, the first part (integrated background rate) is represented by a Poisson model with log-intensity $\log \Lambda_0(\mathcal{X})$, this will be automatically linearised by inlabru. Given that, the integrated background rate is just a scalar and not a summation, and therefore we only need one observation to represent it assuming counts equal 0 and exposures equal 1. Table 4.2 shows that to represent a Hawkes process model having observed n events, we need $1 + \sum_h (B_h) + n$ events with B_h number of bins in the approximation of the expected number of induced events by observation h.

Furthermore, Table 4.2 lists the components that has to be provided by the user, namely the surrogate Poisson models log-intensities. More specifically, the user only needs to create the datasets with counts c_i, exposures e_i, and the information on the events $\mathbf{x}_i$ representing the different log-likelihood components; and, to provide the functions $\log \Lambda_0(\mathcal{X})$, $\log \Lambda_h(b_{i,h})$, and, $\log \lambda(\mathbf{x})$. The linearisation is automatically performed by inlabru as well as the retrieving of the parameters' posterior distribution. Regarding the functions representing integrals, they do not need to be exact, a function performing numerical integration is also fine.

We provide a step-by-step tutorial on how to implement the approximation method described above. The tutorial gives more details on which functions has to be provided by the user, how to construct the binning strategy, how to set different priors for the parameters, and how to pass everything to inlabru to retrieve the posterior distribution of the parameters. The tutorial can be found at https://github.com/Serra314/Hawkes_process_tutorials/tree/main/how_to_build_Hawkes.

Name	Objective	Approximation	Surrogate $\log \lambda_P$	Number of data points	Counts and Exposures
Part I	$\Lambda_0(\mathcal{X})$	$\exp \overline{\log \Lambda_0}(\mathcal{X})$	$\log \Lambda_0(\mathcal{X})$	1	$c_i = 0,\ e_i = 1$
Part II	$\sum_{h=1}^{n} \sum_{i=1}^{B_h} \Lambda_h(b_{i,h})$	$\sum_{h=1}^{n} \sum_{i=1}^{B_h} \exp \overline{\log \Lambda_h}(b_{i,h})$	$\log \Lambda_h(b_{i,h})$	$\sum_h B_h$	$c_i = 0,\ e_i = 1$
Part III	$\sum_{h=1}^{n} \log \lambda(\mathbf{x}_h)$	$\sum_{h=1}^{n} \exp \overline{\log \lambda}(\mathbf{x}_h)$	$\log \lambda(\mathbf{x})$	n	$c_i = 1,\ e_i = 0$

Table 4.2: Hawkes process log-likelihood components approximation

4.2.6 Real Data Example

We provide a practical example of a temporal marked Hawkes process to illustrate the capabilities of our technique. We implement the temporal version of the Epidemic-Type-Aftershock-Sequence model (ETAS, Ogata, 1988), the most popular model to describe the evolution of seismicity in time, and we apply it to the 2016 Amatrice seismic sequence (Marzocchi et al., 2017). Specifically, we have considered 1137 events with a magnitude greater or equal to 3 from 24/08/2016 to 15/08/2017, with longitude in $(42.45, 43.08)$ and latitude in $(12.93, 13.54)$. The temporal evolution of the number of events is illustrated in Figure (4.1). The data is taken from the Italian Seismological Instrumental and Parametric Database (ISIDe, Group, 2007) downloaded from https://doi.org/10.13127/ISIDE.

The example consists of mainly two parts. In the first one, we compare the results of our implementation with the results obtained with the bayesianETAS R-package (Ross, 2021), which provides an automatic MCMC implementation of the temporal ETAS model. The implementations are compared in terms of goodness-of-fit, expected number of events, and expected number of induced events. This is because we use different parameterizations preventing us from directly comparing the posterior of the parameters. We do this to show that our technique provides similar results to the MCMC implementations but in less time. This is relevant because we are working with an approximation method, while the MCMC implementation is exact, and the fact that both implementations provide similar results shows the accuracy of our approximation method.

In the second part of this example, we provide a retrospective daily forecasting experiment in which we compare daily forecasts of seismicity against observed seismicity in terms of number of events per day, for 120 days starting from 24/08/2016, just after the first large earthquake in the sequence. This is done using the inlabru implementation only given the similarity of the results of the MCMC implementation. We use catalog-based forecasts (Savran et al., 2020) for which the forecast for each day is composed of 10000 simulated catalogs. Each simulated catalog is based on a different set of parameters extracted from the posterior distribution.

ETAS model

The ETAS model is the most used Hawkes process to model the evolution of seismicity over time and space (Ogata, 1988; Ogata and Zhuang, 2006; Ogata, 2011). We are going to implement the first version of the model which is a temporal marked Hawkes process model with the event's magnitude as marking variable. The conditional intensity of the ETAS model is given by:

$$\lambda_{\mathrm{E}}(t, m|\mathcal{H}_t) = \left(\mu + K \sum_{h:t_h<t} \exp\{\alpha(m_h - M_0)\}(t - t_h + c)^{-p} \right) \pi(m), \tag{4.27}$$

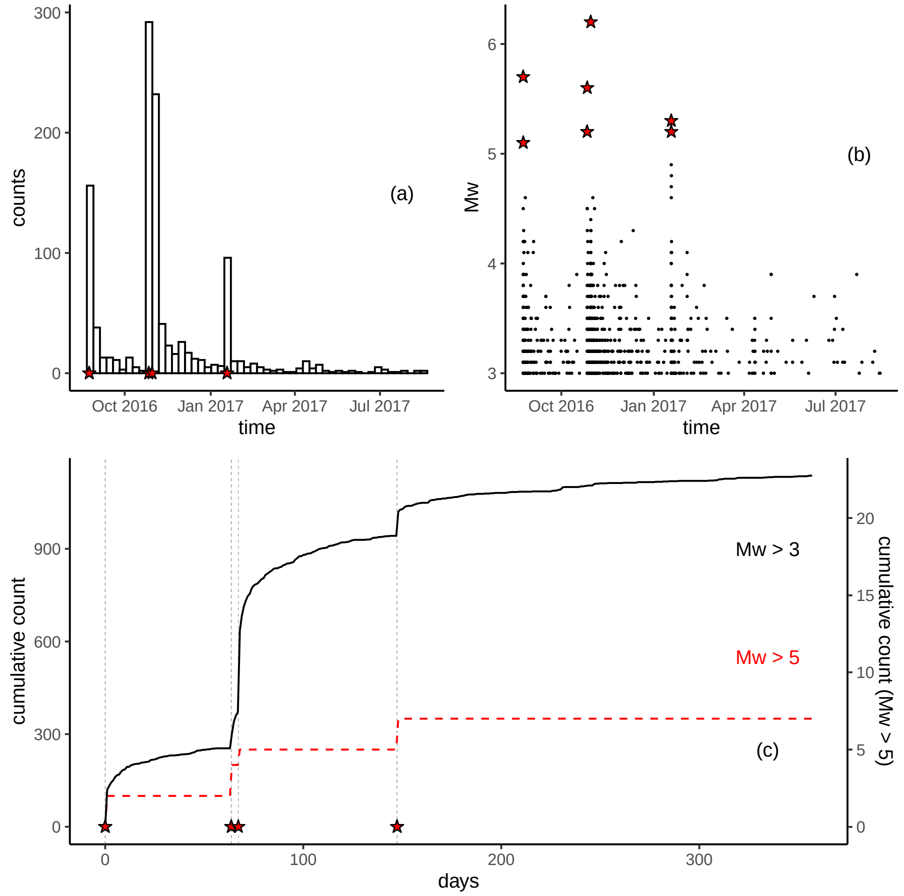


Figure 4.1: Amatrice sequence comprising 1137 events from 24/08/2016 to 15/08/2017, with longitude in $(42.45, 43.08)$ and latitude in $(12.93, 13.54)$. The first event in the catalogue is the magnitude 6.01 which started the sequence. Red stars indicate events with magnitude greater than 5. Panel (a): Histogram reporting the number of events per week; Panel (b): Scatter plot of time versus magnitude; (c) Cumulative number of events as function of the number of days from the first event in the sequence, for events with magnitude greater than 3 (solid black) and for events with magnitude greater than 5 (dashed red).

where, $M_0 \in \mathbb{R}$ is the minimum recorded magnitude, and $\pi(m)$ is the magnitude distribution which is estimated independently from the Hawkes process parameters and assumed to follow a form of Gutemberg-Richter (GR) law (Gutenberg and Richter, 1956). The temporal evolution of the number of points is regulated by 5 parameters $\mu, K, \alpha, c, \geq 0$ and $p \geq 1$. The parameters μ, K, and α are productivity parameters regulating: the number of background events (μ), the number of induced events or aftershocks (K), and how the aftershock productivity scales with magnitude (α, the higher the magnitude the more events are generated). The parameters c and p are the parameters of the Omori's law (Omori, 1894) and regulate the temporal decay of the aftershock activity. The quantity M_0 is a cut-off magnitude such that $m_h \geq M_0$, $\forall h$.

The `bayesianETAS` package implements the ETAS model with a normalized time triggering function to integrate to 1 over $(0,\infty)$. The conditional intensity is given by:

$$\lambda_{\text{bE}}(t,m|\mathcal{H}_t) = \left(\mu + K \sum_{h:t_h<t} \exp\{\alpha(m_h - M_0)\} c^{p-1}(p-1)(t-t_h+c)^{-p} \right) \pi(m). \tag{4.28}$$

With our technique, it is best to work with a different parametrization than the one used in the `bayesianETAS` package. Specifically, we choose the following conditional intensity

$$\lambda_{\text{bru}}(t,m|\mathcal{H}_t) = \left(\mu_b + K_b \sum_{h:t_h<t} \exp\{\alpha_b(m_h - M_0)\} \left(\frac{t-t_h}{c_b} + 1 \right)^{-p_b} \right) \pi(m). \tag{4.29}$$

The parameters of the `inlabru` implementation have the same constraints, and the same interpretation, as in the `bayesianETAS` implementation. The two implementations are equivalent considering

$$K_b = \frac{K(p-1)}{c}, \quad c_b = c, \quad p_b = p. \tag{4.30}$$

However, we are not going to use the above constraint in the example. The only constraints that we impose are $\mu, K, \alpha, c \geq 0$ and $p > 1$.

Priors

Priors are an essential part of the Bayesian approach. The `bayesianETAS` package has fixed priors that cannot be changed. Specifically, they consider,

$$\begin{aligned} \mu &\sim \text{Gamma}(0.1, 0.1) \\ K, \alpha, c &\sim \text{Unif}(0, 10) \\ p &\sim \text{Unif}(1, 10). \end{aligned} \tag{4.31}$$

This set of priors induces a prior on the parameter K_b, using Equation (4.30), with very light tails, highlighting how informative uniform priors may be (Zhu and Lu, 2004). We use the same set of priors except for K_b for which we choose a log-normal distribution matching the 1 and 99% quantiles of the empirical distribution of K_b obtained simulating 1000000 independent samples of K, c, p from the priors in Equation (4.31). We chose a log-normal distribution with mean and standard deviation of the logarithm equal to -1 and 2.03. Table 4.3 reports summary statistics of the `bayesianETAS` prior for K_b and the log-normal prior we chose to replicate it. The full set of priors used to replicate the `bayesianETAS` priors are

$$\begin{aligned} \mu_b &\sim \text{Gamma}(0.1, 0.1) \\ K_b &\sim \text{LogN}(-1, 2.03) \\ \alpha_b, c_b &\sim \text{Unif}(0, 10) \\ p_b &\sim \text{Unif}(1, 10). \end{aligned} \tag{4.32}$$

We use this replicate set of priors to minimize the differences between the implementations which do not depend on the methodology used to find the posterior distribution of the parameters. We refer to this case as `inlabru` replicate case.

We also consider a different set of priors that better reflects the scale of each parameter. For example, for the `inlabru` implementation the parameters, μ and c are on a very different scale than K, α, and p. To reflect this piece of information through the prior, we use

Implementation	Mean	St.Dev	0.01q	0.25q	Median	0.75q	0.99q
`bayesianETAS`	11.854	3583.873	0.004	0.111	0.262	0.758	41.914
`inlabru`	2.887	22.482	0.003	0.094	0.368	1.447	41.367

Table 4.3: Prior distribution summary statistics of parameters K_b in the `bayesianETAS` and `inlabru` implementation. The distribution in the `bayesianETAS` case is obtained sampling independently 1000000 times from $K, c \sim \text{Unif}(0, 10)$, $p \sim \text{Unif}(1, 10)$, and setting $K_b = K(p-1)/c$. The distribution in the `inlabru` case is a log-normal distribution with mean and standard deviation of the logarithm equal to -1 and 2.03 in order to match the extreme quantiles of the `bayesianETAS` case.

Name	Mean	St.Dev	0.01q	0.25q	Median	0.75q	0.99q	Implementation
μ	1	3.162	0.000	0.000	0.006	0.353	15.884	`bayesianETAS`
μ	0.1	0.316	0.000	0.000	0.001	0.035	1.588	`inlabru` - Gamma
K_b	11.854	3583.873	0.004	0.111	0.262	0.758	41.914	`bayesianETAS`
K_b	2	2	0.020	0.575	1.386	2.773	9.210	`inlabru` - Gamma
α	5	2.88	0.1	2.5	5	7.5	9.9	`bayesianETAS`
α	2	2	0.020	0.575	1.386	2.773	9.210	`inlabru` - Gamma
c	5	2.888	0.1	2.5	5	7.5	9.9	`bayesianETAS`
c	0.1	0.316	0.000	0.000	0.001	0.035	1.588	`inlabru` - Gamma
p	5.5	2.598	1.09	3.25	5.5	7.75	9.91	`bayesianETAS`
p	1.2	0.632	1.000	1.000	1.001	1.071	4.177	`inlabru` - Gamma

Table 4.4: Prior distribution summary statistics of ETAS parameters for the `bayesianETAS` implementation and the `inlabru` gamma case which considers $\mu, c \sim \text{Gamma}(0.1, 1)$, $K, \alpha \sim \text{Gamma}(1, 0.5)$, and $p - 1 \sim \text{Gamma}(0.1, 0.5)$.

gamma priors for all parameters with different parameters reflecting the different scales. This information is usually available from previous studies of the same model. We use

$$\begin{aligned}
\mu_b &\sim \text{Gamma}(0.1, 1) \\
K_b &\sim \text{Gamma}(1, 0.5) \\
\alpha_b &\sim \text{Gamma}(1, 0.5) \\
c_b &\sim \text{Gamma}(0.1, 1) \\
p_b - 1 &\sim \text{Gamma}(0.1, 0.5).
\end{aligned} \tag{4.33}$$

Table (4.4) reports a comparison between summary statistics of `bayesianETAS` priors and the gamma priors.

In the remainder of the article, we refer to the `inlabru` implementation considering the priors in Equation 4.32 as `inlabru` replicate and to the `inlabru` implementation with the priors in Equation 4.33 as `inlabru` gamma. Appendix 4.4.1 compares the prior and the posterior distributions for each model and shows the robustness of `inlabru`'s results under change of priors. Furthermore, Appendix 4.4.3 provides a more complete prior sensitivity analysis. In there, we consider all the parameters as having the same log-normal prior, with the logarithmic mean equal 0 and different values of the logarithmic standard deviation.

Copula transformation

The INLA method is designed for Latent Gaussian models and, therefore, all the parameters should have a Normal distribution. This is not the case for the ETAS parameters and the priors illustrated in the previous section. In order to overcome this problem we are going to use a copula transformation. Using this method allows us to represent internally the parameter as free-constraints and normally distributed. The constraints are implemented through the transformation itself.

More formally, we use a transformation method based on the probability integral transform. The probability integral transform can be stated as follows:

Theorem 4.2.1. *Given a continuous random variable X with cumulative distribution function (CDF) $F_X(\cdot)$, then the variable*

$$Y = F_X(X)$$

has a Uniform distribution in (0,1).

The theorem implies also that given $Y \sim \text{Unif}(0, 1)$ then, $X = F_X^{-1}(Y)$.

We apply this theorem by considering each parameter as having a standard normal distribution and then, transforming it to have the target distribution. More formally, assume θ has a starting distribution with CDF $F_\theta(\cdot)$, and that we want to transform it in $\eta(\theta)$ having a target CDF $F_Y(\cdot)$. Applying the transformation

$$\eta(\theta) = F_Y^{-1}\left(F_\theta(\theta)\right), \tag{4.34}$$

the quantity $\eta(\theta)$ is distributed according to F_Y.

This allows us to consider a set of internal free-constraint parameters $\theta_\mu, \theta_K, \theta_\alpha, \theta_c, \theta_p$, representing (respectively) μ, K, α, c, p, with a standard normal prior distribution and then transforming them to have the desired prior distribution. We can incorporate the constraint on the parameter values using appropriate prior distributions. For example, using any distribution with positive support ensures that the transformed parameter is greater or equal to zero.

Goodness-of-fit

We compare the `inlabru` and the `bayesianETAS` implementation in terms of goodness-of-fit. This is due to the use of different parametrizations. Indeed, different parametrizations and different priors make a direct comparison of the posterior of the parameters elusive, because it is hard to determine if the differences in the posterior distributions come from the different parameterizations, the different priors, or the different methodologies. With this section, we want to convince the reader that our approximation provides results similar in terms of goodness-of-fit to MCMC implementations but in less time. This is relevant considering that MCMC is an exact method, with the ability to sample from the true marginal posteriors of the model, while our method is based on a series of approximations. Showing that the `inlabru` implementation provides similar results shows the goodness of the approximation.

We compare the goodness-of-fit of the models using the Random Time Change Theorem (Meyer, 1971). This is a standard technique to measure the goodness-of-fit for Hawkes process models as described in Laub et al. (2021). Below we report the Random Time Change Theorem as stated in Laub et al. (2021) (Theorem 9.1):

Theorem 4.2.2. *Say $\mathcal{H} = \{t_1, ..., t_k\}$ is a realisation over time $[0, T]$ from a point process* *with conditional intensity $\lambda(t|\mathcal{H})$. If $\lambda(t|\mathcal{H})$ is positive over $[0, T]$ and $\Lambda(T) < \infty$ almost surely, then the transformed points $\{\Lambda(t_1), ..., \Lambda(t_k)\}$ form a Poisson process with unit rate.*

Where in our case,

$$\Lambda(t_i|\mathcal{H}) = \int_{M_0}^{\infty} \int_0^{t_i} \lambda(t, m|\mathcal{H}) dt dm. \tag{4.35}$$

In other words, if we calculate the sequence of values $\Lambda(t_1), ..., \Lambda(t_n)$, for observed $t_1, ..., t_n$, using the respective expressions of $\Lambda(t_i)$ for the `bayesianETAS` and `inlabru` implementation, we have to obtain a sequence of points uniformly distributed over the interval $[0, n]$, where n is the number of observed points. For the MCMC method, we consider estimates based on 10000 posterior samples with a burn-in of 5000 samples. The `bayesianETAS` package requires around 9 minutes to generate a total of 15000 posterior samples, while the `inlabru` method only requires around 3 minutes to converge. Section 4.2.7 shows how these times scales increasing the number of observations, while Appendix 4.4.2 illustrates the variation of the `inlabru` computational time for different binning strategies.

Figure 4.2a-c compares the sequences $\Lambda_{\text{bE}}(t_1), ...\Lambda_{\text{bE}}(t_n)$, and $\Lambda_{\text{bru}}(t_1), ...\Lambda_{\text{bru}}(t_n)$ with observed cumulative counts $N(t_1), ..., N(t_n)$. Figure 4.2b-d shows the cumulative counts as a function of $\Lambda(t_h)$ and should look like a straight line if the values are uniformly distributed as expected by the theorem. For both plots, we report 95% posterior intervals for the quantity of interest based on 10000 samples from the posterior of the parameters.

There are small differences between the two `inlabru` implementations, which was expected from the similarity of the posterior distributions provided by the model and reported in Figure 4.9. The differences in the results are greater if we compare the `bayesianETAS` and the `inlabru` implementations. In fact, the `inlabru` implementation estimates a lower background rate (around 1/4 of the MCMC one) and a greater capability of each event of generating aftershocks, which allows the prediction to match the observations in the last part of the sequence. In fact, in Figure 4.2 (d) the dashed line representing the theoretical uniform distribution is outside the `bayesianETAS` boundaries while it is inside the `inlabru` ones. Apart from these small differences, the three implementations provide consistent results.

The main difference between the `bayesianETAS` and `inlabru` implementations is the computational time. The `bayesianETAS` R-package requires around $4, 6, 9$ minutes to generate, respectively, $1000, 5000, 10000$ posterior samples considering 5000 burn-in samples. Our `inlabru` implementations require around 3 minutes to converge for different binning strategy. The minimum convergence time is 2.93 minutes obtained, while the maximum is 3.7. Table 4.6 reports the computational time and iterations needed for convergence for different binning strategy parameters.

Expected number of events and branching ratio

We also compare the `inlabru` and `bayesianETAS` implementations in terms of the expected number of events and branching ratio. This is done because these two quantities are usually relevant in applications. Given a Hawkes process model with conditional intensity $\lambda(t|\mathcal{H}_t)$, the expected number of events in a time interval (T_1, T_2), $0 \leq T_1 < T_2 < \infty$ given the history of the process is given by the integral of the conditional intensity

$$\Lambda(T_1, T_2) = \int_{T_1}^{T_2} \lambda(t|\mathcal{H}_t) dt. \tag{4.36}$$

The number of points has a Poisson distribution with rate $\Lambda(T_1, T_2)$.

Figure 4.3 (right) shows the posterior distributions of $\Lambda(T_1, T_2)$ for the `inlabru` and `bayesianETAS` implementations. We show only the `inlabru` replicate case given that the

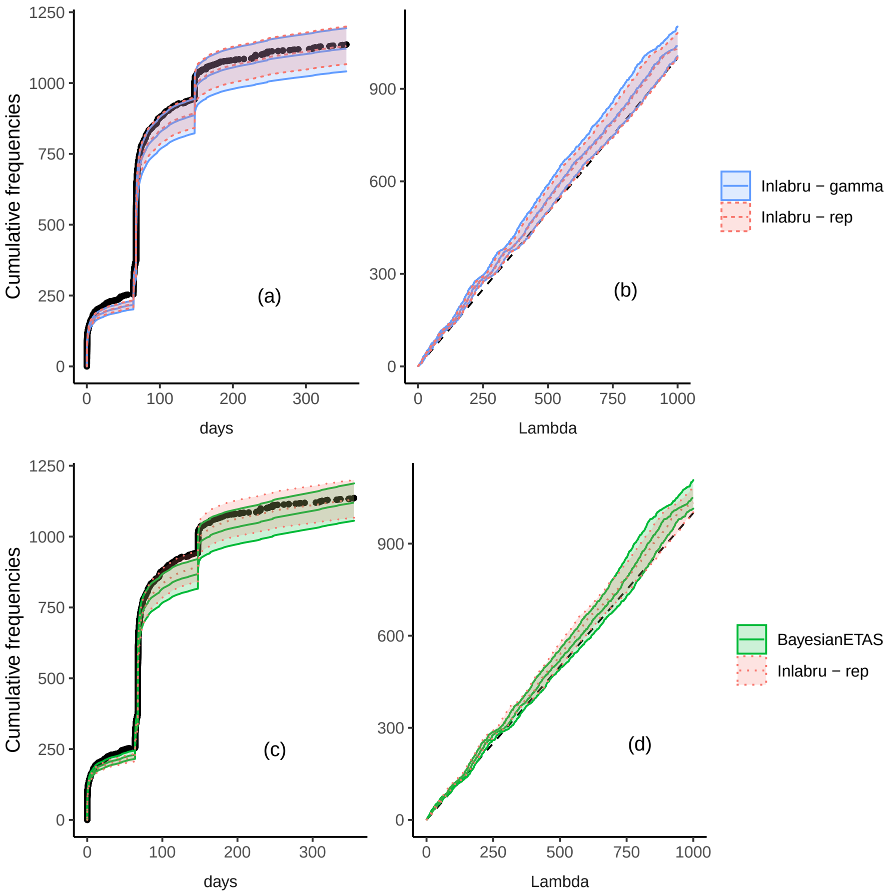


Figure 4.2: Application of the Random Time Change Theorem. Top row (a-b): Compares the `inlabru` replicate and gamma (solid blue) implementations. Panel a-c: Observed cumulative number of events as a function of time (black dots) with the prediction provided by the model; Bottom row (c-d): Compares the `bayesianETAS` (solid green) and `inlabru` replicate (dotted red) implementations; Panel a-c : Cumulative number of events as a function of time. Panel b-c : Cumulative number of events as a function of $\Lambda(t_h)$, the black dashed line represents the uniform case. The shaded region represents the 95% predictive interval for each quantity obtained by sampling 10000 times the posterior of the parameters.

`inlabru` gamma case provides the same results. For the two implementations, the posterior distribution of $\Lambda(T_1, T_2)$ is estimated by calculating the analytical expression of $\Lambda(T_1, T_2)$ for the two approaches using 10000 samples from the posterior distribution of the parameters. The approaches provide coherent results between each other, although the mode of the posterior distribution of $\Lambda(T_1, T_2)$ is closer to the observed number of points (vertical dashed line) in the `inlabru` case.

Another important quantity in analyzing Hawkes process models is the branching ratio BR. The branching ratio is the expected total number of events induced by another event. The branching ratio can be calculated as the integral of the excitation (triggering) function for time differences going from 0 to ∞. In the ETAS case, we have an excitation function that depends also on the magnitude, namely $g : (0, \infty) \times (M_0, \infty) :\to (0, \infty)$ such that

$$g(t - t_h, m_h) = g_t(t - t_h, m_h)\pi(m_h), \tag{4.37}$$

where $\pi(m_h)$ is the magnitude distribution.

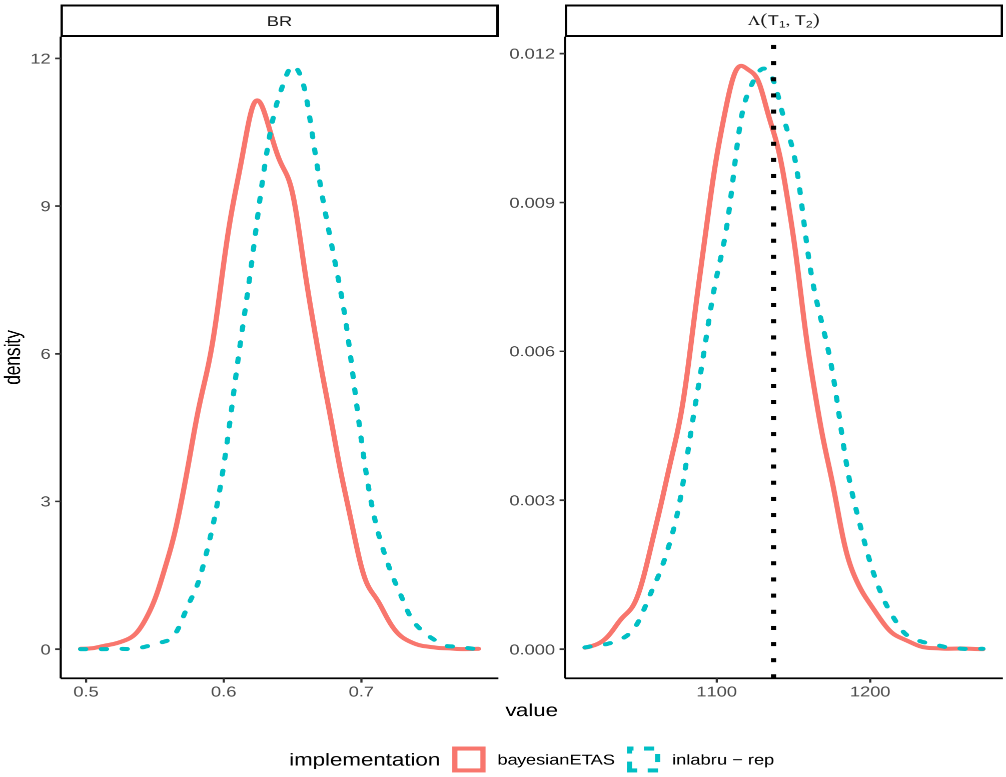


Figure 4.3: Right Panel: Expected number of events $\Lambda(T_1, T_2)$ posterior distribution comparison for the `inlabru` replicate implementation (blue dashed) and the `bayesianETAS` implementation (red solid). The vertical dotted line represents the observed number of points. Left Panel: Branching ratio BR posterior distribution comparison for the `inlabru` replicate implementation (blue dashed) and the `bayesianETAS` implementation (red solid).

In this case, the branching ratio is given by

$$\text{BR} = \int_{M_0}^{\infty} \left(\int_0^{\infty} g_t(s, m) ds \right) \pi(m) dm. \tag{4.38}$$

Therefore, the branching ratio can be seen as the expected value under the magnitude distribution of the expected number of events induced by another. Assuming to have a point in 0, then the number of points induced by that event has a Poisson distribution with rate $\sum_{i=1}^{\infty} \text{BR}^i$. As explained by Laub et al. (2021) in Section 3 the branching ratio should be between 0 and 1 for the process to be stationary and for asymptotic results to be valid (Hawkes, 1971b). We did not set any constraints to ensure this property in the present implementation.

To calculate the branching ratio for a given set of parameters, we calculate analytically the inner integral 10000 times, using samples from the magnitude distribution and we take the mean. This is repeated for 10000 times, using as ETAS parameters samples from the parameters' posterior distribution. In this way, we obtain 10000 samples from the posterior distribution of the branching ratio which can be used to approximate the posterior distribution empirically. Figure 4.3 (left) compares the posterior distributions of the branching ratio for the `inlabru` and `bayesianETAS` implementations. Both posterior distributions only assign a positive probability value between 0 and 1. The one obtained with `inlabru` has a slightly smaller posterior variance and a larger mode. This is due to the smaller background rate estimated by the `inlabru` implementation which in turns imply a higher number of induced events.

Retrospective Forecasting Experiment

We perform a retrospective daily forecasting experiment using the same data used to fit the data on the Amatrice seismic sequence. We choose to do retrospective forecasts and not pseudo-prospective or prospective because we want to check the ability of the model in describing the data on which has been calibrated, which is, in my opinion, the first challenge a model has to pass to be used operationally. Also, the model will be submitted to the next Italy CSEP experiment to be started in 2023 in which the models will be tested prospectively.

For each forecasting period defined by (t_j, t_{j+1}), we simulate 10000 synthetic catalogs assuming known all the events happened strictly before the forecasting period, namely $\mathcal{H}_{t_j}$. If, in the forecasting period (t_j, t_{j+1}) there is an earthquake with magnitude greater than 5.5 with recorded time $t_m : t_j < t_m < t_{j+1}$, then, we consider the forecast for the period t_j, t_m and we start a new daily forecast from $t_m + dt$, for $dt > 0$ (we use $dt = 10^{-6}$ days). This is done to resemble a true forecasting experiment, like the ones performed by the Collaboratory for the Study of Earthquake Predictability (CSEP, Savran et al., 2020, and reference therein), in which the forecasts are updated in presence of large earthquakes.

The results of the retrospective experiment are shown in Figure 4.4. The shaded region represents the 95% forecasting interval of the number of events for each period. The extremes of each interval are the 2.5% and the 97.5% quantiles of the number of events of the synthetic catalogs composing the forecast for each day. Around 90% of the number of events are comprised in the forecasted intervals and the model shows a temporal decay which agrees with the data. The fact that the forecast misses around 5% more days than expected can be explained by the time at which the forecasts are issued. Indeed, we issue a forecast at midnight except for the days with events above magnitude 5.5, for which the forecast is updated to start 1 second after the event. We can observe that the days with the greatest number of events are indeed forecasted correctly. Therefore, updating the forecast more often will provide a better coverage.

4.2.7 Simulation Experiment

We performed a simulation example to compare the robustness of the `inlabru` and `bayesianETAS` approach if applied to different catalogs coming from the same model, and to give an idea of how the computational time scales increasing the amount of data. As data generating model, we use the `inlabru` replicate implementation presented in Section 4.2.6. We generate 10000 synthetic catalogs for the period going from 24/08/2016 to 15/08/2017 (same period used for the Amatrice sequence) using as parameters the posterior median. In simulating the catalogs, we assume as known the 3 events with the greatest magnitude in the Amatrice catalogue recorded, respectively, on the 24/08/2016, 26/10/2016, and 30/10/2016, with magnitudes 5.7, 5.6, and 6.2. This is done to have a high probability of having, at least, 800 events per catalog. From the set of synthetic catalogs we select 5 catalogs corresponding to $900, 1500, 2000, 2500, 3500$ number of events. We use these catalogs to fit 5 different models with the `inlabru` and `bayesianETAS` implementations. For the `inlabru` implementation, we use the same priors and starting points as in the `inlabru` replicate case and binning strategy parameters given in Appendix 4.4.2. For the `bayesianETAS` implementation, we consider 5000 posterior samples with 5000 burn-in samples.

Table 4.5 shows how the computational time scales increasing the number of events in the data for the two implementations. The advantages of the `inlabru` approach are clear,

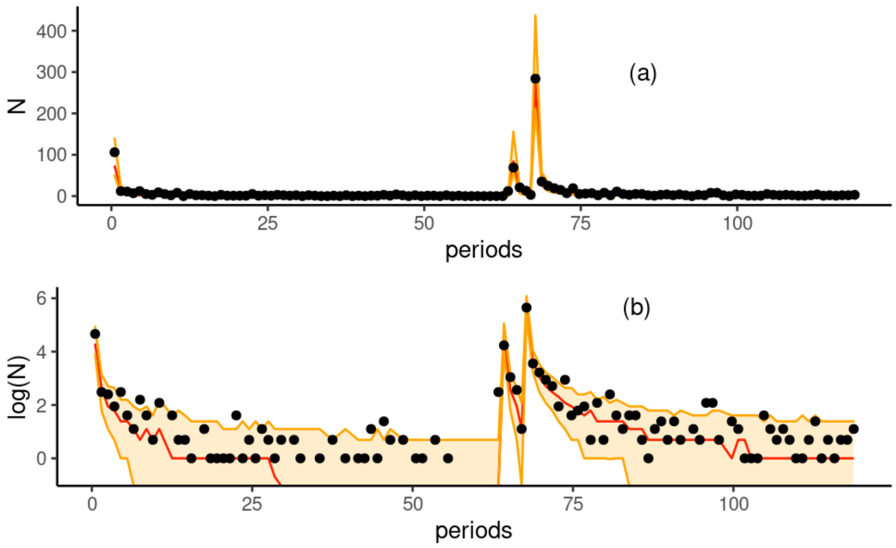


Figure 4.4: Retrospective forecasting experiment results. Black dots represent the observed number of events per forecasting period; the red solid line represents the median of the number of events of the synthetic catalogs per forecasting period; the shaded region represents the 95% forecasting intervals for the number of events of the synthetic catalogs per forecasting period. Panel (a) shows the number of events in the natural scale. Panel (b) shows the logarithm of the number of events, periods with zero events have been omitted.

especially for catalogs with more than 2500 events for which `inlabru` is 10 times faster than `bayesianETAS`. Figure 4.5 (`bayesianETAS`) and 4.6 (`inlabru`) show the posterior of the parameters for the different simulated catalogs. The differences between the posteriors obtained by each approach on different catalogs are expected. For example, the case with 3500 (as well as 900) events can be considered an extreme case and, thus, the posterior distribution would be different from more *common* catalogs. Indeed, the parameters μ, K, α, regulating the number of events, are the ones with more differences in the posteriors for different catalogs, while the parameters p and c regulating the temporal decay of the induced events are more similar. In this regard, the `inlabru` implementation is more stable than the `bayesianETAS` implementation providing posteriors distributions more similar between each other. This is particularly true for parameters μ, c, and p. In addition, the two implementations provide coherent results between each other, for example, analyzing parameter α, for both approaches the parameter's posterior distribution moves to the right as we increase the

N events	`bayesianETAS`	`inlabru`	time ratio
900	3.90 (mins)	2.96 (mins)	1.31
1500	9.75 (mins)	1.56 (mins)	6.21
2002	16.80 (mins)	2.69 (mins)	6.24
2500	30.73 (mins)	2.75 (mins)	11.15
3500	56.09 (mins)	5.22 (mins)	10.72

Table 4.5: Comparison of computational times for the `bayesianETAS` and `inlabru` implementations in minutes. Last column report the ratio between the number of minutes needed by `bayesianETAS` and `inlabru`.

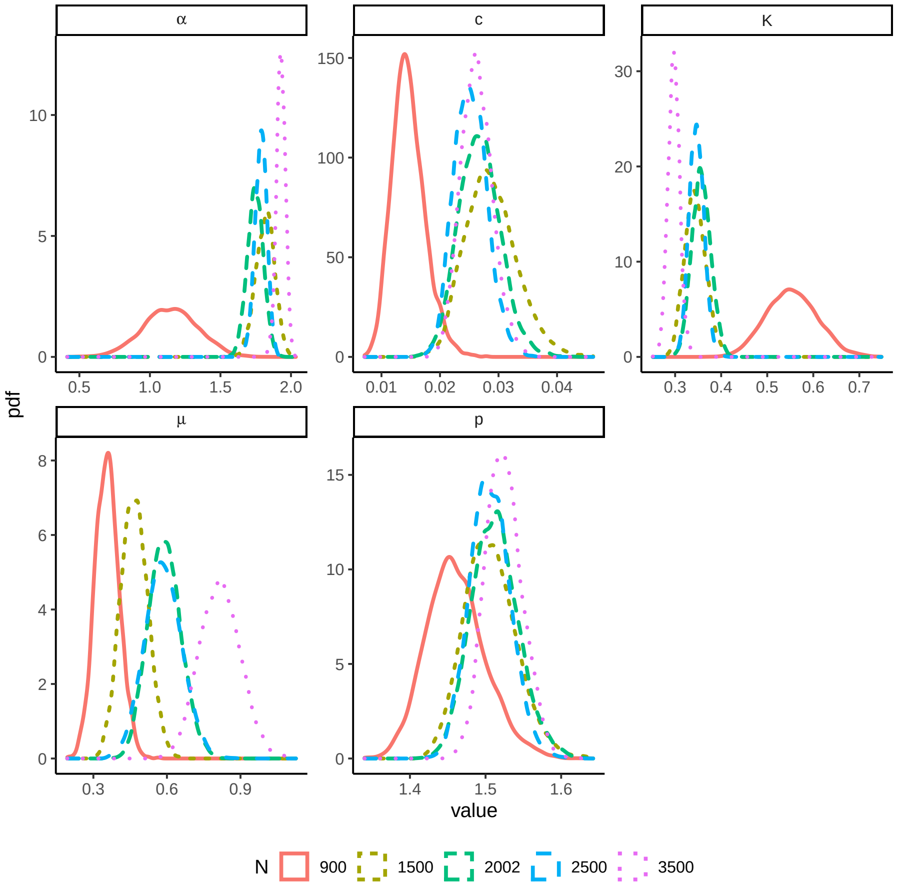


Figure 4.5: ETAS parameters' posterior distribution using the `bayesianETAS` R-package on 5 synthetic earthquake catalogs. The color and the linetype represents the number of events in each synthetic catalog. The synthetic catalogs are simulated using as parameters the median of the posterior distribution of the `inlabru` replicate implementation obtained on the Amatrice seismic sequence.

amount of data, and the opposite happens for parameter K.

The coherence of the results for the two implementations considered illustrates the reliability of our approximation, and, the gain in computational time shows the advantage of our approach. Furthermore, the gain in computational time would be even greater if more complex models are considered. For example, we foresee that the computational gain will increase considering a spatio-temporal model, or, alternatively, considering one of the parameters as temporally varying. This has not to be underrated, in fact, in seismology, many researchers are discouraged to update their models (in an online fashion) or using large catalogs (> 100000 events) by the price to pay in terms of computational time.

4.2.8 Discussion and conclusions

In this paper, we presented a technique to implement Bayesian Hawkes process models based on the INLA algorithm and carried out with the R-package `inlabru`. The proposed technique is new and differs substantially from other Hawkes process implementations. Specifically, we rely on a new Hawkes process log-likelihood approximation technique which allows us to apply the INLA method to Hawkes process models. Our technique provides similar results, in terms of goodness-of-fit, expected number of events, and branching ratio, as an MCMC

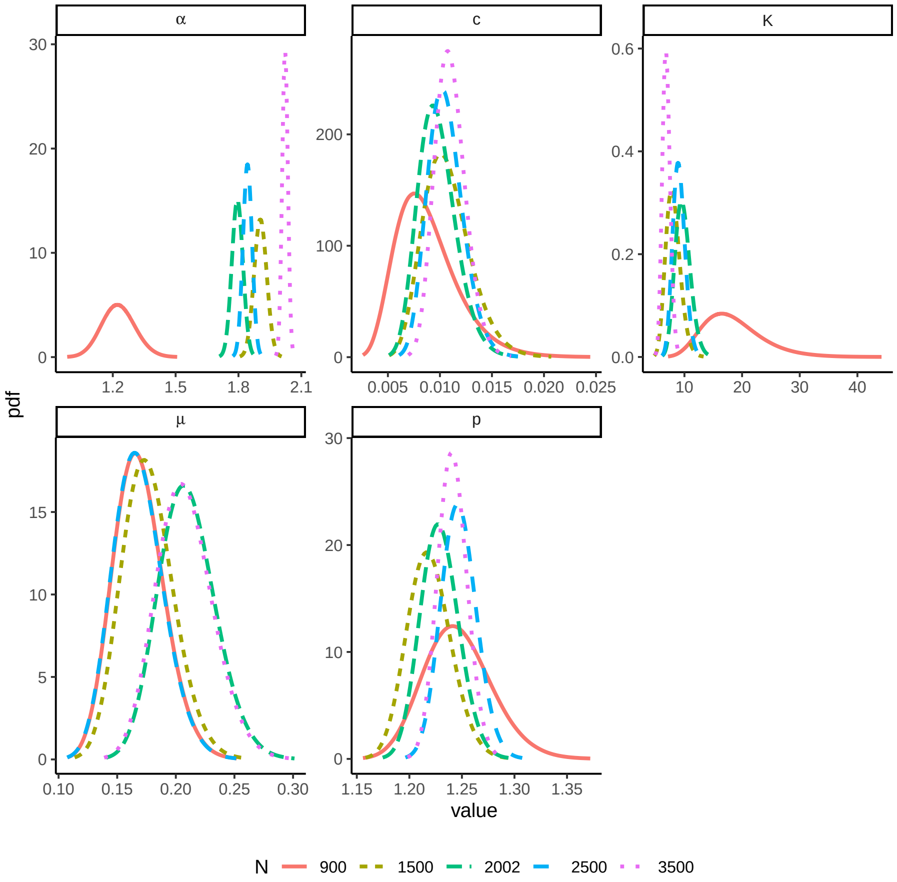


Figure 4.6: ETAS parameters' posterior distribution using the `inlabru` R-package on three synthetic earthquake catalogs. The color and the linetype represents the number of events in each synthetic catalog. The synthetic catalogs are simulated using as parameters the median of the posterior distribution of the `inlabru` replicate implementation obtained on the Amatrice seismic sequence.

technique (Ross, 2021) implemented through the `bayesianETAS` package but requiring less time. Using simulated data, I have shown that although the marginal posterior distributions obtained with the `inlabru` approach from catalogues simulated using the same parameters set do not overlap for parameter α, this happens also with the `bayesianETAS` R-package. However, the posterior distributions obtained with `inlabru` are more stable than the ones obtained with MCMC, which could be seen as an advantage of our approach. In applied contexts, where the interest is on forecasting the probabilities of future events rather than retrieving correctly the parameters value this may not be a problem, given the fact that different parameters sets can yield the same probabilities due to the correlation between parameters.

Regarding the time, the `bayesianETAS` approach requires around double the time required by our technique for catalogs composed of circa 1000 events, and 10 times more for catalogs with more than 2500 events. We believe that in more complex cases (e.g spatio-temporal case, inclusion of covariates, parameters with structured variations) the gain in computational time provided by the `inlabru` approach would be even larger. We have also shown that our technique provides reasonable results in a retrospective forecasting experiment, correctly predicting the number of events per day for most of the considered days. Furthermore, our algorithm is deterministic ensuring the same numerical results if the analysis is repeated on different machines with the same specifics. Moreover, the user does not have to program explicitly the algorithm itself, they only have to provide the functions to be approximated, and the approximation is performed automatically by the `inlabru` R-package. Also, we do not rely on any declustering algorithm assigning the observations to the background rate or the triggered part of the intensity.

An important difference from other algorithms for Hawkes process models is that we offer a general and extendable framework to perform Bayesian analyses of Hawkes process parameters. Indeed, INLA was designed for models comprising covariates and random effects, and to compare them. This allows us to bring the advantages of the Latent Gaussian model world into the Hawkes process world. For example, we can consider the parameters as linear functions of available covariates. Another extension consists of considering the parameters as structured random effects: a parameter assumed to be a Gaussian Markov Random Field (GMRF) varying over space, or time, or both. For example, considering a parameter as an SPDE effect (Lindgren et al. (2011)) we can have spatially (or temporally) varying parameters where the absolute value of the correlation between the parameter's values at different locations (times) is a decreasing function of the distance between locations (times). Given the correlation between the parameter's values and the correlation between different parameters, these models would be difficult to implement using an MCMC technique, which, in case, should be tailored to the specific problem. Hence, we take advantage of the fact that INLA was designed specifically to handle large GMRF and correlated parameters in an efficient manner. Using our method, all the models undergo the same optimization routine making them homogeneous under these aspects. When comparing two models optimized with different routines, it is hard to distinguish whether the differences come from the different models or the different algorithms. Using our technique, researchers may compare models incorporating different hypotheses being sure of no differences, at least, on the optimization part, and thus, any difference in performance comes from the model formulation itself. Comparing results obtained using the proposed technique with results obtained with different ones remains cumbersome.

The limitations of our approach reside in the functional form of the triggering (or excitation) function and the binning strategy. Specifically, we want the triggering function so that the functions to be approximated are as close as possible to be linear. In our experience, the unnormalized version of the triggering functions works best. Also, care has to be taken on the numerical stability of the provided functions which may be eased by linearly approximating them for values of the argument above/below a certain threshold. The binning strategy to further decompose Part II of the log-likelihood is essential to reach convergence. In our experience, a number of bins greater than 3 per observation is required. Also, the width of the bins is essential, considering too large bins prevents the algorithm to converge as shown in Table 4.6. We suggest to regulates the width and number of bins based on the problem at hand. For example, a triggering function decaying slowly with time would need larger bins than a function with a faster decay. With the same rationale, the function decaying slower needs fewer bins to be accurately approximated than one decaying faster.

Future developments will regard the inclusion of covariates and random effects in the model. We think that providing researchers with the freedom of focusing on the hypotheses incorporated in the model, and not on the optimization routine, is essential, especially in applied contexts. To facilitate the use of our technique, we are working on a R-package to automatically fit a Hawkes process model, retrieve information on the parameters' posterior distribution, and produce forecasts. We are planning to start with a R-package focused on the ETAS model and extend it to include different Hawkes process models. Indeed, we have already provided these functions in a tutorial [2]. Specifically, we provided the user with one-line functions to fit the ETAS model used in the real data example on user-specified datasets, retrieve the posterior distributions of the parameters and the number of points, and produce forecasts for a user-specified number of periods and period's length. We have also made publicly available another tutorial[3] illustrating in detail how to build the functions used in the first tutorial. The second tutorial explains which functions have to be provided by the user, how to construct the binning strategy, and how to make them interact with `inlabru` and provides details on the possible difficulties that may be encountered in each step. This can be used as a template to implement Hawkes process models different from the ETAS model.

To conclude, we have shown that the `inlabru` approach is a valuable alternative to MCMC techniques for Hawkes process models, it provides comparable results in terms of quality but in a fraction of the time needed by MCMC. This is particularly relevant in applied contexts, such as seismology, where researchers are discouraged to use Hawkes process models on large datasets (> 100000 observations) by long computational times. On the same line, models used to produce daily forecasts are not updated daily, for the same reasons. The `inlabru` approach softens this burden and allows researchers to fit models on larger datasets in less time. Also, our approach can be extended to consider more complex models which would have needed an ad-hoc implementation if an MCMC technique had to be used. We believe that the `inlabru` approach could make Hawkes models more accessible for a greater number of users, which would have the freedom to make inference on models incorporating different hypotheses without the burden of adapting the methodology.

4.3 Chapter Summary

In this chapter, I have presented our novel approximation technique for Bayesian Hawkes processes and applied it to the temporal ETAS model. The presented technique is the basis on which the R-package `ETAS.inlabru` is built. Here, I have shown that our approach provides results close (if not better) than the one provided by a MCMC alternative (the `bayesianETAS` R-package) in terms of goodness of fit but that our methodology is more efficient in terms of time and in how it scales increasing the number of events per catalogue. This shows that our method can be competitive and has the potential of being used in place of MCMC techniques. In the next chapter, `ETAS.inlabru` package is applied to simulated data to study how characteristics of the data affect the posterior of the parameters and we give some advice on how to avoid potential biases deriving from the quality of the data used in input to estimate the parameters.

4.4 Supplementary material

4.4.1 Parameters posterior distribution

Here, we show the marginal posterior distribution of the ETAS parameters calibrated on the Amatrice sequence comprising 1137 events from 24/08/2016 to 15/08/2017 with latitude in $(42.456, 43.084)$, and longitude in $(12.936, 13.523)$. Below are reported the posterior distribution of the ETAS parameters for the implementations considered in the article. Figure 4.7 shows the posterior distributions obtained using the MCMC implementation provided by the

[2] The tutorial is available at https://github.com/Serra314/Hawkes_process_tutorials/tree/main/how_to_use_Hawkes

[3] The tutorial is available at https://github.com/Serra314/Hawkes_process_tutorials/tree/main/how_to_build_Hawkes

R-package bayesianETAS considering 10000 posterior samples and 5000 burn-in samples. Figure 4.8 shows the posterior distribution of the ETAS parameters for the inlabru replicate case, while Figure 4.9 compares the distribution of the inlabru replicate and gamma implementations. For the latter, we chose to use a logarithmic scale for the comparison to highlight the differences in the prior.

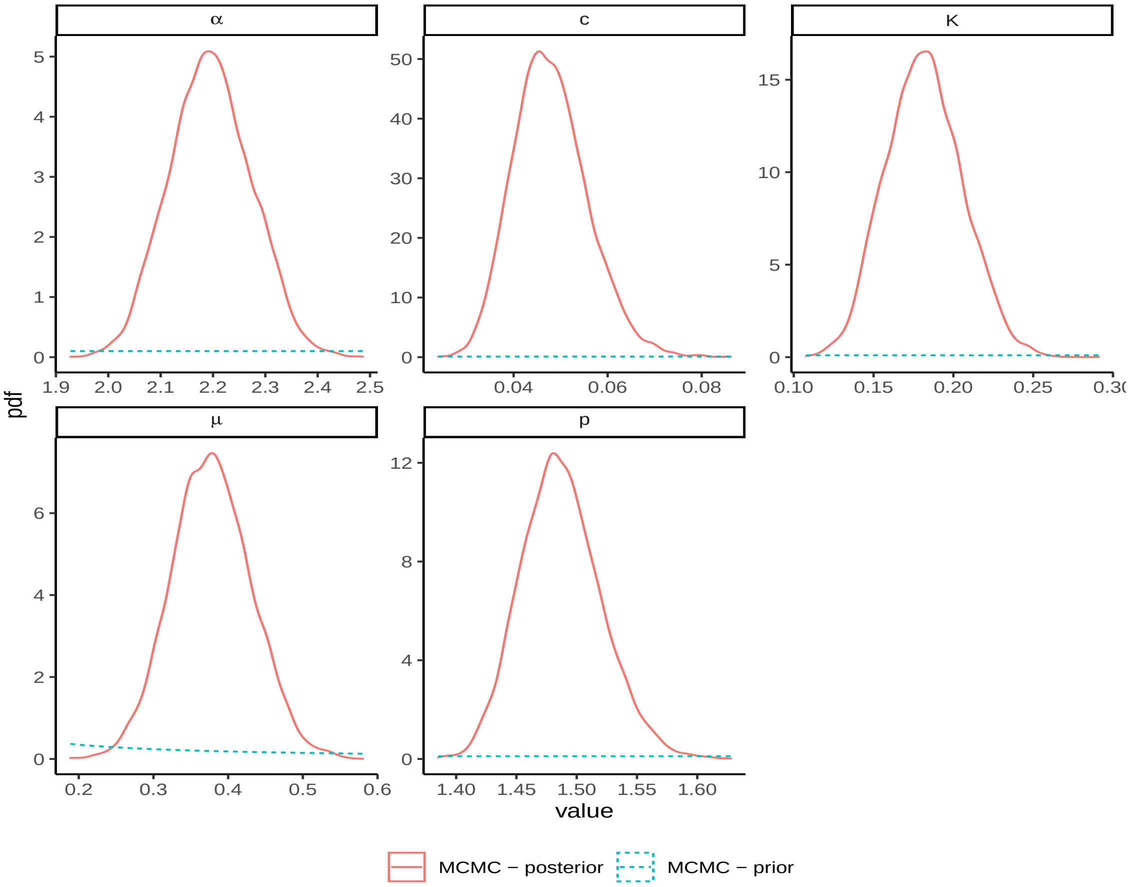


Figure 4.7: Posterior and prior distributions of ETAS parameter using the bayesianETAS package considering 1000 posterior samples and 5000 burn-in samples. The results are based on the Amatrice seismic sequence.

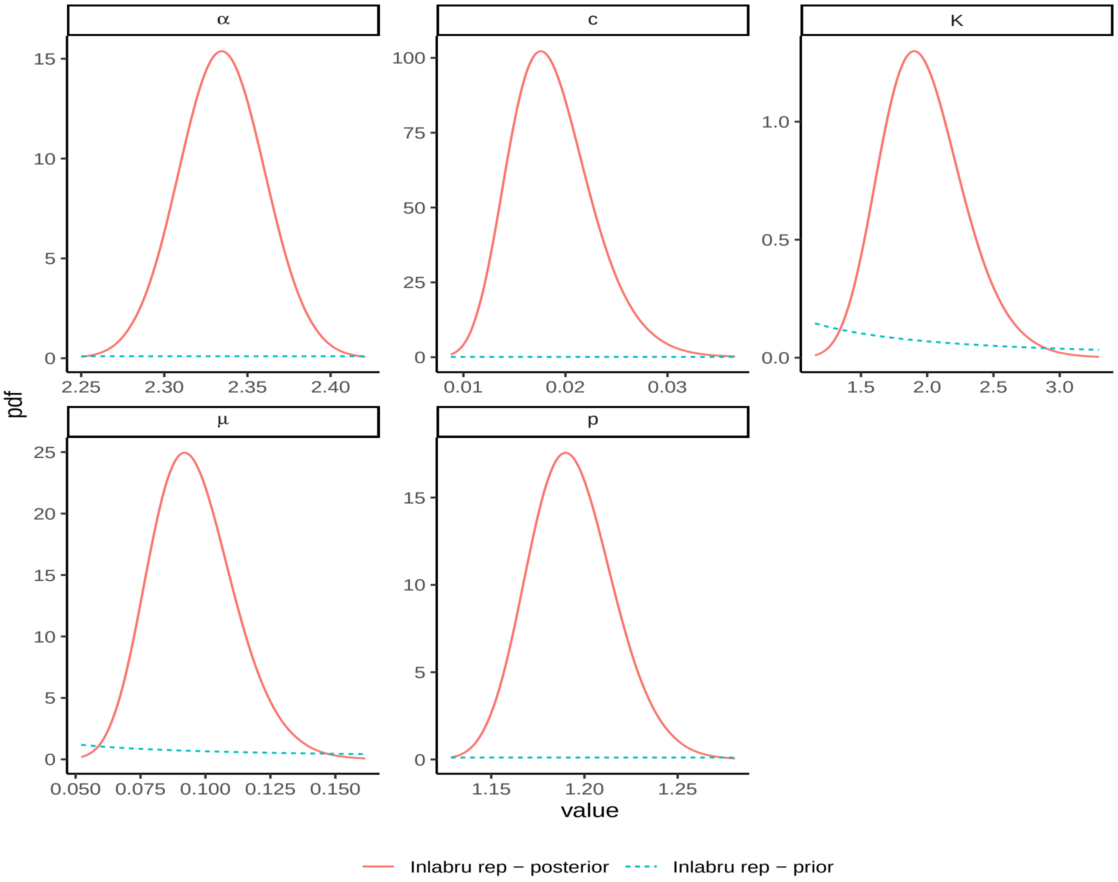


Figure 4.8: Posterior and prior distributions of ETAS parameter for `inlabru` replcate case. The results are based on the Amatrice seismic sequence.

4.4.2 Sensitivity to binning strategy

In our three factors decomposition of the point process log-likelihood, to approximate the second part (the expected number of triggered events Sec 4.2.5), we split the time domain into bins and we approximate the integral in each bin separately. In this paper, we use a different set of bins for each observed point. Specifically, for each arrival time t_h, the bins are defined by the sequence:

$$t_h, t_h + \Delta, t_h + \Delta(1+\delta), t_h + \Delta(1+\delta)^2, ..., t_h + \Delta(1+\delta)^{n_h}, T_2,$$

where n_h is such that $t_h + \Delta(1+\delta)^{n_h} < T_2$ or $n_h < n_{max}$. This binning strategy is defined by three parameters: Δ regulating the length of the first bin, δ regulating the increase in length of each subsequent bin, and n_{max} which regulates the maximum number of bins per observed points $(n_{max} + 2)$.

In this section, we take the `inlabru` replicate implementation and we try different parameters of the binning strategy. Specifically, we consider $\delta = 1, 2, 3, 4, 5, 7$, $\Delta = 0.1, 0.2, 0.5$ and $n_{max} = 3, 10$. The binning strategy affects mostly the ability to converge and the computational time required to reach convergence. Table 4.6 reports the number of iterations needed for convergence (n iter), the computational time (in minutes), and the convergence state for each combination of binning strategy parameters. We set a maximum number of iterations equal to 100 so that if the number of iterations for convergence is equal to 100

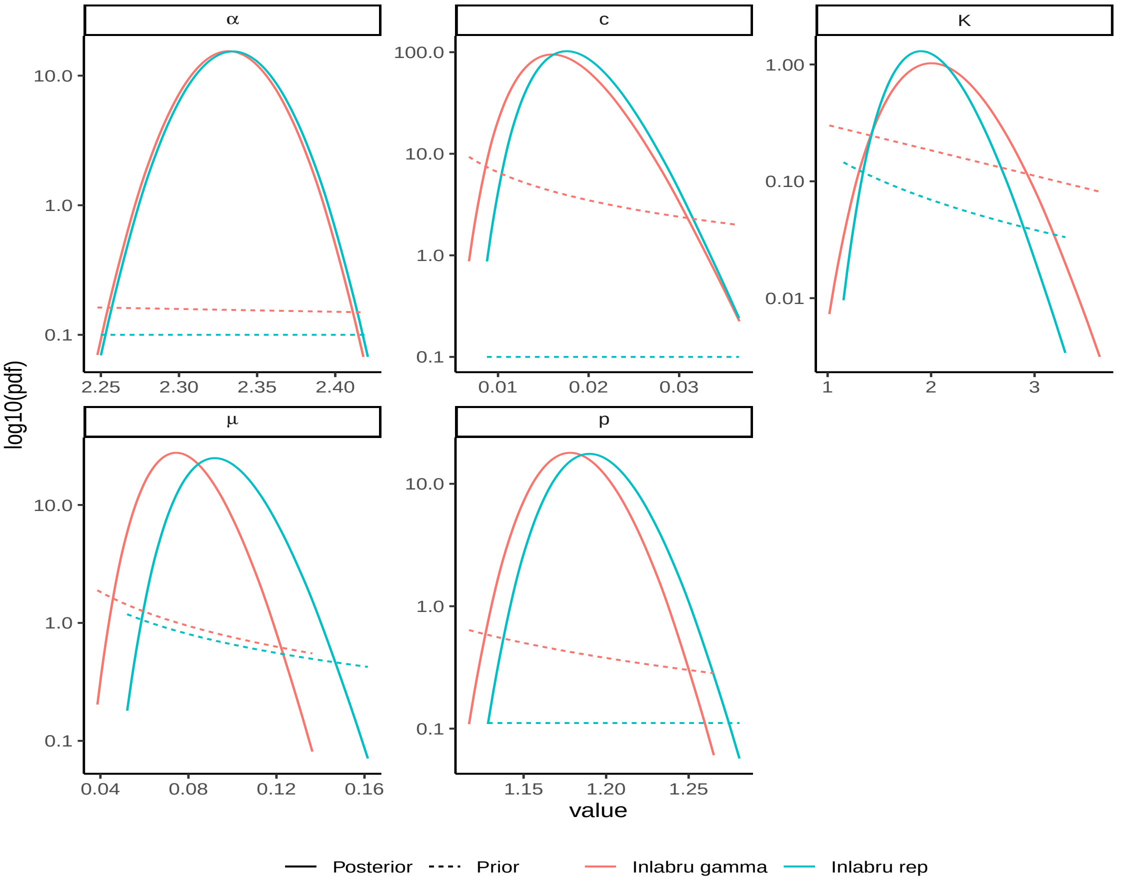


Figure 4.9: Posterior and prior distributions of ETAS parameter for the two `inlabru` implementations considered, namely gamma and replcate. The value of the density is on a logarithmic (base 10) scale to highlight the differences in the prior.

it means that the algorithm has not converged. We checked that the models are not able to converge looking at the posterior modes for each iteration of the algorithm, more detail on how to retrieve these quantities are reported in the tutorial on how to implement Hawkes process models with `inlabru`. The fact that different binning strategies converge in a similar number of iterations highlights the robustness of our approach. The time needed for each iteration changes with different binning strategies.

Examining Table 4.6, models with $\delta = 7, 10$ tend to not converge. This is due to the fact that these binning strategies induce too wide bins (especially close to the observations, where we need a finer partition) which in turn provide an approximation that is not accurate enough. Instead, strategies with $\delta = 2$ behave well and are the fastest to converge. In this paper, we use a binning strategy defined by $\delta = 2$, $\Delta = 0.1$ and $n_{max} = 3$ because it is the fastest to reach convergence.

The binning strategy only affects the distribution of the parameters K, c, and p: the only parameters of the time triggering function, and therefore, we compare the posterior distributions of these parameters only. We show the posteriors distributions for the case $\delta = 2$ which is the one with the lowest computational time. Figure 4.10 shows that there are small differences between the models. Only the implementation with $\Delta = 0.1$ and $n_{max} = 3$ has lighter tails, this is due to having too small/not enough bins.

δ	n_{max}	Δ	n iter	time (mins)	converged
2	3	0.2	63	2.93	TRUE
2	10	0.2	63	2.98	TRUE
2	10	0.1	63	2.99	TRUE
2	3	0.1	63	3.03	TRUE
2	10	0.5	63	3.03	TRUE
5	10	0.1	65	3.06	TRUE
2	3	0.5	63	3.06	TRUE
5	10	0.5	65	3.07	TRUE
3	10	0.1	65	3.08	TRUE
1	10	0.1	63	3.15	TRUE
1	10	0.2	63	3.17	TRUE
1	10	0.5	63	3.19	TRUE
1	3	0.5	63	3.23	TRUE
5	3	0.5	65	3.24	TRUE
1	3	0.2	63	3.26	TRUE
3	3	0.1	64	3.30	TRUE
3	10	0.2	65	3.36	TRUE
5	10	0.2	65	3.37	TRUE
3	3	0.2	65	3.40	TRUE
3	10	0.5	65	3.40	TRUE
1	3	0.1	63	3.41	TRUE
3	3	0.5	64	3.47	TRUE
5	3	0.2	71	3.70	TRUE
10	3	0.2	100	5.41	FALSE
10	10	0.2	100	5.47	FALSE
10	3	0.5	100	5.60	FALSE
5	3	0.1	100	5.72	FALSE
7	10	0.2	100	5.87	FALSE
10	10	0.5	100	5.88	FALSE
10	10	0.1	100	5.94	FALSE
7	3	0.2	100	6.00	FALSE
7	10	0.1	100	6.01	FALSE
10	3	0.1	100	6.20	FALSE
7	3	0.1	100	6.25	FALSE
7	10	0.5	100	6.26	FALSE
7	3	0.5	100	6.35	FALSE

Table 4.6: Number of iterations needed by `inlabru` to converge (n iter) considering 100 maximum possible iterations, computational time needed to reach convergence in minutes, and a true/false column reporting if the model converged or not, for different values of parameters of the binning strategy δ, Δ, and n_{max}.

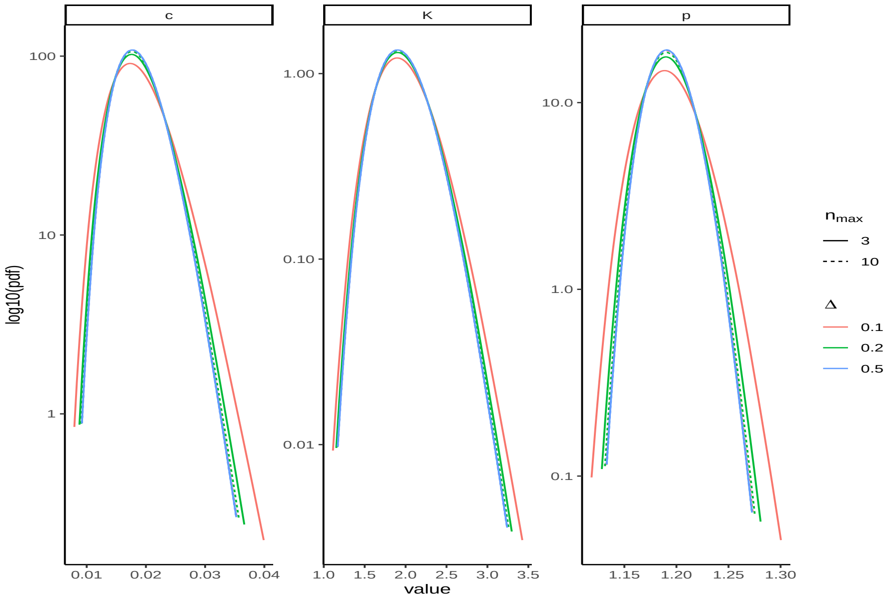


Figure 4.10: Posterior distribution of ETAS parameters for the `inlabru` replicate implementation for different binning strategies. The binning strategies have the same parameter $\delta = 2$, while the others are varying $\Delta = 0.1, 0.2, 0.5$ (color), and $n_{max} = 3, 10$ (line type).

4.4.3 Sensitivity to prior choice

In this section, we explore the sensitivity of our methodology to change of priors mean and standard deviation. For this task, we chose to use the same prior for all the parameters. We use a Log Gaussian prior with logarithm mean equal to 0 and varying the logarithm standard deviation $\sigma_{log} = 1, 1.5, 2, 2.5$. Table 4.7 reports summary statistics of the Log Gaussian distribution for the values of σ_{log} considered in this analysis.

σ_{log}	mean	sd	q0.025	q0.5	q0.975
1.0	1.625	2.197	0.141	1	7.099
1.5	3.137	8.642	0.053	1	18.915
2.0	6.907	43.587	0.019	1	50.397
2.5	28.476	144.870	0.007	1	134.278

Table 4.7: Table reporting summary statistics of Log Gaussian distribution with logarithm mean equal to 0 and logarithm standard deviation $\sigma_{log} = 1, 1.5, 2, 2.5$.

Figure 4.11 shows that the posterior distributions are robust under the considered changes in prior. Specifically, they appear to converge for increasing values of the prior variance which is what we expect to happen.

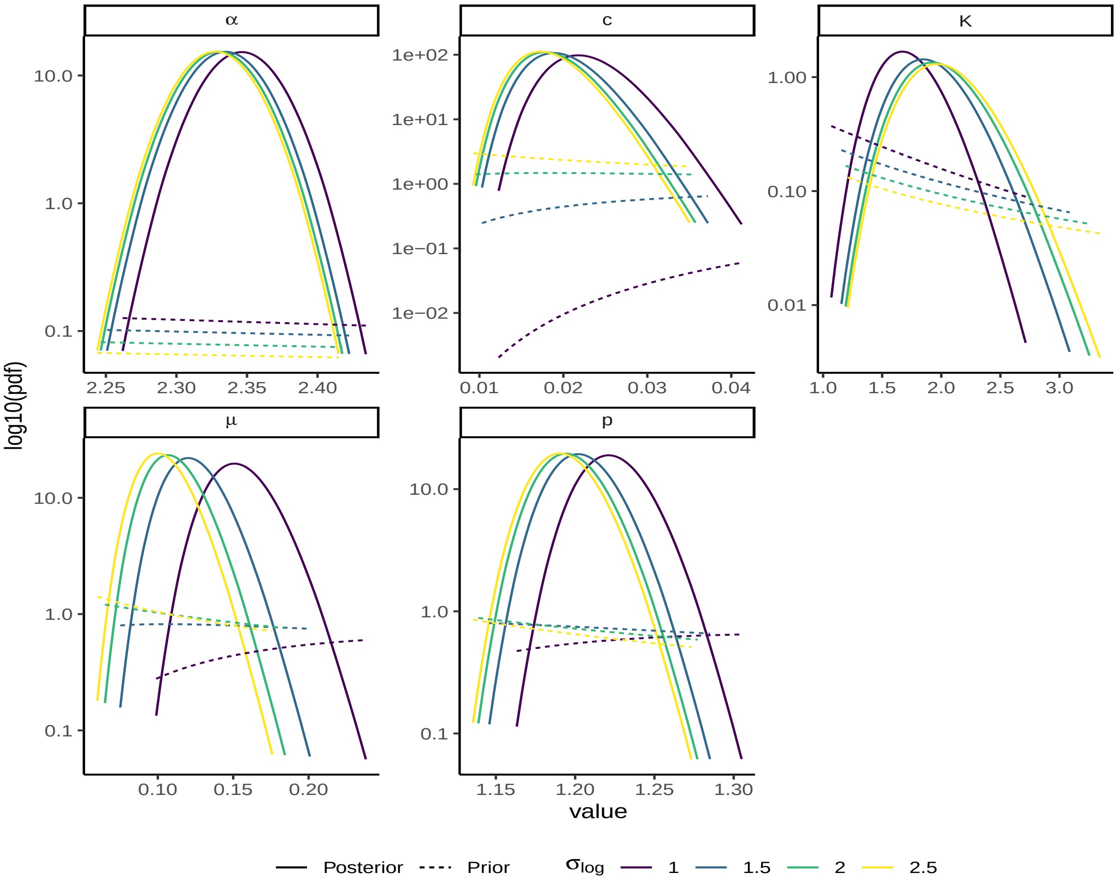


Figure 4.11: Posterior distribution of ETAS parameters changing the prior mean and standard deviation regulated by the parameter σ_{log}, the larger the parameter the higher the prior mean and standard deviation. Specifically, we considered $\mu, K, \alpha, c, p-1 \sim \text{LogN}(0, \sigma_{log})$.

Chapter 5

Bayesian modelling of the temporal evolution of seismicity using the `ETAS.inlabru` R-package

5.1 Introduction

This chapter includes a paper submitted to the Frontiers of Earth Science journal freely available in preprint (Naylor et al., 2022). The authors of the paper are ordered Mark Naylor, Francesco Serafini (me), Finn Lindgren, and Ian Main. I have contributed to the article by providing the code used to produce the results, and writing the most theoretical parts of it.

In this chapter, I explore the capabilities and robustness of the methodology presented in Chapter 4 using synthetic catalogues. All the results are obtained using the `ETAS.inlabru` R-package (Naylor and Serafini, 2023) which provides user-friendly implementation of the proposed technique as well as Rmd notebooks to reproduce the figures. Here, I study the parameter estimates provided by `ETAS.inlabru` on synthetic catalogues (for which we know the value of the parameters that have generated the data) representing scenarios in which ETAS inversion algorithms are known to struggle. I explore the performance of the inversion as a function of the training catalogue length, the impact of large events that happen to occur in the sequence, the consequence of short term incompleteness after large events as well as various `inlabru` model choices.

The results are not new, meaning that is well-known that ETAS inversion algorithms struggle in certain situations. However, they are valuable to us because i) the fact that our algorithm provides similar results to the ones obtained by previous studies with different algorithms shows the reliability of our approach, ii) quantifying the bias in parameters estimates coming from data quality issues is helpful in defining *good practices* that should be followed to avoid those biases. What is new is that the speed of our approach enables a more exploratory approach to considering such issues. I believe that this can lead to a more rigorous estimation of uncertainty, and will enable an improvement in the best practice application of ETAS for seismicity modelling and operational earthquake forecasting, for example by defining what is needed in a representative training data set for the paremeter estimation to be reliable.

5.2 The paper

5.2.1 Abstract

The Epidemic Type Aftershock Sequence (ETAS) model is widely used to model seismic sequences and underpins Operational Earthquake Forecasting (OEF). However, it remains challenging to assess the reliability of inverted ETAS parameters for a range of reasons. For example, the most common algorithms just return point estimates with little quantification of uncertainty. At the same time, Bayesian Markov Chain Monte Carlo implementations remain slow to run, do not scale well and few have been extended to include spatial structure. This makes it difficult to explore the effects of stochastic uncertainty. Here we present a new approach to ETAS modelling using an alternative Bayesian method, the Integrated Nested Laplace Approximation (INLA). We have implemented this model in a new R-Package called `ETAS.inlabru`*, which builds on the R packages R-INLA and* `inlabru`*. Our work has included extending these packages, which provided tools for modelling log-Gaussian Cox processes, to include the self-exciting Hawkes process that ETAS is a special case of. Whilst we just present the temporal component here, the model scales to a spatio-temporal model and may include a variety of spatial covariates. This is a fast method which returns joint posteriors on the ETAS background and triggering parameters. Using a series of synthetic case studies, we explore the robustness of ETAS inversions using this method of inversion using some of the classic scenarios that ETAS can struggle with. We also included runnable notebooks to reproduce the figures in this paper as part of the package's GitHub repository. We demonstrate that reliable estimates of the model parameters require that the catalogue data contains periods of relative quiescence as well as triggered sequences. We explore the robustness under stochastic uncertainty in the training data and show that the method is robust to a wide range of starting conditions. We show how the inclusion of historic earthquakes prior to the modelled domain affects the quality of the inversion. Finally, we show that rate dependent incompleteness after large earthquakes has a significant and detrimental effect on the ETAS posteriors. We believe that the speed of the* `inlabru`*inversion, which include a rigorous estimation of uncertainty, will enable a deeper exploration of how to use ETAS robustly for seismicity modelling and operational earthquake forecasting.*

5.2.2 Introduction

The Epidemic Type Aftershock Sequence model (ETAS) ((Ogata, 1988; Ogata and Zhuang, 2006; Ogata, 2011)) is one of the cornerstones of seismicity modelling. It models evolving seismic sequences in terms of background seismicity and seismicity triggered by previous events. As such, it is a self-exciting point process model which is commonly termed a Hawkes process (Hawkes, 1971b) in the statistical literature. ETAS achieves this by combining several empirical relationships for seismicity. The ETAS model enables us to generate synthetic earthquake sequences and to invert earthquake space-time-magnitude data for the underlying ETAS parameters that characterise both the background and triggering rates. However, the likelihood space for some parameters is notoriously flat and many factors can affect the robustness of the results.

There are many different implementations of the ETAS model. The most common approach for determining ETAS parameters is the maximum likelihood method which returns a point estimate on the ETAS parameters using an optimisation algorithm (e.g Jalilian, 2019).

In some cases uncertainty is quoted using the Hessian matrix. Bayesian alternatives are available; for example the "bayesianETASR̈-package (Ross, 2021) uses the Markov Chain Monte Carlo (MCMC) method to return full posteriors. However MCMC methods are notoriously slow as building the Markov Chain is an inherently linear algorithm requiring many successive samples of the full posterior distribution. A major benefit of Bayesian methods is that they better describe uncertainty. We have developed a new Bayesian ETAS package using the Integrated Nested Laplace Approximation (INLA) instead of MCMC; this is implemented in the R-Package ETAS.inlabru and will be made available through GitHub and the Comprehensive R Archive Network (CRAN) (Naylor and Serafini, 2023). The results presented in this paper are reproducible using this package and a series of Rmd notebooks. Unlike the MCMC implementation of the ETAS model, our method does not internally rely on a latent variable to classify whether events are background or triggered.

The Integrated Nested Laplace Approximation (INLA, Rue et al., 2017) and inlabru (Bachl et al., 2019) offer a fast approach for Bayesian modelling of spatial, temporal and spatio-temporal point process data and have had over 10 years of development. The INLA method is a well-known alternative to MCMC methods to perform Bayesian inference. It has been successfully applied in a variety of fields such as seismic hazard (Bayliss et al., 2020, 2022), air pollution (Forlani et al., 2020), disease mapping (Riebler et al., 2016; Santermans et al., 2016; Schrödle and Held, 2011a,b), genetics (Opitz et al., 2016), public health (Halonen et al., 2015), ecology (Roos et al., 2015; Teng et al., 2022), more examples can be found in Bakka et al. (2018); Blangiardo et al. (2013); Gómez-Rubio (2020).

To date, the main limitation for the application of inlabru to seismicity was that it only addressed log-Gaussian Cox Processes (Taylor and Diggle, 2014), which do not include self-exciting clustering. Serafini et al. (2022a) addressed this specific limitation by showing how the methodology used for log-Gaussian Cox processes could be extended to model self-exciting Hawkes Processes (Hawkes, 1971a,b), using R-INLA and inlabru, when the function form of the triggering function can be integrated. The novelty of our approach resides in the likelihood approximation. We decompose the log-likelihood into the sum of many small components, where each is linearly approximated with respect to the posterior mode using a Taylor expansion. This means that the log-likelihood is exact at the posterior mode and the accuracy of the approximation decreases as we move away from that point. Furthermore, the linear approximation and the optimization routine to determine the posterior mode are internally performed by the inlabru package. In this work, the specific application to the ETAS model was presented. The temporal model provides posteriors on the background rate and all ETAS parameters.

ETAS.inlabru (Naylor and Serafini, 2023) provides the functions to be approximated whilst the user provides the data and specifies the priors. The advantages of our approach are both in terms of computational time and its scalability to include relevant covariates (Bayliss et al., 2020) such as maps of faults, strain rates etc. in addition to earthquake catalogue data, and/or to introduce alternative structures to the parameters (e.g. considering one of them as temporally, or spatially, varying).

Here, we present a broad analysis of how the inlabru inversion performs on synthetic earthquake catalogues where we know all of the controlling parameters. We explore the performance of the inversion as a function of the training catalogue length, the impact of large events that happen to occur in the sequence, the consequence of short term incompleteness after large events as well as various inlabru model choices. These results build on a wealth of literature that explores the challenges in fitting the ETAS model including (Ogata and Zhuang, 2006; Ogata, 2006; Touati et al., 2014; Hainzl, 2016b,c; Touati et al., 2011). Our results are generic and not specific to our implementation of the ETAS model - rather our fast Bayesian model allows us to make a more rapid assessment of potential biases derived from the likelihood function itself. We want the reader to come away with an understanding of when the ETAS model is likely to describe a sequence well, and to be able to identify sources of potential bias, understand how synthetic modelling allows us to explore potential data quality issues, and decide whether fitting an ETAS model is an appropriate way to proceed. We conclude with a demonstration of how the results can be used to develop a temporal Operational Earthquake Forecast.

5.2.3 Method

In this section, we introduce our `inlabru` implementation of the temporal ETAS model and refer the reader to (Serafini et al., 2022a) for a complete description of the mathematical formulation.

The temporal ETAS model

The temporal ETAS model is a marked Hawkes process model, where the marking variable is the magnitude of the events. The ETAS model is composed of three parts: a background rate term, a triggered events rate representing the rate of events induced by past seismicity, and a magnitude distribution independent from space and time. Given the independence between the magnitude distribution and the time distribution of the events, the ETAS conditional intensity is usually the product between a Hawkes process model describing only the location and a magnitude distribution $\pi(m)$.

More formally, the ETAS conditional intensity function evaluated at a generic time point $t \in (T_1, T_2), T_1, T_2 \geq 0, T_1 < T_2$ having observed the events $\mathcal{H}_t = \{(t_h, m_h) : t_h < t, m_h > M_0, \forall h = 1, ..., N(t^-)\}$, where M_0 is the minimum magnitude in the catalogue which needs to be completely sampled, and $N(t)$ is the counting process associated with the Hawkes process representing the number of events recorded up to time t (included), is given by:

$$\lambda_{ETAS}(t, m|\mathcal{H}_t) = \lambda_{Hawkes}(t|\mathcal{H}_t)\pi(m) \tag{5.1}$$

where λ_{Hawkes} is the conditional intensity of a temporal Hawkes process describing the occurrence times only. In our ETAS implementation this is given by:

$$\lambda_{Hawkes}(t|\mathcal{H}_t) = \mu + \sum_{(t_h,m_h)\in\mathcal{H}_t} Ke^{\alpha(m_h-M_0)}\left(\frac{t-t_h}{c}+1\right)^{-p} \tag{5.2}$$

The parameters of the model are $\mu, K, \alpha, c \geq 0$ and $p > 1$. Different parametrisations of the ETAS model exist; we focus on this one because it has proven to be the most suitable parametrisation for our method.

In seismology, the magnitude distribution, $\pi(m)$ is commonly assumed to be independent of space and time for simplicity of analysis. In this work, we take this to be the Gutenberg-Richter distribution with a b-value of 1. In this section, we focus on the Hawkes part of the model assuming the parameters of the magnitude distribution are determined independently. From now on, for ease of notation, where not specified differently we refer to λ_{Hawkes} as simply λ.

Hawkes Process Log-likelihood approximation for `inlabru`

The Hawkes process is implemented in `inlabru` by decomposing its log-likelihood function (Eqn.5.3) into multiple parts, the sum of which returns the exact log-likelihood at the point we expand it about. We linearly approximate the single components with respect to the posterior mode and apply the *Integrated Nested Laplace Approximation* (INLA) method to perform inference on the parameters of the model. Both the linearisation and the optimization, to find the posterior mode, are performed internally by `inlabru`. Our package, `ETAS.inlabru` (Naylor and Serafini, 2023), provides `inlabru` with the ETAS-specific functions representing the log-likelihood components to be approximated. We outline the decomposition below.

Having observed a catalogue of events $\mathcal{H} = \{(t_i, m_i) : t_i \in [T_1, T_2], m_i \in (M_0, \infty)\}$, the Hawkes process log-likelihood is given by:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{H}) = -\Lambda(T_1, T_2) + \sum_{(t_i,m_i)\in\mathcal{H}} \log \lambda(t_i|\mathcal{H}_{t_i}) \tag{5.3}$$

Where $\boldsymbol{\theta}$ is a vector of the model parameters, $\mathcal{H}_{t_i} = \{(t_h, m_h) \in \mathcal{H} : t_h < t_i\}$ is the history of events up to time t_i, and

$$\begin{aligned}
\Lambda(T_1, T_2) &= \int_{T_1}^{T_2} \lambda(t|\mathcal{H}_t)dt \\
&= (T_2 - T_1)\mu + \sum_{(t_i,m_i)\in\mathcal{H}} \int_{T_1}^{T_2} Ke^{\alpha(m_i - M_0)} \left(\frac{t - t_i}{c} + 1\right)^{-p} \mathbb{I}(t > t_i)dt \\
&= (T_2 - T_1)\mu + \sum_{(t_i,m_i)\in\mathcal{H}} Ke^{\alpha(m_i - M_0)} \int_{\max(T_1,t_i)}^{T_2} \left(\frac{t - t_i}{c} + 1\right)^{-p} dt \\
&= (T_2 - T_1)\mu + \sum_{(t_i,m_i)\in\mathcal{H}} Ke^{\alpha(m_i - M_0)} \frac{c}{p-1} \left(\left(\frac{\max(t_i, T_1) - t_i}{c} + 1\right)^{1-p} - \left(\frac{T_2 - t_i}{c} + 1\right)^{1-p}\right) \\
&= \Lambda_0(T_1, T_2) + \sum_{(t_i,m_i)\in\mathcal{H}} \Lambda_i(T_1, T_2)
\end{aligned} \tag{5.4}$$

The above integral can be considered as the sum of two parts, the number of background events $\Lambda_0(T_1, T_2)$ and the remaining summation which is referred as the sum of the number of triggered events by each event t_i, namely $\Lambda_i(T_1, T_2)$. We approximate the integral by linearising the functions $\Lambda_0(T_1, T_2)$ and $\Lambda_i(T_1, T_2)$. Note that it is not enough to be able to evaluate the exact integral; we need the *linearised* log-contributions to have the full degrees of freedom with respect to the model parameters for the iterative update of the modal parameters by `inlabru` to be stable, and this is why the integrals need to be split. Also, having more bins and linearising them separately provides a more accurate approximation than approximation over the whole domain. The resulting approximate integral is the sum of $|\mathcal{H}| + 1$ linear functions of the parameters.

However, we observed that this approximation alone is not sufficiently accurate for the algorithm to converge. To increase the accuracy of the approximation, for each integral $\Lambda_i(T_1, T_2)$, we further consider a partition of the integration interval $[\max(T_1, t_i), T_2]$ in B_i bins, $t_0^{(b_i)}, ..., t_{B_i}^{(b_i)}$ such that $t_0^{(b_i)} = \max(T_1, t_i)$, $t_{B_i}^{(b_i)} = T_2$ and $t_j^{(b_i)} < t_k^{(b_i)}$ if $j < k$. By doing this, the integral becomes,

$$\Lambda(T_1, T_2) = \Lambda_0(T_1, T_2) + \sum_{(t_i,m_i)\in\mathcal{H}} \sum_{j=0}^{B_i - 1} \Lambda_i(t_j^{(b_i)}, t_{j+1}^{(b_i)}) \tag{5.5}$$

In this way, the integral is decomposed in $\sum_i B_i + 1 > |\mathcal{H}| + 1$ terms providing a more accurate approximation. We discuss the options for temporal binning in Section 5.2.3.

Substituting Eqn.5.5 into 5.3, the Hawkes process log-likelihood can be written,

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{H}) = -\Lambda_0(T_1, T_2) - \sum_{(t_i,m_i)\in\mathcal{H}} \sum_{j=0}^{B_i-1} \Lambda_i(t_j^{(b_i)}, t_{j+1}^{(b_i)}) + \sum_{(t_i,m_i)\in\mathcal{H}} \log\lambda(t_i|\mathcal{H}_{t_i}). \tag{5.6}$$

In our approximation we linearise the logarithm of each elements within summations with respect to the posterior mode $\boldsymbol{\theta}^*$. Other choices led to a non-convergent model (Serafini et al., 2022a). In this case, the approximate log-likelihood becomes,

$$\overline{\mathcal{L}}(\boldsymbol{\theta}|\mathcal{H}) = -\exp\{\overline{\log\Lambda}_0(T_1, T_2)\} - \sum_{(t_i,m_i)\in\mathcal{H}} \sum_{j=0}^{B_i-1} \exp\{\overline{\log\Lambda}_i(t_j^{(b_i)}, t_{j+1}^{(b_i)})\} + \sum_{(t_i,m_i)\in\mathcal{H}} \overline{\log\lambda}(t_i|\mathcal{H}_{t_i}) \tag{5.7}$$

where for a generic function $f(\boldsymbol{\theta})$ with argument $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$, the linearised version with respect to a point $\boldsymbol{\theta}^*$ is given by a truncated Taylor expansion,

$$\overline{f}(\boldsymbol{\theta}) = f(\boldsymbol{\theta}^*) + \sum_{k=1}^{m} (\theta_k - \theta_k^*) \frac{\partial}{\partial\theta_k} f(\boldsymbol{\theta}) \bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \tag{5.8}$$

The other key component is the functions for each of the three incremental components in Eqn.5.6 that need to be linearised in this way. These are provided in `ETAS.inlabru` and will be discussed in Section 5.2.3.

Temporal binning

For each event, time binning is used to in Eqn.5.5 to improve the accuracy of the integration of the term describing the sum of the number of triggered events. The binning strategy is fundamental because the number of bins determines, up to a certain limit, the accuracy of this component of the approximation. Considering more bins enhances the accuracy of the approximation but increases the computational time because it increases the number of quantities to be approximated. Also, we cannot reduce the approximation error to zero, and the numerical value of the integral in each bin goes to zero increasing the number of bins which can be problematic in terms of numerical stability. We found that for the ETAS model considered here, having around 10 bins for each observed point is usually enough, and that is best considering higher resolution bins close to the triggering event. In fact, the function

$$g_t(t - t_i, m_i) = Ke^{\alpha(m_i - M_0)} \left(\frac{t - t_i}{c} + 1 \right)^{-p} \mathbb{I}(t > t_i) \tag{5.9}$$

varies the most for value of t close to t_i and become almost constant moving away from t_i. This means that we need shorter bins close to t_i, to capture the variation, and wider bins far from t_i where the rate changes more slowly.

We choose a binning strategy defined by three parameters $\Delta, \delta > 0$, and $n_{max} \in \mathbb{N}^+$. The bins relative to the observed point t_i are given by

$$t_i,\ t_i + \Delta,\ t_i + \Delta(1+\delta),\ t_i + \Delta(1+\delta)^2, ..., t_i + \Delta(1+\delta)^{n_i}, T_2, \tag{5.10}$$

where, $n_i \leq n_{max}$ is the maximum $n \in \{0, 1, 2, 3, ...\}$ such that $t_i + \Delta(1+\delta)^n < T_2$. The parameter Δ regulates the length of the first bin, δ regulates the length ratio between consecutive bins, and the value n_{max} regulates the maximum number of bins.

This strategy presents two advantages. The first is that we have shorter bins close to the point t_i and wider bins as we move away from that point. The second is that the first (or second, or third, or any) bin has the same length for all points. This is useful because the integral in a bin is a function of the bin length and not of the absolute position of the bin. This means that we need to calculate the value of the integral in the first (second, third, or any) bin once time and reuse the same result for all events. This significantly reduces the computational burden.

Functions to be linearized

This section and the next one illustrate what we need to provide to `inlabru` to approximate Hawkes process models. This one focuses on the functions to be provided while the next one on how they are combined to obtain the desired approximation. Regarding the functions to be provided, we remark that those are already present in the `ETAS.inlabru` package, so the user does not have to provide anything apart from the data, the area of interest, and the prior parameters. However, these sections are useful to understand what happens under the curtains and if one wants to extend this approach to more complicated ETAS implementations.

To build an ETAS model, we need to provide functions for each of the components of the likelihood function (Eqn.5.6). The linearisation and the finding of the mode $\boldsymbol{\theta}^*$ are managed automatically by the `inlabru` package. We only have to provide the functions to be linearised. Specifically, we need to provide the logarithm of the functions needed to approximate the integral and the logarithm of the conditional intensity. More formally, for our approximation of the ETAS model (i.e. for each term in Eqn 5.6), `ETAS.inlabru` provides the functions,

$$\log \Lambda_0(T_1, T_2) = \log(T_2 - T_1) + \log(\mu), \tag{5.11}$$

$$\begin{aligned}\log \Lambda_i(t_j^{(b_i)}, t_{j+1}^{(b_i)}) = \log(K) + \alpha(m_i - M_0) + \log\left(\frac{c}{p-1}\right) \\ + \log\left(\left(\frac{t_j^{(b_i)} - t_i}{c} + 1\right)^{1-p} - \left(\frac{t_{j+1}^{(b_i)} - t_i}{c} + 1\right)^{1-p}\right),\end{aligned}$$

and

$$\log \lambda(t|\mathcal{H}_t) = \log\left(\mu + \sum_{(t_h, m_h) \in \mathcal{H}_t} K e^{\alpha(m_h - M_0)} \left(\frac{t - t_h}{c} + 1\right)^{-p}\right). \tag{5.12}$$

For full details see Serafini et al. (2022a).

Implementation Details: The Poisson Count model trick

Our implementation in `inlabru` works by combining three INLA Poisson models on different datasets. The use of the INLA Poisson model here is related to computational efficiency purposes; it does not have any specific statistical meaning. Specifically, we leverage the internal log-likelihood used for Poisson models in R-INLA (and `inlabru`) to obtain the approximate Hawkes process log-likelihood as part of a computational trick.

More formally, INLA has the special feature of allowing the user to work with Poisson counts models with exposures equal to zero (which should be improper). A generic Poisson model for counts $c_i, i = 1, ..., n$ observed at locations $\mathbf{t}_i, i = 1, ..., n$ with exposure $E_1, ..., E_n$

Name	Objective	Approximation	Surrogate $\log \lambda_P$	Number of data points	Counts and Exposures
Part I	$\Lambda_0(\mathcal{X})$	$\exp \overline{\log \Lambda}_0(\mathcal{X})$	$\log \Lambda_0(\mathcal{X})$	1	$c_i = 0,\ e_i = 1$
Part II	$\sum_{h=1}^{n} \sum_{i=1}^{B_h} \Lambda_h(b_{i,h})$	$\sum_{h=1}^{n} \sum_{i=1}^{B_h} \exp \overline{\log \Lambda}_h(b_{i,h})$	$\log \Lambda_h(b_{i,h})$	$\sum_h B_h$	$c_i = 0,\ e_i = 1$
Part III	$\sum_{h=1}^{n} \log \lambda(\mathbf{x}_h)$	$\sum_{h=1}^{n} \exp \overline{\log \lambda}(\mathbf{x}_h)$	$\log \lambda(\mathbf{x})$	n	$c_i = 1,\ e_i = 0$

Table 5.1: Hawkes process log-likelihood components approximation

with log-intensity $\log \lambda_P(\mathbf{t}) = f(\mathbf{t}, \boldsymbol{\theta})$, in `inlabru` has log-likelihood given by:

$$\mathcal{L}_P(\boldsymbol{\theta}) \propto -\sum_{i=1}^{n} \exp\{\overline{f}(\mathbf{t}_i, \boldsymbol{\theta}, \boldsymbol{\theta}^*)\} * E_i + \sum_{i=1}^{n} \overline{f}(\mathbf{t}_i, \boldsymbol{\theta}, \boldsymbol{\theta}^*) * c_i. \tag{5.13}$$

Each Hawkes process log-likelihood component (Eqn. 5.6) is approximated using one surrogate Poisson model with log-likelihood given by Eqn. 5.13 and an appropriate choice of counts and exposures data. Table 5.1 reports the approximation for each log-likelihood component with details on the surrogate Poisson model used to represent it. For example, the first part (integrated background rate) is represented by a Poisson model with log-intensity $\log \Lambda_0(\mathcal{X})$, this will be automatically linearised by `inlabru`. Given that, the integrated background rate is just a scalar and not a summation, and therefore we only need one observation to represent it assuming counts equal 0 and exposures equal 1. Table 5.1 shows that to represent a Hawkes process model having observed n events, we need $1 + \sum_h (B_h) + n$ events with B_h number of bins in the approximation of the expected number of induced events by observation h.

Furthermore, Table 5.1 lists the components needed to approximate the ETAS log-likelihood which will be internally considered as surrogate Poisson log-intensities by `inlabru`. More specifically, we only need to create the datasets with counts c_i, exposures e_i, and the information on the events $\mathbf{x}_i$ representing the different log-likelihood components; and, to provide the functions $\log \Lambda_0(\mathcal{X})$, $\log \Lambda_h(b_{i,h})$, and, $\log \lambda(\mathbf{t})$. The linearisation is automatically performed by `inlabru` as well as the retrieving of the parameters' posterior distribution.

More detail on how to build the functions in the `ETAS.inlabru` package can be found at https://github.com/Serra314/Hawkes_process_tutorials/tree/main/how_to_build_Hawkes.

Prior specification

We have to set the priors for the parameters. The INLA method is designed for Latent Gaussian models, which means that all the unobservable parameters have to be Gaussian. This seems in contrast with the positivity constraint of the ETAS parameters μ, K, α, c, p, but we have a solution.

Our idea is to use an internal scale where the parameters have a Gaussian distribution and to transform them before using them in the log-likelihood components calculations. We refer to the internal scale as INLA scale, and to the parameters in the INLA scale as $\boldsymbol{\theta}$. In practice, all parameters have a standard Gaussian prior in the INLA scale and they are transformed to be distributed according to a target distribution in the ETAS scale. Specifically, assuming that θ has a standard Gaussian distribution with cumulative distribution function (CDF) $\Phi(\theta)$, and calling F_Y^{-1} the inverse of the CDF of the target distribution for the parameter, we can switch between the Gaussian and the target distributions using,

$$\eta(\theta) = F_Y^{-1}(\Phi(\theta)), \tag{5.14}$$

where $\eta(\theta)$ has a distribution with CDF $F_Y(\cdot)$.

The `ETAS.inlabru` R-package uses the following default priors in the ETAS scale,

$$\begin{aligned} \mu_b &\sim \text{Gamma}(\text{shape} = 0.5, \text{rate} = 0.5) \\ K_b &\sim \text{LogNormal}(\text{mean}(\log(K)) = -1, \text{sd}(\log(K)) = 0.5) \\ \alpha_b &\sim \text{Unif}(\alpha_{min} = 0, \alpha_{max} = 10) \\ c_b &\sim \text{Unif}(c_{min} = 0, c_{max} = 1) \\ p_b &\sim \text{Unif}(p_{min} = 1, p_{max} = 2), \end{aligned} \tag{5.15}$$

however, they can be changed to different distributions that better describe the available prior information.

The package `inlabru` provides a function to easily implement such transformation. The function is called `bru_forward_transformation` and takes in input the quantile function of the target distribution and its parameters. Below we report three examples of transformations such that the parameters in the ETAS scale have a Gamma, Uniform, or Log-Gaussian distribution. We show the empirical density obtained by transforming the same sample of values from a standard Gaussian distribution.

The prior for μ is the one that will most commonly need to be modified as it changes with the size of the domain being modelled. We choose Gamma(shape $= a_\mu$, rate $= b_\mu$)) prior for μ. The mean of the distribution is given by $a_\mu / b_\mu = 1$ event/day, the variance is $a_\mu / b_\mu^2 = 2$ and the skewness $2/\sqrt{\alpha} = 2.5$. One strategy for setting these parameters is to estimate an upper limit on the rate by dividing the duration of the catalogue total number of events; this is likely an overestimate as it combines the triggered and background events. One might choose to pick a mean rate that is half of this which defines the ratio of a_μ and b_μ. There is then some trade-off in the variance and skewness parameters.

Samples drawn from the priors used in this paper are shown in Figure 5.1, including lines showing the initial and true values that will be used through the majority of the results. The sensitivity to the choice of initial values is the first part of the results section. Please note how broad these priors are as this is helpful to see how much more informative the posteriors we generate are from these initial distributions.

The ETAS parameters themselves are not easy to interpret given that it is their combination within the right hand term of Eqn.5.2 that is important.

The Omori decay anaylses the magnitude independent decay,

$$\lambda_{Omori} = K \left(\frac{t - t_h}{c} + 1 \right)^{-p}. \tag{5.16}$$

Whilst the full triggering function also includes a magnitude dependent productivity term,

$$\lambda_{triggering} = K e^{\alpha(m_h - M_0)} \left(\frac{t - t_h}{c} + 1 \right)^{-p}. \tag{5.17}$$

We draw 1000 samples from the priors to generate samples of the Omori decay, the triggering function for an M4 event and the triggering function of an M6.7 event (Fig.5.2). We see that these priors produce a wide range of behaviour, including unrealistically large productivity compared to real earthquake process. This information is useful for comparison with the triggering functions derived from sampling posteriors later in the paper.

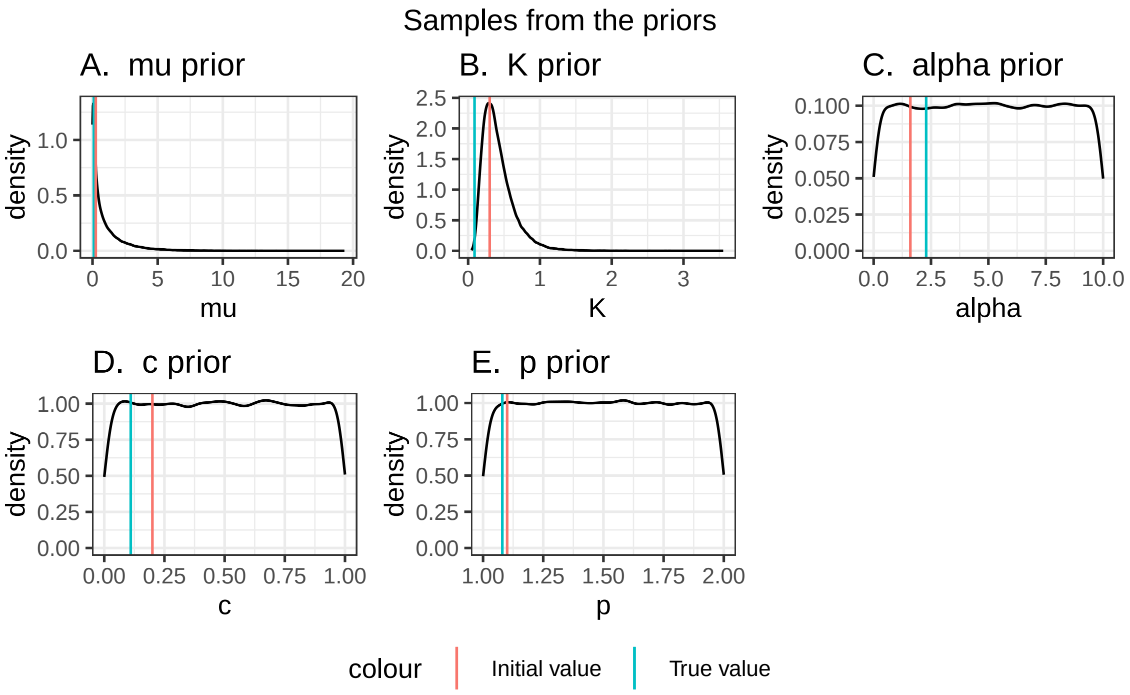


Figure 5.1: Plot showing samples from the priors on the ETAS scale that we use throughout this paper. They are intentionally broad. The red line shows the initial value used for the majority of the analyses in this paper and the green line shows the true value.

Fitting the model

The function `Temporal.ETAS.fit(list.input)` performs the ETAS inversion. The `list.input` object is a structured list containing the raw catalogue, the catalogue formatted for `inlabru`, definition of the model domain, an initial set of trail parameters on the ETAS scale, the link functions used to transform from the internal scale to the ETAS scale, parameters to set each of the priors, parameters to generate the time binning, and a series of runtime parameters that control the behaviour of `inlabru`There is a complete description of the parameters in Table 5.2 which cross-references to the section of the paper that describes their role.

In the results section, we vary the catalog, start times, and the initial set of trial parameters. To achieve this, we create a default `list.input` object and then modify these inputs by hand - the notebooks provided demonstrate how to do this.

Once we call `Temporal.ETAS.fit(list.input)`, the iterative fitting of the model parameters is handled automatically by `inlabru` until they converge or *max_iter* iterations have occurred. For a comprehensive discussion of the underlying mathematical framework, we refer the reader to Serafini et al. (2022a).

The iterative process is illustrated in Figure 5.3 and outline each of the steps below.

Step 1: Initialise/update trial ETAS parameter set We start with a set of trial ETAS parameters $\boldsymbol{\theta}_0 = (\mu_0, K_0, \alpha_0, p_0, c_0)$ which will be used as the linearisation point for the linear approximation. These initial values should lie within their respective priors. They could be sampled from the priors, but it is possible that a very unrealistic parameter combination

Parameter and Type	Default Value	Further information
Data		Catalogue of event times and magnitudes
catalogue [$t, M(, ...)$]		The input catalog as it is provided with at least a set of times and magnitudes
catalogue.bru list([ts, magnitude, idx.p])		The input catalog in the format needed for inlabru. For each event we have a [time, magnitude, id]
Domain Definition		Time domain varied in Sections 5.2.4 and 5.2.4
time.int		The provided start and end date in string format
T12 double [T1, T2]		The start and end date as number of days from the provided starting date
lat.int double	[-90,90]	Min and max latitude bounds for filtering the catalogue
lon.int double	[-180,180]	Min and max longitude bounds for filtering the catalogue
M0 double	2.5	Minimum magnitude for the model domain
Initial trial paras		Varied in Section 5.2.4
mu.init double	0.3	Initial guess for the background rate, μ
K.init double	0.1	Initial guess for the, K
alpha.init double	1	Initial guess for, α
c.init double	0.2	Initial guess for, c
p.init double	1.1	Initial guess for, p
Link functions		A list of functions used to transform the parameters from the internal scale to the ETAS scale
Priors		See Section 5.2.3 for definition
a_mu double	0.5	Gamma distribution shape parameter
b_mu double	0.5	Gamma distribution rate parameter
a_K double	-1	log-Normal distribution mean
b_K double	0.5	log-Normal distribution standard deviation
a_alpha double	0	min of a uniform distribution
b_alpha double	10	max of a uniform distribution
a_c double	0	min of a uniform distribution
b_c double	1	max of a uniform distribution
a_p double	1	min of a uniform distribution
b_p double	2	max of a uniform distribution
Time binning paras		See Section 5.2.3
Nmax int	8	value of the parameter n_{max} in Eqn.5.10
coef.t double	1.0	value of the parameter δ in Eqn. 5.10
delta.t double	0.1	value of the parameter Δ in Eqn. 5.10
bru.opt.list		See bru documentation
bru.verbose int	3	type of visual output from `inlabru`
bru_max_iter int	100	maximum number of `inlabru` iterations
num.threads int	5	number of cores used in each `inlabru` iteration
inla.mode string	'experimental'	type of approximation used by INLA
bru.inital: th.mu, th.K, th.alpha, th.c, th.p	list[double[5]]	Initial trial parameters on the internal scale. These are calulated using the inverse of the copula transformation functions in `ETAS.inlabru`
Runtime paras		
max_iter int	100	maximum number of iterations for the inlabru algorithm. The number of iterations will be less than this number if the algorithm have converged
max_step	NULL	this parameters refers to how far the parameter value can jump from one iteration to another. The greater the value the greater the potential jump. Setting a value different from NULL prevents the `inlabru` algorithm to check for convergence and the algorithm will run exactly the number of iterations specified in *max_iter*.

Table 5.2: Description of the model definition contained in list.input. This information will be passed to `ETAS.inlabru` to start the inversion. Each analysis in the results section is initialised by adjusting this list.

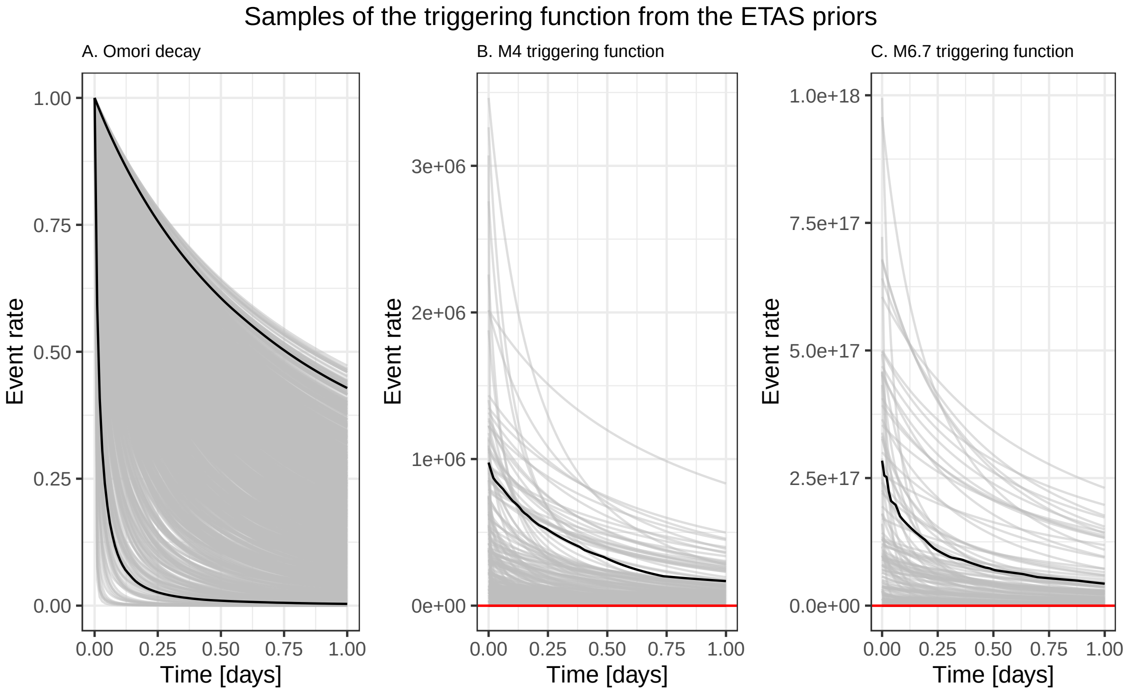


Figure 5.2: Plot samples of the Omori decay (Eqn5.16)and the triggering functions (Eqn5.17) drawn from the priors. The red lines show the 95% credibility intervals of the background event rate samples and visually allow the user to assess whether the triggering function has decayed below this.

might be chosen. These parameters will be updated each loop of the `inlabru` algorithm. In general, extreme parametrisations (e.g. parameters smaller than 10^{-5} or greater than 20) should be avoided. Usually, setting all the parameters to 1 (expect p which could be set to 1.1) is a safe choice. Another approach could be to use the maximum likelihood estimate.

Step 2. Integrated Nested Laplace Approximation `ETAS.inlabru` contains the ETAS functions that will be internally linearised (see Section 5.2.3) about an arbitrary point and then integrated. The nested integration is performed by R-INLA, but this is managed by `inlabru` so we never need to call it directly. The R-INLA output returns a comprehensive output, including the joint posteriors (LINK TO R-INLA output doc).

Step 3. Extract the ETAS posteriors and their modes From the R-INLA output, we extract the modes of the approximated posteriors $\boldsymbol{\theta}_1^*$. In early iterations, this point is usually far away from the true mode posterior, this depends on the point $\boldsymbol{\theta}_0$ used as starting point. The approximate posterior mode tends to the true one as the iterations run.

Step 4. Line search to update modal parameters At this point we have the initial set of trial ETAS parameters $\boldsymbol{\theta}_0$ that were used as the linearisation point, and the posterior modes derived from R-INLA $\boldsymbol{\theta}_1^*$. The value of the linearisation point is updated to $\boldsymbol{\theta}^* = \alpha\boldsymbol{\theta}_0 + (1-\alpha)\boldsymbol{\theta}_1^*$, where the scaling α is determined by the line search method described here `https://inlabru-org.github.io/inlabru/articles/method.html`.

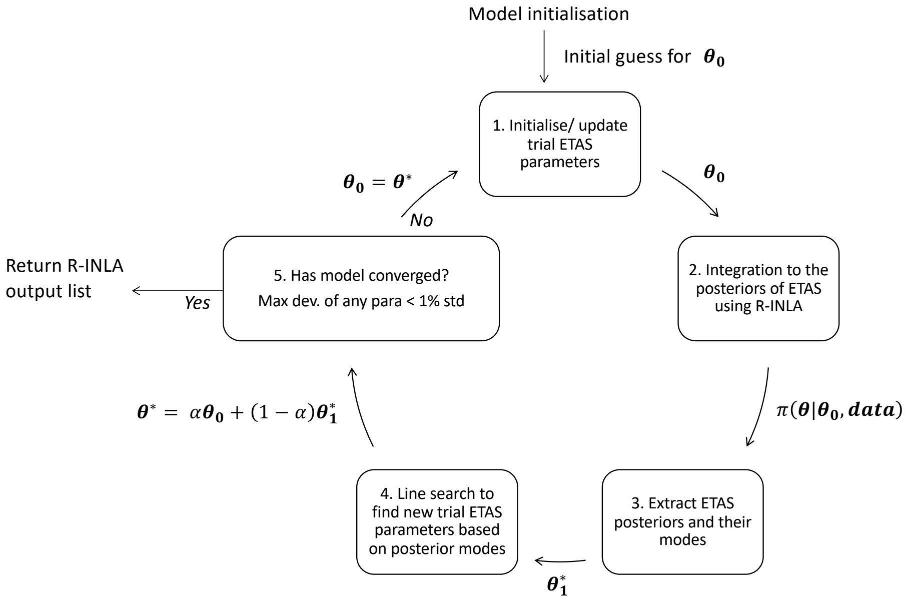

Model initialisation
Initial guess for $\boldsymbol{\theta_0}$
1. Initialise/ update trial ETAS parameters
$\boldsymbol{\theta_0}$
2. Integration to the posteriors of ETAS using R-INLA
$\pi(\boldsymbol{\theta}|\boldsymbol{\theta_0}, \boldsymbol{data})$
3. Extract ETAS posteriors and their modes
$\boldsymbol{\theta_1^*}$
4. Line search to find new trial ETAS parameters based on posterior modes
$\boldsymbol{\theta}^* = \alpha\boldsymbol{\theta_0} + (1-\alpha)\boldsymbol{\theta_1^*}$
5. Has model converged? Max dev. of any para < 1% std
Yes
Return R-INLA output list
No
$\boldsymbol{\theta_0} = \boldsymbol{\theta}^*$

Figure 5.3: Schematic diagram showing the `inlabru` workflow which iteratively updates a set of trial ETAS parameters.

Step 5. Evaluation of convergence Convergence is evaluated by comparing $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_0$. By default, convergence is established when there is a difference between each parameter pair is less than a 1% of the parameter standard deviation. The value 1% can be modified by the user. If convergence has not been achieved and the maximum number of iterations have not occurred, we set $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^*$ and return to step 1 using the new linearisation point as the set of trial parameters.

Generation of synthetic catalogues

The final component of this paper is the production of synthetic catalogues to be analysed. The synthetics are constructed leveraging the branching structure of the ETAS model. Specifically, for temporal models, background events are selected randomly in the provided time window with a rate equal to μ. Then, the offsprings of each background event are sampled from an inhomogeneous Poisson process with intensity given by the triggering function. This process is repeated until no offsprings are generated in the time frame chosen for the simulation.

Using `ETAS.inlabru`, we generate catalogues with a duration of 1000 days with a background rate of $\mu = 0.1$ events per day and ETAS triggering parameters of $c = 0.11$, $p = 1.08$, $\alpha = 2.29$ and $K = 0.089$. We take a b-value of 1 for the Gutenberg-Richter magnitude distribution. The lower magnitude threshold $M_0 = 2.5$ which is motivated by catalogues such as those in the Appenines of Italy or for Ridgecrest in California. Since we choose to study catalogues which start with quiet periods, we do not include a preparatory phase for the synthetics; this will be required when developing full forecasts including uncertainty in the preparatory phase.

We also use two different scenarios, a seeded version of these catalogues where we impose an extra M6.7 event on day 500, and an unseeded catalogue where the events are purely generated by the ETAS model. This leads to catalogues which are relatively active in the former case and relatively quiet in the latter case. Using these scenarios, we can generate different stochastic samples of events to produce a range of catalogues consistent with these parameterisations.

The R Markdown notebooks in the GitHub repository Naylor and Serafini (2023) allow the reader to see how we have implemented these catalogues for the range of models investigated in the results.

5.2.4 Results

We present the performance of the `inlabru` ETAS inversion across a range of synthetic case studies motivated by various challenges of analysing real earthquake catalogues. We are interested in the accuracy and precision of the inversion compared to the original ETAS parameterisation, understanding sources of systematic bias derived from differences in the catalogues being modelled, and the computational efficiency of the method.

All of the analyses start with one or more catalogues of 1000 days in length, generated using a constant background rate of $\mu = 0.1$ events/day above a constant magnitude threshold of $M_0 = 2.5$, and true ETAS parameters listed on the top row of Table 5.3. We choose to use this minimum magnitude as it is equivalent to the real case study examples of California and L'Aquila, Abruzzo, Italy we will return to this in the discussion section.

A consequence of choosing the 1000 day window is that there will not be a fixed number of events when comparing different samples, as some samples contain large events whilst others are relatively quiet. We make this choice because we believe that it represents the closest analogy to the data challenge faced by practitioners. However, we will be explicit in exploring the implications of this choice.

Within the sequences, there are three different timescales or frequencies that inter-relate. The duration of the synthetic catalogues, the background rate and the rate at which aftershocks decay (e.g. Touati et al., 2009). A short catalogue would be one which only samples several background events or perhaps a single mainshock aftershock sequence, or less. A long catalogue would contain periods dominated by small background events and also separate periods containing relatively isolated mainshock aftershock sequences. Clearly there is scope for a whole range of behaviour in between. Given that the accuracy of the ETAS inversion is conditional on the catalogue, we should therefore expect factors such as the catalogue duration, rate of background events and the presence of large events to influence the ability of the algorithms to find accurate solutions.

Impact of varying the initial trial ETAS parameter set, θ_0

Where algorithms require an initial set of trial starting parameters, it is important to test whether the results are robust irrespective of the choice of starting conditions. We explore the influence of the initial conditions by generating two catalogue (See Fig.5.4) using the ETAS model with the true parameters given on the top row of Table 5.3. They are both 1000 days long and the second catalogue has a M6.7 event seeded on day 500 to produce a more active sequence (Fig.5.4(B)). We then invert each catalogue using the different sets of trial ETAS parameters also given in Table 5.3; the third set of initial parameters includes the true solution.

Parameter set	μ	K	α	c	p
True parameters	0.1	0.089	2.29	0.11	1.08
Trial parameter set 1	0.05	0.01	1.	0.05	1.01
Trial parameter set 2	5.0	1.	5.	0.3	1.5
Trial parameter set 3	0.1	0.089	2.29	0.11	1.08
Trial parameter set 4	0.3	0.1	1.	0.2	1.01

Table 5.3: Table showing the true ETAS parameters and the 4 sets of different initial conditions used in analysing the catalogues in Figure 5.4 to produce the ETAS posteriors in Figure 5.5.

The first catalogue is relatively quiet and has only 217 events (Fig. 5.4(A)). All four sets of initial trial parameters find the same posteriors (Fig.5.5(A)). `inlabru` provides a good estimate of the background rate μ for this catalogue. The other posteriors are the parameters that govern the rate of self-exciting triggering. These posteriors are all skew and some of the posteriors are strongly influenced by the their priors, for example the posterior for p spans the entire range of its prior (compare Fig.5.1 for the priors and Fig 4A for the posteriors). The posteriors for the triggering parameters are relatively broad because the data is not sufficient to produce a narrow likelihood function.

In the second catalogue we have seeded a M6.7 event on day 500 (Fig. 5.4(B)). This catalogue has a well defined aftershock sequence and therefore contains significantly more events; 2530 events in total. Again, all four sets of initial trial parameters find the same posteriors (Fig.5.5(B)). The posterior for the background rate, μ, remains well resolved and there is no reduction in its standard deviation; this indicates that both catalogues have sufficient information to resolve the background rate, even though they are dominated by aftershocks. All of the posteriors for the triggering parameters are significantly narrower than for the first unseeded catalogue. This is down to two factors, firstly there are many more events in the seeded catalogue, and secondly the well resolved aftershock sequence makes it much easier to constrain the triggering parameters.

All of these models find similar posteriors irrespective of the initial trial ETAS parameters set. It is important that the priors are set broad enough to allow the potential for the posteriors to resolve the true value. This is particularly evident for the quieter model where the posteriors on the triggering parameter's rely on more information from the priors.

In Supplemental Material, we have included plots of samples from the joint posteriors derived from these two catalogues. We observe similar trends to the posteriors from other approaches, for example in the Supplemental Material of Shcherbakov (2021).

In real catalogues, the prior for the background rate needs to be set with care because, when considering a purely temporal model, it will vary depending upon the spatial extent being considered; i.e. the background rate of a temporal model when considering a global dataset would be significantly higher than just California. Further, changing the lower magnitude limit significantly changes the total number of background events.

It is difficult to interpret the ETAS posteriors directly, so we explore the triggering function by sampling the parameter posteriors 100 times, calculating the triggering functions for these posterior samples, and plotting the ensemble of triggering functions (Fig.5.6). The first column shows the Omori function which is the temporal decay of the triggering function, but without the magnitude dependent term - and is therefore also independent of the ETAS α parameter. The larger uncertainty in the posteriors for the unseeded quieter catalogue (Fig.5.4(A)) propagate through to much larger variability in the Omori decay (Fig.5.6(A)) than when the sequence is seeded with the large event in the sequence (Fig.5.6(B)). It is reassuring to see that the Omori decay from the sequence seeded with the M6.7 event lies within the confidence intervals of that derived from the quieter unseeded sequence. This implies that the prior is sufficiently broad to capture these extremes in catalogue type.

When incorporating the magnitude dependence, any bias or uncertainty in α becomes important. The figures show the triggering functions for magnitude 4.0 and 6.7 events over 24 hours. Whilst the triggering functions for the M4 events are nested similar to the Omori sequences, the triggering functions for the M6.7 event are systematically different. The posteriors from the training catalogue seeded with an M6.7 event result an initial event rate 50% higher and the two distributions barely overlap.

We conclude that the choice of training data could have a significant effect on the forecasts of seismicity rate after large events. In the next section we explore the robustness of these results to stochastic uncertainty in the training catalogues.

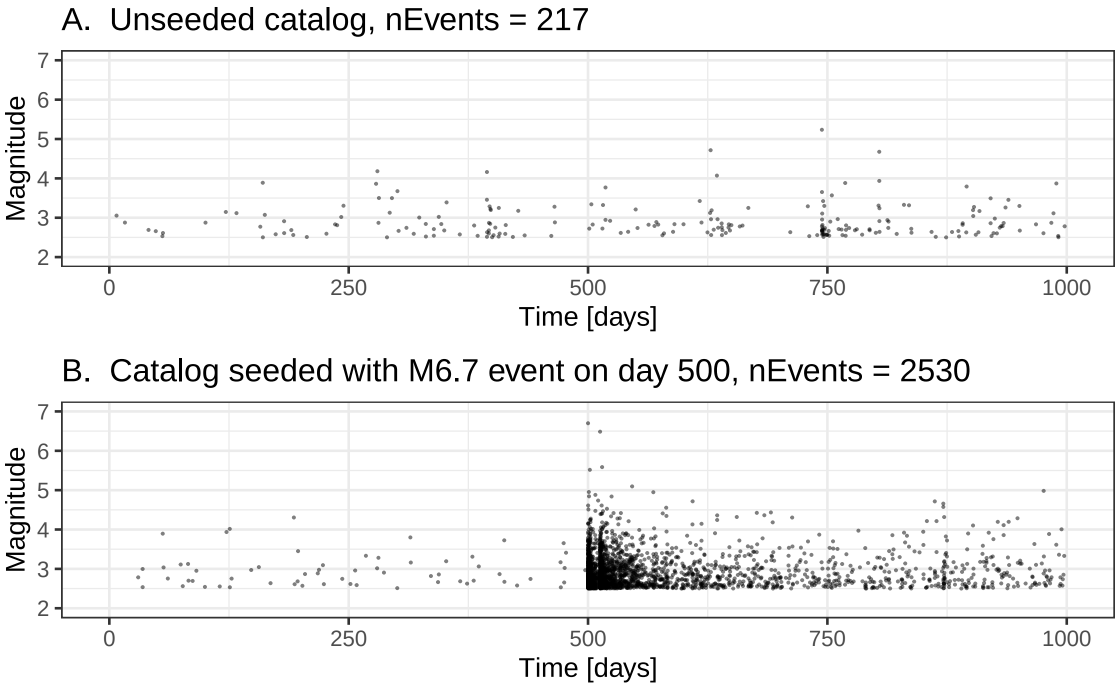


Figure 5.4: The two catalogues we use when varying the starting point for the ETAS parameters.

Impact of stochastic variability

We extend the analysis of the previous section to explore the impact of stochastic variability. We produce 10 synthetic catalogues for both the unseeded and M6.7 seeded catalogues and compare the parameters posterior distribution.

In the family of catalogues where we did not seed large event (Fig.5.7), we see posteriors of the background rate, μ, that are distributed about the true background rate (Fig.5.9(A)) and capture it well. In contrast, we mostly see very large uncertainty in the posteriors for the triggering parameters. However, the true values generally still lie within these posteriors. The very broad posteriors correspond to catalogues that had very few triggered sequences in them. Such broad posteriors illustrate how the Bayesian approach enables us to see where the data did not have sufficient power to narrow the priors significantly; this is useful in evaluating

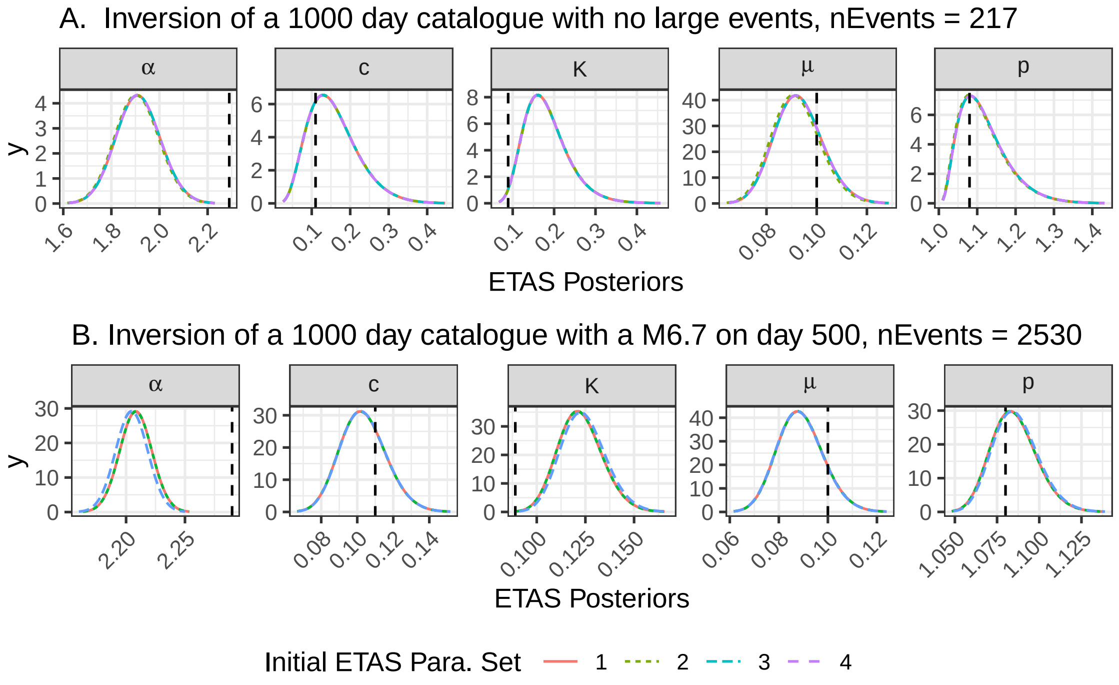


Figure 5.5: The posteriors for the inversion of the catalogues in Fig.5.4 given 4 different starting points. The vertical black line shows the true values used when generating the synthetic catalogues. Note the very different scales on the x-axes.

the robustness of a fit. Moreover, the large posterior uncertainty on the parameters would propagate through to large uncertainty in the triggering function if used within a forecast with a rigorous quantification of uncertainty. Synthetic catalogues 3 and 6 (Fig.5.7) contain the largest number of events (1842 and 930 events respectively) as a result of the events triggered by a large random event; these cases have correspondingly tighter and more accurate posteriors for the triggering parameters in (Fig.5.9(A)). Similarly, catalogues 1 and 9 have the next highest number of events (265 and 245 events respectively), and these also have the next most informative posteriors. Catalogues 2 and 5 have the fewest events (117 and 128 respectively) and produce posteriors that are significantly informed by the priors, as can be seen the the range of values being explored.

Considering the 10 seeded catalogues (Fig.5.8), we see a complementary story in the posteriors (Fig.5.9(B)). Again, the posteriors for the background rate are distributed about the true value, and show a similar spread to the unseeded case. All of the triggering parameters have much tighter posteriors. Even though some of the triggering parameter posteriors do not contain the true value, the percentage error remains small. This is due to the stochastic variability of these catalogues and this bias should decrease for longer catalogues.

There is always trade-off between α and K which is difficult to resolve. K describes the magnitude independent productivity seen in the Omori law and α describes how the full triggering function productivity varies with magnitude; consequently one requires many sequences from parents of different magnitudes to resolve K and α well.

Studies which are seeking to assign a physical cause to spatial and/or temporal variability of the background and triggering parameters should ensure that the variability cannot be explained by the stochastic nature of finite earthquake catalogues. The methods presented here provide one possible tool for doing this.

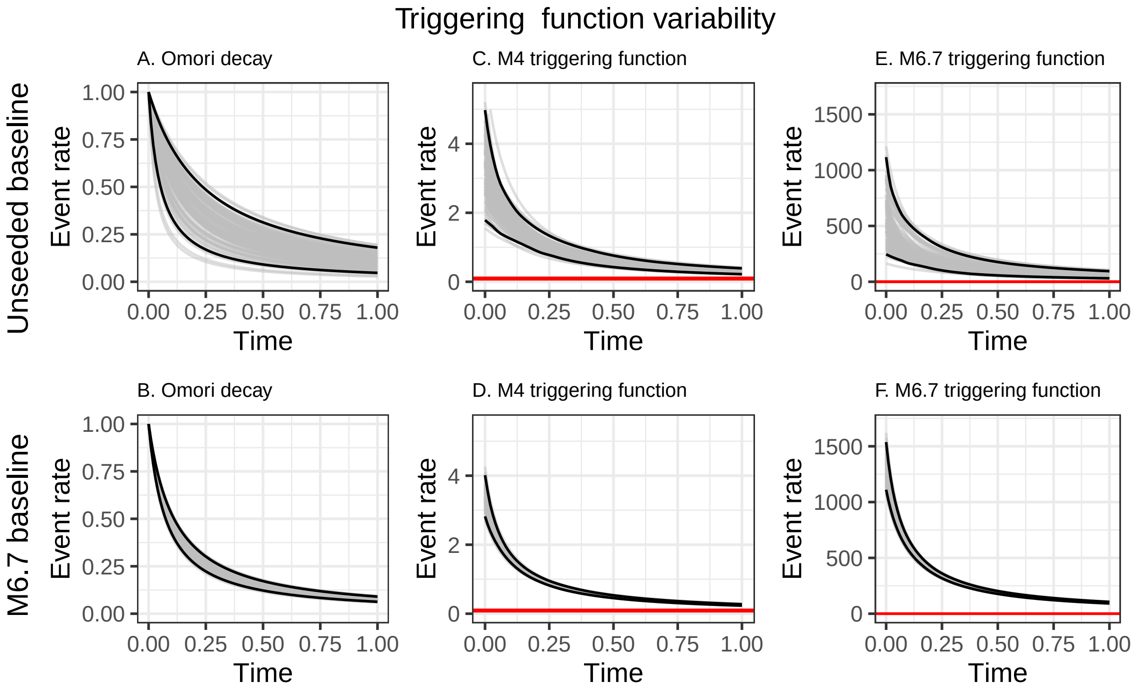


Figure 5.6: Propagation of ETAS parameter uncertainty on the triggering functions. We take 100 samples of the ETAS posteriors for the 1000 day quiescent baseline (top row) and the 1000 day catalogue with an M6.7 on day 500 (bottom row) and use these samples to explore variability in the Omori decay (Eqn.5.16), the time-triggering function (Eqn.5.17) following an M4 event, and the time-triggering function following an M6.7 event. The red lines show the 95% credibility intervals of the background event rate samples and visually allow the user to assess whether the triggering function has decayed below this.

Each of the 20 stochastic catalogues generated for this section have a different number of events. We timed the runtime for each analysis and have plotted it in Fig.5.10 as a function of the number of events in the training catalogue. We find that not only is our `inlabru` method 10 times faster than "`bayesianETAS`̈for catalogues of more than 2500 events - but also that it scales relatively linearly with the number of events. We inverted a catalogue with 15000 events in 70 minutes and it is likely this can be speeded up further using the the high performance sparse matrix solver `pardiso` package.

The inversions of synthetic data presented here show that the stochastic variability in the training catalogues produces understandable variability in the posteriors. More data and sequences containing both triggered sequences and background allow us to resolve all parameters well. Better resolution of α and K would require aftershock sequences from parents of different sizes. We see that only having lower magnitude events leads to broad posteriors on the triggering parameters. The following section explores the impact of reducing the amount of background data on the resolution of μ for the seeded sequences.

Importance of a representative sample

The motivation for applying the ETAS model is sometimes the presence of an 'interesting' feature, such as an evolving or complex aftershock sequence following a notable event. In this section we explore whether it is important to have both quiet periods as well as the aftershock

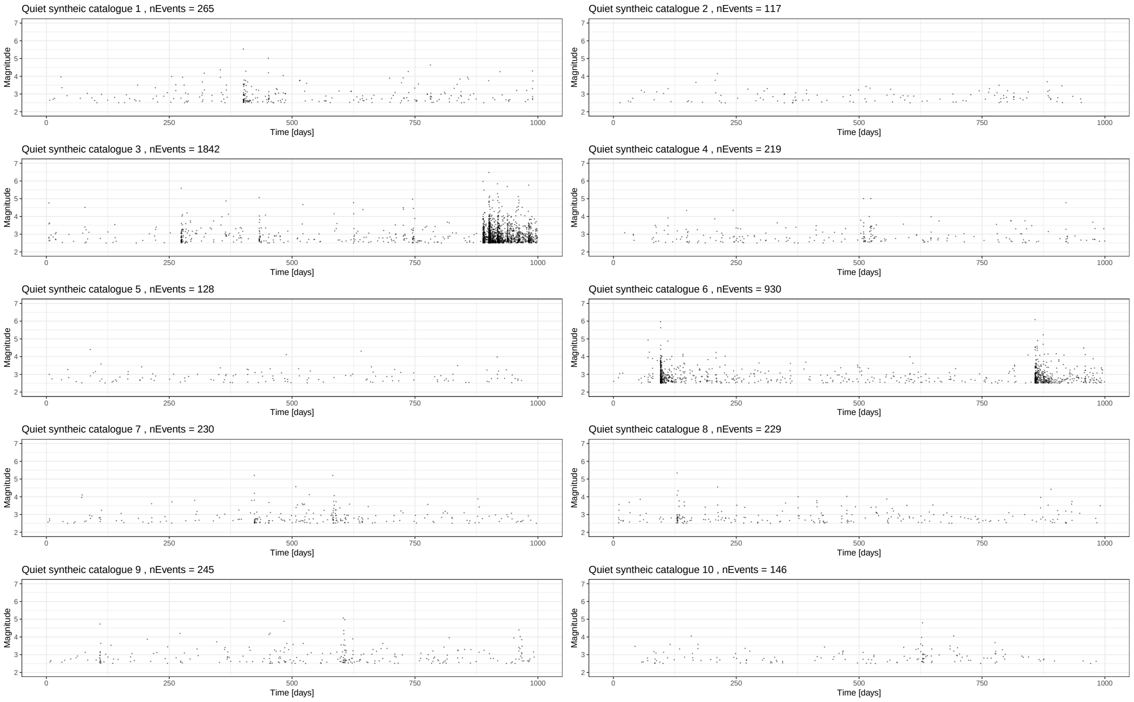


Figure 5.7: 10 synthetic catalogues based on the baseline model of 1000 days with background events but no seeded large event. All parameters are the same between the runs and these just capture the stochastic uncertainty. These are all inverted using inlabru and the family of posteriors is presented in Fig.5.12.

sequence itself for accurately recovering the true parameterisation. This motivates defining what a representative sample looks like; evaluating this in practice is non-trivial, but we can outline what is insufficient.

We start with the a 1000 day catalogue including a M6.7 event seeded on day 500. We then generate catalogue subsets by eliminating the first 250, 400, 500 and 501 days of the catalogue (Start dates of subcatalogues shown as vertical dashed lines in Fig.5.11(A)) and rerun the inlabru ETAS inversion on these subsets. Since the initial period is relatively quiet, we do not remove a large proportion of the events - however, we are removing events from the period where the background events are relatively uncontaminated by triggered events. In doing this, we explore what the necessary data requirements are for us to expect that inlabru can reliably estimate both the background and triggering parameters. When we remove 501 days, we are also removing the seeded mainshock from the subcatalogue.

First, we consider what happens to the posterior of the background rate, μ, as the length of the sub-catalogues is shortened. With 500 days of background before the seeded event, we resolve μ accurately. As the quiet background is progressively removed, the model estimate of μ systematically rises. When there is between 250-100 days of background data, the mode overestimates the data-generating parameter by around 30% but it still lies within the posterior distributions (turquoise and brown curves for μ in Fig.5.11(B)). When there is no background period, the overestimation of μ (blue curve for μ in Fig.5.11(B)) is on the order of a factor of 2.5. The estimate corresponds to the level of seismicity at the end of the model domain which has not decayed back to the background rate. From an operational perspective, it is much easier to extend the start date of training data back before the sequence of interest started than to wait until the background rate has been recovered.

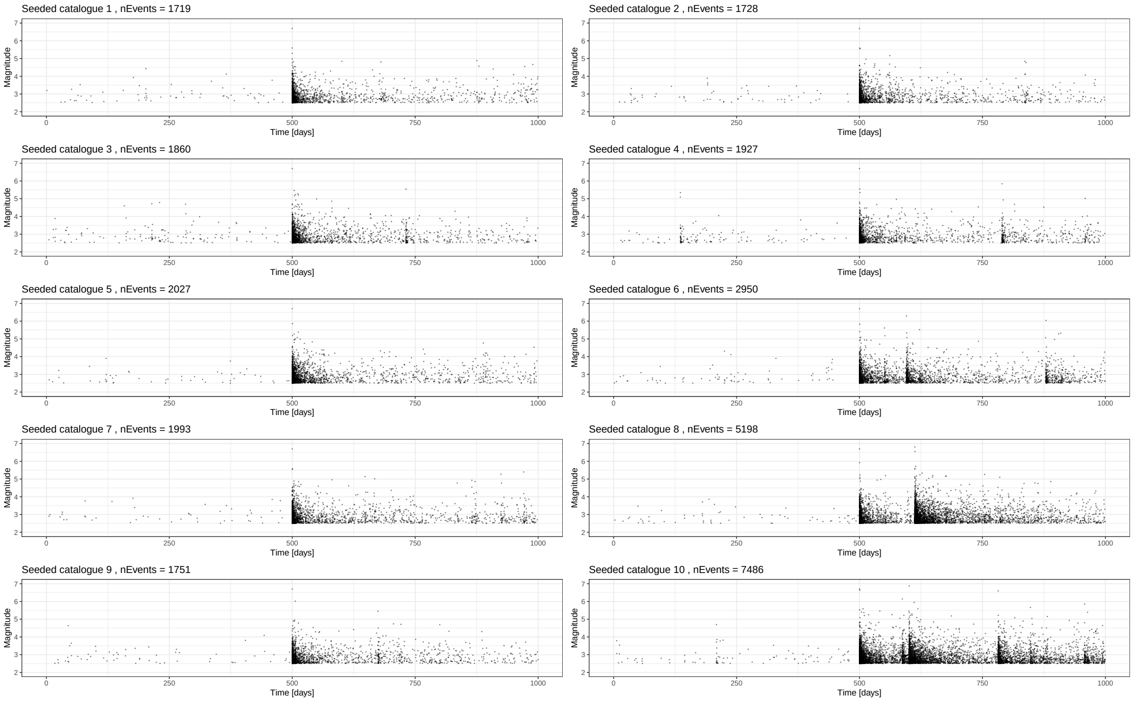


Figure 5.8: 10 synthetic catalogues based on the baseline model of 1000 days with background events and a M6.7 event on day 500. All parameters are the same between the runs and these just capture the stochastic uncertainty. These are all inverted using `inlabru` and the family of posteriors is presented in Fig.5.12.

We should therefore expect an analysis looking for time varying background rate during the sequences carries a risk of bias by this effect.

All of the models, apart from the one starting on day 501, contain the M6.7 event and 500 days of its aftershocks (Fig.5.11A). In these cases, the triggering parameters are well described by the posteriors (Fig.5.11B). However, where the model domain starts on day 501 we lose the M6.7 event and its aftershocks on the first day. This results in significant bias in all the triggering parameters as well as the background rate (Pink curve in Fig.5.11B). Modelling of specific sequences needs particular care to be taken in the choice of model domain and exclusion of the mainshock from the analysis can pose a major problem in conditioning the ETAS parameters.

The results already presented in Fig.5.9 showed that the inversion scheme struggles to recover the triggering parameters when there is no significant sequence in the dataset. Combined with the results for having no background period in this section, we argue that a representative sample should include periods of activity and inactivity if both the background rate and triggering parameters are to be estimated reliably. We also suggest mainshocks of different magnitudes would help for resolving α. By running synthetics such as the ones presented here, one can gain insight into the data requirements in specific case studies.

The model where the mainshock was not part of the subcatalogue was particularly biased (pink curve in Fig.5.11B). In the next section, we explore whether we can correct for this by including the triggering effects of events that occurred prior to temporal domain being evaluated.

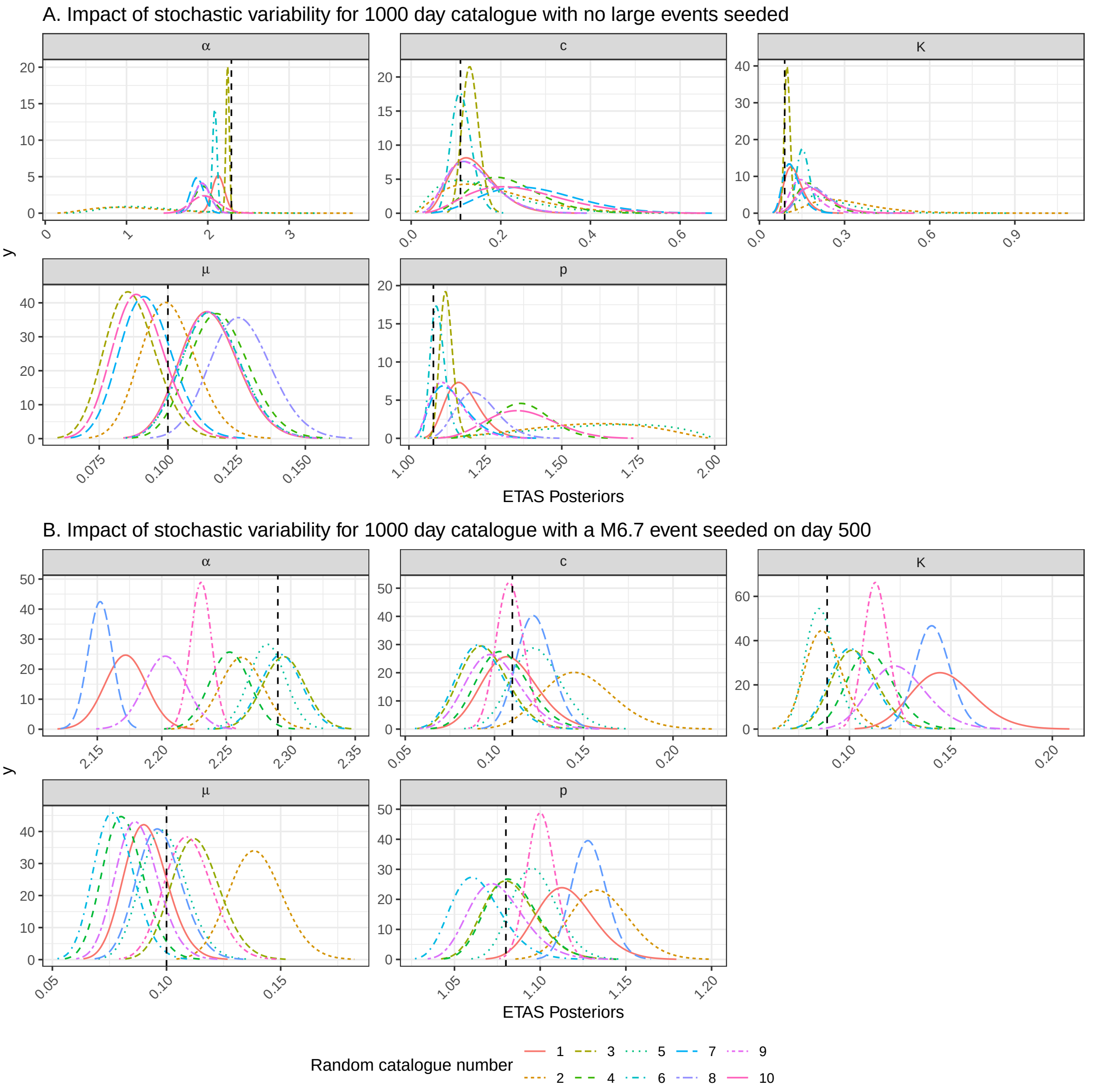


Figure 5.9: Posteriors that explore the impact of stochastic variability on the inverted ETAS parameters using `inlabru`. These are based on the relatively quiet catalogues in Figs.5.7 and those with a large event seeded on day 500 in Fig.5.8. The catalogue numbers can be cross-referenced between the figures.

Impact of historic run-in period

In the previous section, we explicitly cropped out subcatalogues and ran the analysis on that subset of the data, effectively throwing the rest away. We demonstrated the consequences removing the M6.7 mainshock from the sub-catalogue being analysed; the posteriors on the triggering parameters and background became significantly biased (Fig.5.11(B)). This example talks to the wider need for the intensity function to be conditioned on historic events prior to the start of the model domain (Ogata, 2006). This is a common issue in modelling regions that have experienced the largest earthquakes.

In `ETAS.inlabru`, we have another option when analysing catalogue subsets. Rather than cropping out the data, we can provide an extended catalogue and specify a model domain that is smaller than the whole dataset. This allows us to fit the ETAS model over data in the

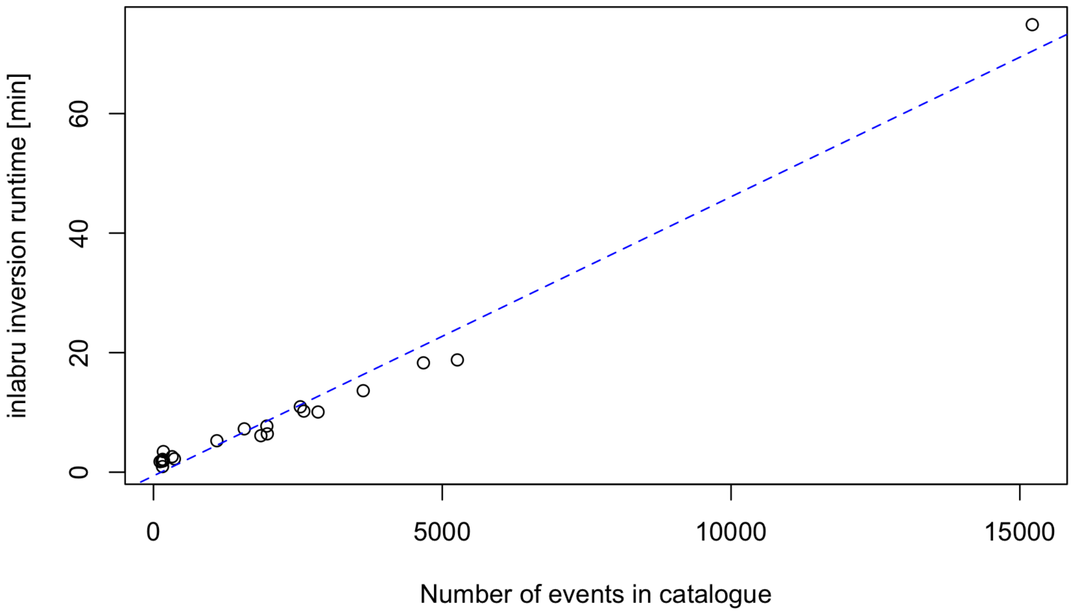


Figure 5.10: The 20x1000 day synthetic catalogues presented in Figs.5.7 and 5.8 all have different numbers of events in them because of their stochastic nature. This figure plots the time taken for `inlabru` to invert each of these catalogues as a function of the number of events in the catalogues.

interval $[T_1, T_2)$, whilst pre-condition on events, particularly large ones, that occured in an earlier preparatory phase $[T_0, T_1)$ where $T_0 < T_1$. Events later than T_2 are discarded due to causality. This defines two histories, H_0 with events between T_0 and T_2, and H_1 with events between T_1 and T_2 such that $\mathcal{H}_1 \subset \mathcal{H}_0$. Including this pre-conditioning, the log-likelihood evaluated over the interval $[T_1, T_2)$ becomes,

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{H}_1, \mathcal{H}_0) = -\mu(T_2 - T_1) - \sum_{(t_i,m_i)\in\mathcal{H}_0} \int_{\max(T_1,t_i)}^{T_2} g_t(t - t_i, m_i)dt + \sum_{(t_i,m_i)\in\mathcal{H}_1} \log\lambda(t_i|\mathcal{H}_{0,t}), \tag{5.18}$$

where $g_t(t - t_i, m_i)$ is given by equation 5.9 and $H_{0,t} = \{(t_i, m_i) \in H_0 : t_i < t\}$. In this way, the events between T_0 and T_1 contribute to the expected number of events in $[T_1, T_2)$ through the first summation, and also contribute to conditional intensity through the third term. Notice that the second summation is only over the target events within T_1 to T_2. For pre-conditioning periods containing large events, this leads to more robust estimates of the parameters as shown in Figure 5.11(B-C) and reduces the effect the choice of the starting date.

In practice, this complicates the implementation of the time binning because events occurring prior to the start of the model domain only need to be evaluated from ‘T1‘ onward. The breaking of similarity of the time bins has a penalty in the speed of the implementation.

The results of conditioning the inversion using the historic events can be seen in Fig.5.11(C) and should be compared to the equivalent results for the cases where the subcatalogues did not have this preconditioning (Fig.5.11(B)). As the start date increases, the inclusion of small background events in the history has little effect on the results because their triggering

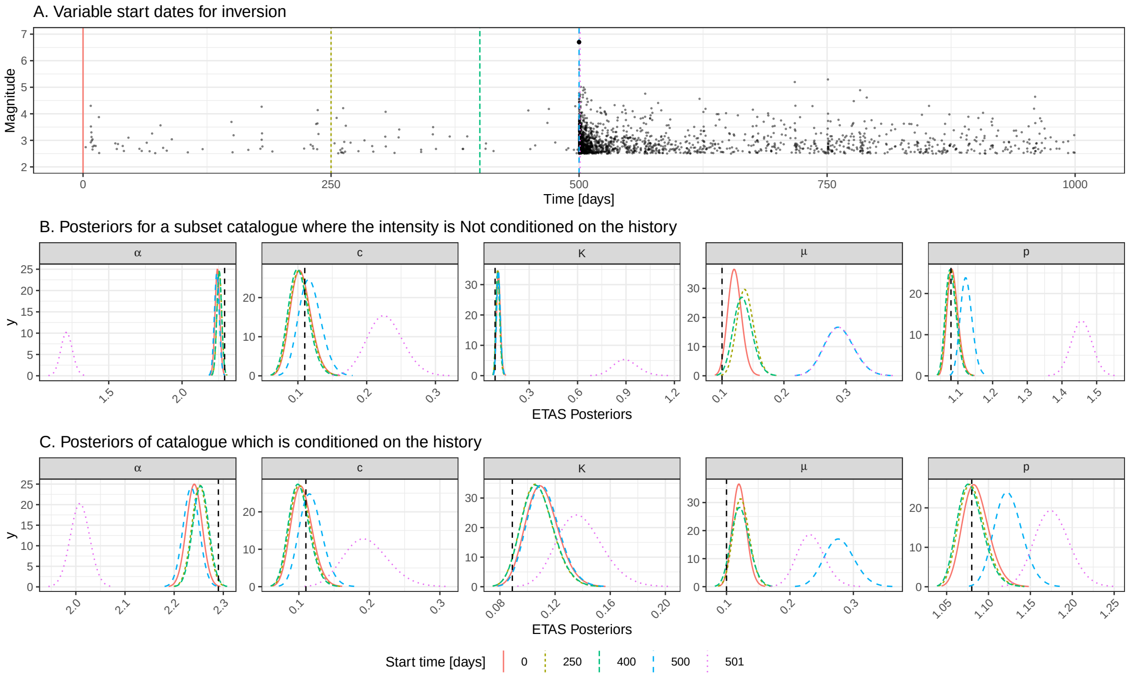


Figure 5.11: (A) Catalogue used to explore the concept of a representative sample and history conditioning. The baseline case on the top row has 500 days of background and a M6.7 event on day 500 with the sequence being recorded until day 1000. We vary the start date for the analysis to remove the first 250 days, 400 days, 500 days and 501 days. In the last 2 cases there is no background period remaining and the large event is also prior to the catalogue subset for the final case. (B) Posteriors of the ETAS parameters for each of the catalogue subsets when we crop the sub catalogue and do not use the preceding data to condition the model. (C) Rather than cropping out sub-catalogues, we now retain the events preceding the start of the model domain and use these when estimating the triggering function. This produces a notably improved performance for the start date on day 501 when the large event is no-longer within the model domain.

effect is small. However, a significant improvement in the estimated posteriors is seen when the M6.7 mainshock is removed from the model domain (c.f pink lines for each parameter in Fig.5.11(B) and (C)). The historic pre-conditioning improves the estimation of all the triggering parameters when the mainshock is missing.

In the analysis of real catalogues, this effect will be particularly relevant when there have been very large past earthquakes which are still influencing today's rates.

Impact of short term incompleteness

Finally, we explore the effect of short term incompleteness after large mainshocks on the inverted parameters (Zhuang et al., 2017; Hainzl, 2016b,c; Helmstetter et al., 2006b). This rate dependent incompleteness occurs because is hard to resolve the waveforms of small earthquakes when overprinted by many larger events, yet the effect is short lived.

We take a 1000 day catalogue with a M6.7 event seeded on day 500 and then introduce a temporary increase in the completeness threshold after the M6.7 event using the functional form suggested by Helmstetter et al. (2006b),

$$M_c(t) = M_i - G - H \log_{10}(t - t_i), \tag{5.19}$$

where, M_i and t_i are the magnitude and occurrence time of the event we are modelling the incompleteness for, t is the time we wish to evaluate the new completeness threshold for and G, H are parameters of the model. We do not address here how these parameters should be determined in a real dataset and, informed by (van der Elst, 2021a), we set them to 3.8 and 1.0 respectively for our synthetic study. Further, in this exploratory analysis we do not include incompleteness effects for other events in the sequence - so it should be considered a relatively conservative analysis.

We perform the inversion on the original catalogue and the one where short term incompleteness has removed a number of events (Fig.5.12) and compare the posteriors (Fig.5.13).

The complete catalogue contained 1832 events and the incomplete catalogue contains 1469 events (Fig.5.12). This is difficult to see on the event time plot as most of these event are very close in time to the mainshock, so we have also plotted the magnitudes as a function of event number; here we see that after the mainshock (red dashed line) there is a transient in the completeness threshold.

All of the parameter estimates in the incomplete catalogue are now notably more biased and their standard deviation has not increased to compensate for this so the true values lie significantly outwith the posteriors (Fig.5.13), and are therefore biased. The incomplete catalogue underestimates the background rate as there are fewer events in the same time period. Propagating the triggering parameter posteriors through to compare the triggering functions, we see extremely different triggering behaviour (Fig.5.14). The Omori decay for the incomplete catalogue is longer lasting but the productivity, driven by α and K, is systematically lower.

The bias in the predicted triggering functions arising from short term incompleteness is significant and cannot be ignored within an OEF context. Solving this problem within `inlabru` is beyond the scope of this paper, however it would be possible to handle it with `inlabru` knowing the functional form of the rate at which the events are missing.

5.2.5 Discussion

Having analysed a range of synthetic datasets, we now emphasise the lessons learned we should carry forward for the analysis of real sequences. In both examples below, the completeness is often assumed to be 2.5 and this compares well with our baseline synthetics.

Reliable inversions can only result from data that is representative of the processes the model is trying to capture. This means that datasets need to contain both productive sequences and periods that resolve the background without being overprinted by triggered events. The main difference between α and K is that the former describes how the productivity varies with the magnitude of the triggering event whilst the latter is a magnitude insensitive productivity. If we are to resolve these uniquely, the training data would need relatively isolated sequences that are triggered by mainshocks of different sizes; this will be challenging in many use cases.

Interpreting the results of an ETAS inversion is non-trivial. We advocate the routine analysis of synthetic catalogues to understand what it is possible to resolve in principle. In the simulation example, we have seen that the posterior of the parameters may vary widely depending on the catalogue even though the catalogues comes from the same parameters set. We have already observed this in Chapter 4 in Figure 4.5 and 4.6, and in this chapter in Figure 5.9, where some of the posterior distributions do not overlap. I believe this is not

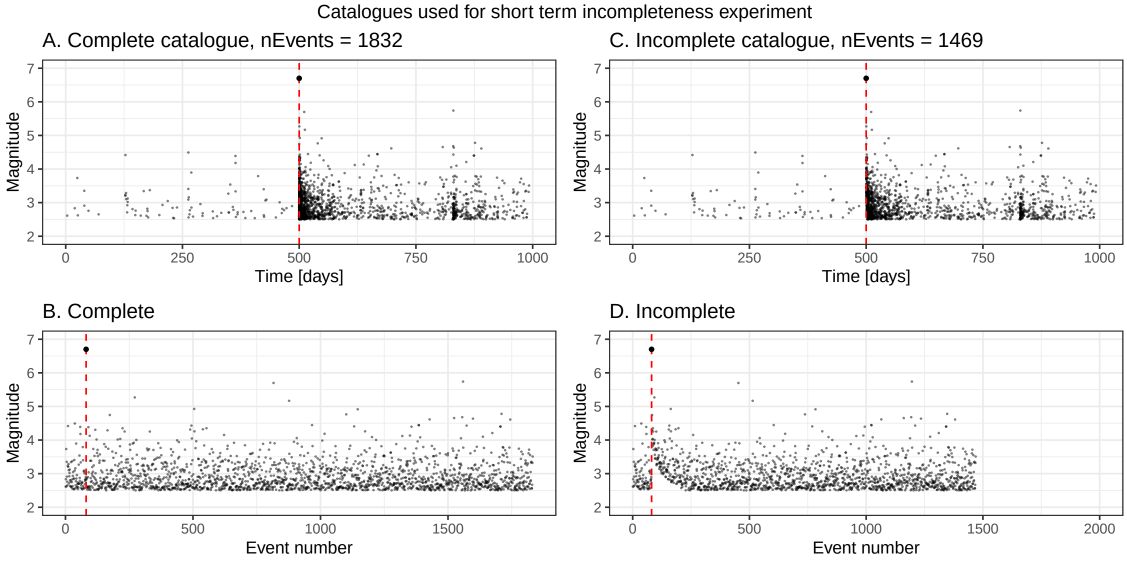


Figure 5.12: Plots of the complete baseline catalogue and the catalogue with incompleteness artificially introduced using the functional form suggested by (Helmstetter et al., 2006b). The complete catalogue contains 1832 events and the incomplete has 1469 events. The time magnitude plot does not present this incompleteness well because it occurs in the very short term after the M6.7 event. The plot of magnitude as a function of event number clearly highlights the temporally varying incompleteness just after the large event.

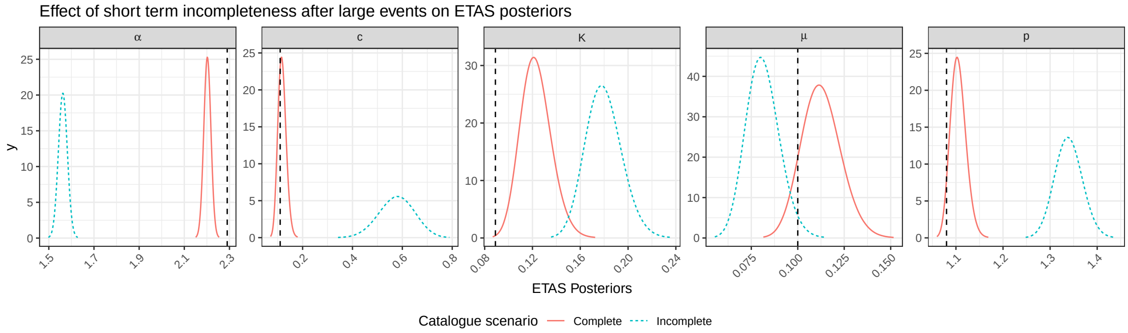


Figure 5.13: Plot comparing the posteriors of the complete and incomplete catalogues presented in Fig.5.12. The true parameters are shown with the black dashed lines.

a problem of our approximation because it happens also using MCMC which is exact. On the contrary, in Chapter 4 the posterior distributions obtained with `inlabru` overlap more than the ones obtained using MCMC which could be considered a further advantage of our approach. I believe the problem stems from the multimodality of the likelihood which may be dominated by different modes using different catalogues. Moreover, this may not be problematic in applications where the focus is on estimating the probability of future events, and, given that the same probability can be obtained with different parameters sets, this is not crucial for the estimation of the probabilities of interest.

Preconditioning models using large historic events can be significant. By considering samples of the triggering functions once can pre-determine the magnitude of events that need to be included as a function of time.

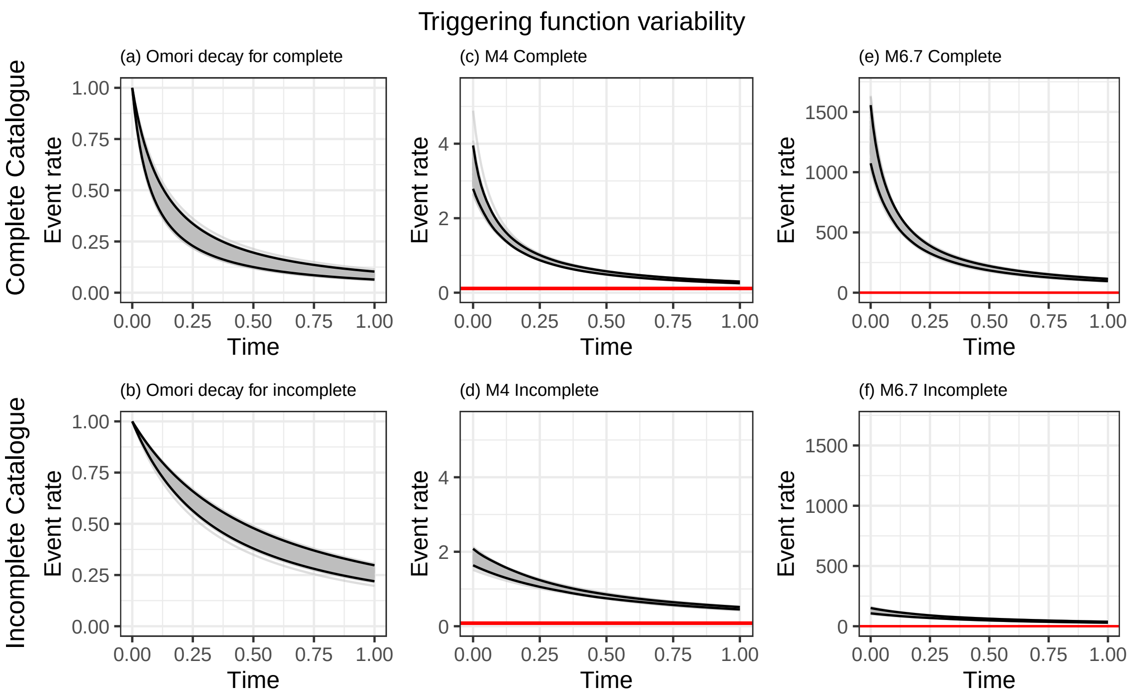


Figure 5.14: Propagation of ETAS parameter uncertainty on the triggering functions. We take 100 samples of the ETAS posteriors for the complete 1000 day catalogue with an M6.7 on day 500 (top row) and for the temporally incomplete version of this catalogue as described in the text. We then use these samples to explore variability in the Omori decay (Eqn.5.16), the time-triggering function (Eqn.5.17) following an M4 event, and the time-triggering function following an M7.6 event.
.

The use of synthetic modelling should be particularly important if time varying background rates are being inferred from the inversion of catalogue subsets in moving windows using the ETAS model.

The impact of short term incompleteness following large events is very significant and needs to be addressed either by raising the magnitude of completeness or formal modelling of the incompleteness. Resolving this for `ETAS.inlabru` is beyond the scope of this paper.

These are some of the considerations we explored here, but different use cases will present other modelling challenges that can be effectively explored through similar suits of synthetic modelling.

5.2.6 Conclusions

`ETAS.inlabru` is a fast and reliable tool for approximate Bayesian inference of the temporal ETAS model (Serafini et al., 2022a; Naylor and Serafini, 2023) . The advantage of INLA over MCMC-based methods is that it is much faster. For large models, INLA finds a solution where MCMC methods would take far too long. For smaller problems, the speed of computation allows us to take a more exploratory and interactive approach to model construction and testing (Wang et al., 2018).

The exploratory approach illustrated here can be used to identify and understand sources of uncertainty and bias in the ETAS parameter posteriors that are derived from the structural and stochastic nature of the training data. We identify the need for a representative sample to contain periods of relative quiescence as well as sequences with clear triggering behaviour if all parameters are to be well resolved.

Where studies focus solely on active sequences, the background rate can be erroneously estimated to be several times larger than the real background rate and the triggering parameters erroneously imply more rapidly decaying sequences than the true underlying parameterisation would. This implies that caution is needed in studies that allow the parameters to vary in time using windowing methods. Whilst one cannot conclusively rule out that background rates and triggering behaviors may vary, we advocate that a stationary model with constant parameters should be adopted unless there is compelling evidence independent of the ETAS inversion, Colfiorito being a case in point (e.g. Touati et al., 2014).

Rate dependent incompleteness severely degrades the accuracy of the ETAS inversion and needs to be addressed directly.

The use of synthetic modelling, as presented here, provides a basis for discriminating when variations in the posteriors of ETAS parameters can be explained by deficiencies in the training data and when there is likely a robust and potentially useful signal. Such exploration requires a fast method for performing the inversion. The interpretation is easier when full posteriors can be compared, as opposed to just having point estimates. `inlabru` is particularly well suited to this task.

5.3 Chapter Summary

In this chapter, I have shown how the `ETAS.inlabru` R-package can be used to explore potential biases in parameter estimates deriving from the quality of the data. This would have required much more effort if done using an MCMC technique which is noticeably slower than our method. In general, this type of analyses based on simulated data may have great value in determining a set of *good practices* to be followed before fitting an ETAS model and to interpret the posterior distribution of the ETAS model parameters being conscious of potential problems coming from the data used for the inversion. The next chapter generalizes the method applied in the `ETAS.inlabru` R-package to perform inference on the parameters to the spatio-temporal case and shows how covariates can be accounted for when modelling the expected number of aftershocks.

5.4 Supplementary Material

In this section, we show the marginals and pairwise joint posterior distributions of the temporal ETAS model parameters obtained from 10000 samples from the full joint posterior distribution of the parameters. The results are in line with the ones presented by Shcherbakov (2014) with a different parametrization. Figure 5.15 shows the marginal and pairwise joint distribution obtained fitting the model on the unseeded synthetic catalogue shown in Figure 5.4, while Figure 5.16 shows the same with the catalogue with the 6.7 magnitude event in Figure 5.4

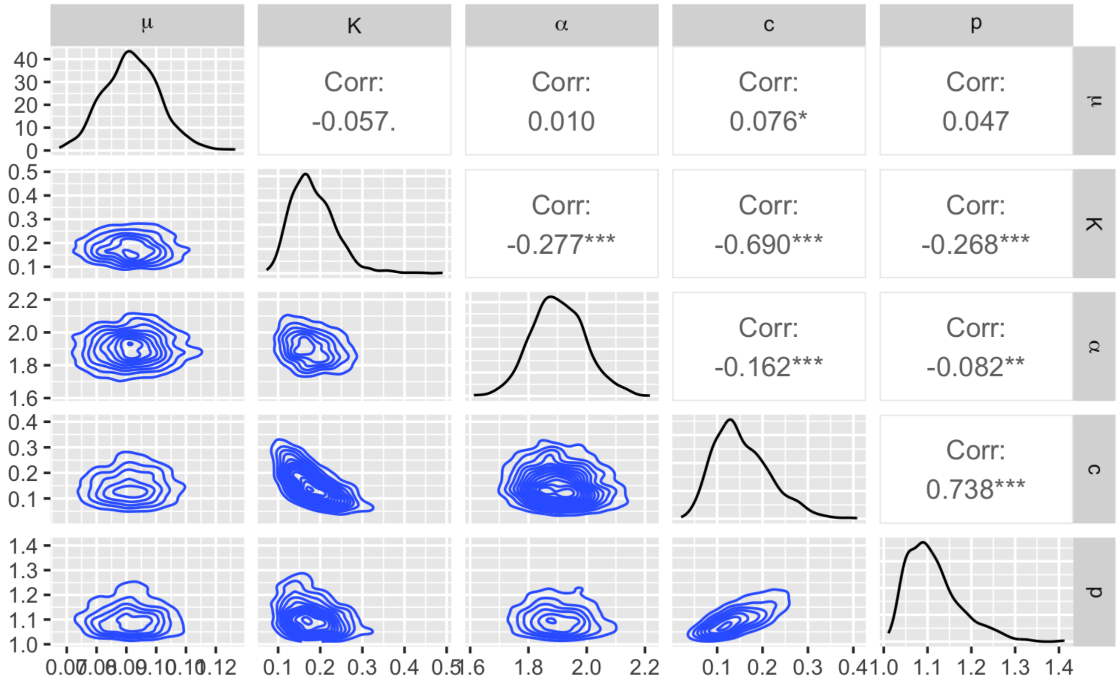


Figure 5.15: Example of a pairs plot sampling from the full posterior distribution of the unseeded model presented in Figure 5.4.

.

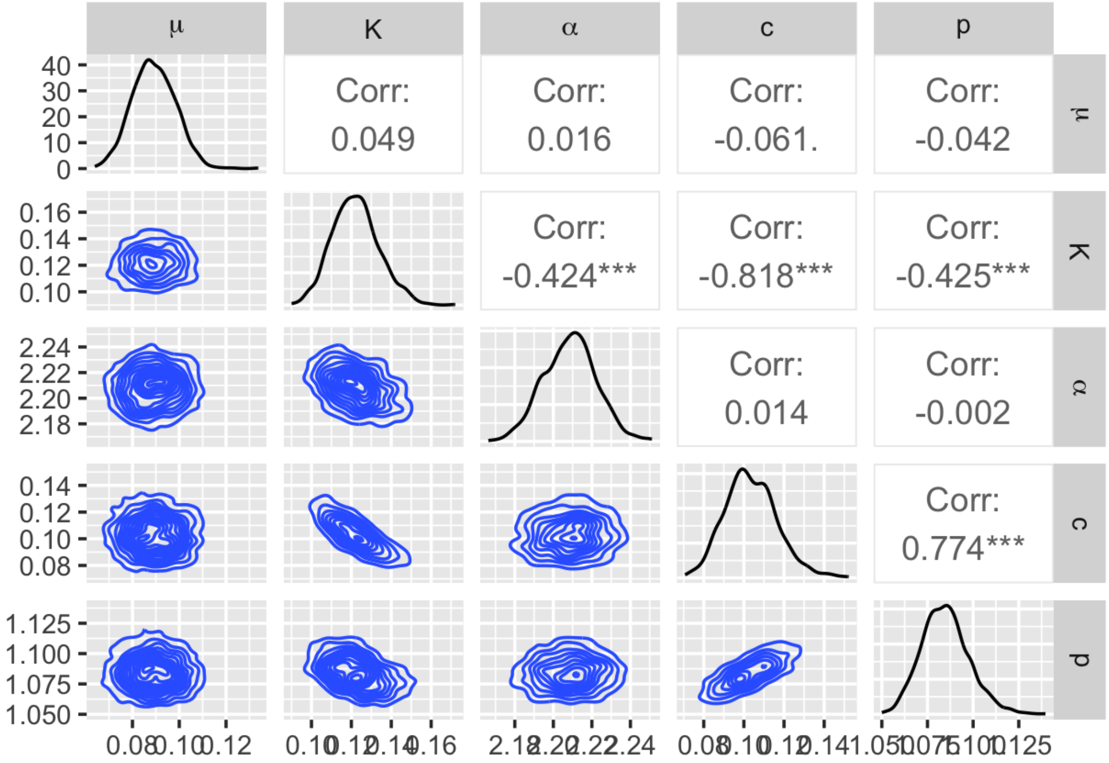


Figure 5.16: Example of a pairs plot sampling from the full posterior distribution of the M6.7 seeded model presented in Figure 5.4.

.

Chapter 6

Spatio-temporal application with covariates

6.1 Introduction

This chapter generalizes the methodology presented in chapters 4 and 5 for the temporal ETAS model to the spatio-temporal case which allows us to relax some of the assumptions within the classical ETAS model.

First, I include a spatially varying background rate and show how spatial covariates, such as fault maps or strain rate maps, can be used to develop more complex spatial models within the Bayesian framework.

Further, in the classical ETAS model, the productivity of an event only depends on its magnitude (Eq. 6.5). This means that events with the same magnitude generate aftershock sequences governed by the same conditional intensity. We relax this assumption by taking inspiration from other fields facing similar restrictions (Meyer et al., 2014; Reinhart and Greenhouse, 2018) as has been implemented in Adelfio and Chiodi (2021) and Chiodi et al. (2021). The solution is to introduce covariates in the model so that it resembles the Generalized Linear Model (GLM) framework. This provides a method for varying aftershock productivity, where the productivity may vary depending on additional characteristics of the event (e.g, depth) or to external covariates (e.g. fault information, heatflow, strain rate).

In this chapter, I compare the performance of these generalised models using data from two Italian seismic sequences that both occurred within the Apennines so they can be considered similar in style, namely the 2009 L'Aquila and 2016 Amatrice seismic sequences. The chapter is structured as follows: Section 6.2 introduces the classic spatio-temporal ETAS model and describes the spatially varying background rate and the extensions proposed Adelfio and Chiodi (2021) to include covariates; Section 6.3 describes the data on the Italian seismic sequences used to perform the comparison and the covariates; Section 6.4 illustrates the results of the comparison, while Section 6.5 discusses the results and describes possible extensions.

6.2 Models

In the spatio-temporal seismicity model, earthquake events are represented by points of the form $\mathbf{x} = (t, \mathbf{s}, m)$, where $t \in [T_1, T_2], T_1 < T_2$ is the time, $\mathbf{s} \in W \subset \mathbb{R}^2$ is the 2D space location of the epicenter, and $m \in (M_0, \infty)$ is the magnitude. The domain is the product of the domains on the different dimensions $\mathcal{X} = [T_1, T_2] \times W \times (M_0, \infty)$. Assuming that N events have been observed within the domain $\mathcal{X}$ such that the history of the process can be described by $\mathcal{H} = \{\mathbf{x}_h \in \mathcal{X}, h = 1, ..., N\}$, and that $\mathcal{H}_t = \{\mathbf{x}_h = (t_h, \mathbf{s}_h, m_h) \in \mathcal{H} : t_h < t\}$ is the history of the process up to time t, then the conditional intensity of the spatio-temporal ETAS model for a generic point $\mathbf{x} \in \mathcal{X}$ is given by,

$$\lambda(\mathbf{x} = (t, \mathbf{s}, m)|\mathcal{H}_t) = \mu + \sum_{h:\mathbf{x}_h \in \mathcal{H}_{t_h}} g(t - t_h, \mathbf{s} - \mathbf{s}_h, m_h), \tag{6.1}$$

where $g(t - t_h, \mathbf{s} - \mathbf{s}_h, m_h)$ is the triggering or *excitation* function.

Under this model, each point $\mathbf{x}_h$ induces an aftershock sequence which is an inhomogeneous Poisson process with conditional intensity

$$\lambda_{\text{after}}(\mathbf{x}) = g(t - t_h, \mathbf{s} - \mathbf{s}_h, m_h)\mathbb{I}(t > t_h), \tag{6.2}$$

where $\mathbb{I}(t > t_h)$ is an indicator function assuming value 1 if $t > t_h$ and 0 otherwise.

The form of the triggering function used in the classical ETAS model is equivalent to

$$g(t - t_h, \mathbf{s} - \mathbf{s}_h, m_h) = K \exp\{\alpha(m_h - M_0)\}g_t(t - t_h)g_s(\mathbf{s} - \mathbf{s}_h), \tag{6.3}$$

where the function $g_t(t - t_h)$ is the Omori's law describing the temporal decay of aftershock activity and is usually either in form

$$g_t(t - t_h) = (t - t_h + c)^{-p} \quad \text{or} \quad g_t(t - t_h) = \left(\frac{t - t_h}{c} + 1\right), \tag{6.4}$$

while the function $g_s(\cdot)$ is a function of the distance between $\mathbf{s}$ and $\mathbf{s}_h$, examples are given in Chapter 4 Section 4.2.4. The spatial kernel $g_s(\mathbf{s} - \mathbf{s}_h)$ describes the spatial distribution of aftershocks induced by an event in $\mathbf{s}_h$. In this chapter, we consider an isotropic spatial kernel given by Gaussian density with a correlation coefficient equal to zero and the same variance on both axes. This assumption is clearly erroneous because the aftershocks of large events does not distribute isotropically, however, isotropic kernels are used in many studies (Ogata, 2011; Ebrahimian et al., 2022; Chiodi et al., 2021; Molkenthin et al., 2022), and I believe it is a nice starting point to generalize our approach to the spatio-temporal case.

The expected number of triggered earthquakes generated within the domain $\mathcal{X} = [T_1, T_2] \times W \times (M_0, \infty)$ by an event $\mathbf{x}_h = (t_h, \mathbf{s}_h, m_h)$ is given by

$$\begin{aligned}\Lambda_{\text{after}}(\mathbf{x}_h, T_1, T_2, W) &= \int_{\max(t_h, T_1)}^{T_2} \int_W K \exp\{\alpha(m_h - M_0)\}g_t(t - t_h)g_s(\mathbf{s} - \mathbf{s}_h)d\mathbf{s}dt \\ &= K \exp\{\alpha(m_h - M_0)\}I_t(t_h, T_1, T_2)I_s(\mathbf{s}_h, W),\end{aligned} \tag{6.5}$$

where

$$I_t(t_h, T_1, T_2) = \int_{\max(t_h, T_1)}^{T_2} g_t(t - t_h)dt \tag{6.6}$$

$$I_s(\mathbf{s}_h, W) = \int_W g_s(\mathbf{s} - \mathbf{s}_h)d\mathbf{s}. \tag{6.7}$$

The quantities I_t and I_s are, respectively, the integral of the time and space components of the triggering function.

Under this model, the expected number of aftershocks generated by $\mathbf{x}_h$ is influenced only by the time of the event t_h through $I_t(t_h, T_1, T_2)$, the spatial location $\mathbf{s}_h$ through $I_s(\mathbf{s}_h, W)$ and by the magnitude m_h through $\exp\{\alpha(m_h - M_0)\}$. Regarding the latter, the model assumes that the logarithm of the expected number of aftershock scales linearly with the magnitude of the parent event.

6.2.1 Spatially varying background rate

The first modification we consider is to include a spatially varying background rate instead of a constant one. Specifically, we consider a spatially varying background rate given by

$$\mu(\mathbf{s}) = \mu\nu(\mathbf{s}), \tag{6.8}$$

where $\mu \in (0, \infty)$ has the same role as in the temporal model presented in Chapters 4 and 5, and $\nu(\mathbf{s}) : \mathbb{R}^2 \rightarrow (0, \infty)$ represents the spatial variation of the background rate.

We further assume that

$$\int_W \nu(\mathbf{s}) d\mathbf{s} = 1, \tag{6.9}$$

so that the expected number of background events in $[T_1, T_2) \times W$ is equal to

$$\Lambda_0 = \int_{T_1}^{T_2} \int_W \mu(\mathbf{s}) d\mathbf{s} dr = (T_2 - T_1)\mu. \tag{6.10}$$

In our implementation, the quantities μ and $\nu(\mathbf{s})$ are estimated independently. Specifically, μ is estimated along all the other ETAS parameters, while $\nu(\mathbf{s})$ is estimated separately. To estimate $\nu(\mathbf{s})$ we fit a LGCP model (see Section 3.3.1) with intensity $\mathcal{E}(\mathbf{s})$ and set

$$\nu(\mathbf{s}) = \frac{\mathcal{E}(\mathbf{s})}{\int_W \mathcal{E}(\mathbf{s}) d\mathbf{s}}, \tag{6.11}$$

so that it integrates to 1 over the spatial domain W.

The advantage of this approach is that it is straightforward to include covariates in the expression of the background rate. Following the approach of Bayliss et al. (2020, 2022), and described in Section 3.4, to build time-independent models for seismicity, we can include them by considering

$$\log \mathcal{E}(\mathbf{s}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{z}(\mathbf{s}) + u(\mathbf{s}), \tag{6.12}$$

where $\beta_0 \in \mathbb{R}$ is the intercept, $\mathbf{z}(\mathbf{s})$ is the vector of covariates with coefficients $\boldsymbol{\beta} \in \mathbb{R}^d$ with d being the number of available covariates. The quantity $u(\mathbf{s})$ is a GMRF with zero mean and Matérn covariance function estimated using the SPDE approach (see Section 3.3.2). The `inlabru` R-package provides a function to fit an LGCP model of this type automatically. The user can also choose to estimate the spatial variation of the background rate using the whole catalogue or a declustered one.

All the models below consider a spatially varying background rate. For simplicity, we consider the spatial variation of the background rate as composed by only the intercept and the GMRF $u(\mathbf{s})$ and we use the whole catalogue to estimate them. This produces a spatial variation of the background rate which is essentially a spatial smoothing of the observations in the provided catalogue. The difference with commonly used spatial smoothing is that with this approach, the parameters of the smoothing (e.g. length scale) are determined from the data and have a posterior distribution as the other parameters of the model, while in more classical approaches they need to be imposed by the user or determined by cross-validation.

6.2.2 Aftershock productivity as a function of covariates

We follow the approach proposed by Adelfio and Chiodi (2021) to include covariates in the number of expected aftershocks generated by each earthquake. Following equation 6.5 the logarithm of the expected number of aftershocks generated by an event $\mathbf{x}_h = (t_h, \mathbf{s}_h, m_h)$ in the spatio-temporal region $(T_1, T_2] \times W$ is given by

$$\log \Lambda_{\text{after}} = \log K + \alpha(m_h - M_0) + \log I_t(t_h, T_1, T_2) + \log I_s(\mathbf{s}_h, W), \tag{6.13}$$

Focusing on the magnitude-dependent productivity term, we can replace $\alpha(m_h - M_0)$ with a linear predictor $\eta(\mathbf{x}_h)$ such that

$$\eta(\mathbf{x}_h) = \boldsymbol{\beta}^T \mathbf{z}_h, \tag{6.14}$$

where $\boldsymbol{\beta}$ is a vector of coefficients, and $\mathbf{z}_h$ is a vector of covariates relative to $\mathbf{x}_h$. If $\boldsymbol{\beta} = \alpha$ and $\mathbf{z}_h = (m_h - M_0)$, we recover the classic ETAS formulation. The conditional intensity of the modified ETAS model becomes

$$\lambda(\mathbf{x}|\mathcal{H}_t) = \mu(\mathbf{s}) + \sum_{h: x_h \in \mathcal{H}_t} K \exp\{\eta(\mathbf{x}_h)\} g_t(t - t_h) g_s(\mathbf{s} - \mathbf{s}_h). \tag{6.15}$$

The main advantage of this approach is to allow the modelling of the effect of covariates on the expected number of aftershocks as in a Generalized Linear Model (GLM) framework with similar interpretations of the coefficients. In practice, it allows consideration of the expected number of aftershocks as an additive function of functions of the covariates. Here, I consider a linear function of the covariates, and therefore, assume that the expected number of aftershocks scales linearly with each covariate. This can be generalized to more complex functions.

In this chapter, I consider two covariates: the depth and the mean strike associated with the nearest fault. Specifically, four models are compared (see Table 6.1), the first one considers the depth d_h and the linear predictor $\eta(\mathbf{x}_h)$ is given by

$$\eta_{\text{depth}}(\mathbf{x}_h) = \alpha(m_h - M_0) + \beta_d d_h. \tag{6.16}$$

I refer to this model as the depth model.

The second model is based on associating each event with a fault and considering the mean strike associated with the fault. For each fault, the mean strike is obtained as the mean of the minimum and maximum expected strike, and events are associated with the nearest fault polygon, Section 6.3 describes the fault data. The linear predictor for this model is

$$\eta_{\text{strike}}(\mathbf{x}_h) = \alpha(m_h - M_0) + \beta_{\text{ms}} \text{ms}_h, \tag{6.17}$$

where ms_h is the mean strike of the fault associated with the $\mathbf{x}_h$ observations. I refer to this model as the strike model.

I consider a third model which includes both covariates and has a linear predictor given by

$$\eta_{\text{full}}(\mathbf{x}_h) = \alpha(m_h - M_0) + \beta_{\text{ms}} \text{ms}_h + \beta_d d_h. \tag{6.18}$$

I refer to this model as the full model.

Finally, I also consider a fourth model which does not include any covariate, and the linear predictor is just $\alpha(m_h - M_0)$. I refer to this model as the basic model. In the next Sections, I am going to compare the four models described here, which are the depth, the strike, the full, and the basic model.

For all models, we consider a spatially varying background rate as described in Section 6.2.1 and an isotropic spatial kernel given by a bivariate Gaussian density with zero correlation and equal variances.

Name	$\eta(\mathbf{x})$	parameters
full	$\alpha(m - M_0) + \beta_d d + \beta_{\text{ms}}\text{ms}$	$\mu, K, \alpha, c, p, \sigma, \beta_d, \beta_{ms}$
depth	$\alpha(m - M_0) + \beta_d d$	$\mu, K, \alpha, c, p, \sigma, \beta_d$
strike	$\alpha(m - M_0) + \beta_{\text{ms}}\text{ms}$	$\mu, K, \alpha, c, p, \sigma, \beta_{ms}$
basic	$\alpha(m - M_0)$	$\mu, K, \alpha, c, p, \sigma$

Table 6.1: Models for the comparison. For each model, the name, the linear predictor $\eta(\mathbf{x})$, and the set of parameters to be estimated are reported.

6.2.3 Priors

For all the models, we consider the same set of priors. For the standard ETAS parameters μ, K, α, c, p these are,

$$\begin{aligned}\mu &\sim \text{Gamma}(0.3, 0.6)\\ K &\sim \text{logN}(-3, 1.5)\\ \alpha &\sim \text{logN}(0.5, 1.5)\\ c &\sim \text{logN}(-1, 1.5)\\ p - 1 &\sim \text{logN}(-1, 1.5),\end{aligned}$$

where $\text{logN}(a, b)$ is a log normal distribution, such that if $X \sim \text{logN}$ implies that $\log X \sim N(a, b^2)$.

Regarding the spatial triggering function, being a bivariate Gaussian density centered at the observations, it is determined by just one parameter (the variance) σ^2 regulating the scale of the kernel. Here, we consider an exponential prior for σ^2,

$$\sigma^2 \sim \text{Exp}(0.5).$$

Regarding the coefficients of the linear predictor β_d and β_{ms} both have a standard normal prior

$$\beta_d, \beta_{\text{ms}} \sim N(0, 1).$$

All the priors are implemented using the transformation method illustrated in Section 4.2.6.

6.2.4 Spatial Grid for the log-likelihood approximation

Our technique to fit the spatio-temporal ETAS model builds on the log-likelihood approximation for the temporal model outlined in Chapters 4 and 5. The technique requires that the spatio-temporal domain is discretised into non-overlapping bins in order to provide a sufficiently accurate approximation of the integral, where for sufficiently accurate we intend sufficiently for the algorithm converges. We have illustrated the binning on the time domain in Appendix 4.4.2 and in Section 5.2.3, here I describe the binning strategy used to partition the spatial domain.

As with the temporal case, each point (earthquake event) has its own set of bins used to approximate the spatial distribution of the expected number of aftershocks generated by that point. We use a 2D spatial grid based on three parameters n_l, Δ_s, and, $\min_\Delta$. The n_l regulates the maximum number of layers minus 1, where each layer is composed of 4 polygons. The first layer is composed of 4 square bins with edge length equal to Δ_s. The second layer is composed of 4 rectangular with the shorter edge long Δ_s and the longer $3\Delta_s$, the third layer is composed of four polygons with shorter edge long Δ_s and longer $5\Delta_s$, in general, the j-th layer is composed of 4 polygons with shorter edge always equal to Δ_s and longer $(2j+1)\Delta_s$. The last layer is composed of 4 different rectangles with minimum edge length $\min_\Delta$. If the n_l is such that the minimum edge of the last layer is smaller than $\min_\Delta$ we remove the last layer. So that the total number of bins for each point is $\min(n_l, n_{max})$ where n_{max} is the maximum number of layers such that the minimum edge length of the polygons in the last layer is greater or equal than $\min_\Delta$ given a value of Δ_s. Figure 6.1 shows an example of a spatial grid obtained with this method for a point in the center of the domain considering $n_l = 5$, $\Delta_s = 20$km and $\min_\Delta = 1$km for a total of 24 spatial bins.

The advantage of using this grid is that similarly to the grid used for the time domain, the integral in each bin belonging to the same layer is the same in the case of an isotropic kernel. Furthermore, the integral on a bin of a layer is the same for all the points. This represents a great computational advantage because for N points we only have to calculate maximum $4N + n_l$ integrals instead of maximum $4N(n_l + 1)$. In the case of an anisotropic spatial kernel, there still could be some advantages given the symmetry of the Gaussian distribution. The only limitation is that this grid works only for a rectangular domain, but this is not very restrictive, in fact, any polygon can be embedded in a rectangle. The grid is only used to estimate the parameters, while the calculations on the abundance or the generation of synthetic catalogues can be done using domains with different shapes. In this chapter, for all the models, we consider a grid defined by $n_l = 8, \Delta_s = 0.1$km, and $\min_\Delta = 0.1$. We have also tried different combinations with $\Delta_s = 1$, $\Delta_s = 8$, and $n_l = 15$. The results are robust for the setting considered and we chose the parameters requiring the smallest computational time.

6.3 Data

For the model comparison, we use data from the new Italian Homogenized instrumental seismic catalogue (HORUS, Lolli et al., 2020). The HORUS catalogue reports the events from 1960 to present and covers the whole Italy region. The magnitudes have been homogenized to the moment magnitude scale, and the estimated magnitude of completeness is 4 for the period 1960-1980, 3 for the period 1981-1989, 2.5 for the period 1990-2002, 2.1 for the period 2002-2005, and 1.8 from 2005 to present. The uncertainty around the estimated magnitude is around 0.07 magnitude units and decreases over the last decade. For each event, the HORUS catalogue reports the occurrence time (year, month, day, hour, minute, second), the longitude and latitude of the epicenter, the depth, and the moment magnitude.

We focus on two seismic sequences: the 2009 L'Aquila seismic sequence and the 2016 Amatrice seismic sequence. For both of them, we consider the same spatial region of interest (orange square in figure 6.2). I have chosen to use the orange region as spatial domain because Chapter 5 showed that considering an extended domain with quiet areas provides more accurate estimates of the background rate. Regarding the L'Aquila sequence, we consider events from 01-01-2009 to 01-01-2010 with a magnitude threshold of 2.5 for a total of 1047 events. For the Amatrice sequence, we consider data from 01-01-2016 to 01-01-2018 and a magnitude threshold of 3 for a total of 1369. Figure 6.2 shows the spatial

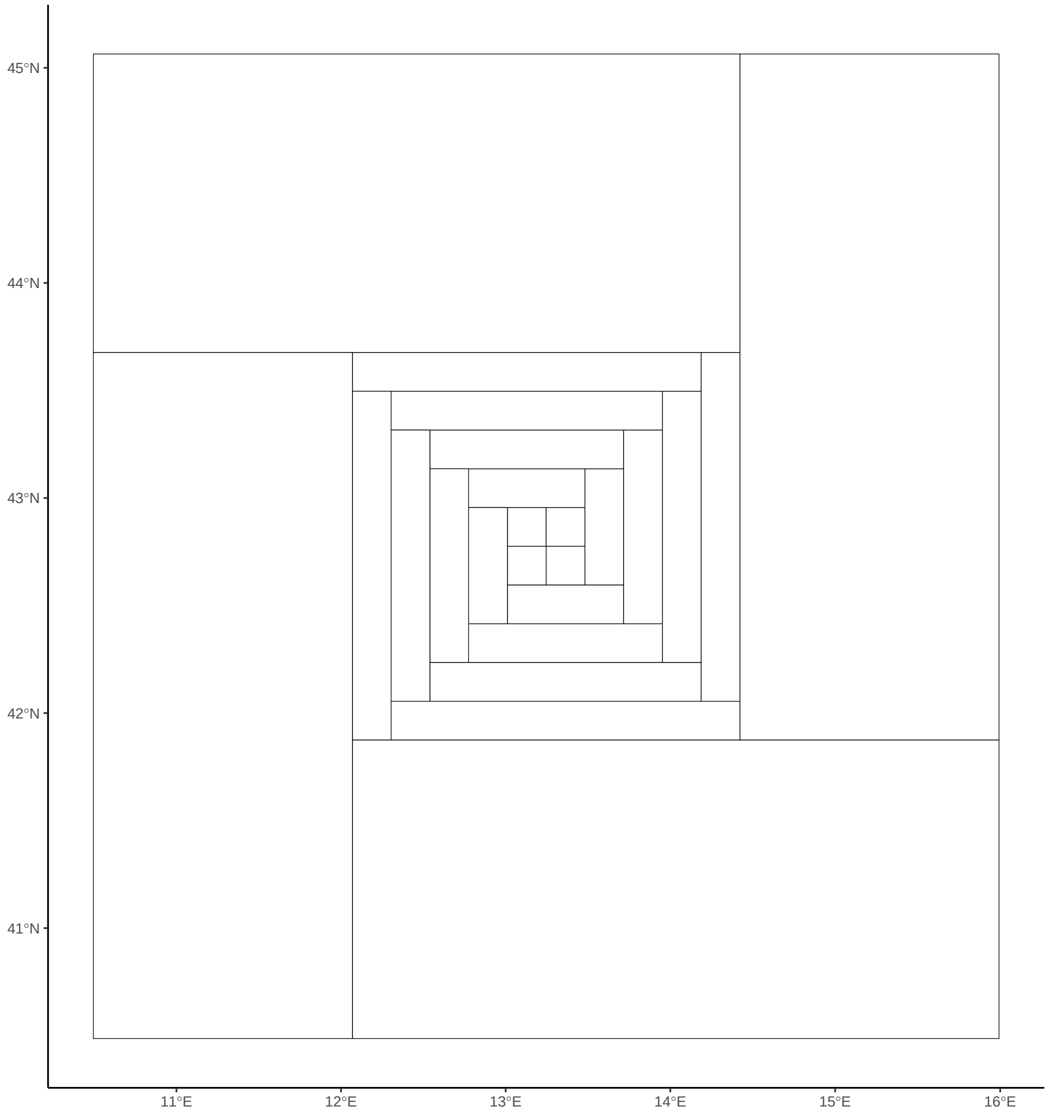


Figure 6.1: Example of a spatial grid used to approximate the integral of the conditional intensity in the log-likelihood approximation for a point located at the center of the domain considering $n_l = 5, \Delta_s = 20$km and $\min_\Delta = 1$km.

distribution of the events for the L'Aquila (panel a-c) and Amatrice (panel b-d) sequences. Figure 6.3 shows the temporal evolution of the cumulative number of events and the scatter plot of time versus magnitude. Both the spatial and temporal domain contains quiet periods (and areas).

6.3.1 Covariates

The depth of each event is provided by the HORUS catalogue, Figure 6.4 shows a histogram of the observed depth for the two sequences under study. The binwidth is equal to 0.1 km. Given the resemblance with a Gaussian distribution, I do not suspect there are significant artifacts in the depth distribution other than possibly the small local peak at 10km, likely to be the starting depth for the depth estimation algorithm. The fact this residual artifact is so small is a consequence of having a dense seismic network above the event to provide a better triangulation for the depth estimation from several nearby stations

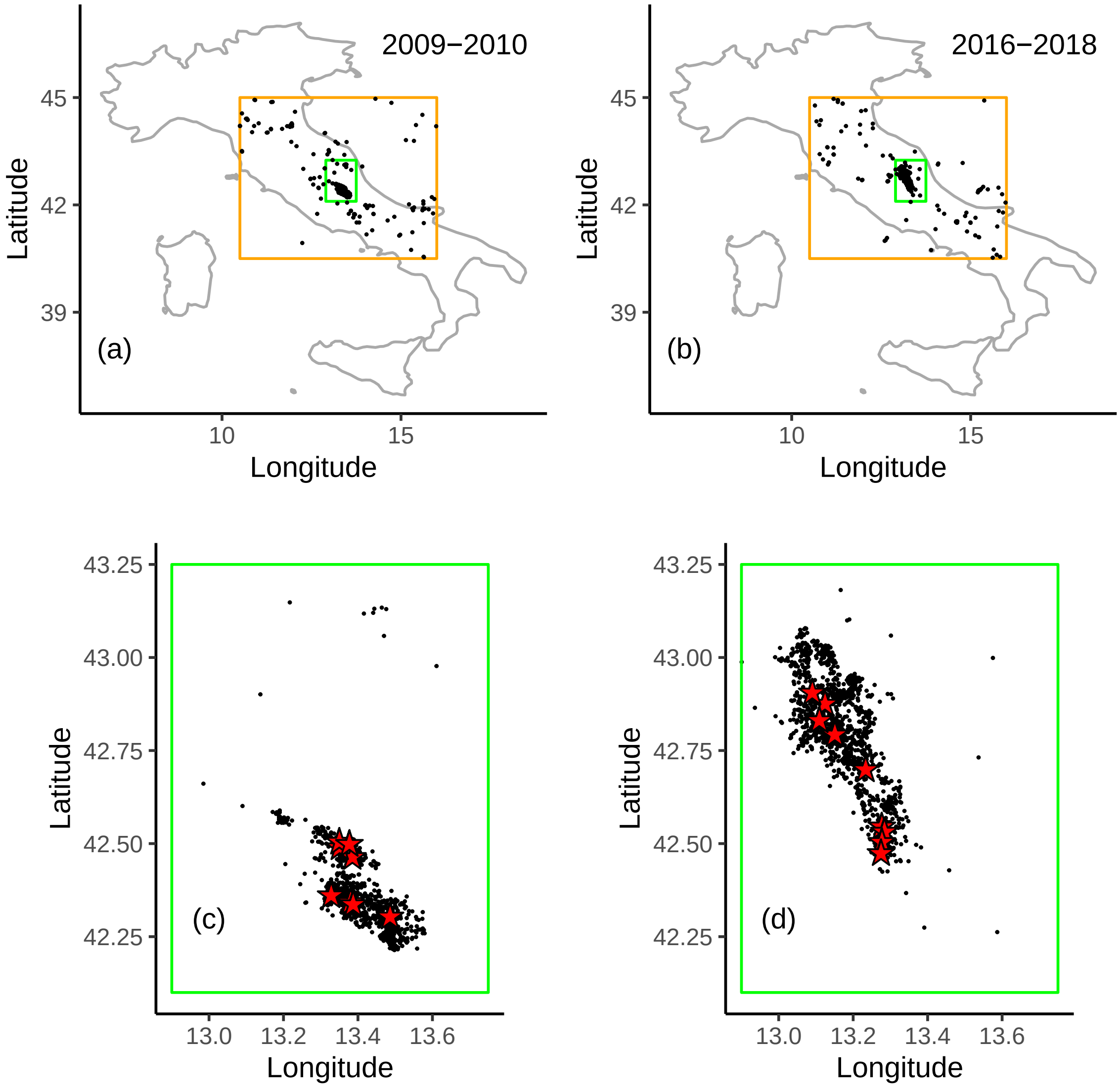


Figure 6.2: Panel a-b: area of interest (square orange polygon) considered for both sequences with respect to the Italian territory, panel (a) is for the L'Aquila sequence, and panel (b) is for the Amatrice sequence. Panel c-d: Zoom of the spatial distribution of events (green polygon) for the L'Aquila (c) and Amatrice (d) sequences. Red stars indicate events with magnitude above 5. For the L'Aquila sequence, we considered all events between 01-01-2008 and 01-01-2009 with a magnitude above 2.5 (1047 events), while for Amatrice we consider events between 01-01-2016 and 01-01-2018 with magnitude above 3 (1369 events).

Regarding the fault data, we used data from the Italian Database of Individual Seismogenic Sources (DISS 3.3.0, Basili et al., 2021). DISS provides a 2D representation of the fault network in which each fault is represented by a polygon (Figure 6.5). The polygons represent the projection on the surface of the faults and overlapping polygons indicate faults at different depths. Figure 6.6 shows the vertical section of the faults in the area of the L'Aquila and Amatrice sequences. The DISS provides also information about the 3D orientation of each fault, indeed, for each fault, we have access to its minimum and maximum depth, strike, dip, and rake.

In this chapter, we use only the mean strike to demonstrate the type of analysis that

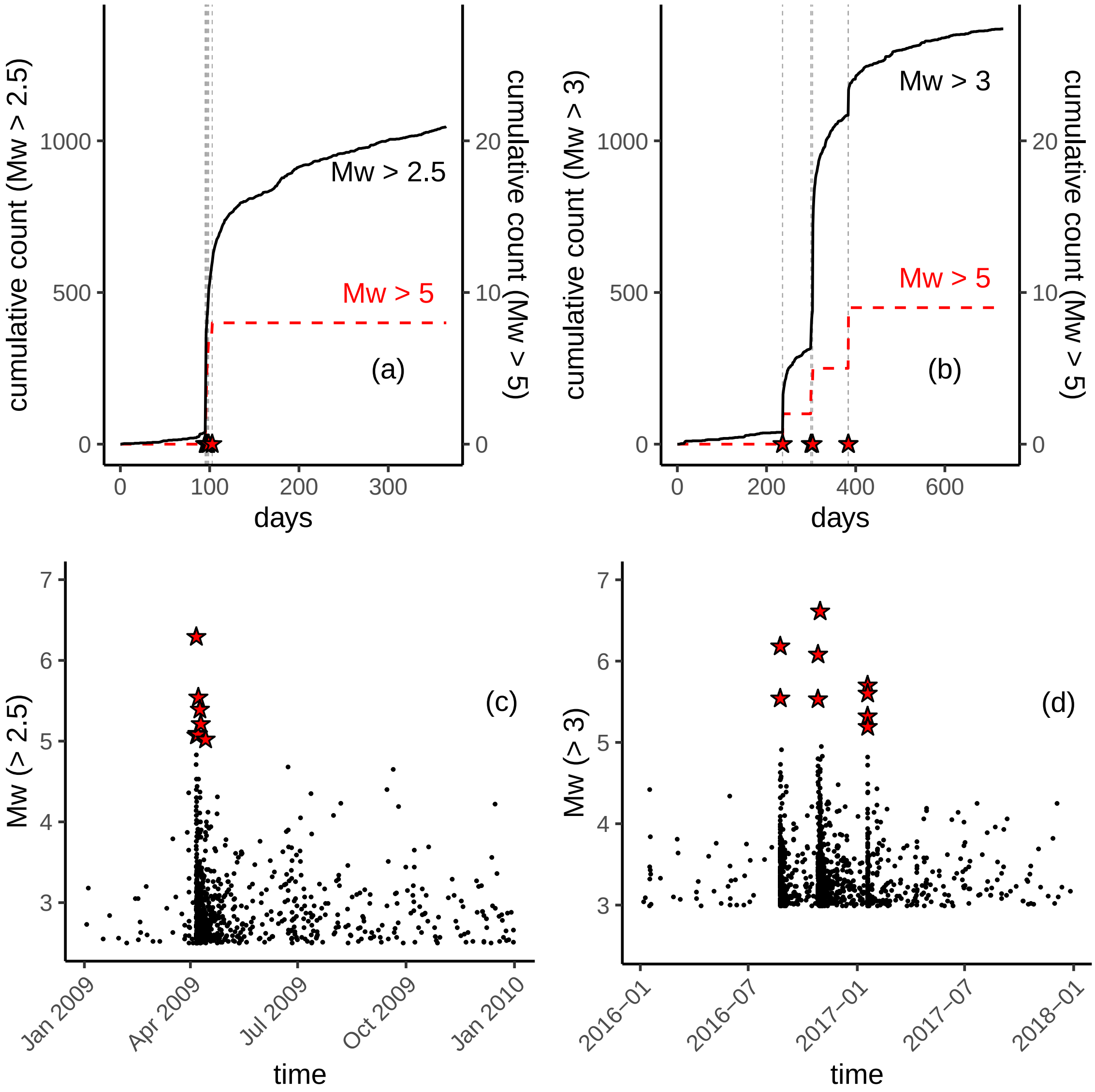


Figure 6.3: Panel a-b: temporal evolution of the cumulative number of events as a function of the number of days from the 01-01-2008 and 01-01-2016 for the L'Aquila and Amatrice sequences respectively. Black solid line represents the cumulative number of events of magnitude above 2.5 (L'Aquila, panel a) and 3 (Amatrice, panel b), while the red dashed line represents the cumulative number of events of magnitude above 5. Panel c-d: scatter plot of time versus magnitude for the L'Aquila and Amatrice sequences respectively. Red stars indicate events with magnitude above 5.

could be carried out with this approach. The value of the mean strike for each event is the mean strike of the nearest fault. Figure 6.7 shows the L'Aquila (panel a) and Amatrice (panel b) sequences and the active faults, where for active faults we intend faults with at least one event associated with them (colored in figure 6.7).

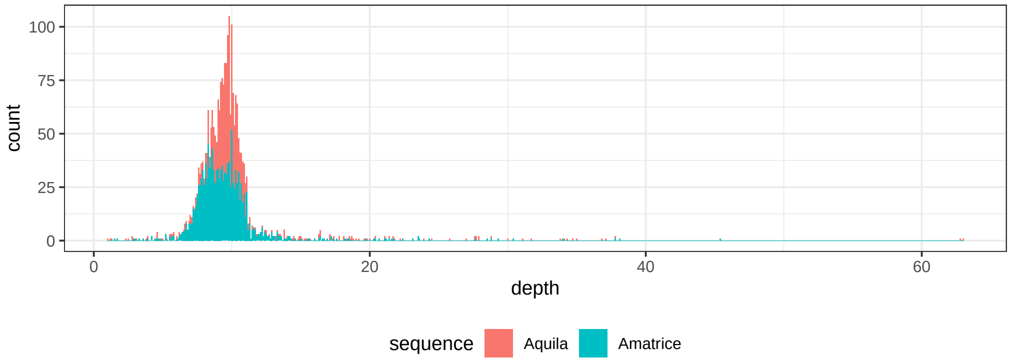


Figure 6.4: Histogram of observed depth for the 2009 L'Aquila (red) and the 2016 Amatrice (blue) seismic sequences, the binwidth is equal to 0.1 km.

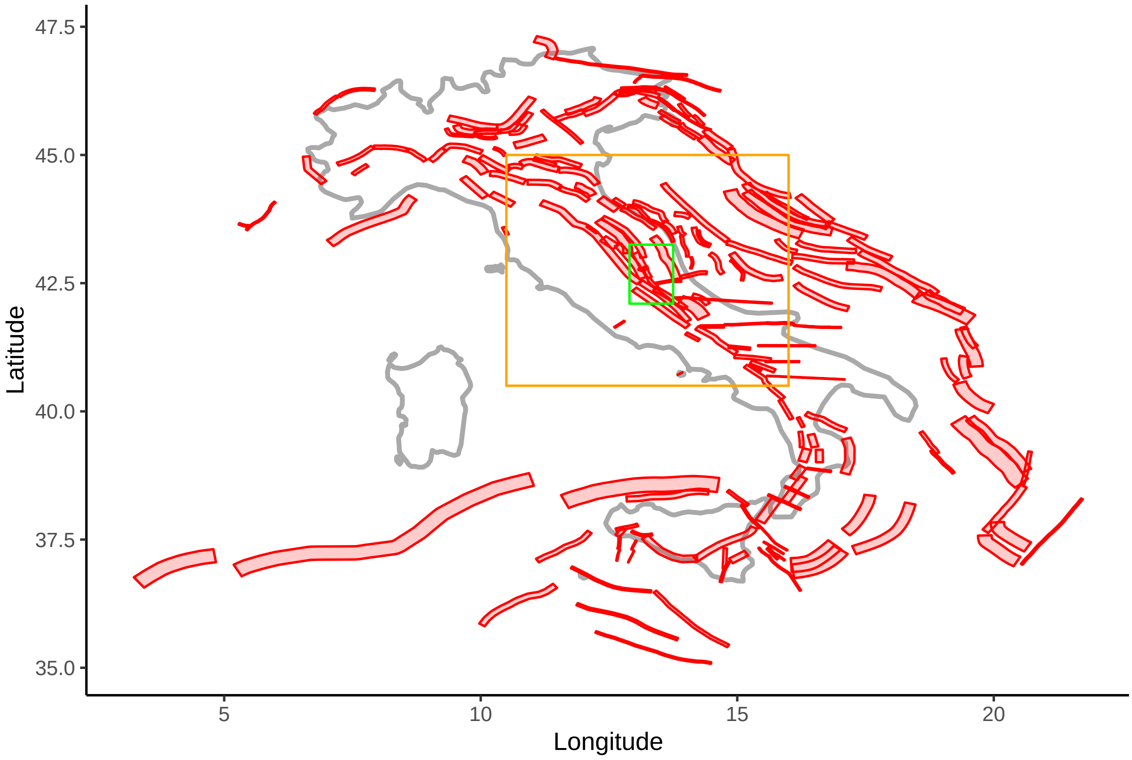


Figure 6.5: Fault network as provided by the Italian Database of Individual Seismogenic Sources DISS 3.3.0

6.4 Results

The two sequences considered in this chapter present different spatial and temporal distributions as illustrated by figure 6.2 and 6.3. For example, the L'Aquila sequence is more concentrated in time with all major events happening in a span of a few days, while the Amatrice sequence covers a longer period of time with three major identifiable clusters of large

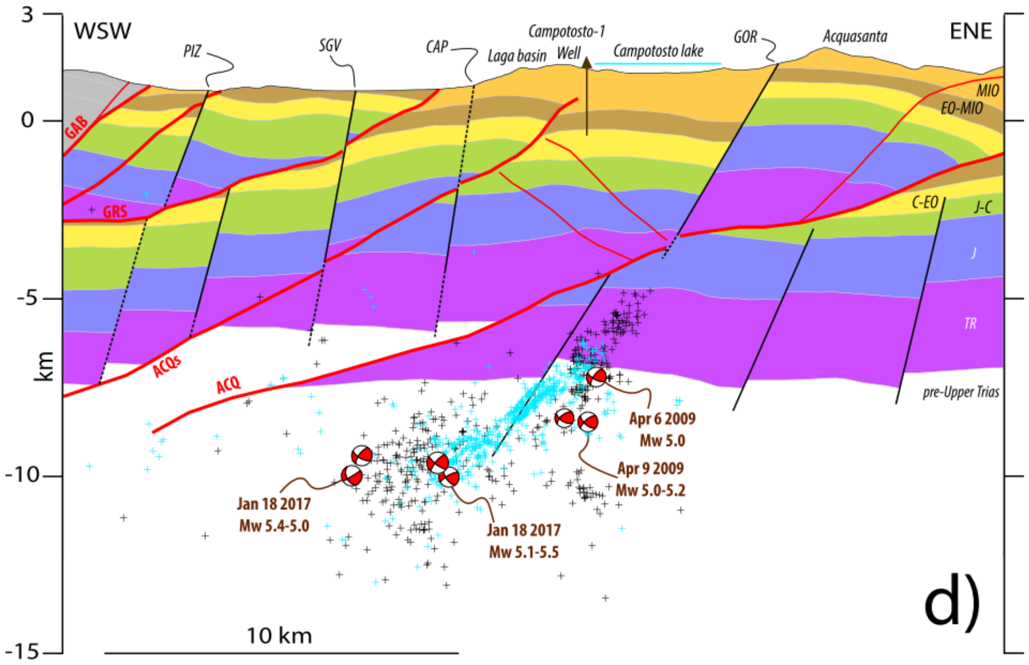


Figure 6.6: Vertical section of the fault network in the area of the 2009 L'Aquila (light blue dots) and 2016 Amatrice (black dots) seismic sequences. The image is taken from Figure 9 of Buttinelli et al. (2021), the color represents lithographic units reported in Figure 3 and Table 1 of the latter.

events. Consequently, we first compare the posteriors of the spatially varying background field and parameters $\mu, K, \alpha, c, p, \sigma$ for the two sequences. Then, within each sequence, we compare the models based on different covariates combinations to verify which one is *best*. For this task, we use the Akaike information criterion (AIC, Akaike, 1974) as it is usually done to rank competing models in the GLM framework.

6.4.1 Background field posterior distribution

We start showing the estimates of the unnormalized spatial variation $\mathcal{E}(\mathbf{s})$ which determines the spatial variation of the background rate $\nu(\mathbf{s})$ through equation 6.11. In our example, we determine $\mathcal{E}(\mathbf{s})$ by fitting an LGCP model with intensity equal to $\mathcal{E}(\mathbf{s})$ where this is given by

$$\begin{aligned} \log \mathcal{E}(\mathbf{s}) &= \beta_0 + u(\mathbf{s}) \\ u(\mathbf{s}) &\sim GMRF(0, \mathbf{Q}^-1(\mathbf{s})) \\ \beta_0 &\sim N(0,1) \end{aligned}$$

where $\mathbf{Q}^-1(\mathbf{s})$ is the precision matrix such that the covariance between location i and j is given by $\mathbf{Q}_{ij} = r(\mathbf{s}_i, \mathbf{s}_j)$ which is the Matérn covariance function as defined in by equation 3.18.

The R-package `inlabru` provides user-friendly functions to estimate the posterior distribution of the unnormalized spatial variation $\mathcal{E}(\mathbf{s})|\mathcal{H}$ given a catalogue $\mathcal{H}$ for all locations $\mathbf{s}$ in the spatial domain W. Figure 6.8 shows the mean and standard deviation of the posterior distribution of $\log \mathcal{E}(\mathbf{s})|\mathcal{H}$ for the L'Aquila (left) and Amatrice (right) sequences. The estimated mean field is essentially a spatial smoothing of the observations which is due to

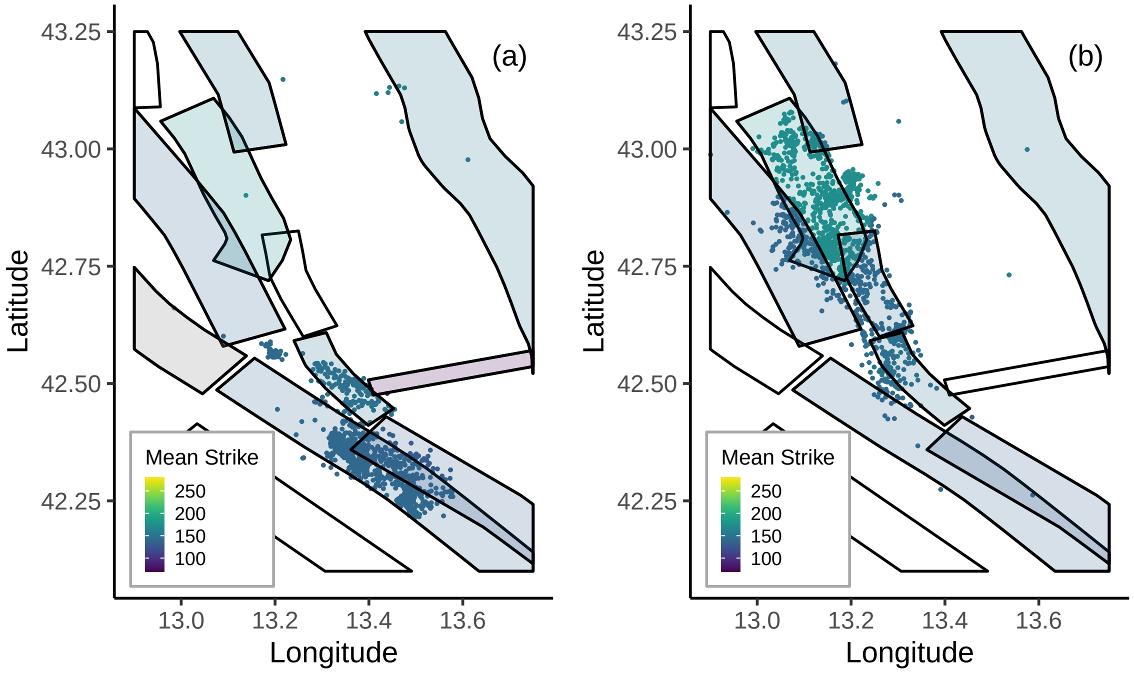


Figure 6.7: Panel a: zoom of a map of the L'Aquila sequence and the fault network. Panel b: zoom of a map of the Amatrice sequence and the fault network. The color of the points and the polygons represents the mean strike.

the fact that we have not included (at this stage) any covariate in the model. The standard deviation is higher in regions with a low number of observations and lower in regions with more observations.

The estimates of the remaining ETAS parameters are conditional on an estimate of the spatial variation of the background field $\nu(\mathbf{s})$. We consider the following estimator of the spatial variation

$$\tilde{\nu}(\mathbf{s}) = \frac{\mathbb{E}[\mathcal{E}(\mathbf{s})|\mathcal{H}]}{\sum_{j=1}^{M} \mathbb{E}[\mathcal{E}(\mathbf{s}_j^{(m)})|\mathcal{H}]\omega_j} \tag{6.19}$$

where $\mathbb{E}[\mathcal{E}(\mathbf{s})|\mathcal{H}]$ is the posterior mean of $\mathcal{E}$ given a catalogue $\mathcal{H}$, and $s_1^{(m)}, ..., s_M^{(m)}$ is a set of mesh nodes defining a triangulation of the space W with weights ω_j. Defined in this way, the numerator is a numerical approximation of the posterior expectation of the integral of $\int_W \mathcal{E}(\mathbf{s})d\mathbf{s}$ over the domain ensuring that the integral of $\tilde{\nu}(\mathbf{s})$ over the domain is equal 1. For both sequences, we have used the same mesh with 1397 nodes.

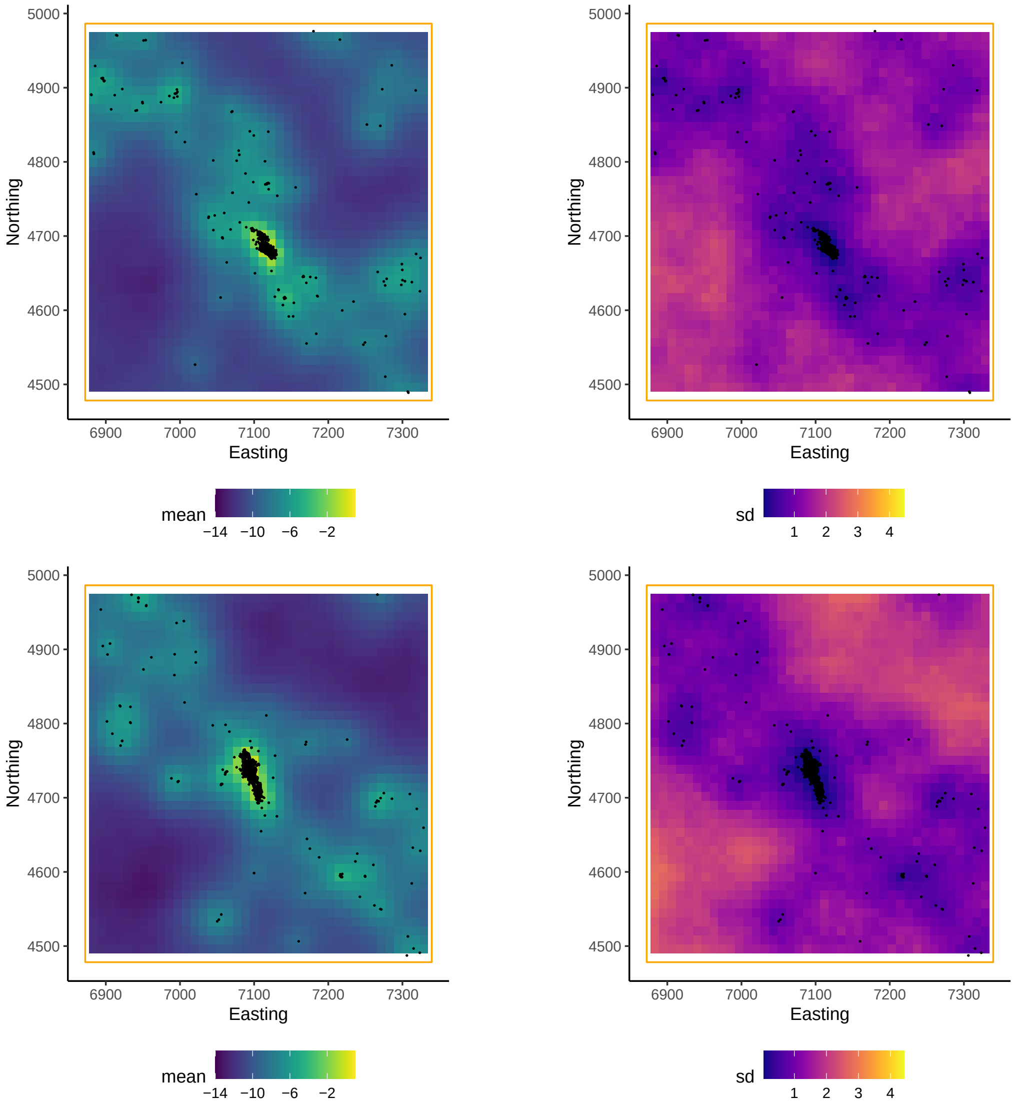


Figure 6.8: Posterior mean (left) and standard deviation (right) of the unnormalized spatial variation of the background field estimated using the 2009 L'Aquila sequence (top) and the 2016 Amatrice sequence (bottom).

6.4.2 ETAS parameters marginal and pairwise joint posterior distributions

I start by comparing the marginal posterior distribution of the parameters μ, K, α, c, p, and σ^2 which are common to all the considered models. I refer to these as basic ETAS parameters because they are the only parameters of the models without covariates. Figure 6.9 shows the marginal posterior distributions of basic ETAS parameters when no covariates are considered for the Amatrice (solid red) and the L'Aquila (dashed light blue) sequences. The differences in the parameters' marginal posterior distributions reflect the differences in the sequences shown by figure 6.2 and 6.3. Indeed, the L'Aquila sequence is more concentrated in space and time and therefore presents lower values of σ^2 and higher values of p. Also, the Amatrice sequence presents higher K and α which regulates the aftershock productivity and, in fact, it presents a lower background rate.

We compare the marginal posterior distributions of the parameters for all four models considered in this chapter (Table 6.1) using the L'Aquila (figure 6.10) and the Amatrice

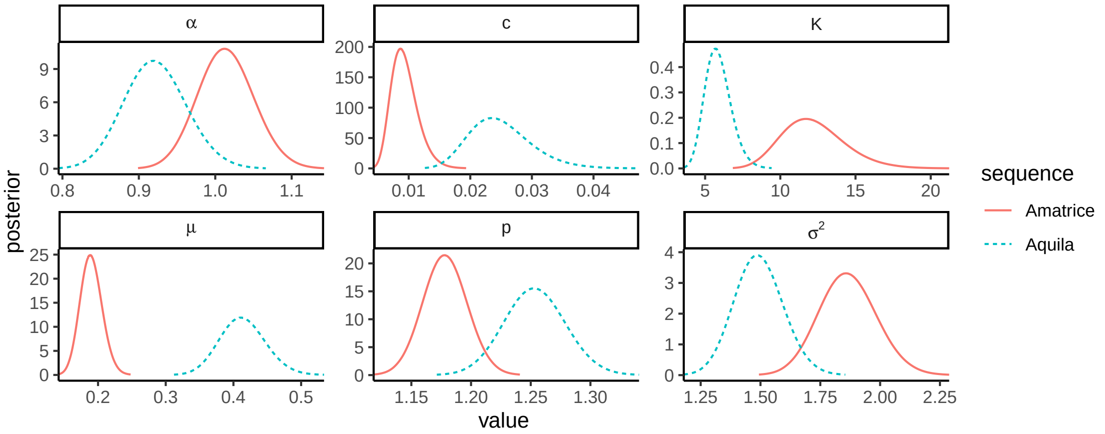


Figure 6.9: Marginal posterior distributions of ETAS parameters $\mu, K, \alpha, c, p, \sigma^2$ for the basic model with no covariates for the 2009 L'Aquila (dashed blue) and the 2016 Amatrice (solid red) seismic sequences.

(figure 6.11). For both of them, the only parameters affected by including covariates are the aftershock productivity parameters K, α, β_d, and β_{ms}, while the parameters μ, c, p, and σ^2 regulating the number of background events, and the spatio-temporal distribution of aftershocks are the same. For both sequences, the depth coefficient β_d is significantly negative, meaning that the 95% credibility interval does not include zero. This means that the depth has an apparent negative effect on the number of aftershocks produced by an event, the deeper the event the smaller the number of expected aftershocks. On the other hand, the mean strike does not have a coefficient significantly different from zero except for the case of the L'Aquila case when we do not include the depth, in which case is negative. However, the β_{ms} posterior credibility interval includes zero when considering also the depth, which may indicate that part of the variation explained by the mean strike may be just due to the correlation with the depth and its effect is not significantly different from zero when also the latter is included.

An advantage of the Bayesian approach is that it allows the study of the correlation between parameters by analyzing samples from the joint posterior distribution using the R-package `inlabru`, which provides functions to sample the joint posterior. The ability to sample the joint posterior of the parameters is also essential to produce catalogue-based forecasts incorporating the epistemic uncertainty around the values of the parameters, as well as to estimate the posterior distribution of functions of the parameters in a Monte Carlo fashion. Figure 6.12 and 6.13 show the pair plots of parameters K, α, β_d, and β_{ms} obtained from 10000 samples from the joint posterior distribution for the L'Aquila and Amatrice sequence, respectively. In both cases, K is negatively correlated with all the others, which explains the differences in the K posterior distribution between different models. The correlation coefficient between the other parameters is always below 0.2 and in some cases (for example between α and β_{ms}) changes sign from one sequence to the other.

6.4.3 Models comparison - Akaike information criterion

A popular way to compare models is by comparing their likelihood. Basically, the likelihood of a model measures *how likely* is to observe the data (that has been observed) under the model. Therefore, the likelihood can be seen as a measure of the goodness-of-fit of a model to the observed data. However, the likelihood per se does not account for model complexity,

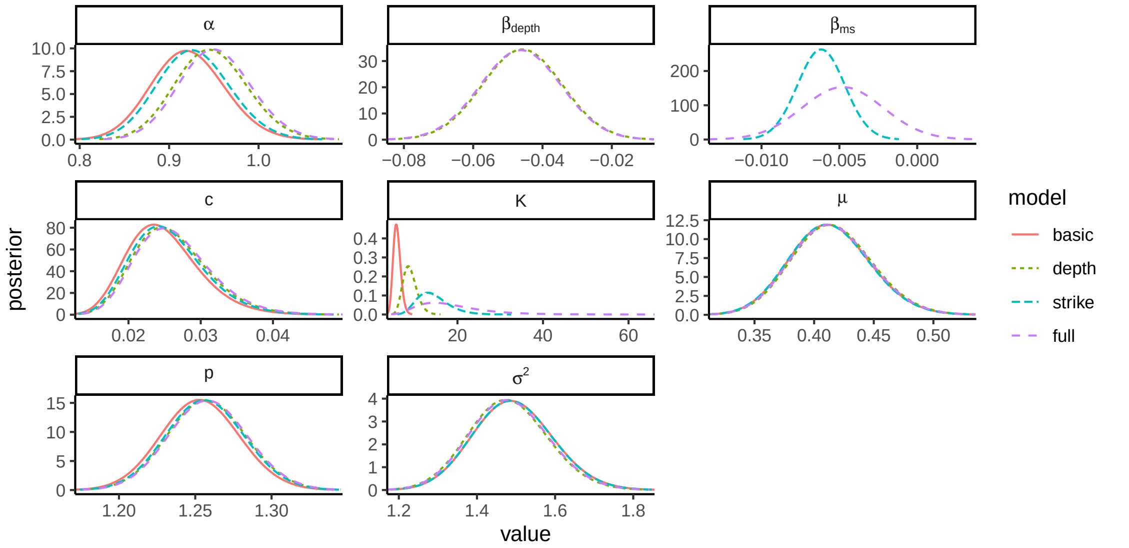


Figure 6.10: Marginal osterior distributions of the parameters $\mu, K, \alpha, c, p, \sigma^2, \beta_d$ and β_{ms} estimated using the 2009 L'Aquila sequence. Color and line type indicate the different models: full, depth, strike, and basic.

computational cost, or overfitting and, indeed, it is common to add penalties to account for these model characteristics. This gave rise to a multitude of model selection criteria based on different penalties, among the most noticeable examples are the Akaike Information Criterion (AIC, Akaike, 1974), the Bayesian Information Criterion (BIC, Schwarz, 1978) and the Widely Applicable Information Criterion (WAIC, Watanabe, 2013). We focus on the AIC because it penalizes the models according to the number of parameters which is the main difference between the models we are considering here.

Given a model with likelihood $\mathcal{L}$ and parameters $\boldsymbol{\theta}$ the AIC is defined as

$$\text{AIC} = 2|\boldsymbol{\theta}| - 2\log \mathcal{L} \tag{6.20}$$

where $|\boldsymbol{\theta}|$ is the number of elements of the vector $\boldsymbol{\theta}$. The AIC is a decreasing function of the likelihood, and therefore, the model with the lowest AIC should be selected. The AIC applies a linear penalty on the number of parameters and, consequently, between two models with the same likelihood, the one with fewer parameters has the lowest AIC. In this sense, the AIC is based on a parsimony principle, for which we should select the model with the lowest number of parameters (complexity) explaining the data with a certain level of likelihood.

As with any other function of the parameters, the AIC has a posterior distribution. The R-package `inlabru` offers an easy way to calculate functions of the parameters and extract summary statistics of the posterior distribution. Table 6.2 and 6.3 show some summary statistics (mean, standard deviation, 2.5%, 50%, and 97.5% quantiles) of the AIC posterior distribution for the different models for the L'Aquila and Amatrice sequences respectively. The models considered are the full model which includes both the depth and the mean strike, the depth and strike model which include only one of the covariates, and the basic model which does not include any covariate. All the models uses the same spatially varying background rate depending on the sequence used to fit the model.

For both sequences, the model with no covariates (basic) is the worst-performing one according to AIC. Between the models having one covariates, the model including the depth performs better than the model including the mean strike. Regarding the model with both

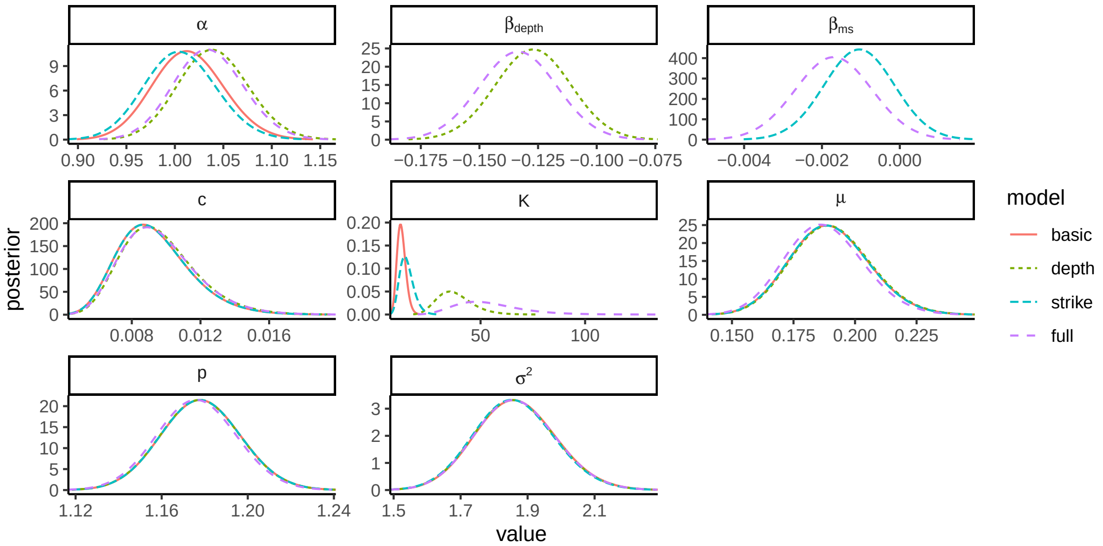


Figure 6.11: Marginal posterior distributions of the parameters μ, K, α, c, p, σ^2, β_d and β_{ms} estimated using the 2016 Amatrice sequence. Color and line type indicate the different models: full, depth, strike, and basic.

Model	Mean	SD	q0.025	q0.5	q0.975
full	2744.517	103.859	2918.409	2757.751	2562.445
depth	2751.358	109.135	2960.686	2755.869	2555.678
strike	2772.409	101.118	2961.416	2768.233	2586.899
basic	2783.697	99.737	2990.288	2785.814	2592.922

Table 6.2: Summary statistics of the posterior distribution of the Akaike Information Criterion (AIC) estimated using the 2009 L'Aquila sequence. The rows indicate the different models: full, depth, strike, and basic, while the columns show the mean, standard deviation, 2.5% quantile, median, and 97.5% quantile of the AIC posterior distribution.

Model	Mean	SD	q0.025	q0.5	q0.975
full	5114.495	110.975	5350.797	5105.715	4929.224
depth	5084.409	122.490	5315.215	5087.508	4890.211
strike	5112.818	118.416	5345.241	5112.157	4923.160
basic	5120.768	122.108	5385.752	5123.404	4896.675

Table 6.3: Summary statistics of the posterior distribution of the Akaike Information Criterion (AIC) estimated using the 2016 Amatrice sequence. The rows indicate the different models: full, depth, strike, and basic, while the columns show the mean, standard deviation, 2.5% quantile, median, and 97.5% quantile of the AIC posterior distribution.

covariates (full), it is difficult to express a preference (or not). In the L'Aquila case, the full model performs better than the strike model, but the posterior mean and median provide different rankings with respect to the depth model. In the Amatrice case, ranking the models using the posterior mean lead to expressing a preference for the strike model over the full

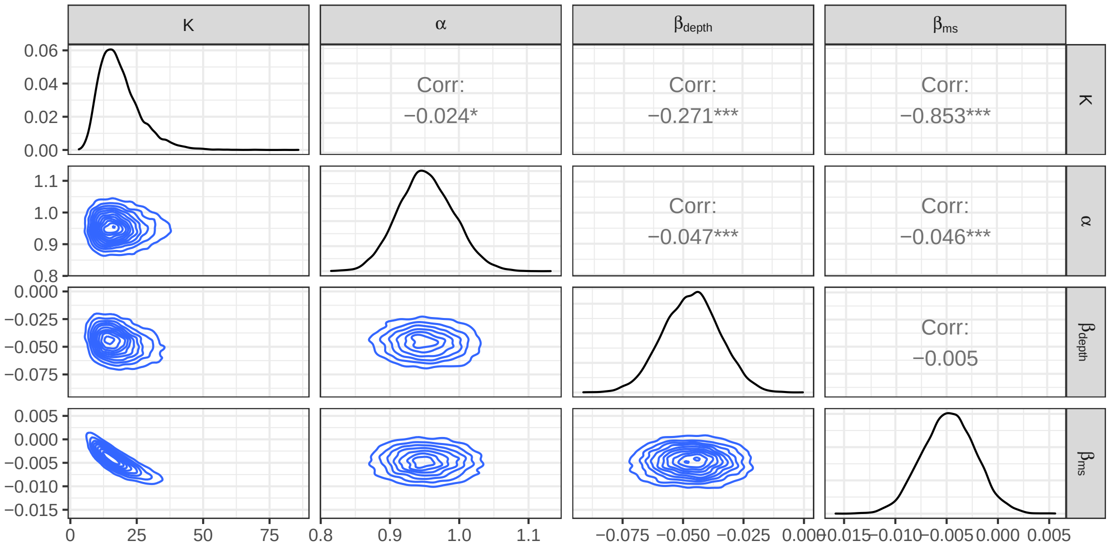


Figure 6.12: Joint bivariate and marginal posterior distributions of parameters K, α, β_d, and β_{ms} for the full model estimated on the 2009 L'Aquila sequence. The distributions are estimated using 10000 samples from the full joint posterior distribution of the parameters

one, however, this is not the case looking at the median for which the full model is better than the mean one and the depth one is better than the full. Further, the median provides the same ranking (depth, full, strike, basic) for both sequences, while the ranking changes from one sequence to another using the mean.

In general, we can express a preference for the depth model over the strike model, because both the AIC posterior mean and median is lower for the depth model than for the strike model in both cases. The depth model should be preferred to the full model because (a) the median is more trustworthy than the mean, being less influenced by the tail of the distribution, and (b) when the difference is not clear it is good practice to choose the simpler model (parsimony principle).

6.4.4 Abundance

Here, we analyze the posterior distribution of the total number of events and how these events are distributed in space. The analysis is retrospective so we expect a good fit between the model and the data. The expected number of events (or *abundance*) in a space-time-magnitude region is given by the integral of the conditional intensity over that region. Given the form of the intensity, this can be divided into the number of background events and the number of aftershocks. Under the ETAS model, the number of points in a region follows a Poisson distribution with rate equal to the abundance, and the same is true for the number of background events and aftershocks. The abundance is a function of the parameters, and, given that in the Bayesian framework, the parameters are random variables equipped with a posterior distribution, the abundance is also a random variable with its own posterior distribution. The posterior distribution of the abundance describes the information that we have on the expected number of earthquakes and allows us to better describe the uncertainty around this quantity. The abundance is a random variable, so the number of points in an area is a Poisson distribution with a rate parameter which is itself a random variable. This induces a posterior distribution on the space of the possible Poisson distribution of the number of events. Considering this extra layer of variability instead of looking at the posterior

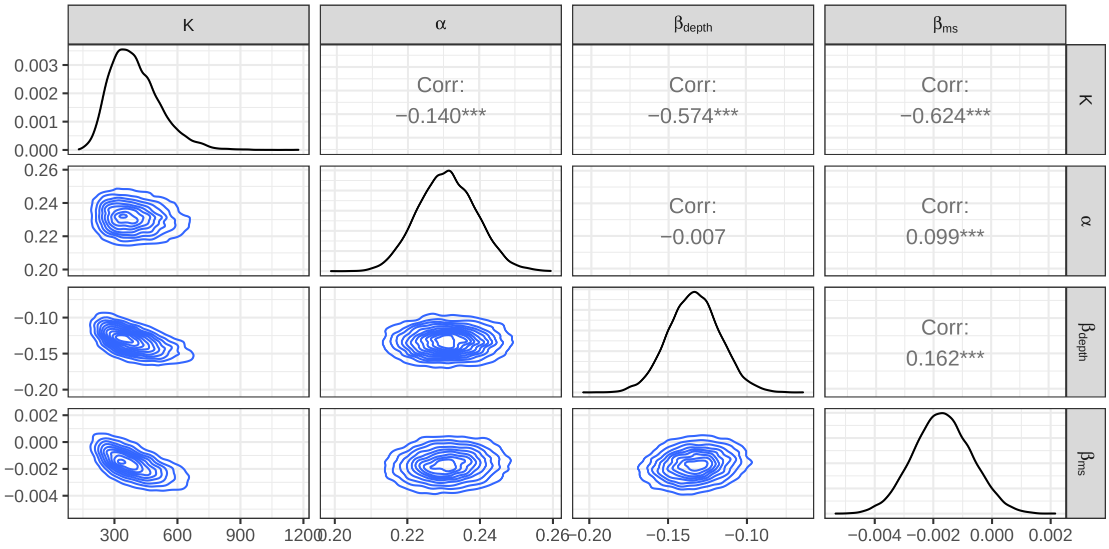


Figure 6.13: Joint bivariate and marginal posterior distributions of parameters K, α, β_d, and β_{ms} for the full model estimated on the 2016 Amatrice sequence. The distributions are estimated using 10000 samples from the full joint posterior distribution of the parameters

distribution of the abundance is crucial to represent fairly the uncertainty on the number of events.

Figure 6.14 shows the posterior median of the distribution of distributions of the number of events, number of background events, and number of aftershocks for the L'Aquila (left) and Amatrice (right) sequences under all the models considered. The vertical lines represent the observed number of events. For all models, the observed number of events lies close to typical estimators such as the posterior mean, median, or mode. The background events are only the around the 10% of the expected number of events despite the fact that we estimated the spatial variation of the background field on the same data used to estimate the other parameters which could have led to assigning all the events to the background.

Figure 6.15 and 6.16 show the spatial distribution of the number of events for the L'Aquila and Amatrice sequences, respectively. The red stars on both figures indicate events with magnitude greater than five while grey bins indicate that the logarithm of the number of events is smaller than -10. More formally, it shows the logarithm of the posterior median of the abundance calculated for each bin b of a regular grid covering the area W, namely

$$\Lambda(T_1, T_2, b) = \int_{T_1}^{T_2} \int_b \lambda(\mathbf{x}|\mathcal{H}) d\mathbf{x}. \tag{6.21}$$

There is a good fit between the observed number of events (bottom right panel of Figure 6.15 and 6.16) and the expected one (bottom left panel of Figure 6.15 and 6.16). The model represents the clustering behavior well and all the models we have considered provide very similar maps for the same earthquake sequence. This is due to the fact that the background rate and the parameter σ^2 determining the spatial kernel are the same across different models.

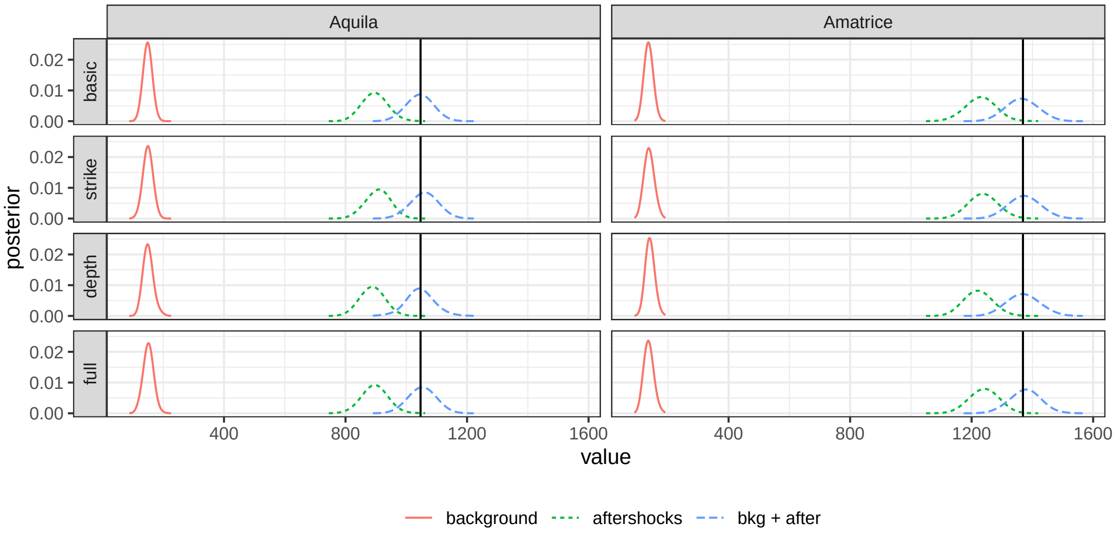


Figure 6.14: Posterior distribution of the expected number of events factorized as background number of events (solid red) number of aftershocks (dotted green) and total number of events (blue dashed); the vertical black lines represent the observed number of events. The rows represent the different models: full, depth, strike, and basic, while the columns represent the different sequences: the 2009 L'Aquila sequence (left) and the 2016 Amatrice sequence (right).

6.4.5 Effect of covariates

Here, we analyze the effect of covariates on the expected number of aftershocks. For the model considered, we will consider the quantity

$$g(t,\eta) = K\exp\{\eta\}\left(\frac{t}{c}+1\right)^{-t}, \tag{6.22}$$

where η is the linear predictor which depends on the model.

Equation 6.22 is the temporal intensity of the sub-process generated by an event with linear predictor η and occurred at time 0. The space-triggering function has been omitted because, in absence of boundary conditions, it integrates to one for the models considered in this chapter. This allows us to study the effect of covariates on the temporal distribution of aftershocks and produce Figures similar to Figures 5.14 and 5.6, and with the same interpretation. Below, I describe separately the effect of the depth and the strike. The posterior distribution of the parameters obtained with the full model which includes both covariates is used.

Effect of depth

To study the effect of depth on the aftershock distribution we fix the magnitude $m = 6$ and the mean strike to the mean value for each sequence which is around 147 for L'Aquila and 165 for Amatrice. In this way, η is a function of the depth and the coefficient β_d only and I study how $g(t,\eta)$ defined in Equation 6.22 changes as the depth changes. The change depends on the coefficient β_d of the model which is negative for both sequences, therefore, we expect that increasing the depth produces smaller values of $g(t,\eta)$. Also, the L'Aquila sequence presents higher values of p than the Amatrice sequence, which implies a shorter clustering in time. This is shows by Figure 6.17 in which the median of the posterior of the time distribution of aftershocks is shown for different values of depth. The range of depth values contains the 99% of observed depth.

For each value of the depth the temporal intensity of aftershocks $g(t, \eta)$ is determined by the coefficient β_d. The coefficient is considered a random variable and therefore also $g(t, \eta)$ is a random variable. Figure 6.18 shows summary statistics of the posterior distribution of $g(t, \eta)$ for different levels of depth. I considered three values: 5 km, 10 km, and 15 km which roughly correspond to the 5%, 50%, and 95% of the empirical distribution of depth. The posterior summary statistics of the temporal intensity are obtained by calculating the temporal intensity using 10000 samples from the joint distribution of the model parameters. The figures show that by increasing the depth the aftershock intensity decreases, especially close to zero which is where the event generating the aftershocks occurred. This in turn implies that under this model, shallow earthquakes produce more earthquakes with a higher degree of clustering in time than deep earthquakes. This is particularly visible in the Amatrice case. The variance of the posterior distribution of $g(t, \eta)$ is not affected by changes in the depth.

The difference between the L'Aquila and Amatrice case is due to the difference in the estimates of the parameters and how influent is the depth in determining the temporal intensity of aftershocks. We can measure the importance of the depth with the quantity

$$\gamma_d(d) = \frac{|\beta_d d|}{|\alpha(m - m_0)| + |\beta_d d| + |\beta_{\text{ms}}\text{ms}|}. \tag{6.23}$$

By definition the quantity $\gamma_d(d) \in [0, 1]$ for any value of d. Values of $\gamma_d(d)$ close to zero indicate a scarce influence while values close 1 indicate a strong influence of the depth on the temporal intensity of aftershocks. As expected, the two sequences present very different values of $\gamma_d(d)$. Specifically, the depth component accounts for more than the 50% of the linear predictor in the Amatrice case while this is less than the 25% for the L'Aquila case for values of the depth between 5 km and 15 km. As a consequence the temporal intensity of aftershocks changes more in the Amatrice case than in the L'Aquila case as the depth changes.

Effect of mean strike

I now replicate the analysis in the previous section but for the mean strike instead of the depth. The magnitude is fixed at 6 and the depth at 10, the latter is close to the average observed depth for both catalogues. Figure 6.20 shows the posterior median of the temporal intensity of aftershocks for different values of mean strike between 100 and 300 which is the observed range of values of the mean strike. In this case, the effect of the mean strike is stronger in the L'Aquila sequence than in the Amatrice sequence. This can also be appreciated from Figure 6.21 which shows summary statistics and 500 samples from the posterior distribution of the temporal intensity of aftershocks.

I use the same technique used in the previous section to calculate the influence of the mean strike on the linear predictor. Specifically, we define the quantity

$$\gamma_{ms}(\text{ms}) = \frac{|\beta_{\text{ms}}\text{ms}|}{|\alpha(m - m_0)| + |\beta_d d| + \beta_{\text{ms}}\text{ms}|}, \tag{6.24}$$

which has the same interpretation of $\gamma_d(d)$ defined in Equation 6.23. Figure 6.22 shows that the mean strike is more influential in the L'Aquila case than in the Amatrice one in accordance with previous Figures. Comparing γ_d and γ_{ms} it is evident that for the Amatrice case the depth has a greater influence than the mean strike on the linear predictor. Regarding the L'Aquila example, instead, the two covariates have similar levels of influence on the linear predictor with the mean strike being slightly higher than the depth.

6.4.6 Example Application: Daily forecasts for real-time loss forecasting

I am involved in a joint project exploring the feasibility of real-time financial loss forecasting (Nievas et al., 2023) using synthetic data. The idea is to provide an example of how the information provided by building-specific sensors measuring ground shaking may be used to improve preparedness during a seismic sequence. Our approach is employed to generate daily forecasts of seismicity, which then propagate into estimates of ground shaking at individual buildings. This information is used to calculate the expected loss for each building offering a more detailed description of the expected loss in near real-time. The forecasts take the form of 10000 synthetic catalogues, each one is simulated using as parameters a different sample from the joint posterior distribution. This allows a better estimation of future earthquake rates and their uncertainty at different magnitudes than simply forecasting the rate and its uncertainty directly.

We used the basic spatio-temporal ETAS model with a Gaussian isotropic spatial kernel for the time and spatial location of the events and the tapered GR law with a corner magnitude equal to 7 for the frequency-magnitude distribution. This magnitude distribution is the same as that described in Section 2.3.1, the choice of the corner magnitude is based on the expected maximum magnitude for the central Italy region (Petricca et al., 2019). For each forecast, we generate synthetic catalogues conditional on all the events that occurred before the forecasting date. For the project, only events with magnitude above 4 are retained, however, in this section, we show all the events with magnitude above a cutoff magnitude m_0 which is $m_0 = 2.5$ for the L'Aquila sequence and $m_0 = 3$ for the Amatrice sequence. This is done because these are the magnitude cutoffs chosen for the two sequences, which influence the estimates of all the parameters.

For each sequence, we issue a forecast at midnight of each day containing at least one event with a magnitude greater than 5, and a forecast one second after the event. This resulted in generating 12 forecasts for the L'Aquila sequence and 13 for the Amatrice sequence. The forecasting dates for the L'Aquila and Amatrice sequences are reported, respectively, in the first column of Table 6.4 and 6.5. The parameters of the ETAS model and the spatially varying background field are estimated on the same data we wish to forecast which makes the forecasts retrospective.

Table 6.4 and 6.5 report the observed number of events in the 24 hours starting from the reported dates in the first column along with the mean, the 2.5% quantile, the median, and the 97.5% quantile of the number of events provided by the forecasts. There is a noticeable difference between the mean and the median which implies that the distribution of the number of simulated events per catalogue is highly skewed and therefore not Poisson. As expected, the model fails to forecast large earthquakes underestimating the number of events before a large earthquake occurs as shown by the forecasts issued at midnight. The number of forecasted events is more similar to the observed one at the different date ranges and times shown in Table 6.4 and 6.5, but, the model underpredicts the number of events in most cases. This is expected because we selected only the days just after a large earthquakes which are usually difficult to forecast given that the model has to reflect the occurrence of earthquakes for longer periods of time and these selected days are quite different from the *average* day in the catalogue.

Figures 6.23 and 6.24 show the logarithm of the observed number of events and the logarithm of the mean, the 2.5% quantile, the median, and the 97.5% quantiles of the

Date	Obs	Mean	q0.025	Median	q0.975
2009-04-06 00:00:00	315	4.5420	0	2	26.000
2009-04-06 01:32:41	321	26.2727	4	21	81.000
2009-04-06 02:37:05	278	83.2304	9	63	275.000
2009-04-06 23:15:37	67	122.1694	13	95	393.000
2009-04-07 00:00:00	56	125.4003	13	96	397.050
2009-04-07 09:26:29	44	110.2385	10	85	372.000
2009-04-07 17:47:38	35	107.2333	8	76	365.000
2009-04-09 00:00:00	56	80.6339	3	55	313.000
2009-04-09 00:53:00	57	87.8700	2	58	332.025
2009-04-09 19:38:17	31	59.3091	1	31	275.000
2009-04-13 00:00:00	16	20.2517	0	2	143.025
2009-04-13 21:14:25	30	26.0996	0	10	159.025

Table 6.4: Number of observed events in the 24 hours after the reported date (columns one) along with the mean, the 2.5% quantile, the median, and the 97.5% quantile of the forecasted number of events for the day for the 2009 L'Aquila seismic sequence.

Date	Obs	Mean	q0.025	Median	q0.975
2016-08-24 00:00:00	125	1.686	0	0	0.000
2016-08-24 01:36:33	124	24.390	1	14	105.000
2016-08-24 02:33:29	92	76.853	2	37	368.000
2016-10-26 00:00:00	47	1.044	0	0	0.000
2016-10-26 17:10:37	79	17.824	0	9	78.025
2016-10-26 19:18:08	73	46.166	1	27	198.000
2016-10-30 00:00:00	277	72.952	0	0	447.000
2016-10-30 06:40:18	309	92.957	3	29	463.000
2017-01-18 00:00:00	84	0.717	0	0	0.000
2017-01-18 09:25:41	86	12.128	0	5	51.000
2017-01-18 10:14:10	82	26.052	1	14	123.000
2017-01-18 10:25:24	73	56.325	3	34	238.000
2017-01-18 13:33:37	36	88.427	3	48	402.000

Table 6.5: Number of observed events in the 24 hours after the reported date (columns one) along with the mean, the 2.5% quantile, the median, and the 97.5% quantile of the forecasted number of events for the day for the 2016 Amatrice seismic sequence.

forecasted number of events in the region highlighted in green in Figure 6.2. Red stars indicate events with a magnitude above or equal 5 that occurred in the 24 hours *before* the forecasting date. This is to show the influence of these events on the forecasts. The Figures show the ability of the model in providing seismicity patterns evolving with time and that these match the observed evolution of seismicity well (top-row). The time evolution and the influence of past events are particularly noticeable when inspecting the mean of the logarithm of the simulated number of events per pixel (second row). Furthermore, from a visual inspection, the median provides seismicity patterns resembling the observed ones, and the 97.5% quantiles usually cover the area where earthquakes occurred which is a sign that the model is able to capture some of the aspects of the spatial distribution of aftershocks.

The code used to generate the forecasts described in this section is available at https://github.com/edinburgh-seismicity-hub/spatio_temporal_ETAS_for_OEF. There, beyond the code relative to the L'Aquila and Amatrice seismic sequences, there is also an example of the 1997 Colfiorito earthquake which also occurred in central Italy. This example is relevant because in there I provide an animation [1] showing the temporal evolution of the probability of activity (i.e. probability of observing at least one earthquake with magnitude above 3) provided by the basic ETAS model used also for the forecasts reported here. I also report the spatial variation of the probability of activity for 20 weeks covering the period of the Colfiorito sequence. This is another nice example of the outputs that can be produced using our approach to estimate ETAS model parameters.

6.5 Discussion and conclusions

In this chapter, I have extended our methodology for the ETAS model depicted in Chapters 4 and 5 to the spatio-temporal case with spatially varying background rate and the possibility of introducing covariates. I compared the models only using the AIC because forecasts produced with this approach will be evaluated prospectively in the next Italy CSEP experiment. Furthermore, catalogue-based forecasts produced with the basic model will be used in a study about real-time loss forecasting.

The approach proposed in this chapter expands the type of covariates analyses proposed by Bayliss et al. (2020) and Bayliss et al. (2022), and briefly reported in Section 3.4, for time-independent models to time-dependent models. With the proposed framework, the covariates can be used in two ways: for the background rate and for the expected number of aftershocks. In the first case, it is possible to build an LGCP model with covariates and use it as spatial variation of the background rate once normalized to integrate to one. This offers a flexible approach to model the background rate and to incorporate many different sources of information. In principle, the output from other models can also be incorporated as a covariate. This ability increases the number of possible models that can be produced with this approach. For example, if the output of a time-dependent model as the Coulomb state-and-rate (Mancini et al., 2020) model is used as covariate, then the background rate would be spatio-temporally varying. In the same way, PSHA maps can be used as spatial variation of the background rate providing a way to incorporate that information into OEF models.

The proposed approach also allows the inclusion of covariates in modelling the expected number of aftershocks through a linear predictor. This provides a bridge between Hawkes process models and the Generalized Linear Models (GLMs) framework. Indeed, the coefficients associated with the covariates have the same interpretation as in GLMs and the same

[1] https://github.com/edinburgh-seismicity-hub/spatio_temporal_ETAS_for_OEF/blob/main/Colfiorito_Example/utilities/Fore_activity.gif

model selection techniques can be applied. With this approach, we find out that incorporating the depth of the events is beneficial in terms of AIC and that deeper earthquakes seem to generate fewer aftershocks according to parameter estimates. This is in accordance with the analysis of Chiodi et al. (2021) who also used the depth as a covariate. A limitation of this approach is that it assumes the depth to be known for each event, or fully determined by the location as the mean strike, as it is done in simple regression models. A better model would consider the depth as a random variable, as it is done for the magnitude. However, doing so brings a series of difficulties that are beyond the scope of this paper. For example, which depth distribution should be used? Is the depth distribution independent of space, time, and magnitude? These questions do not have a clear answer, however, the proposed approach can be extended to consider also the depth as random. This in fact can be done by considering a 1D grid on the depth domain as it is done with the time if the depth distribution is assumed to be independent of time and space. If one considers a depth distribution that depends on the location then the spatial grid described in Section 6.2.4 should be extended to be three-dimensional, and the computations can be carried out in the same way as described in this thesis.

I also used the mean strike of the nearest fault as a covariate which also provides better AIC than the basic model, however, the sign of the coefficient is uncertain. The uncertainty may be larger than what it should be due to the fact that we have associated each event with the nearest fault and we use a 2D representation of the fault. Using a 3D model to assign each event to one fault and maybe be beneficial. Also using a different estimator of the strike or different characteristics of the fault may provide deeper insights into the use of the fault information. Another interesting covariate is the material of the lithosphere where the event happened as shown in figure 6.6. In the same way, the use of temporally varying covariates such as displacement data can be incorporated. Furthermore, structured and unstructured random effects in the form of GMRF may be included in the model given the high level of efficiency shown by `inlabru` in dealing with such effects. The approach can be extended further by considering also other ETAS parameters as linear predictors which would enable studying the effect of covariates on other aspects of the earthquake-generation process and not only on the expected number of aftershocks.

An assumption of the model shown in this chapter that may limit the effect of the covariates is that we consider an isotropic kernel. Many studies have criticized this hypothesis which is particularly weak for large earthquakes (Hainzl et al., 2013; Grimm et al., 2022a). In fact, the aftershock region usually reflects the rupture area induced by an event, and using an isotropic spatial kernel means assuming that the rupture area is always circular with the event in the center. This assumption can be relaxed by considering a more flexible spatial kernel that accounts for the characteristics of the specific event. For example, we can consider a Gaussian kernel with a correlation coefficient equal to zero (isotropic) if an event is below a certain magnitude and different otherwise. Many other options are possible in this framework and we intend to further extend our approach to allow for anisotropic spatial kernels.

The basic ETAS model with Gaussian isotropic kernel and spatially varying background rate is used to produce forecasts that will be used in a collaborative study investigating the advantages of building-specific sensors when performing real-time loss forecasting. The forecasts are described in Section 6.4.6 which shows that the model in general underestimates the number of events for the forecasting periods considered in the experiment but is able to capture the spatial evolution of the number of earthquakes. This is encouraging because the model considered here is among the *most* simple spatio-temporal ETAS models we could consider and we can expect that models considering the information provided by available covariates or models including an anisotropic spatial kernel would improve the forecasts. The basic ETAS model considered in this chapter as well as models making use of available covariates would be submitted to the next Italian CSEP experiments and will be evaluated prospectively against future data.

In conclusion, we have developed a bayesian framework to study the effect of covariates on the number of aftershocks generated by an event in the spatio-temporal ETAS framework. The approach can be used to build more complex models of seismicity and to include in the model different sources of information. I plan to extend the approach to include models with random effects in the form of GMRF and to include anisotropic spatial kernels. I am working on including the spatio-temporal case in our `ETAS.inlabru` R-package, for the time being, the code used is available at https://github.com/edinburgh-seismicity-hub/spatio_temporal_ETAS_for_OEF.

This chapter concludes the part on modelling of seismicity with `inlabru`. The next chapter is dedicated to the study of a fundamental statistical property that scoring functions used to rank competing forecasts must have to provide trustworthy results. Being able to express a preference toward a model (or hypothesis) is fundamental to enhancing our ability to forecast future earthquakes. In fact, not only do we need methodologies facilitating the process of including different hypotheses in a model such as the one proposed in this chapter, but also reliable methods to validate these hypotheses against observed data and select the ones providing *better* forecasts. We will see in the next chapter that models can be ranked using scoring rules, and that different scoring rules apply different penalties and provide different rankings. Nevertheless, any scoring rule used for this task has to have some statistical properties to ensure that the validation is fair and the results trustworthy. One of the most fundamental of these properties is for a score to be *proper* and I focus on this property in the next chapter providing analytical and visual techniques to check if a score is proper or not.

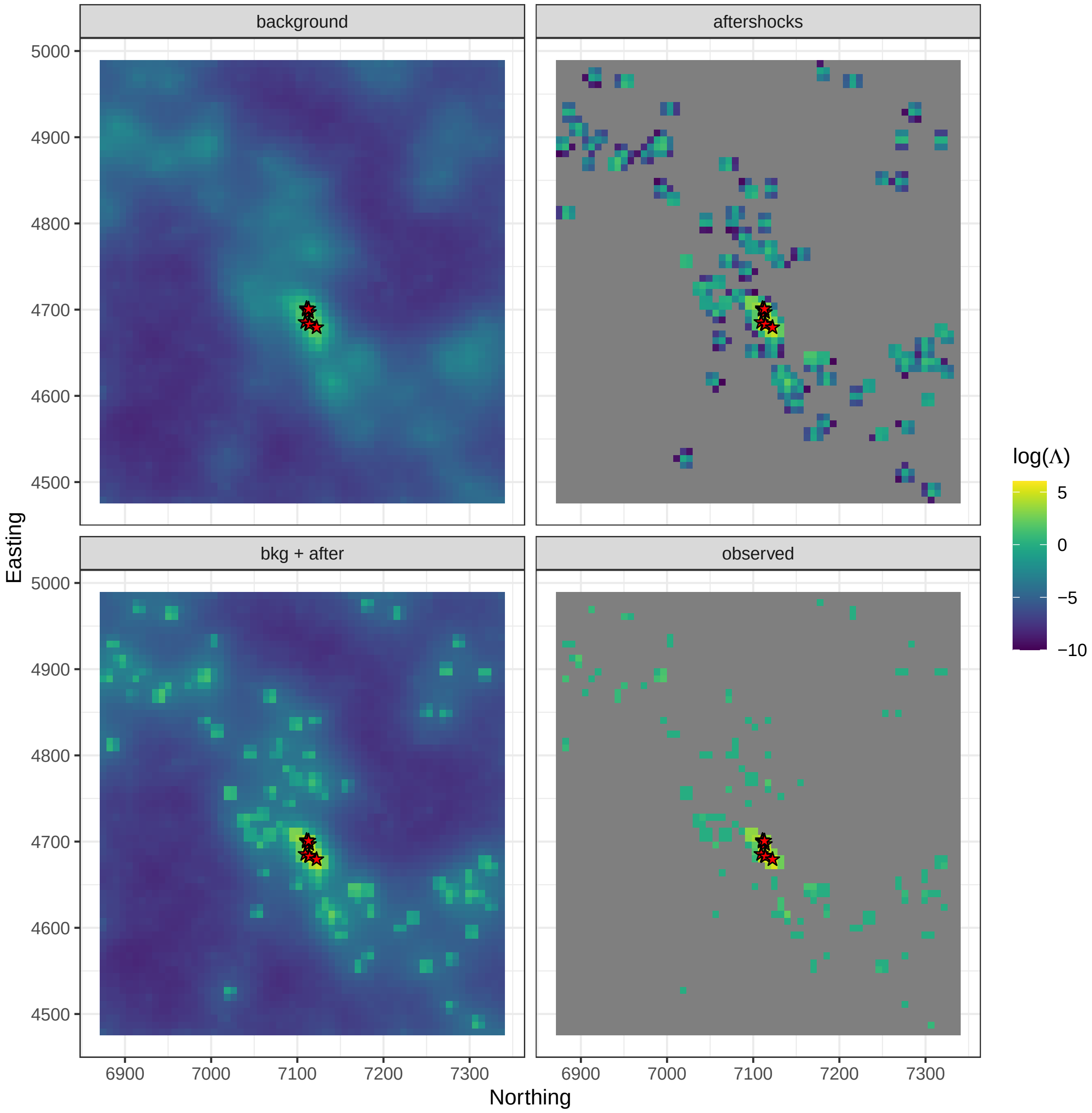


Figure 6.15: Map reporting for each pixel the posterior median of the logarithm of the expected number of background events (top-left), the expected number of aftershocks (top-right), the expected total number of events (bottom-right), and the observed total number of events (bottom-left) for the 2009 L'Aquila sequence obtained using the depth model. Grey pixels indicate values below −10, while red stars indicate the locations of events with magnitude above 5.

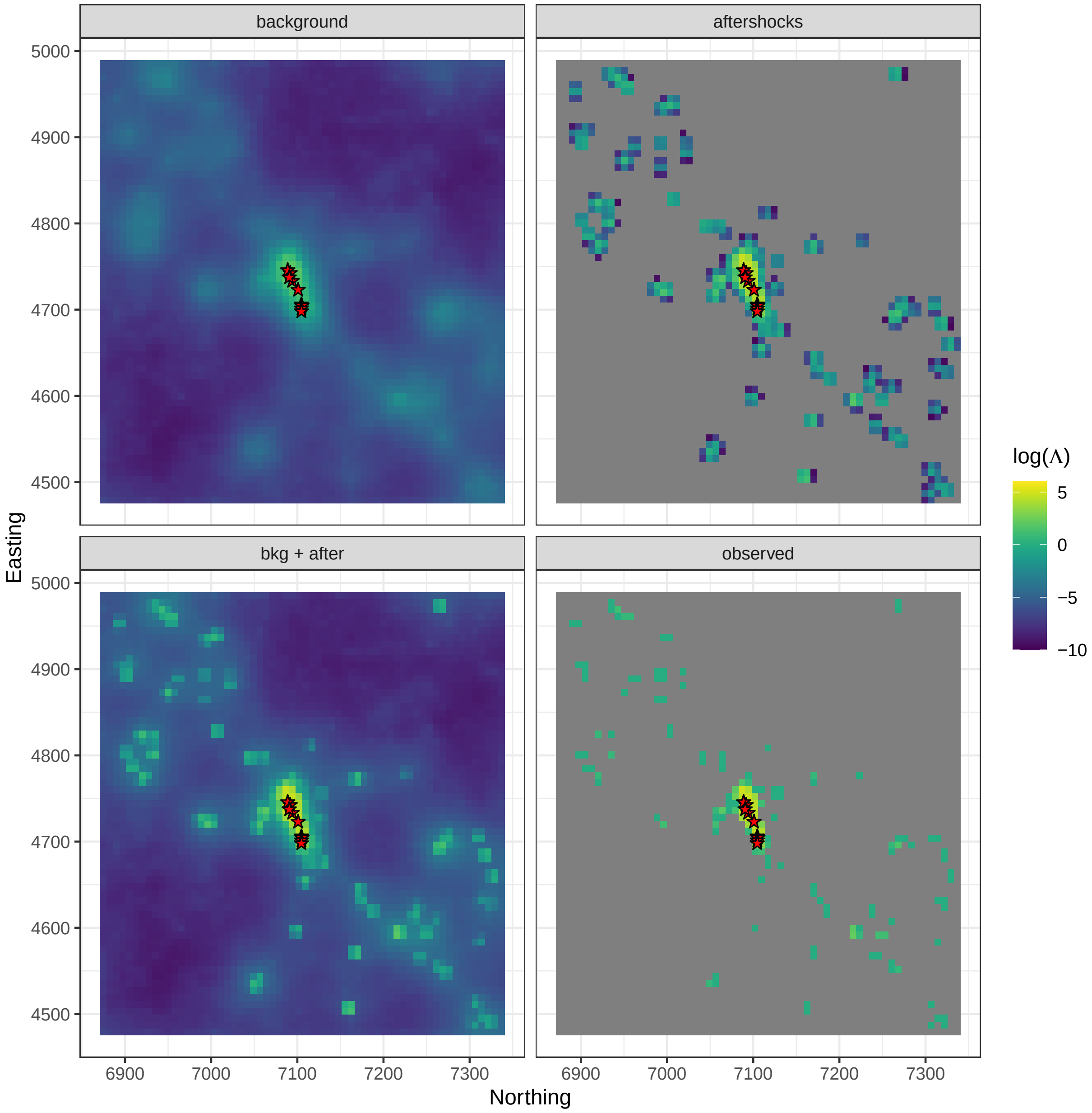


Figure 6.16: Map reporting for each pixel the posterior median of the logarithm of the expected number of background events (top-left), the expected number of aftershocks (top-right), the expected total number of events (bottom-right), and the observed total number of events (bottom-left) for the 2016 Amatrice sequence obtained using the depth model. Grey pixels indicate values below −10, while red stars indicate the locations of events with magnitude above 5.

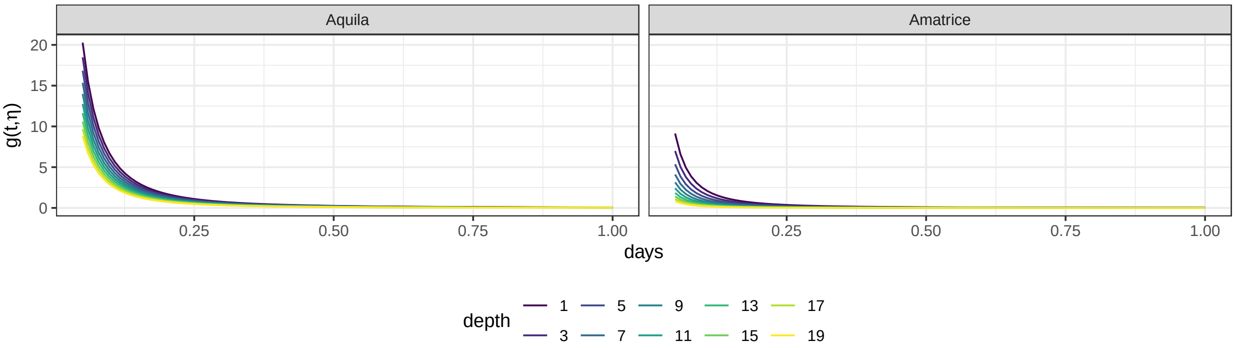


Figure 6.17: Median of the temporal intensity of aftershocks induced by an event of magnitude 6 for different levels of depth for the 2009 L'Aquila sequence (left) and the 2016 Amatrice sequence (right).

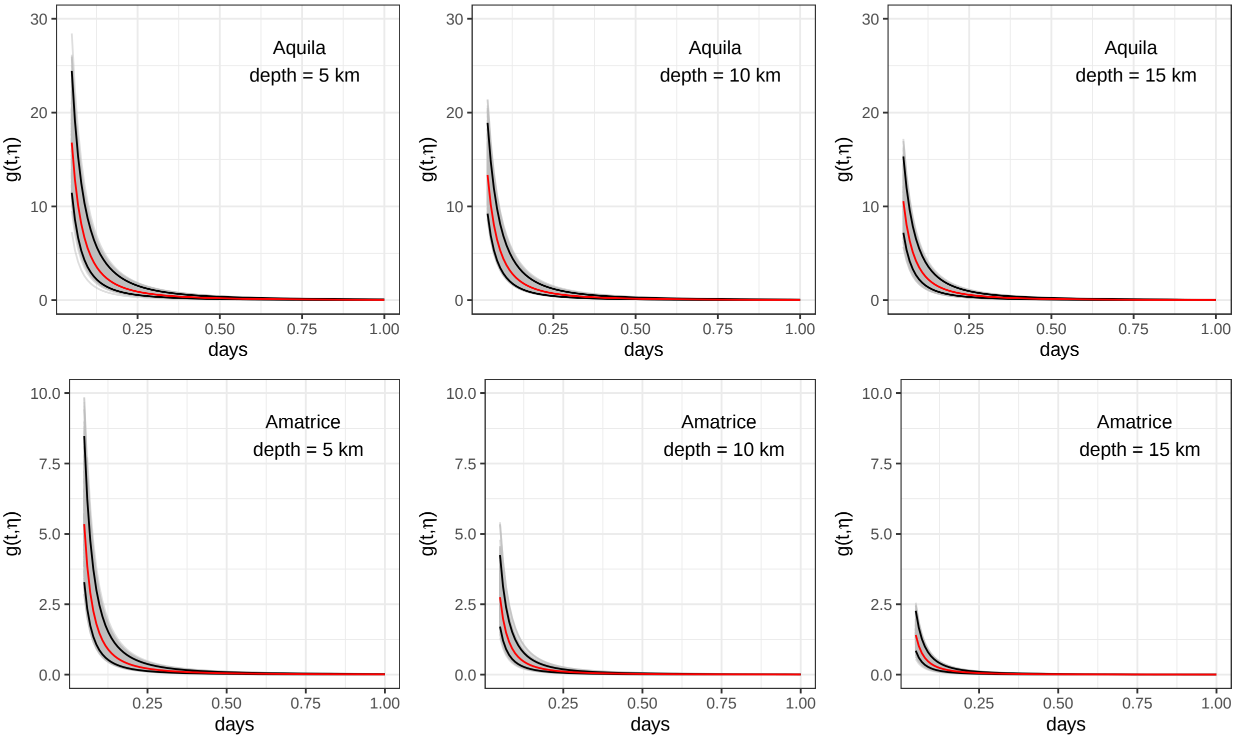


Figure 6.18: Temporal intensity of aftershocks induced by an event of magnitude 6 for the 2009 L'Aquila sequence (top-row) and the 2016 Amatrice seismic sequence (bottom-row). The red line represents the median of the posterior distribution of the temporal distribution of aftershocks, while the black lines represent. the 2.5% and the 97.5% quantiles of the posterior distribution. The grey lines represent a sample of 500 elements from the posterior. The plot show the posterior distribution for three levels of depth: 5 km (first column), 10 km (second column) and 15 km (third column).

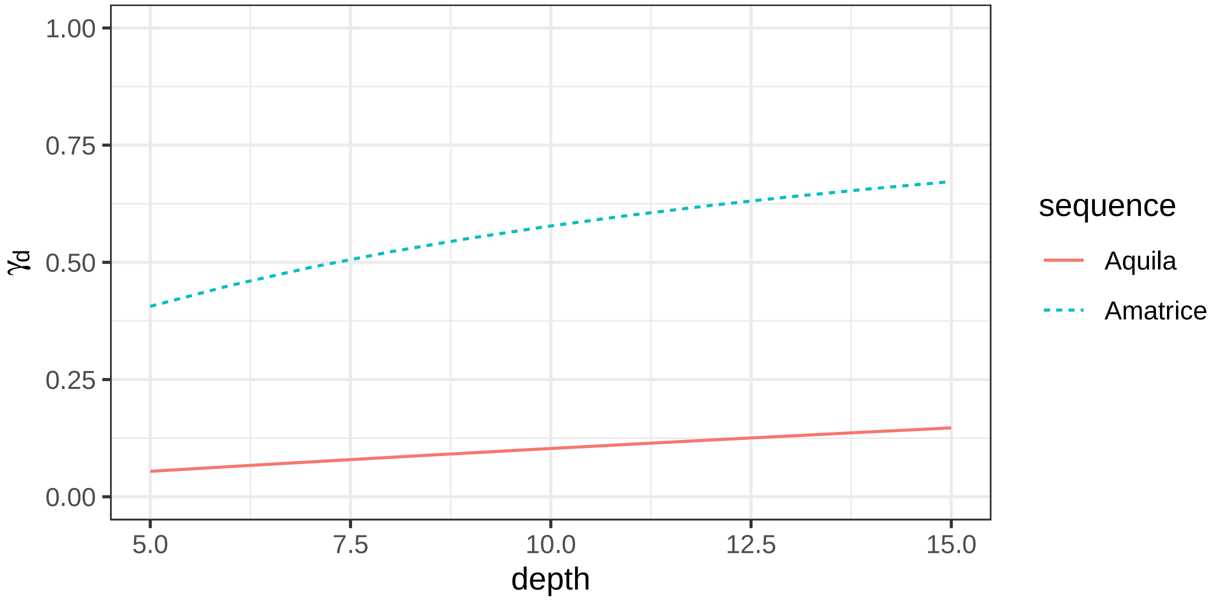


Figure 6.19: Influence of the depth on the linear predictor as a function of the depth, namely $\gamma_d(d)$, for the 2009 L'Aquila sequence (solid red) and the 2016 Amatrice sequence (dashed blue).

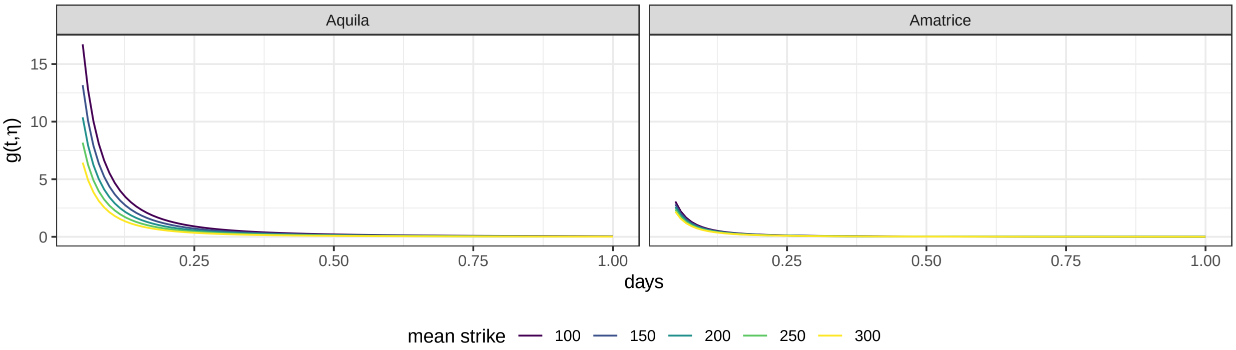


Figure 6.20: Median of the temporal intensity of aftershocks induced by an event of magnitude 6 for different levels of mean strike for the 2009 L'Aquila sequence (left) and the 2016 Amatrice sequence (right).

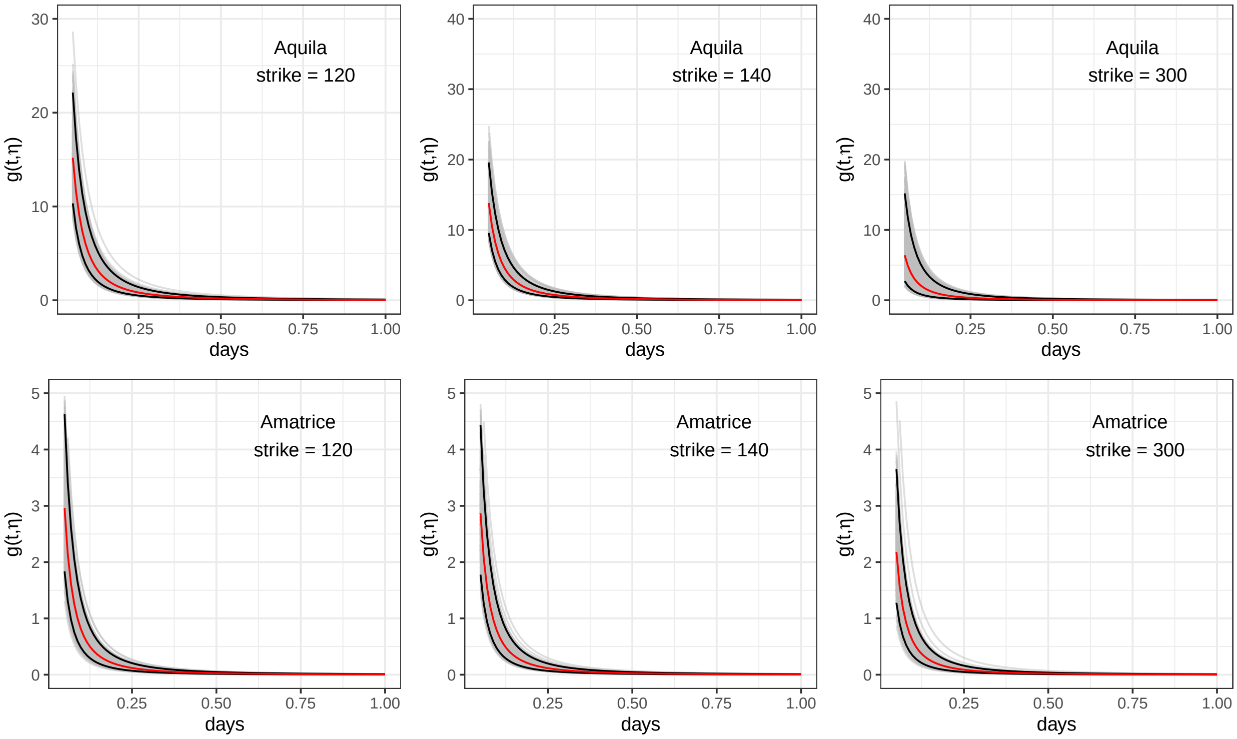


Figure 6.21: Temporal intensity of aftershocks induced by an event of magnitude 6 for the 2009 L'Aquila sequence (top-row) and the 2016 Amatrice seismic sequence (bottom-row). The red line represents the median of the posterior distribution of the temporal distribution of aftershocks, while the black lines represent. the 2.5% and the 97.5% quantiles of the posterior distribution. The grey lines represent a sample of 500 elements from the posterior. The plot show the posterior distribution for three levels of mean strike: 120 (first column), 140 (second column) and 300 (third column).

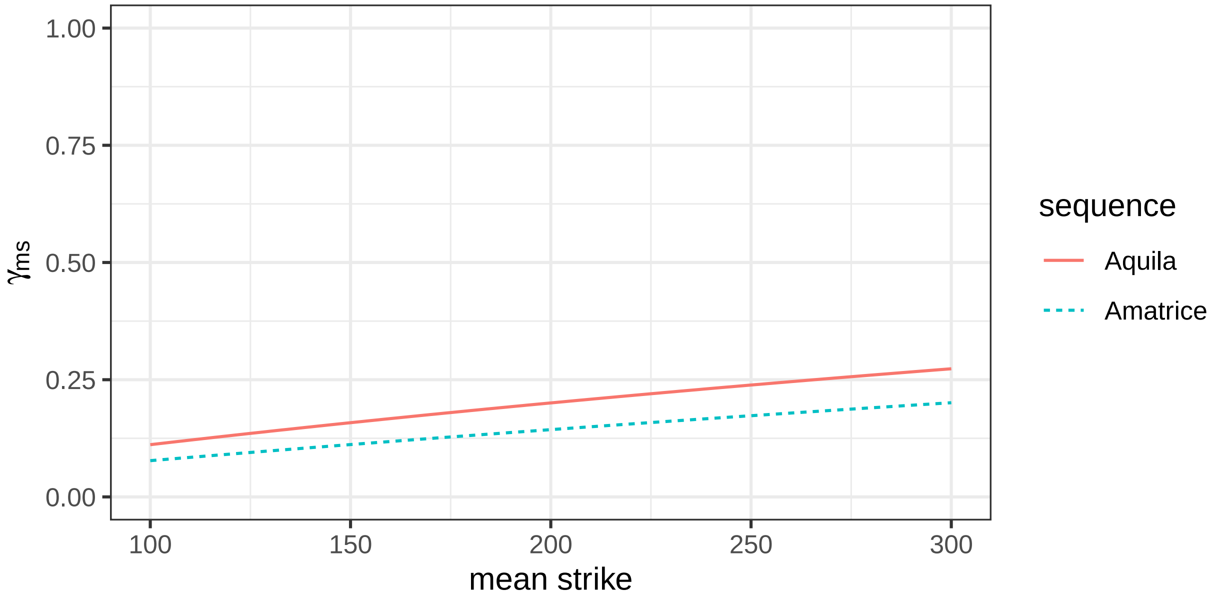


Figure 6.22: Influence of the mean strike on the linear predictor as a function of the mean strike, namely γ_{ms}(ms), for the 2009 L'Aquila sequence (solid red) and the 2016 Amatrice sequence (dashed blue).

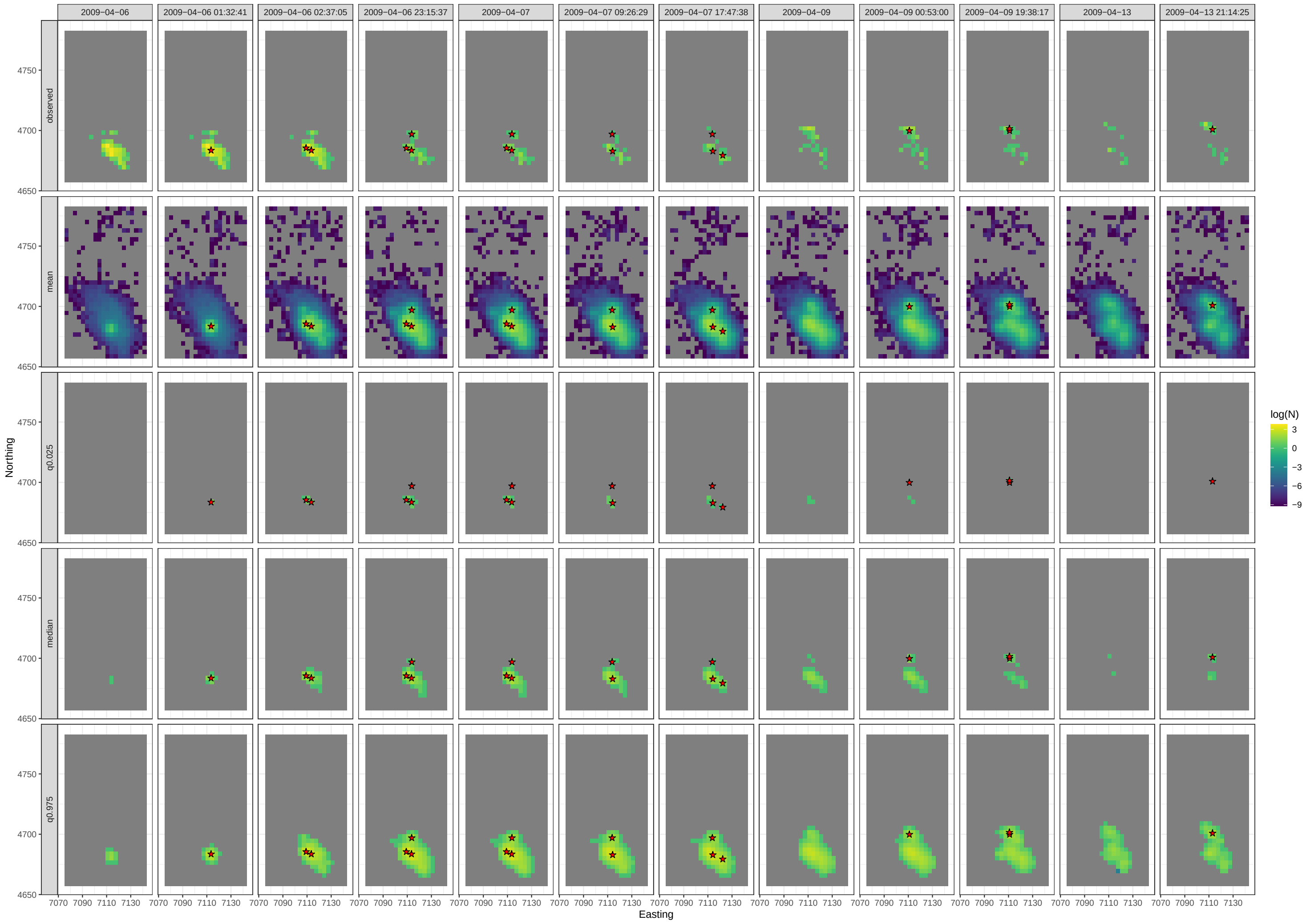


Figure 6.23: Spatial distribution of the logarithm of the observed number of events (top-row), and the logarithm of the mean, the 2.5% quantile, the median, and the 97.5% quantile of the forecasted number of events for the 2009 L'Aquila seismic sequence. The reported numbers are relative to the 24 hours after the dates reported in Table 6.4 and above each column. The red stars indicate events with a magnitude above 5 in the 24 hours *before* the forecasting date.

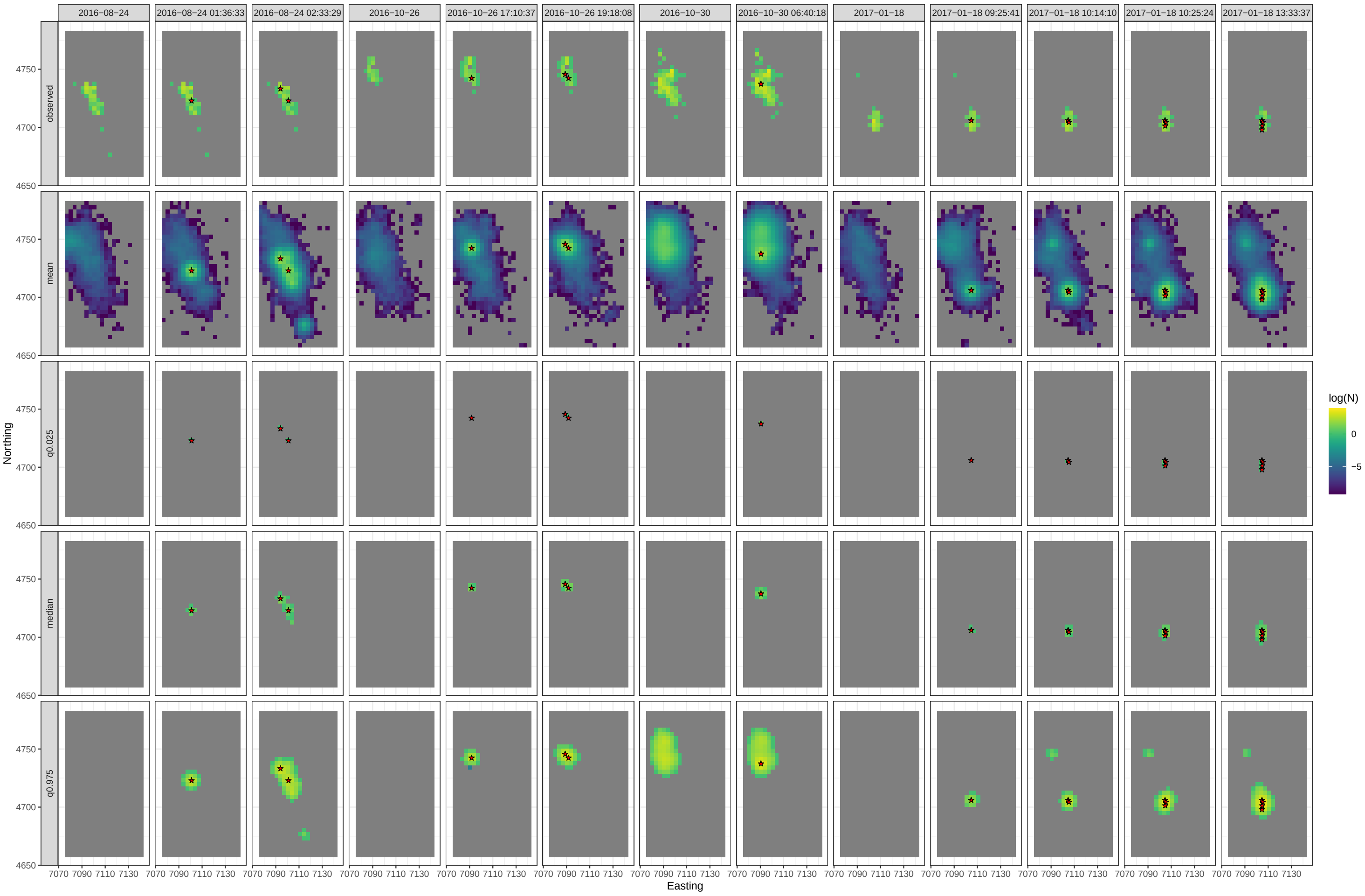


Figure 6.24: Spatial distribution of the logarithm of the observed number of events (top-row), and the logarithm of the mean, the 2.5% quantile, the median, and the 97.5% quantile of the forecasted number of events for the 2016 Amatrice seismic sequence. The reported numbers are relative to the 24 hours after the dates reported in Table 6.5 and above each column. The red stars indicate events with a magnitude above 5 in the 24 hours *before* the forecasting date.

Chapter 7

Ranking earthquake forecasts using proper scoring rules: Binary events in a low probability environment

7.1 Introduction

This chapter includes a paper published on the Geophysical Journal International (Serafini et al., 2022b) made in collaboration with the University of Bristol. The authors of the paper are Francesco Serafini, Finn Lindgren, Mark Naylor, Maximilian Werner, and Ian Main. As the first author, my contribution was to write the paper and the computer code needed for the analysis and lead the review process until the final stage of the publication process.

The only way to increase our knowledge of the earthquake generation process is to build models incorporating different hypotheses and test them against each observed data and against each other. In principle, if a model incorporating a hypothesis does not *work better* than its simpler version, we can reject the hypothesis. This can be done by running forecasting experiments as the ones organized by CSEP (see Section 2.5). Each competing model produces a forecast for a certain spatio-temporal region, then the models are ranked based on observed data. The ranks are usually obtained via positively (or negatively) oriented scores, where a score is simply a function of the forecast and the data. The higher (lower) the score the better the model that produced the forecast. Different scores provide different rankings based on different properties of the forecast. There can be scores accounting only for the spatial, or temporal variability of earthquake occurrences, on the magnitude distribution, or all of them together. Independently on the properties accounted by a score, for the ranking to be trustworthy, the score needs to have some statistical properties. One of the most important properties that a score needs to have is to be *proper*, where being proper means that, on average, the model *closer* to the data-generating model receives the higher (lower) score. Different scores employ different meanings of closer, which in turn determines the ability of a score to distinguish between forecasts. Improper scores may favour models far from the data-generating one, and we can not trust their rankings.

In this paper, the authors explore the notion of *properness* and the consequences of using an improper score. We take as an example the parimutuel gambling score and we prove that it is proper only when two forecasts are compared against each other, and it is improper in all other situations. We prove that analytically and visually using simulated data and comparing the rankings obtained with the parimutuel gambling score with two proper alternatives: the Brier and Logarithmic scores. Using simulated data also allows the retrieval of crucial information on the score performances and the ability to distinguish between models. We also show how to take into account uncertainty around the observed score value and design a three-options (preference for the first model, preference for the second model, not a preference) decision rule based on the confidence interval of the score difference. This allows us, with simulated data, to calculate the probability of expressing a preference for any given score, which can be seen as the power of a statistical test based on the score under analysis. I believe that the study of these probabilities is crucial to understand the capabilities of a score and can provide useful information on the experimental design (e.g. amount of data needed to have a probability of expressing a preference above a certain threshold, the number of bins to be considered if the space-time region is discretized, how the score behaves in a low probability environment). We illustrate this for probabilistic forecasts of binary events (e.g. probability of having more than one event with certain properties for each spatio-temporal bin), but the simulation technique can be easily generalized to other cases. This paper provides a further step in defining sanity checks for scoring rules, such that, if a score is not reliable in simulated simple settings, it is not safe to use it on real data.

7.2 The paper

7.2.1 Abstract

Operational earthquake forecasting for risk management and communication during seismic sequences depends on our ability to select an optimal forecasting model. To do this, we need to compare the performance of competing models in prospective experiments, and to rank their performance according to the outcome using a fair, reproducible, and reliable method, usually in a low-probability environment. The Collaboratory for the Study of Earthquake Predictability (CSEP) conducts prospective earthquake forecasting experiments around the globe. In this framework, it is crucial that the metrics employed to rank the competing forecasts are 'proper', meaning that, on average, they prefer the data generating model. We prove that the Parimutuel Gambling score, proposed, and in some cases applied, as a metric for comparing probabilistic seismicity forecasts, is in general 'improper'. In the special case where it is proper, we show it can still be used improperly. We demonstrate the conclusions both analytically and graphically providing a set of simulation based techniques that can be used to assess if a score is proper or not. They only require a data generating model and, at least two forecasts to be compared. We compare the Parimutuel Gambling score's performance with two commonly-used proper scores (the Brier and logarithmic scores) using confidence intervals to account for the uncertainty around the observed score difference. We suggest that using confidence intervals enables a rigorous approach to distinguish between the predictive skills of candidate forecasts, in addition to their rankings. Our analysis shows that the Parimutuel Gambling score is biased, and the direction of the bias depends on the forecasts taking part in the experiment. Our findings suggest the Parimutuel Gambling score should not be used to distinguishing between multiple competing forecasts, and for care to be taken in the case where only two are being compared.

7.2.2 Introduction

Probabilistic earthquake forecasts are used to estimate the spatial and/or temporal evolution of seismicity and have potential utility during earthquake sequences, including those following notable earthquakes. For example, they have been applied to forecast (pseudoprospectively) the seismicity that followed the Darfield earthquake and in turn led to the 2011 Christchurch earthquake (Rhoades et al., 2016), and to monitor induced seismicity at Groningen (Bourne et al., 2018). In Italy, earthquake probabilistic forecasts and ground-motion hazard forecasts are produced on a regular basis by the Instituto Nazionale di Geofisica e Vulcanologia (INGV) to inform the Italian government on the risk associated with natural hazard (Marzocchi et al., 2014). INGV is working to use probabilistic forecasts as a basis for modelling important quantities for operational loss forecasting such as the number of evacuated residents, the number of damaged infrastructure, the number of fatalities (Iervolino et al., 2015). A wider uptake requires further demonstrations of the operational utility of the forecasts, and in presence of multiple alternative models, a fair and rigorous method to express a preference for a specific approach is needed. The Collaboratory for the Study of Earthquake Predictability (CSEP, see Jordan 2006; Zechar et al. 2010b; Schorlemmer et al. 2018) is a global community initiative that seeks to make earthquake research more rigorous and open-science. This is done by comparing forecasts against future data in competition with those from other models through prospective testing in pre-defined testing regions. In this paper, we focus on comparing different forecasts that can be made from such competing models in the light of observed data.

In statistics, a common approach to compare probabilistic forecasts is the use of scoring rules (Gneiting and Raftery, 2007). Scoring rules have been widely applied in many fields of science to measure the quality of a forecasting model and to rank competing models based on their consistency with the observed data and the degree of uncertainty around the forecast itself. Much of the underlying methodology and concepts (such as what it means to be a "good" forecast) have been developed for weather forecasts (Murphy, 1993; Jolliffe and Stephenson, 2003). A positively oriented scoring rule, to be effective, has to be *proper*, which simply means that the highest score is achieved, on average, by the forecasting model "closer" to the distribution that has generated the observations. Various meaning of "closer" can be used depending on the context and the use that will be made of the forecasting model under evaluation, thus, a variety of proper scoring rules exists. Proper scoring rules are mathematically appealing for a range of different tasks: they can be used as utility function tailored to the problem at hand, they can be used as loss functions in parametric estimation problems and they can be used to rank competing models based on different aspects of the phenomenon under analysis (Rosen, 1996; Hyvärinen and Dayan, 2005; Hernández-Orallo et al., 2012).

CSEP aims to compare the predictive performance of diverse earthquake forecasts in a rigorous and reproducible fashion. The forecasts themselves are generated by underlying physical, stochastic or hybrid models using a variety of input data such as past seismicity, deformation rates, fault maps, etc (Field et al., 2014; Steacy et al., 2014; Bayliss et al., 2020). The two most widely used types are alarm-based forecasts and probabilistic forecasts. The first class of forecasts is usually expressed as a binary statement ("alarm" or "not alarm") based on the value of a precursory alarm function. In contrast, probabilistic forecasts, as intended in past CSEP experiments (Schorlemmer and Gerstenberger, 2007a), provide a distribution for the number of earthquakes. They can be expressed as grid-based forecasts (providing the expected number of events in each space-time-magnitude bin) or as catalogue-based (providing a number of simulated catalogues, Savran et al. 2020). The forecasts are variously compared using a suite of community-endorsed tests. Depending upon the forecasts at hand, three common challenges are the need for a reference model, how to handle bins (or regions) for which the forecaster didn't provide a forecast and, the need to specify a likelihood. The latter has been partially solved by the possibility of considering a pseudo-likelihood (Savran et al., 2020).

Molchan diagrams (Zechar and Jordan, 2008) and the area-skill score (Zechar and Jordan, 2010) do not need a likelihood and can be used to compare both alarm-based and probabilistic forecasts together. However, they need a reference model for assessing the significance of the results. This can be problematic because specifying a credible reference model is a difficult task (Stark, 1997; Luen et al., 2008; Marzocchi and Zechar, 2011). Likelihood-based tests (Schorlemmer et al., 2007a; Zechar et al., 2010a; Rhoades et al., 2011; Schneider et al., 2014) allow for pairwise comparison without the need of a reference model, but can only be applied to probabilistic forecasts. Further, methods for grid-based forecasts rely on the Poisson assumption, which has been observed to be not realistic (Werner and Sornette, 2008). Moreover, pairwise comparison may lead to paradoxical results like model A is preferred to model B which is preferred to model C which is preferred to model A (Zechar et al., 2013). Bayesian methods have been proposed (Marzocchi et al., 2012) but they also rely on the Poisson assumption. Catalogue-based forecasts, can be evaluated using a pseudo-likelihood approach (Savran et al., 2020) which does not rely on the Poisson assumption and enable information gains and likelihood ratios to be used. However, the latter are unbounded and sensitive to low-probability events, meaning that they can be unduly influenced by a few observations (Holliday et al., 2005; Zechar and Zhuang, 2014). Lastly, in past experiments such as the Regional Earthquake Likelihood Models (RELM) (Field, 2007), forecasters did not provide a forecast for all bins, some of them were left as missing value; for the methods outlined above, making a comparison is complex, given that considering only the overlapping space-time-magnitude volume may be too restrictive and introduce unfairness in the evaluation.

Zhuang (2010) and Zechar and Zhuang (2014) tried to overcome the difficulties outlined above by introducing the parimutuel gambling score, which provides a framework to evaluate different types of forecasts, with no need to explicitly specify a reference model or a likelihood, and with the ability to handle missing values in an intuitive way. This approach is based on the idea that alarm-based forecasts could be imagined as gamblers engaged in a game called the seismic roulette, where Nature controls the wheel (Main, 1997; Kossobokov, 2004, 2006). In this framework, the forecasters are the gamblers, a forecast consists of a collection of probabilities for observable events (bets) like *observing at least one earthquake in a specified space-time-magnitude bin*. Each bin represents a bet and the probability assigned by the forecaster represents the amount of money wagered. The observations consist of binary variables taking value 1 if the event occurs and zero otherwise. The forecaster gets a reward depending on the forecasted probability and the actual observation of an event or not. The forecasts are ranked based on their rewards. In this sense, the parimutuel gambling score is a positively oriented score (the higher, the better) for binary probabilistic forecasts. In this paper, we prove analytically and graphically that the parimutuel gambling score is not proper in general but only in a specific situation and we compare its performance with two proper alternatives: the Brier (Brier, 1950) and the logarithmic (Good, 1952) score.

The parimutuel gambling score has not been used systematically in CSEP, but it has been used to evaluate global forecasts (Taroni et al., 2016) and forecasts for Italy (Taroni et al., 2014, 2018) in situations where the score is improper. In the context of Italy, it has also been used in combination with two other scores to weight different source and ground motion models in the new Italian seismic hazard ensemble model (MPS19, Meletti et al. 2021). Furthermore, the parimutuel gambling score was mentioned by Schorlemmer et al. (2018) as a new method for evaluating earthquake forecasts without any warning about possible biases. We use the parimutuel gambling score to illustrate different techniques to assess if a score is proper, both analytically and graphically. We find that the parimutuel gambling score is proper only in a specific situation, and event then, it can be used improperly. This finding is emblematic of how much care should be taken in checking if (and when) a metric is proper.

To fairly compare the performance of the scores in a realistic framework, we use simulated data from a known model and we compare it with alternative models. In doing that, it is crucial to account for the uncertainty in the observed score difference. In fact, properness ensures that, at least on average, the scoring rule provides the correct ranking. However, the score calculated from any finite set of observations could be far from its average and, therefore, we need to account for uncertainty. In this paper, we show how to express a preference towards a model using confidence intervals for the expected score difference. This method introduces the possibility of not expressing a preference. Considering this outcome is potentially useful because it indicates that, for a scoring rule, the forecasts have similar performances, or the data are not enough to distinguish between models, or the bins' dimension is offset (too large or too small).

In summary, the main goal of this article is to present the notion of proper scoring rules for probabilistic forecasts of binary events: why is it crucial for a scoring rule to be proper? How can we verify if a score is proper or not? And, how do different scores penalize the same forecast differently? In Section 7.2.3 we define a proper and a strictly proper scoring rule, introduce the Brier and log scores as examples, and give a brief proof of their propriety. We also show how differently forecasts close to zero are penalized by the two scores. In Section 3, we introduce the parimutuel gambling score and analytically explore its improperness in the context of a forecast for a single bin. If a score is proper for single bins, then, the average score of different bins is also a proper score (Gneiting and Raftery, 2007). In Section 7.2.5, we generalise to the case where we have multiple bins but with the same probability, *Multiple Bins Single Probability*. This case is equivalent to considering the activity rate in each bin as independent and identically distributed. This is a significant assumption but allows us to calculate analytically the confidence intervals and the probability of expressing a preference for a given model. We generalise further to the case in which we have multiple bins but with a different probability for each bin, *Multiple Bins Multiple Probabilities*. In this case, we do not have analytical results and we are required to use approximate confidence intervals and simulations to calculate the probability of expressing a preference. These simulations are now close to a real forecast scenario. We illustrate this case using simulations from the time-independent 5 year adaptively-smoothed forecast for Italy (Werner et al., 2010). We choose this model because the adaptively-smoothed approach performed well across multiple metrics in the RELM experiment (Zechar et al., 2013) and, as a result, was incorporated into the California seismic hazard map produced by the third Uniform California Earthquake Rupture Forecast model (UCERF3, Field et al. 2014).

7.2.3 Proper Scores

Scoring rules quantify the quality of probabilistic forecasts, allowing them to be ranked. The quality depends on both the predictive distribution, produced by the model in true prospective mode, and on the subsequent observations. A scoring rule is a function of the forecast and the data measuring two factors: the consistency between predictions and observations and the sharpness of the prediction. Consistency assesses the calibration of the model, how well the forecast and the data agree, and is a joint property of the forecast and the data. Sharpness is a measure of the forecast uncertainty and is a property of the forecast only. Different scoring rules measure the consistency and the sharpness of a forecast differently. As in (Gneiting and Raftery, 2007), we call $S(P|x)$ the score for forecast P given the observation x. In general, we use capital letters for random variables, lowercase letters for scalar quantities such as realizations of a random variable (everything that is not random) and bold letters represents vectors. The only exception is N which represents the number of bins.

Thus, a scoring rule, given a forecast P, is a function of the observation only $S(P|\cdot) : \mathcal{X} \to [-\infty, \infty]$ where $\mathcal{X}$ is the set of all possible values of x. For consistency, we will use a *positively orientated* convention, where a larger score indicates a better forecast. Assuming that the observations are samples from a random variable $X \in \mathcal{X}$ with true distribution Q, the score $S(P|X)$ is a random variable itself, since it is a function of the random variable X. We define $S^E(P|Q)$ as the expected value of the scoring rule under the true distribution Q:

$$S^E(P|Q) = \mathbb{E}_Q[S(P|X)]. \tag{7.1}$$

A positively oriented scoring rule S is said to be *proper* if, for any forecast P and any true distribution Q, $S^E(Q|Q) \geq S^E(P|Q)$ holds. It is said to be *strictly proper* if $S^E(Q|Q) = S^E(P|Q)$ if and only if $P = Q$. Propriety is essential, as it incentivises the assessor to be objective and to use the forecast P "closer" to the true distribution Q. Different scoring rules rely on different meanings of closer. Also, proper scores can be used as loss functions in parameter estimation; in fact, since the likelihood assigned by a model to the observations can be seen as a proper scoring rule, the maximum likelihood estimator can be viewed as optimizing a score function (Huber, 1992). Investigating the ability of a score of distinguishing between different instances of the same model (with different parameters values) may bring insight regarding parameters identifiability.

Here, we are interested in scoring rules for binary variables, in which the variable X can be only 0 or 1, namely $X \in \{0, 1\}$. Grid-based earthquake forecasts divide the region of interest into regular space-time-magnitude bins (e.g. the spatial region is divided in bins of 0.1×0.1 degrees, the magnitude by 0.1 magnitude units, and the time is one 5-year bin), and the forecasters estimate the expected number of earthquakes per bin. In this case, for example, the binary variable might be 0 for empty bins and 1 if at least one event occurs. The forecasts may be ranked based on the average score across different bins (Zechar et al., 2013).

Considering a single bin, for grid-based binary forecasts, where both the forecast P and the true distribution Q are specified by just one number: the probability of X being 1. We call p the probability assigned to the event $X = 1$ by the forecaster, and p^* denotes the true probability. Thus, the expectation is given by

$$S^E(P|Q) = S^E(p|p^*) = p^* S(p|1) + (1 - p^*)S(p|0). \tag{7.2}$$

A scoring rule of this type is proper if, for any $p \in [0, 1]$ and any $p \in [0, 1]$, we have

$$S^E(Q|Q) \geq S^E(P|Q).$$

The properness of a score ensures that given two models p_1, p_2, the model with the greatest expected score $S^E(p_i|Q)$ is the closest to the true p^*. This notion can be generalized to rank a set of k forecasts $p_1, ..., p_k$ according to their expected scores.

Two of the most widely used strictly proper scoring rules, for binary data, are the Brier (or quadratic) score (Brier, 1950) and the logarithmic score (Good, 1952). These are good candidates for evaluating this class of earthquake forecasts. Here we give the definitions of these two scores, including brief proofs of their propriety.

Brier Score

The positively oriented Brier score (Brier, 1950) for a categorical variable X (the binary case is obtained considering only two possible outcomes) can be defined by:

$$S_B(P|x) = -\sum_{z \in \mathcal{X}} [p(z) - \mathbb{I}(z = x)]^2, \tag{7.3}$$

where $\mathcal{X}$ is the set of possible outcomes, $p(z)$ is the forecasted probability of the event $X = z$, and $\mathbb{I}(z = x)$ is an indicator function assuming value 1 if $z = x$ and 0 otherwise. This definition differs from the original only in the sign, since the original Brier score is negatively oriented.

The ordinary Brier score for binary events is the special case $\mathcal{X} = \{0, 1\}$, with $p = p(1)$ and $1 - p = p(0)$:

$$S_B(p|x) = -[(1-p) - (1-x)]^2 - (p-x)^2 = -2(p-x)^2 \quad = \begin{cases} -2(p-1)^2, & x = 1, \\ -2p^2, & x = 0, \end{cases}$$

which has expectation

$$S_B^E(p|p^*) = -2p^*(p-1)^2 - 2(1-p^*)p^2 \tag{7.4}$$

under the true event probability p^*. Taking the derivative with respect to p and imposing it equal zero, we find that the value $p = p^*$ uniquely maximizes the function $S_B^E(p|p^*)$ which proves that the Brier score is strictly proper.

Logarithmic Score

The logarithmic (log) score for binary event forecasts is defined as

$$S_L(P|x) = \ln p_P(x). \tag{7.5}$$

For $\mathcal{X} = \{0, 1\}$, the expectation is

$$S_L^E(p|p^*) = p^* \ln(p) + (1-p^*) \ln(1-p), \tag{7.6}$$

which, once differentiated with respect to p and set equal zero to identify the maximum, proves that also the log score is strictly proper.

Score Comparison

Given an observation x, to express a preference between two forecasts p_1 and p_2, an important quantity is the score difference Δ.

$$\Delta(p_1, p_2, x) = S(p_1|x) - S(p_2|x) = \begin{cases} S(p_1|0) - S(p_2|0) & \text{with prob} \quad 1 - p^*, \\ S(p_1|1) - S(p_2|1) & \text{with prob} \quad p^*. \end{cases}$$

For example, in the case of the Brier score we have

$$\Delta_B(p_1, p_2, x) \begin{cases} -2(p_1^2 - p_2^2) & \text{when} \quad x = 0, \\ -2[(1-p_1)^2 - (1-p_2)^2] & \text{when} \quad x = 1, \end{cases} \tag{7.7}$$

while in the case of the log score

$$\Delta_L(p_1, p_2, x) \begin{cases} \log(\frac{1-p_1}{1-p_2}) & \text{when} \quad x = 0, \\ \log(\frac{p_1}{p_2}) & \text{when} \quad x = 1. \end{cases} \tag{7.8}$$

In principle, if the expected value of Δ is positive we tend to prefer the first forecast, vice versa if it is negative. Considering the observation as a Bernoulli random variable $X \sim \text{Ber}(p^*)$, the difference $\Delta(p_1, p_2, X)$ is also a binary random variable, assuming the values

$\Delta_0 = \Delta(p_1, p_2, 0)$, $\Delta_1 = \Delta(p_1, p_2, 1)$ with probabilities $1 - p^*$ and p^*. The distribution of $\Delta(p_1, p_2, X)$ is therefore completely determined by the distribution of X:

$$\Delta(p_1, p_2, X) = X\Delta_1 + (1 - X)\Delta_0 = \Delta_0 + X(\Delta_1 - \Delta_0). \tag{7.9}$$

It follows that the expected value and variance of $\Delta(p_1, p_2, X)$ are determined by the properties of X:

$$\mathbb{E}[\Delta(p_1, p_2, X)] = \Delta_0 + \mathbb{E}[X](\Delta_1 - \Delta_0) = \Delta_0 + p^*(\Delta_1 - \Delta_0) \tag{7.10}$$

$$\mathbb{V}[\Delta(p_1, p_2, X)] = \mathbb{V}[X](\Delta_1 - \Delta_0)^2 = p^*(1 - p^*)(\Delta_1 - \Delta_0)^2 \tag{7.11}$$

We can give an alternative definition of the properness based on the random variable $\Delta(p_1, p_2, X)$. In fact, a scoring rule S is said to be proper if $\mathbb{E}_X[\Delta(p, p^*, X)] \leq 0$ when $p \neq p^*$, no forecast have an expected score higher than the data generating model p^*. However, they can achieve the same score. S is strictly proper if $\mathbb{E}_X[\Delta] = 0$ if and only if $p = p^*$, the highest score, on average, is achieved only by the data generating model. The definition implies, also, that proper scoring rules are invariant under linear transformations, in the sense that, a linear transformation of a proper score yields another proper score and the operation does not change the ranking.

Figure 7.1 reports the expected score difference between a candidate forecast p and the true value $p^* = 0.001$ using the Brier and the log score. The value $p^* = 0.001$ was chosen to be comparable to the estimated probability of having an event with magnitude greater than 5.5 calculated the days before the L'Aquila earthquake in the neighbourhood of where it struck (Fig. 4 in Marzocchi and Lombardi 2009). To enable a visual comparison, the expected Brier score values have been normalized to match the curvature of the log score when $p = p^*$. This is done by multiplying the expected Brier score values by the ratio of the second derivatives of the two expected scores calculated at $p = p^*$. The proper score scale invariance ensures that the ranking obtained using the original and normalized version of the Brier score is unchanged.

Both expected score differences are uniquely maximized at $p = p^*$ which means that the forecast matching the true probability has the highest expected score. This is an easy way to assess if a scoring rule for binary outcomes is proper or not. Furthermore, Figure 7.1 offers an example of how different scores penalize differently the same forecasts. The log score is asymmetric and takes into account the relative differences between the forecasts (equation 7.8), and if $p^* \neq 0$ the expected score for $p = \{0, 1\}$ is $-\infty$. The log score is analogous to a likelihood score, and brings the same properties: a model which is correct in all the bins but one for which it provides zero probability will have the worst possible score. The Brier score, instead, considers the absolute difference between forecasts (equation 7.7) resulting in a symmetric distribution For example, using the Brier score, a forecast $p = 0$ will be preferred to any forecast in $(2p^*, 1)$, for any $p^* < 1/2$.

The choice of score, and consequently the style of penalty, should reflect the task at hand. Predicting $p = 0, 1$ means that we are absolutely certain about the outcome of X. If the forecasts under evaluation are planned to be used in an alarm based system, for which an alarm is broadcasted if the probability is above or below a certain threshold, being overconfident may put lives at risk and perhaps the log score would be the right choice in this situation. On the other hand, the Brier score may be suitable when such a strict penalty is not desirable (e.g., to calculate the weights of an ensemble model as done by Taroni et al. 2018 and Meletti et al. 2021). This example illustrates the flexibility of proper scores and how important it is to choose the right one depending on the purposes of the forecast under evaluation.

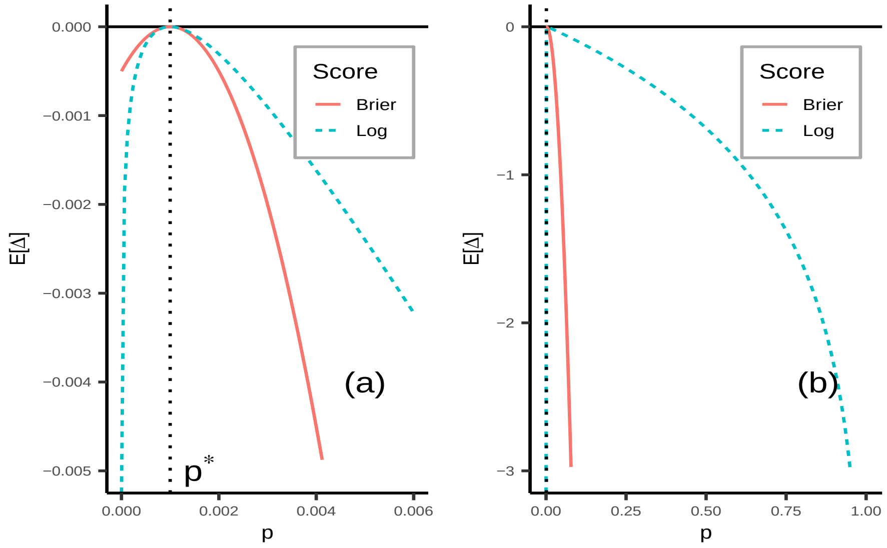


Figure 7.1: Differences in the scores expected value for a generic value of the forecast p and the optimal forecast $p = p^*$, namely $\mathbb{E}[\Delta] = S^E(p|p^*) - S^E(p^*|p^*)$, in the case $p^* = 0.001$. Panel (a) $p \in (0, 0.006]$, Panel (b) $p \in (0, 1)$. The expected Brier score have been normalised to match the curvature of the log score when $p = p^*$.

7.2.4 Improper scores

Scores which are not proper are called improper. Being improper means that a model may exist with expected score greater than the data generating model. In the specific case of probabilistic forecasts for binary events, a score is improper if it is biased towards models which systematically under/overestimate the true probability p^*. In the context of earthquake forecasting experiments we do not know the true value of p^*. Therefore, it is crucial to use proper scoring rules for which we are sure that, at least on average, they will prefer the closest model to the data generating one. Improper scoring rules do not have this property, which implies that the smallest or the largest (or any other) forecast, on average, could achieve the highest score. This is in clear contrast with the aim of any forecasting experiment. Below, we demonstrate that the parimutuel gambling score (Zhuang, 2010; Zechar and Zhuang, 2014) is an example of a scoring rule which is proper only in a specific situation and not in general.

Definition of the parimutuel gambling score

The parimutuel gambling score was designed to rank forecasting models for binary events and was applied to rank earthquake forecasting models in CSEP experiments (Taroni et al., 2018; Zechar and Zhuang, 2010). Initially, it was used to compare models against a reference model (Zhuang, 2010), which is improper. Later, it was generalized to compare models against each other simultaneously (Zechar and Zhuang, 2014), the case with only two players is the special case for which the score is proper, all the others are not. The score is based on a gambling scheme in which the forecasting models play the role of the gamblers and, for each observation, they obtain a reward proportional to the probability assigned by the gambler to the event occurring. In particular, it is a zero-sum game, in the sense that bids and rewards in each bin sum to zero, which makes the parimutuel gambling score relative to one forecast dependent on the other forecasts.

In contract to the Brier and log scores, it is not possible to define the parimutuel gambling score using the form $S(p|x)$ because it needs at least two forecasts to be evaluated and is a function of them all. Given a set of k forecasts $\mathbf{p} = (p_1, ..., p_k)$, we define $S_G(\mathbf{p}|x)$ as the vector such that the i-th component, $S_{G,i}(\mathbf{p}|x)$, is given by the parimutuel gambling score of the i-th forecast, given x has been observed. In the case of the Brier and log score the components of the vector $S(\mathbf{p}|x)$ are defined independently, in the case of the parimutuel gambling score they have to be defined jointly. Let $\bar{p}$ be the average probability involved in the gambling scheme, namely $\bar{p} = \sum_{i=1}^{k} p_i/k$. The parimutuel gambling score relative to the i-th forecast is defined as

$$S_{G,i}(\mathbf{p}|x) = \begin{cases} \frac{p_i}{\bar{p}} - 1, & x = 1, \\ \frac{1-p_i}{1-\bar{p}} - 1, & x = 0. \end{cases}$$

The above expression is a zero-sum game, meaning that $\sum_i S_{G,i}(\mathbf{p}|x) = 0$, therefore the rewards may be positive or negative. Each gambler obtains a positive reward if and only if they assign a greater probability to the observed event than the average gambler involved in the game. Vice versa, the reward is negative if the probability is smaller.

The expected value with respect the true probability p^* is given by

$$\begin{aligned} S^E_{G,i}(\mathbf{p}|p^*) &= p^* \left(\frac{p_i}{\bar{p}} - 1 \right) + (1 - p^*) \left(\frac{1 - p_i}{1 - \bar{p}} - 1 \right), \\ &= \frac{p^* p_i}{\bar{p}} + \frac{(1 - p^*)(1 - p_i)}{1 - \bar{p}} - 1, \\ &= \frac{(p_i - \bar{p})(p^* - \bar{p})}{\bar{p}(1 - \bar{p})}. \end{aligned} \tag{7.12}$$

Equation (7.12) is the same as equation (5) in (Zechar and Zhuang, 2014). The denominator involves all the probabilities in the game which demonstrates the interdependence with all other forecasts and complicates the study of the derivatives. However, it is still possible to prove that the gambling score is strictly proper when $k = 2$. In this case, $\mathbf{p} = (p_1, p_2)$, and

$$\begin{aligned} 4\bar{p}(1 - \bar{p})S^E_{G,1}(\mathbf{p}|p^*) &= 4 \left(p_1 - \frac{p_1 + p_2}{2} \right) \left(p^* - \frac{p_1 + p_2}{2} \right), \\ &= (p_1 - p_2)(2p^* - p_1 - p_2), \\ &= -[(p_1 - p^*) - (p_2 - p^*)][(p_1 - p^*) + (p_2 - p^*)], \\ &= (p_2 - p^*)^2 - (p_1 - p^*)^2. \end{aligned}$$

The expected reward of the first modeler is non-negative when $|p_1 - p^*| \leq |p_2 - p^*|$, implying that p_1 is favoured over p_2 if it is closer to the true probability p^*. In fact, if $p_2 = p^*$ then, $S^E_{G,1}(\mathbf{p}|p^*) \leq 0$, with the equality verified only for $p_1 = p^*$. Furthermore, the expected gambling score in this case is proportional to the expectation of the corresponding Brier score differences $\Delta_B = S^E_B(p_1|p^*) - S^E_B(p_2|p^*)$, thus, they produce, on average, the same rankings.

Improper use of proper score

When comparing forecasting models, ensuring that the score is proper may not be sufficient. It also has to be used properly. The gambling score with $k = 2$ offers a nice example of this situation. We have demonstrated that the parimutuel gambling score is proper when $k = 2$, however, the dependence of the score value on all the forecasts involved in the comparison is a source of bias. In fact, $S^E_{G,1}(\mathbf{p}|p^*) \geq S^E_{G,2}(\mathbf{p}|p^*)$ when p_1 is closer to p^* than p_2, however,

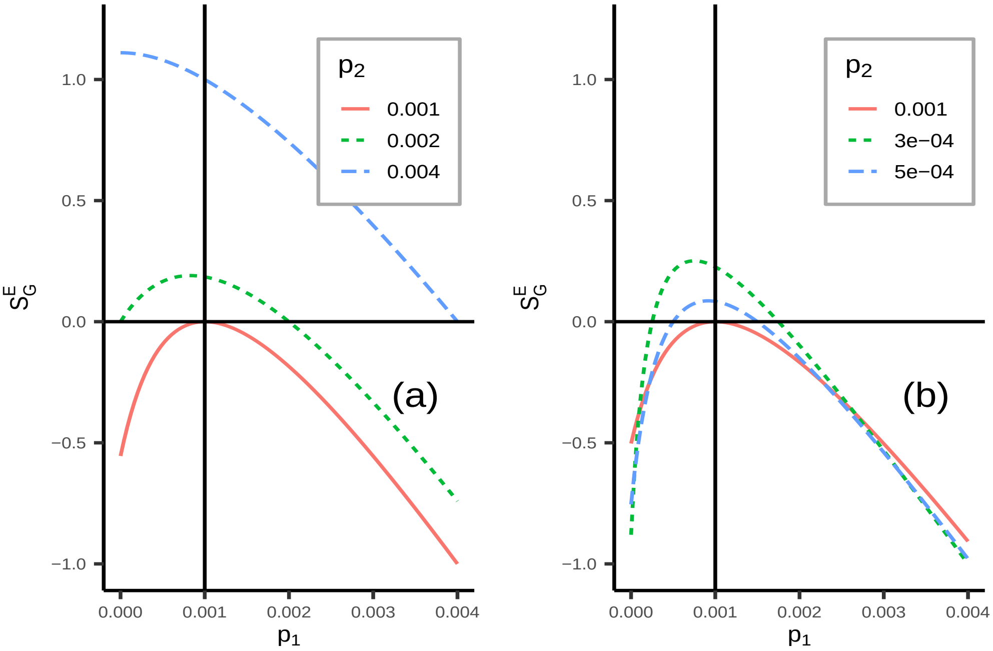


Figure 7.2: Expected value of the parimutuel gambling score (k = 2), $S^E_{G,1}(\mathbf{p}|p^*)$, varying $p_1 \in (0, 0.004)$, $p_2 = \{p^*, 2p^*, 4p^*\}$ (a) and $p_2 = \{p^*, p^*/2, p^*/4\}$(b). The solid vertical line represents the true probability $p^* = 0.001$. The expected scores have been normalized so that their minimum is equal to -1.

$p_1 = p^*$ does not maximize $S^E_{G,i}(\mathbf{p}|p^*)$ as shown in Figure 7.2. This means that the score becomes biased when we rank forecasts based on the score difference against a reference model.

Formally, we are considering pairwise vectors $\mathbf{p}_1 = (p_1, p_0)$, $\mathbf{p}_2 = (p_2, p_0)$, etc., where p_0 is the reference model. For each of these we can estimate pairwise comparison score vectors $S_G(\mathbf{p}_1|x)$, $S_G(\mathbf{p}_2|x)$, and so on. The first component of each vector, namely $S_{G,1}(\mathbf{p}_1|x)$, $S_{G,1}(\mathbf{p}_2|x)$, etc, represents the score of p_1 and, respectively, p_2 against the reference model p_0. At this point, one would be tempted to rank the models based on $S_{G,1}(\mathbf{p}_1|x)$ and $S_{G,1}(\mathbf{p}_2|x)$, and this is the approach taken in Taroni et al. (2014) in which the official national time-independent model (Gruppo di Lavoro, 2004b) is used as reference model.

If the parimutuel gambling score is used to rank forecasts based on the score difference relative to a reference model, it will *not* reliably favour the model closest to the true one, and the size of the bias will depend on the choice of the reference model. For example, in Figure 7.2a, for $p_2 = 0.004$, the gambling score is maximized at $p_1 = 0$. This means that if the reference model is $p_0 = 0.004$, the overconfident forecast $p_1 = 0$ would be favoured by the ranking even if another forecast is perfect, e.g. $p_3 = p^*$. This problem can be particularly relevant in operational seismology where it is common for candidate forecasts to be compared against a reference model which is known to be based on simplistic assumptions (for example a homogeneous Poisson process).

Hereon, the term pairwise gambling score refers to the comparison against a reference model as described in this section, while the term full gambling score will refer to the case where the forecasts compete directly against each other as we describe in the next section. Using this terminology, the full gambling score with $k = 2$ is the only proper score.

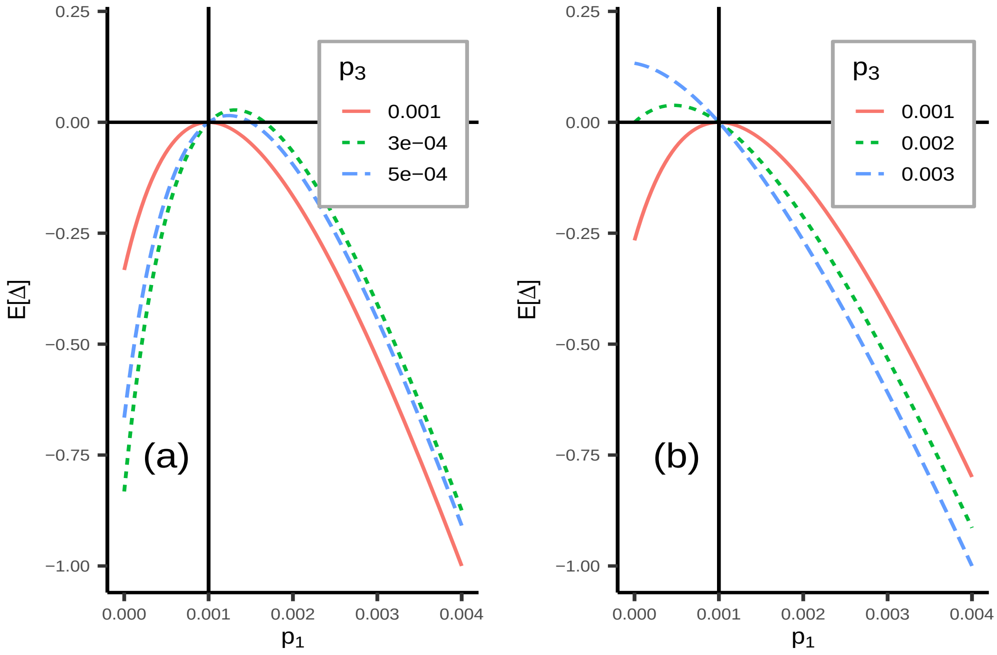


Figure 7.3: Expected gambling score differences (k = 3) between p_1 and p_2, $S^E_{G,1}(\mathbf{p}|p^*) - S^E_{G,2}(\mathbf{p}|p^*)$, as a function of $p_1 \in (0, 0.004)$, $p_2 = p^* = 0.001$ (vertical line), and $p_3 \in \{p^*, p^*/2, p^*/3\}$ (a) and $p_3 \in \{p^*, 2p^*, 3p^*\}$ (b) .

Improperness of the multi-forecast gambling score for $k \geq 3$

The generalized version of the full parimutuel gambling score, as presented in Zechar and Zhuang (2014), for $k \geq 3$ is improper. For example, when $k = 3$ and $p_2 = p^*$, following equation 7.12 the difference between the expected score for p_1 and p_2 is given by

$$\begin{aligned} 3\bar{p}(1-\bar{p})[S^E_{G,1}(p_1, p^*, p_3|p^*) - S^E_{G,2}(p_1, p^*, p_3|p^*)] &= 3(p_1 - p^*)(p^* - \bar{p}) \\ &= (p_1 - p^*)(2p^* - p_1 - p_3), \end{aligned}$$

with both sides scaled by the common factor $3\bar{p}(1-\bar{p})$. This means that when $2p^* - p_3 \geq p_1$, the forecast p_1 will have a positive score for any $p_1 \geq p^*$. Any value of $p_1 \in [p^*, 2p^* - p_3]$ will be preferred to p_2 that is equal to p^*. When $2p^* - p_3 \leq p_1$, with the same reasoning, p_1 is preferred over $p_2 = p^*$ in the interval $[2p^* - p_3, p^*]$.

In Figure 7.3 we consider $k = 3$, $p^* = p_2 = 0.001$ and report the difference between the expected scores of p_1 and p_2, namely $S^E_{G,1}(\mathbf{p}|p^*) - S^E_{G,2}(\mathbf{p}|p^*)$, for different values of p_3. The expected score difference is not maximize at $p_1 = p^*$, which means that the score is biased, and the "direction" of the bias depends on p_3 being greater than or equal to p^*.

Consider $k > 3$ gamblers who propose probabilities $\mathbf{p} = \{p_1, ..., p_k\}$. It is helpful to consider the vector of probabilities that excludes the first component; we name this $\mathbf{p}_{-1} = \mathbf{p}/\{p_1\}$ and its mean $\bar{p}_{-1}$. Assuming $p_2 = p^*$, thus $\mathbf{p} = (p_1, p^*, ..., p_k)$, we have that

$$\begin{aligned} k\bar{p}(1-\bar{p})[S^E_{G,1}(\mathbf{p}|p^*) - S^E_{G,2}(\mathbf{p}|p^*)] &= k(p_1 - p^*)(p^* - \bar{p}), \\ &= (p_1 - p^*)[kp^* - p_1 - (k-1)\bar{p}_{-1}], \end{aligned}$$

from which we conclude that the expected score difference is positive when $p_1 \in [kp^* - (k-1)\bar{p}_{-1}, p^*]$ or $p_1 \in [p^*, kp^* - (k-1)\bar{p}_{-1}]$, depending on if $p^* \gtrless \bar{p}_{-1}$. Specifically, when

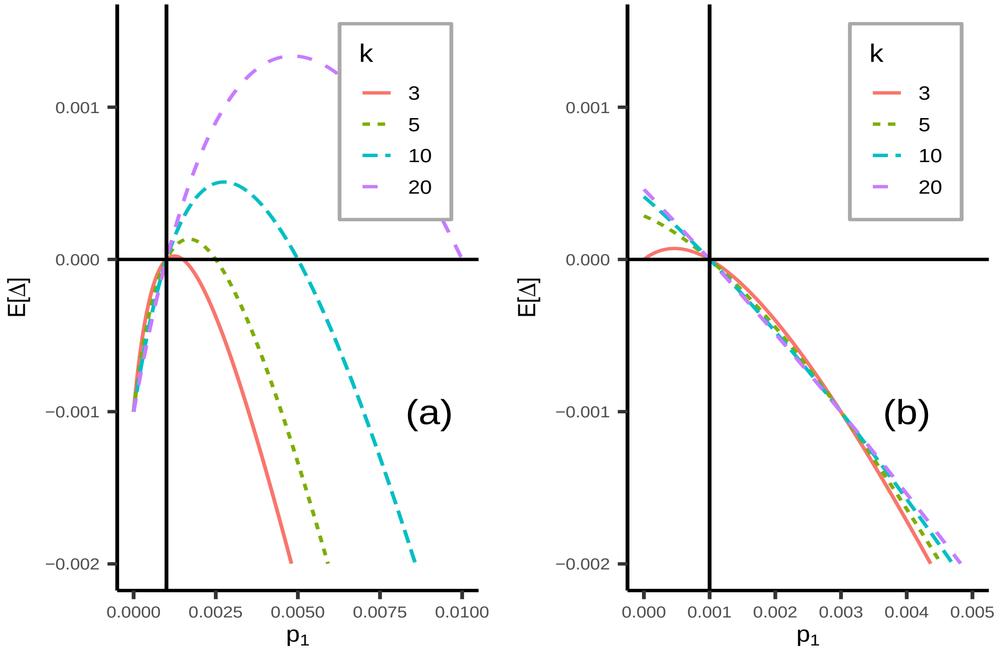


Figure 7.4: Expected gambling score differences ($k \in \{3, 5, 10, 20\}$) between p_1 and p_2 as a function of p_1 considering $p_2 = p^* = 0.001$ (vertical line). The averageof the forecast probabilities (excluding the first forecast) is the constant $\bar{p}_{-1} = p^*/2$ (a) and $\bar{p}_{-1} = 2p^*$ (b).

$p^* > \bar{p}_{-1}$ (Figure 7.3a) the first gambler is encouraged to bet on a $p_1 > p^*$ and vice versa when $p^* < \bar{p}_{-1}$ (Figure 7.3b). Furthermore, p^* is always an extreme of the interval where the expected score difference is positive. Considering $\bar{p}_{-1}$ as fixed, the length of the interval is an increasing function of the number of gamblers k (Figure 7.4) which means that the size of the set of forecasts capable of obtaining a score value higher than the data generating model is an increasing function of the number of forecasts involved in the comparison. It is clear that the multi-forecast parimutuel gambling score favours models that are contrary to the average of the other forecasts. This could be particularly dangerous when evaluating the performance of earthquake forecasting models. For example, the trigger for an alarm being broadcast (or not) is often defined when the probability of having an earthquake above a certain magnitude exceeds a specified threshold. Using a model chosen looking at the full parimutuel gambling score could therefore lead to broadcasting alarms when they are not needed ($p \gg p^*$, 'crying wolf', Figure 7.4a) or not broadcasting an alarm when needed ($p \ll p^*$, providing 'false reassurance', Figure 7.4b).

The root of the problems with this score is that the score, relative to a candidate forecast, explicitly depends on the other forecasts. This design brings two problems: (i) the score, even in the special case when it is proper, can be used improperly and (ii) the score is never proper when considering more than two models. The Brier and log scores do not suffer from the same problem since the score of a forecast depends only on the forecast and the observation. Furthermore, the improperness demonstrated here can be expressed in terms that show that the gambling metaphor is part of the problem: If the outcome $x = 0$ is likely (e.g. $p^* = 0.001$) and the majority of the forecasts have too large probabilities, then the expected gain is higher for an overconfident forecast, $p \ll p^*$, since that will give the forecaster a larger share of the total payout.

7.2.5 Forecasting across multiple bins

Until now, we have analysed the expected score for a single bin, here we analyse the ability to express a preference between two forecasts using the average score across multiple bins. We assume to have access only to one observation $x_i \in \{0, 1\}$ per bin . We analyse extensively the case in which the probability of observing $x_i = 1$, is the same for each bin, $p_i^* = p^*$ for any $i = 1, ..., N$. We refer to this as the Multiple Bins Single Probability case; the only quantity of interest is p^* and a forecast is represented by a single value p. Even though, the Multiple Bins Single Probability case is clearly unrealistic in practice, it builds the basic concepts we will then use to explore the Multiple Bins Multiple Probabilities case where the probability of observing $x_i = 1$ is potentially different for each bin.

Considering multiple bins, we observe a realization of the random variable $X_i \sim \text{Ber}(p_i^*)$ for $i = 1, ..., N$. A forecast is given by the vector $\mathbf{p} = (p_1, ..., p_N)$ specifying the probability of $X_i = 1$ for each bin. Following the terminology in the literature regarding Bernoulli random variables, the event $X_i = 1$ is referred to as a *success*. The quantity $X_S = \sum_i X_i$ is therefore referred as the sum of the observations or the number of successes or the number of active bins.

Given an arbitrary scoring rule $S(p|X)$, the average score associated with the forecast $\mathbf{p}$ is given by:

$$S(\mathbf{p}|\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} S(p_i|X_i).$$

The quantity $S(\mathbf{p}|\mathbf{X})$ is a random variable itself, because it is a function of random variables $\mathbf{X}$. To compare two forecasts $\mathbf{p}_1$ and $\mathbf{p}_2$, we study their score difference:

$$\begin{aligned}\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X}) &= \frac{1}{N}\left(\sum_{i=1}^{N} S(p_{1i}|X_i) - \sum_{i=1}^{N} S(p_{2i}|X_i)\right),\\ &= \frac{1}{N}\sum_{i=1}^{N} \Delta(p_{1i}, p_{2i}, X_i).\end{aligned}$$

The quantity $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})$ is also a random variable as it too depends on the vector of random variables $\mathbf{X}$. If $S(p|X)$ is a proper scoring rule, and if the expected value of the score difference is positive, namely $\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})] > 0$, the forecast $\mathbf{p}_1$ is "closer" to the true $\mathbf{p}^*$ than the alternative forecast $\mathbf{p}_2$. The expected value should be considered with respect the distribution of the observations $\mathbf{X}$. However, we do not observe the full distribution - we only observe a sample (i.e. we observe the quantity $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x})$ which is a realization of the random variable $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})$). Even if the expected score difference $\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})]$ is positive, which means that we should express a preference for the first forecast, the observed score difference $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x})$ may be negative and lead to the opposite conclusion. To avoid this problem we need to account for the uncertainty around the observed $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x})$ which is the point estimate of the expected score difference $\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})]$.

The distribution of score differences - Multiple Bins Single Probability

For the Multiple Bins Single Probability case, the observation in each bin is a binary random variable $X_i \sim \text{Ber}(p^*)$, $i = 1, ..., N$. Given an arbitrary scoring rule $S(p|X)$ and two candidate forecasts p_1 and p_2, the score difference for the i-th bin is a discrete random variable with distribution:

$$\Delta(p_1, p_2, X_i) = \begin{cases} \Delta_0 = S(p_1|0) - S(p_2|0) & \text{with probability} \quad 1 - p^*, \\ \Delta_1 = S(p_1|1) - S(p_2|1) & \text{with probability} \quad p^*. \end{cases}$$

The forecasts are ranked based on the average score difference across all bins:

$$\begin{aligned} \Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X}) &= \frac{1}{N}\sum_{i=1}^{N} \Delta(p_1, p_2, X_i), \\ &= \frac{1}{N}\sum_{i=1}^{N} \left(\Delta_0 + X_i(\Delta_1 - \Delta_0)\right), \\ &= \Delta_0 + \frac{X_S}{N}(\Delta_1 - \Delta_0), \end{aligned}$$

where, $X_S = \sum_i X_i$ is the sum of all observations or, equivalently, the total number of successes. By definition, X_S is the sum of N (assumed to be) independent and identically distributed Bernoulli trials X_i. Therefore, X_S has a Binomial distribution with size parameter N, the number of bins, and probability parameter p^*. When we observe a sample $x_1, ..., x_N$, the observed score difference is given by:

$$\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x}) = \Delta_0 + \frac{\mathrm{x}_S}{N}(\Delta_1 - \Delta_0),$$

where x_S is a realization of the random variable X_S. The observed score difference depends on the observations only through the quantity x_S/N. Thus, it is enough to study the quantity x_S/N to make inference about the expected value of the score difference. The quantity x_S/N it is said to be *sufficient* (Fisher, 1922) with respect to the expected score difference because it contains all the information provided by the observations $x_1, ..., x_N$ on the parameter of interest (in this case, the expected score difference $\mathbb{E}[\Delta(p_1, p_2, X)]$). For an introduction to statistical inference and the theory behind we refer to Schervish (2012); Hastie et al. (2009).

Confidence Intervals for the Expected Score Difference

A way to account for the uncertainty around the observed score difference is to consider an interval estimate of the expected value of the score difference. Once a sample $\mathbf{x} = x_1, ..., x_N$ has been observed and the confidence interval calculated, if the entire interval lies above zero we express a preference towards p_1, alternatively if it lies below zero we express a preference towards p_2. If the interval contains the value zero we conclude that the observed sample does not contain enough information to express a preference. It is important to consider the latter case as a possible outcome because it is an indication that we need to collect more data or that the forecasts perform similarly (as measured by the score) and provide an additional information than the pure rankings.

We are considering the confidence interval for the expected value of the score difference:

$$\begin{aligned} \mathbb{E}[\Delta(p_1, p_2, \mathbf{X})] &= \Delta_0 + \frac{\mathbb{E}[X_S]}{N}(\Delta_1 - \Delta_0), \\ &= \Delta_0 + p^*(\Delta_1 - \Delta_0). \end{aligned} \tag{7.13}$$

Having an observation $\mathbf{x} = x_1, ..., x_N$ per bin, the point estimate of $\mathbb{E}[\Delta(p_1, p_2, X)]$ is the observed score difference:

$$\Delta(p_1, p_2, \mathbf{x}) = \Delta_0 + \hat{p}(\Delta_1 - \Delta_0), \tag{7.14}$$

where $\hat{p} = x_S/N$ is the observed probability of success. Comparing the equations 7.13 and 7.14, the point estimate of the score difference is retrieved plugging in the point estimate of the probability of success $\hat{p}$ in place of p^*. In the same way, to retrieve an interval estimate of the expected score difference is sufficient to retrieve an interval estimate of the probability of success p^*.

Therefore, we need the confidence interval of level α for the true probability p^* given observations $x_1, ..., x_N$ from a $\text{Ber}(p^*)$, namely $CI_{p^*}(\alpha) = (\hat{p}_L, \hat{p}_U)$, and plug those values into expression 7.14 to obtain a confidence interval for $\mathbb{E}[\Delta(p_1, p_2, X)]$, namely $CI_\Delta(\alpha) = (\Delta_L, \Delta_U)$. Various methods have been found to estimate $\hat{p}_L$ and $\hat{p}_U$, most of them relying on a Gaussian approximation. However, this approximation is not reliable for small sample sizes (number of bins N) and for values of p^* close to zero or one, as in our case (Wallis, 2013).

Hereafter, we use the Clopper-Pearson confidence interval (Clopper and Pearson, 1934). This method is referred to as *exact* because it relies on cumulative binomial probabilities rather than an approximation and is therefore more efficient and accurate than simulation based methods. The confidence interval with level α for $\hat{p}$ is given by:

$$\begin{aligned} p_L(\alpha) &= \text{BetaQ}(\frac{\alpha}{2}; x_S, N - x_S + 1), \\ p_U(\alpha) &= \text{BetaQ}(1 - \frac{\alpha}{2}; x_S + 1, N - x_S), \end{aligned}$$

where the function $\text{BetaQ}(q; a, b)$ is the q-th quantile of a Beta distribution with parameters a and b. We can construct confidence intervals for $\mathbb{E}[\Delta(p_1, p_2, \mathbf{X})]$ as follows:

$$\begin{aligned} \Delta_L &= \Delta_0 + \hat{p}_L(\Delta_1 - \Delta_0), \\ \Delta_U &= \Delta_0 + \hat{p}_U(\Delta_1 - \Delta_0). \end{aligned}$$

The obtained confidence interval for p^* depends on the data only through the sum of the observations x_S, which is a sufficient statistic for the problem. Similarly, the confidence interval for $\mathbb{E}[\Delta(p_1, p_2, \mathbf{X})]$ depends on the data through the value of the sufficient statistic, x_S.

Figure 7.5 shows the confidence interval for the score difference as a function of the sum of observations x_S considering two competing forecasts $p_1 = 0.001$, $p_2 = p_1/3$, a reference model for the pairwise gambling score $p_0 = 5p_1$ and $N = 10,000$ bins. Here, we do not need to choose a value for p^*. Indeed, the confidence interval is determined solely by the forecast and observation. The Brier, log and full gambling score(Figure 7.5a, 7.5c, 7.5d) all express a preference for p_1 if we observe $x_S > 12$, while they express a preference for p_2 when $x_S < 2$. This result is expected because $p_1 > p_2$, which means that $x_S > 12$ is much more probable under p_1 than p_2. In fact, the average number of successes using p_1 is $Np_1 = 10$ while $Np_2 = 3.34$. The same reasoning applies when we express a preference for p_2 ($x_S < 2$).

The pairwise gambling score (Figure 7.5 (b)), instead, requires $x_S > 24$ to express a preference for p_1 and $x_S < 9$ to express a preference for p_2. It is heavily biased toward the forecast closer to zero. In fact, when p_1 is the true probability, the probability of observing $x_S > 24$ is less than 0.0001 and the probability of observing $x_S < 9$ is 0.33. Therefore, we are more likely to express a preference for p_2 than for p_1, even when $p_1 = p^*$. This reinforces the problems with employ improper scores introduced in Section 7.2.4.

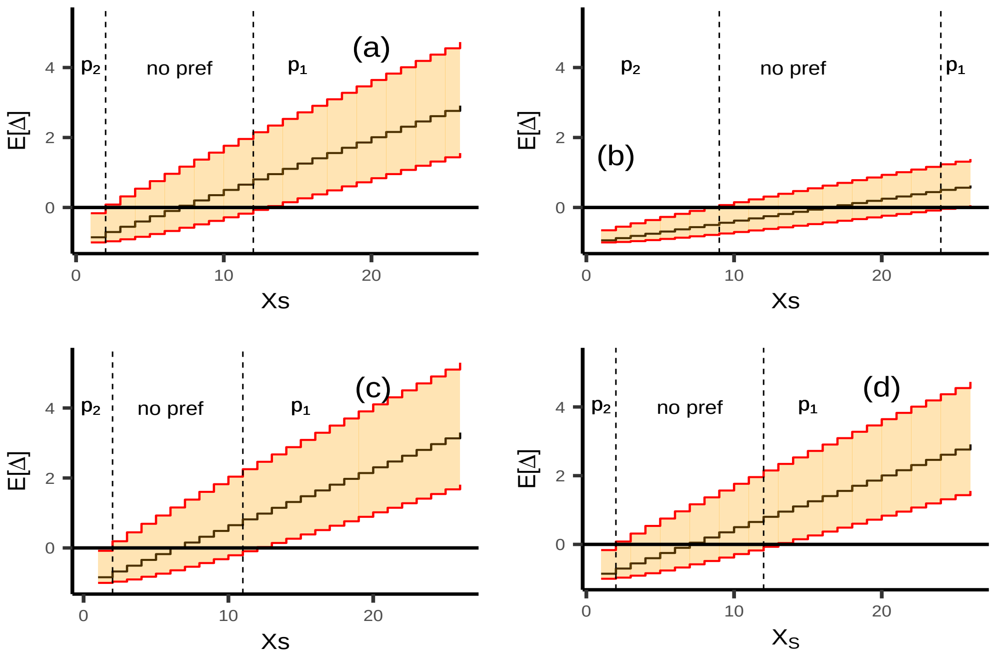


Figure 7.5: Confidence interval (shaded area) and point estimate (black solid line) for $\mathbb{E}[\Delta]$ as a function of the number of observed successes x_S considering $p_1 = 0.001$, $p_2 = p_1/3$, $p_0 = 5p_1$ and N = 10000. In each plot shows a different score: (a) Brier score; (b) pairwise gambling score; (c) logarithmic score; (d) full gambling score. Black solid line represents the observed score difference while the orange area represents the confidence interval. The black vertical dashed lines represent the interval of values of x_S for which we do not express a preference

Table 7.1: Multiple Bins Single Probability case: table reporting the values x_{min} and x_{max} for the Brier, log, pairwise gambling (PG) and full gambling (FG) score. The reported values refers to the case where $N = 10,000$, $p_1 = 0.001$, $p_2 = p_1/3$, and do not depend on p^*.

Score	x_{min}	x_{max}
Brier	2	12
Log	2	11
PG	9	24
FG	2	12

Preference Probabilities

The confidence interval for the expected score difference, $CI_\Delta(\alpha)$, is a function of the competing forecasts, the scoring rule and depends on the data only through the sum of the observations $x_S = \sum_i x_i$. In particular, there are a range of values (between the dashed lines in Figure 7.5) of x_S for which we are not able to express a preference. We refer to this interval as (x_{min}, x_{max}). With respect to the sum of the observations x_S there are only three possible outcomes:

$$\begin{aligned} x_S < x_{min} &\longrightarrow \text{preference for } p_2, \\ x_{max} \leq x_S \leq x_{max} &\longrightarrow \text{no preference}, \\ x_S > x_{max} &\longrightarrow \text{preference for } p_1. \end{aligned}$$

The values x_{min} and x_{max} are determined solely by p_1, p_2, the number of bins N and the scoring rule. Table 7.1 reports the values of x_{min} and x_{max} for the scoring rules depicted in Figure 7.5. These values can be used to compute the preference probabilities once a value for p^* is assumed. Indeed, in the Multiple Bins Single Probability case, the distribution of X_S is a Binomial distribution, $X_S \sim \text{Bin}(N, p^*)$. Table 7.2 reports the probabilities of i) no preference; ii) Preference for p_1; iii) Preference for p_2. The probabilities are calculated considering alternatively p^* equal to p_1 (first half of the table) or p_2 (second half of the table).

The similarity among the values x_{min} and x_{max} for the proper scores lead to similar preference probabilities. The proper scores always assign the greatest probability to the case in which we are not able to express a preference, however, when $p^* = p_1$ it is unlikely to express a preference for p_2. Vice versa when $p^* = p_2$. There is a slightly difference between the Brier and the log score coming from the different penalty applied to forecasts close to zero. The log score penalises more heavily forecasts close to zero and, in fact, when $p_1 = p^*$ chances to express a preference for p_1 are higher than using the Brier score. The full gambling score for p_1 against p_2 is proportional to the Brier score difference between p_1 and p_2 (see Section 7.3.2) and thus, their preference probabilities coincide. Considering the pairwise gambling score the probability of expressing a preference for p_1 is always very close to zero, even when $p^* = p_1$. This shows again that it is possible to find a combination of p_1, p_2 and p_0 such that the model providing the smallest forecast obtains the highest reward with probability over 0.9, even when the other forecast is equal to the true p^*.

Figure 7.6 shows the preference probabilities as a function of $p^* \in (10^{-6}, 10^{-2})$ which is the range of values of the 5-year adaptively-smoothed forecast for Italy (aggregating over the magnitude bins) used later to illustrate the Multiple Bins Multiple Probabilities case. The Brier score behaves as expected. The probability of expressing a preference for p_2 increases as p^* goes to zero, which is what we expect given $p_2 < p_1$. On the other hand, the probability

Table 7.2: Multiple Bins Single Probability case: table reporting for each score (row) the probabilities of expressing (or not) a preference using either the Brier, log-, pairwise gambling (PG) or full gambling (FG) score. The probabilities are calculated considering $N = 10,000$, $p_1 = 0.001$, $p_2 = p_1/3$ and considering two cases: $p^* = p_1$ and $p^* = p_2$.

Score	No pref	Pref p_1	Pref p_2
$p^* = p_1$			
Brier	0.7912	0.2083	0.0005
Log	0.6963	0.3032	0.0005
PG	0.6672	0.0000	0.3327
FG	0.7912	0.2083	0.0005
$p^* = p_2$			
Brier	0.8454	0.0000	0.1545
Log	0.8453	0.0000	0.1545
PG	0.0073	0.2083	0.9927
FG	0.8454	0.0000	0.1545

of preferring p_1 increases when p^* increases, because $p_1 > p_2$. Finally, the probability of not being able to express a preference is higher when $p_2 < p^* < p_1$ (Figure 7.6a). The pairwise gambling score, instead, does not behave as expected. The probability of preferring p_1 is almost zero in the range of values of p^* considered in the example. The two most probable outcomes are: expressing a preference for p_2 or not expressing a preference at all.

Probability of expressing a preference

It is interesting to study how the probability for each case changes as a function of p^* for different numbers of bins N and different ratios between p_1 and p_2. To do that it is useful to focus only on two possible outcomes: expressing a preference and not expressing a preference. The probability of expressing a preference is given by the probability of observing a sample such that the sum of the observations x_S is greater than x_{max} or smaller than x_{min}. We refer to this probability as β which is given by

$$\beta = 1 - \Pr[\mathrm{x}_{min} \leq X_S \leq \mathrm{x}_{max}].$$

This probability depends on the scoring rule, the forecasts p_1 and p_2,the number of bins N and the true probability p^*. We study β as a function of p^* for different numbers of bins. In this artificial case, to increase the number of bins we are considering additional bins with the same probability, we are explicitly not splitting any bin; this is analogous to increase the data at hand applying the model to a larger spatio-temporal region.

Figure 7.7a considers only the Brier score. The region of p^* presenting low values for β shrinks when the number of bins increase which simply means that the more data we have, the more chances of expressing a preference. Moreover, β is at the minimum when $p^* \in (p_2, p_1)$, which is reasonable because if the distances $|p^* - p_1|$ and $|p^* - p_2|$ are similar the probability of no preference should be high. The $N = 2000$ can be explained considering $p_1 = p^*$. In this case, the expected sum of observations is $Np_1 = 2$ and it is more probable to observe $X_S < 2$ than $X_S > 2$. Given that $Np_2 < Np_1$, the probability of not expressing a preference is high.

Figure 7.7 presents the probability β as function of p^* for different scores with a fixed

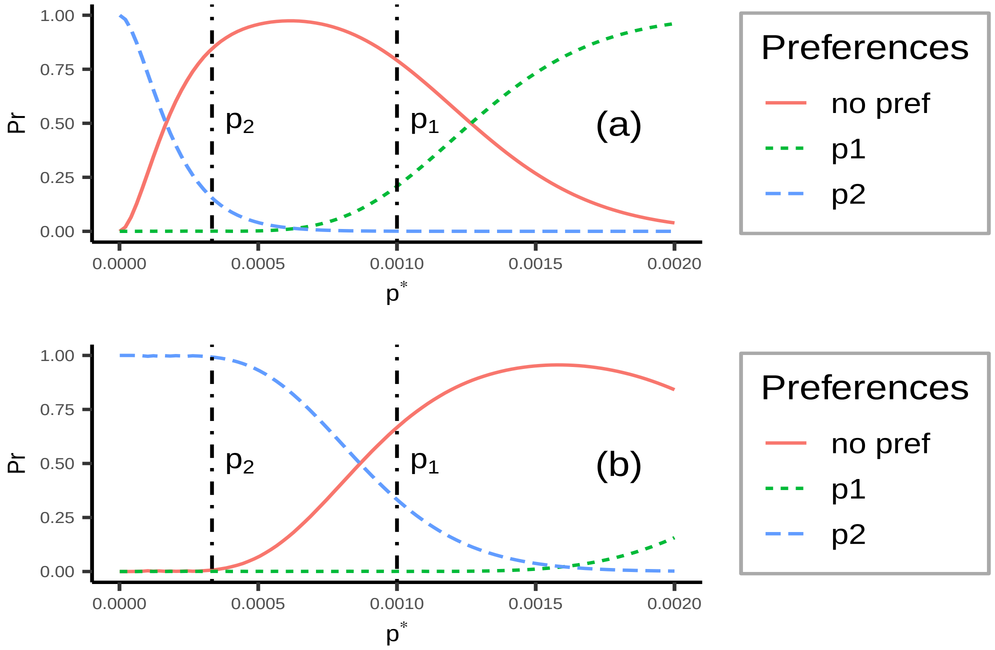


Figure 7.6: Multiple Bins Single Probability case: each plot shows the probability of each possible outcome (solid lines no preference, dotted lines preference for p_1, dashed lines preference for p_2) as a function of p^* using the Brier score (a) and the pairwise gambling score (b) considering $p_1 = 0.001$ and $p_2 = p_1/3$ (vertical lines), $p_0 = 5p_1$ and $N = 10,000$. The true probability p^* varies in $(10^{-6}, 2 \cdot 10^{-2})$ which is a realistic range of values in Italy.

number of bins $N = 5000$ (b) and $N = 20000$ (c). For $N = 5000$, the proper scores (Brier, log and full gambling score) present the same values of β for any value of p^*. For $N = 20000$, the proper scores start to behave differently. The Brier and full gambling score still coincide, while the log score is slightly different. Specifically, the log score presents higher β values when $p^* = p_1$, and lower when $p^* = p_2$. This depends on the different penalties applied to forecasts close to zero. The log score presents greater chances of expressing a preference for p_1 when $p^* = p_1$ because the other forecast p_2 is smaller than p_1 and, therefore, penalized. On the other hand, when $p^* = p_2$ the log score presents smaller β values than the Brier score.

In contrast to the proper scores, the pairwise gambling score reaches its minimum β value for $p^* > p_1$. Here, the pairwise gambling score tends to express a preference for the smaller forecast even when the other one is closer to p^*. This leads to higher values of β when $p^* \in (p_2, p_1)$ because the pairwise gambling score will likely express a preference for p_2. Only when $p^* > p_1$ the probability of no preference grows and the value of β decreases accordingly.

Given that p_1 and p_2 are scalars, we can consider β as a function of the ratio between p_1 and p_2, $\omega = p_2/p_1$, for a fixed p^*. In principle, we expect that β is an increasing function of ω. We assume that the first forecast and the true probability are identical $p_1 = p^* = 0.001$. The reference model for the pairwise gambling score is $p_0 = 5p_1$ and we consider different numbers of bins $N \in \{2000, 5000, 10000, 20000\}$. The ratio $\omega = p_2/p_1$ varies in the interval $(0.1, 4)$. We expect low β values when ω is around one (similar forecast) and high β values otherwise.

Figure 7.8a shows that, as expected, for $N > 2000$, β has its minimum when $\omega = 1$. Considering ω as fixed, β is an increasing function of the number of bins. Figure 7.8b-c

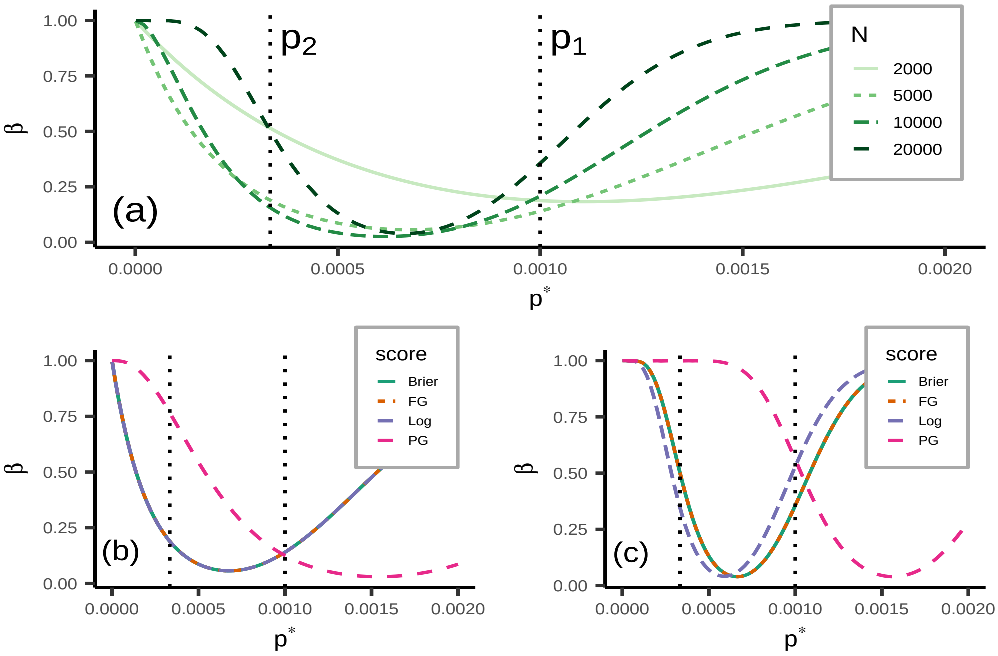


Figure 7.7: Multiple Bins Single Probability case: (a) Brier score preference probability as a function of p^* for different numbers of bins $N \in \{2000, 5000, 10000, 20000\}$. (b-c) Probability of expressing a preference as a function of p^*. Colors represent the different scores: Brier, log , pairwise gambling (PG), and full gambling (FG) score. The Brier and FG scores coincide. The number of bins is fixed to N = 5000 (b) and N = 20000 (c). We set $p_1 = 0.001$, $p_2 = p_1/3$ (vertical lines), $p_0 = 5p_1$, and $p^* \in (10^{-6}, 2^{-3})$ which is a realistic range of values in Italy.

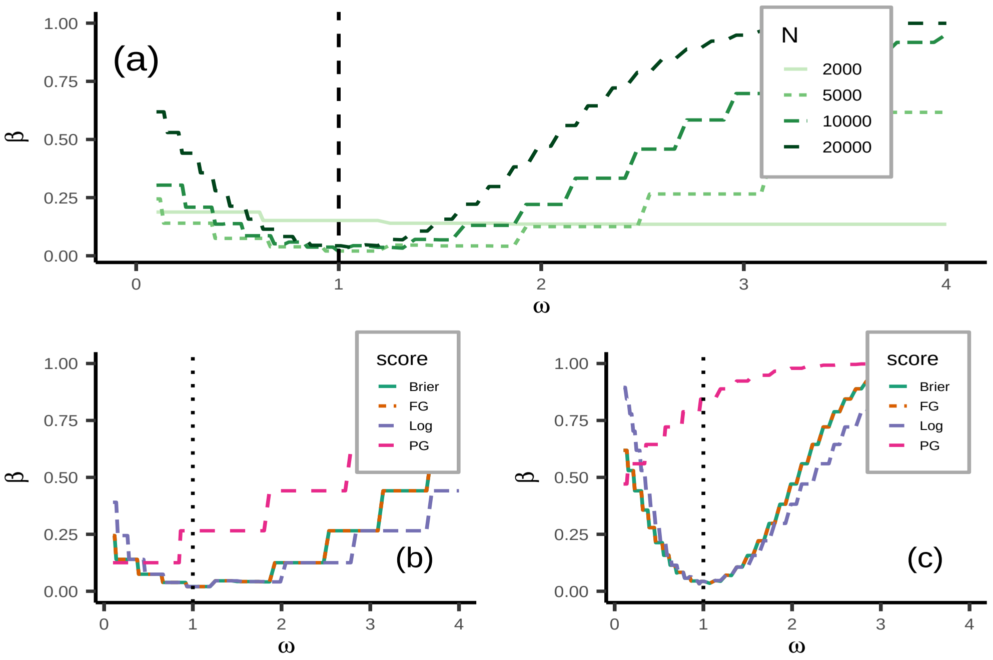


Figure 7.8: Multiple Bins Single Probability case: (a) Probability of expressing a preference using the Brier score as a function of $\omega = p_2/p_1 \in (0.1, 4)$ for different numbers of bins $N \in \{2000, 5000, 10000, 20000\}$. We set $p_1 = p^* = 0.001$, and the reference model is $p_0 = 5p_1$. (Bottom) Probability of expressing a preference as a function of ω. Colors represent the different scores: Brier, log, pairwise gambling (PG), and full gambling (FG) score. The number of bins is fixed to N = 5000 (b) and N = 20000 (c).

compares the β values relative to different scores for a fixed number of bins, $N = 5000$ (b) and $N = 20000$ (c). The Brier and full gambling score coincide, whilst the log score presents slightly different β values. As before, this is due to the different penalties applied to forecasts close to zero.

The pairwise gambling score is not consistent with the trends in the proper scores. Using this score and considering $N = 20000$ (Figure 7.8c), the probability β is consistently greater than 0.5 for the considered values of ω. Considering that $p_1 = p^*$, the quantity ω is also the ratio between p_2 and p^*. This implies that regardless of ω, we will erroneously express a preference for p_2 with a probability above 0.5.

Importantly, these sanity checks of a proposed scoring procedure can be done before looking at the observations. It is possible to check if forecasts can, in principle, be distinguished in light of the amount of expected data. We recommend the use of such exploitative figures when introducing a new scoring rule whose performance have not been tested. If the proposed scoring rule does not behave acceptably in this simple scenario, it is unlikely that it would behave acceptably in a real application.

Score difference distribution - Multiple Bins Multiple Probabilities

The Multiple Bins Multiple Probabilities case generalizes the Multiple Bins Single Probability case, and is much more similar to a real earthquake forecasting experiment. For example, the forecasts involved in the first CSEP experiments (Field, 2007; Schorlemmer and Gerstenberger, 2007b; Zechar et al., 2013; Michael and Werner, 2018) were mostly grid-based forecasts providing for each space-time-magnitude bin, the expected number of earthquakes.

Then, the number of events in each bin is modelled using a Poisson distribution with intensity equal to the number of events provided by the forecasts and the probability of observing at least one event is calculated accordingly. In this scenario, we do not have analytical results for the score difference distribution and we need to recur to simulations.

We now want to specify a true model which has more realistic probabilities. Since we do not actually know these in reality, we choose to work with one of the CSEP models that was submitted to the 2010 Italy experiment (Taroni et al., 2018). We choose to simulate synthetic data from the 5-year adaptively-smoothed forecast for Italy (Werner et al., 2010) and explore the ability of the scoring rules to discriminate between linearly scaled versions of this true model. This means that we are considering only one time bin of size 5 years, while the space-magnitude domain is divided in multiple regular bins. The spatial domain is represented by the coloured area in Figure 7.9 and it is divided in 0.1×0.1 longitude-latitude bins. The magnitude domain ranges from 4.95 to 9.05 magnitude units and is divided in bins of length 0.1. The forecast is relative to the period from January 1, 2010, to December 31, 2014.

The adaptively-smoothed forecast provides the expected number of earthquakes in each space-magnitude bin. For each bin, to calculate the probability of observing at least one earthquake, in accordance with the methodology in the 2010 Italy CSEP forecast experiment, we consider a Poisson distribution for the number of events with intensity given by the predicted number of events. Assuming independence in the magnitude bins, we can aggregate the probabilities over magnitude bins and, for each space bin, obtain the probability of observing at least an earthquake in the period of interest with magnitude greater, or equal, to 4.95. Figure 7.9 shows the forecasted log-probability for each spatial bin used as data generating model.

The Italian adaptively-smoothed forecast reported in Figure 7.9 is the vector of true probabilities $\mathbf{p}^* = p_1^*, ..., p_N^*$, where $N = 8993$. As in the previous sections, we compare two forecasts $\mathbf{p}_1 = \mathbf{p}^*$ and $\mathbf{p}_2 = \omega\mathbf{p}^*$. We will be ignoring the spatial configuration. The average bin score difference is given by

$$\begin{aligned}\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X}) &= \frac{1}{N}\sum_{i=1}^{N}(\Delta_{0,i} + X_i(\Delta_{1,i} - \Delta_{0,i})),\\ &= \bar{\Delta}_0 + \frac{1}{N}\sum_{i=1}^{N} X_i(\Delta_{1,i} - \Delta_{0,i}),\end{aligned}$$

where, $\Delta_{0,i} = \Delta(p_{1i}, p_{2i}, 0)$ and $\Delta_{1,i} = \Delta(p_{1i}, p_{2i}, 1)$ are, respectively, the score difference in the i-th bin in case we observe $X_i = 0$ (no earthquake at or above magnitude 4.95 during the 5 years) or $X_i = 1$ (at least one earthquake above magnitude 4.95 during the 5 years). The quantity $\bar{\Delta}_0$ is the average $\Delta_{0,i}$. The observations $X_i \sim \text{Ber}(p_i^*)$ follow a Bernoulli distribution, each bin has a potentially different parameter $p_i^* \neq p_j^*$ for any $i \neq j$. The expected value of the score difference is given by

$$\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{X})] = \bar{\Delta}_0 + \frac{1}{N}\sum_{i=1}^{N} p_i^*(\Delta_{1,i} - \Delta_{0,i}).$$

Given that we are considering $\mathbf{p}_1 = \mathbf{p}^*$ and $\mathbf{p}_2 = \omega\mathbf{p}^*$, the expected score difference is non-negative if a proper scoring rule is used while it could be negative if the scoring rule is improper. Specifically, we show that it is possible to find a reference model $\mathbf{p}_0$ such that, if used in combination with the parimutuel gambling score to rank the forecasts, the expected score difference is negative. As before, the Brier, log and full gambling score are used for

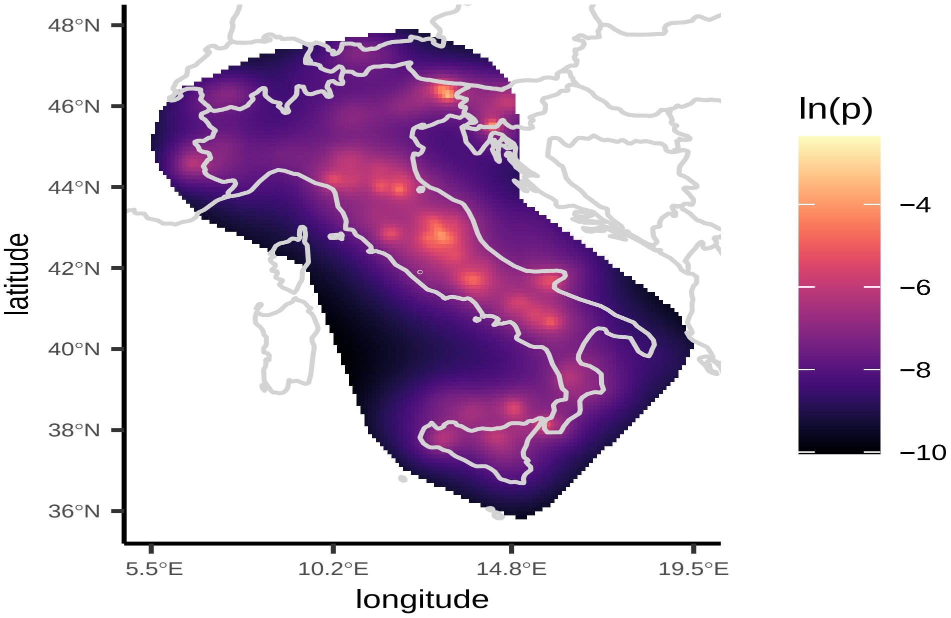


Figure 7.9: 5-year adaptively-smoothed forecast for Italy (Werner et al (2010)). The figure shows for each spatial bin the natural logarithm of the probability of observing at least one earthquake at or above magnitude 4.95 in the period from January 1, 2010, to December 31, 2014.

Table 7.3: Expected score difference considering $\mathbf{p}^*$ equal to the 5 year Italy adaptively-smoothed forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$ and reference model for the pairwise gambling score $\mathbf{p}_0 = 5\mathbf{p}^*$. The scores considered are: the Brier score, the log score, the pairwise gambling (PG) score and the full gambling (FG) score.

Score	$\mathbb{E}[\Delta]$
Brier	0.0000026
Log	0.0003137
PG	-0.0000900
FG	0.0002422

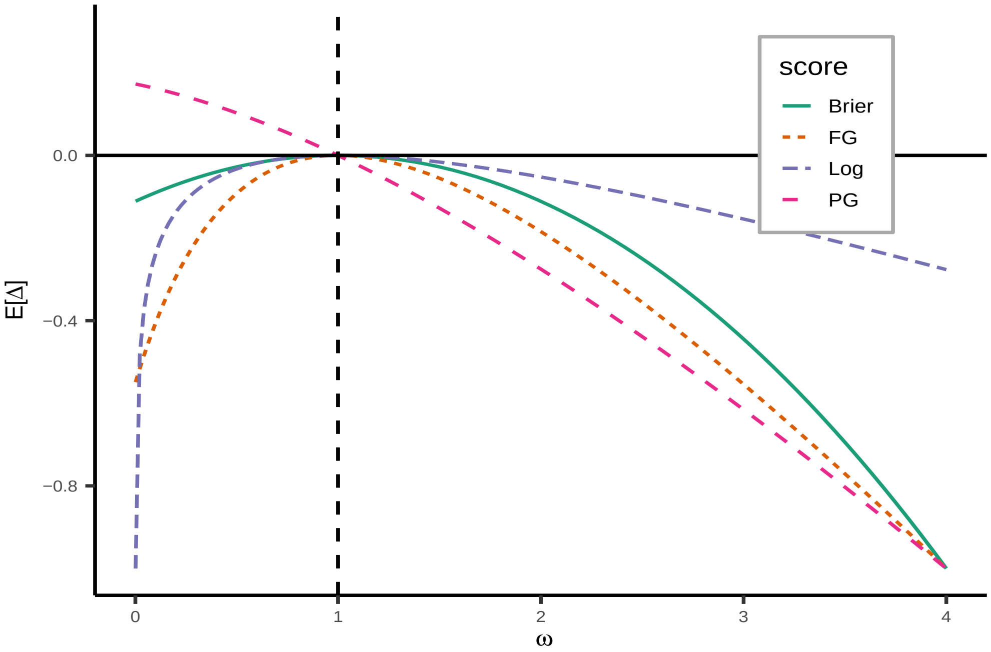


Figure 7.10: Expected score difference between $\mathbf{p}_1$ and $\mathbf{p}_2$ as a function of $\omega = \mathbf{p}_1/\mathbf{p}_2$ for $\omega \in (10^{-3}, 4)$. We set $\mathbf{p}^*$ equal to the 5-year adaptively-smoothed Italy forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$, and reference model for the pairwise gambling score $\mathbf{p}_0 = 5\mathbf{p}^*$.

comparison. In Table 7.3 we report the expected score differences considering different scores. As expected, they are all positive except for the pairwise gambling score.

In Figure 7.10 is showed the expected score difference as a function of the forecasts ratio $\omega \in [10^{-3}, 4]$. The results are similar to the ones reported in Figure 7.1 and 7.2. The Brier, log and full gambling scores behave suitably, while the pairwise gambling score does not. The Brier and full gambling score are bounded and they prefer a forecast $\mathbf{p} = 10^{-3}\mathbf{p}^*$ to $\mathbf{p}' = 4\mathbf{p}^*$. Indeed, in Figure 7.10 the left hand side is greater than the right hand side. That is because the penalty is based on the absolute difference between a forecast and the data generating model, therefore, a forecast $\mathbf{p} = 10^{-3}\mathbf{p}^*$ is preferred to $\mathbf{p}' = 4\mathbf{p}^*$, because $\|10^{-3}\mathbf{p}^* - \mathbf{p}^*\| \leq \|4\mathbf{p}^* - \mathbf{p}^*\|$. On the other hand, the log score is unbounded and is based on the relative difference. With the log score, a forecast $\mathbf{p}' = 4\mathbf{p}^*$ is preferred to $\mathbf{p} = 10^{-3}\mathbf{p}^*$ because $\|\mathbf{p}^*/10^{-3}\mathbf{p}^*\| > \|\mathbf{p}^*/4\mathbf{p}^*\|$. The pairwise gambling score, instead, is heavily biased towards zero.

We can extend the comparison by considering $k = 3$ forecasts. In this case, we consider

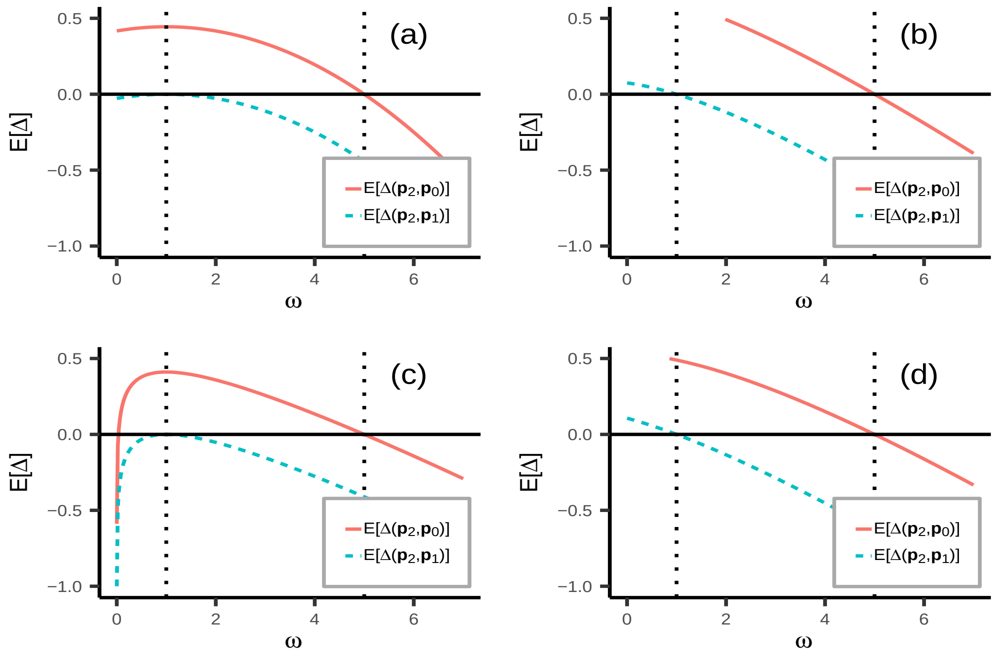


Figure 7.11: Expected score difference between $\mathbf{p}_2$ and $\mathbf{p}_1$ (blue dashed) and $\mathbf{p}_2$ and $\mathbf{p}_0$ (red solid) as a function of $\omega = \mathbf{p}_1/\mathbf{p}_2$ for $\omega \in (10^{-3}, 7)$. We set $\mathbf{p}^*$ equal to 5-year adaptively-smoothed Italy forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$ and $\mathbf{p}_0 = 5\mathbf{p}^*$. Vertical lines represent $\omega = 1$ and $\omega = 5$. We consider the Brier (a), pairwise gambling (b), log (c), and full gambling (d) scores.

the reference model $\mathbf{p}_0 = 5\mathbf{p}^*$ as third competitor. Figure 7.11, for each scoring rule, shows the expected score differences $\mathbb{E}[\Delta(\mathbf{p}_2, \mathbf{p}_1, \mathbf{X})]$ (dashed blue) and $\mathbb{E}[\Delta(\mathbf{p}_2, \mathbf{p}_0, \mathbf{X})]$ (solid red), representing the expected score difference between $\mathbf{p}_2$ and $\mathbf{p}_1$, and the expected score difference between $\mathbf{p}_2$ and $\mathbf{p}_0$. Given that $\mathbf{p}_1$ is equal to the true probabilities, the score differences have to be negative for any value of $\omega \neq 1$ in order for the scoring rule to be effective. Indeed, this is the case for the Brier and log score (Figure 7.11 (a), (c)). On the other hand, both the pairwise and full gambling score (Figure 7.11 (b), (d)) are improper and prefer $\mathbf{p}_2$ over $\mathbf{p}_1$ when $\omega \in (0, 1)$. Moreover, all the scores prefer $\mathbf{p}_0$ to $\mathbf{p}_2$ when $\omega > 5$. However, the log score prefers $\mathbf{p}_0$ to $\mathbf{p}_2$ also when ω approaches zero. This shows, again, how different scoring rules apply different penalties to the forecasts.

We note that the pairwise and full gambling score present almost the same expected score difference between $\mathbf{p}_2$ and $\mathbf{p}_1$. This is because both scoring procedures implicitly assume a reference model given by the average forecast. If the average forecast in a bin is greater than p_i^*, a forecaster will obtain a positive reward each time they submit a value smaller than $\mathbf{p}_i^*$ and $X_i = 0$ occurs. Therefore, given that we are in a low probability environment for which $\Pr[X_i = 0] > 0.99$, the smallest forecast is likely to be preferred. The bias depends on the relationship between the reference model and the true probabilities. In the gambling metaphore, the reference model plays the role of the house (or banker) which determines the returns, and against which all forecasts are competing. Considering equation 7.12, if $p^* < p_0$, the player has a positive reward forecasting $p < p_0$. If the number of forecasts is large enough that changing a forecast does not affect significantly the average, then, the smaller the forecast the higher the reward, and the forecaster is encouraged by the score to provide $p = 0$. The same reasoning applies if $p^* > p_0$.

Confidence Interval and preference probabilities - Multiple Bins Multiple Probabilities

Also in the Multiple Bins Multiple Probabilities case it is crucial to account for the uncertainty around the observed score difference. The binomial formulation used before to retrieve confidence intervals no longer holds, and we need an alternative methodology. One approach to calculate confidence intervals for the expected score difference relies on a Gaussian approximation of the score difference distribution (Rhoades et al., 2011). The score difference in each bin, $\Delta(p_{1i}, p_{2i}, X_i)$ for $i = 1, ..., N$, are assumed to be independent draws from a Gaussian distribution with expected value $\mathbb{E}[\Delta(\mathbf{p}_1, \mathbf{p}_2, X)]$ and variance σ^2. When we observe a sample $\mathbf{x} = x_1, ..., x_N$, the point estimate of the expected score difference is the observed score difference $\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x})$ and the $(1-\alpha)\%$ confidence interval is given by:

$$\Delta(\mathbf{p}_1, \mathbf{p}_2, \mathbf{x}) \pm t_{1-\alpha/2, N-1} \frac{s}{\sqrt{N}},$$

where s^2 is an estimate of the variance σ^2 and $t_{1-\alpha/2,N-1}$ is the $1-\alpha/2$ percentile of a t-student distribution with $N-1$ degrees of freedom. The reliability of such interval estimates is determined by the accuracy of the Gaussian approximation, which, in turns, depends on the amount of data (the more the better) and on the correlation between the score difference in each bin (the more the worst). We analyse the reliability of this approximation in Appendix A: Reliability of the gaussian confidence intervals and conclude that it can be used with the log, pairwise gambling and full gambling score but not with the Brier score.

Figure 7.12 shows the evolution of the preference probabilities varying the forecasts ratio, $\omega = \mathbf{p}_2/\mathbf{p}_1$. It is quite similar to Figure 7.6 and the same problems with the pairwise gambling score are evident; i.e. it favours forecast smaller than the true probability when the average forecast is greater than the latter. On the other hand, the log score probability of preferring $\mathbf{p}_1$ increases rapidly when $\omega \to 0$, while the full gambling score is not able to distinguish between $\mathbf{p}_1$ and $\mathbf{p}_2$ for $\omega < 2.5$. The latter remark suggests a potential problem with the use of the full gambling score given that, in real forecasting experiments, the competing forecasts tends to be quite similar, in which case there is an high probability of no preference.

This concludes our analysis on the use of proper scoring rules to rank earthquake forecasting models.

7.2.6 Discussion

The parimutuel gambling score was introduced as a general scoring rule to compare, within a unified framework, earthquake forecasts of different kinds (e.g. alarm-based forecast and probabilistic forecast). It overcomes two limitations common to other forecast comparison techniques: i) the need to define a reference model, and ii) to allow forecasts defined on different space-time-magnitude regions to be compared. We showed that the parimutuel gambling score is proper only when two forecasts are compared directly against each other. In the other cases (multi-forecast comparison and comparison against a reference model), the parimutuel gambling score is improper. Consequently, we discourage its use in multi-model comparisons such as CSEP and encourage researchers and practitioners to re-consider rankings obtained using this score.

Specifically, the parimutuel gambling score tries to avoid the need to pre-define a reference model by using the average forecast. Therefore, for each bin, a positive reward means that the model is *better* than the average forecast, vice versa if the reward is negative. This allows to produce a map of the parimutuel gambling rewards from which to infer the bins where the forecast is better than the average forecast, and the bins where it is not. Since the parimutuel gambling score is proper only when $k = 2$, any map obtained by computing

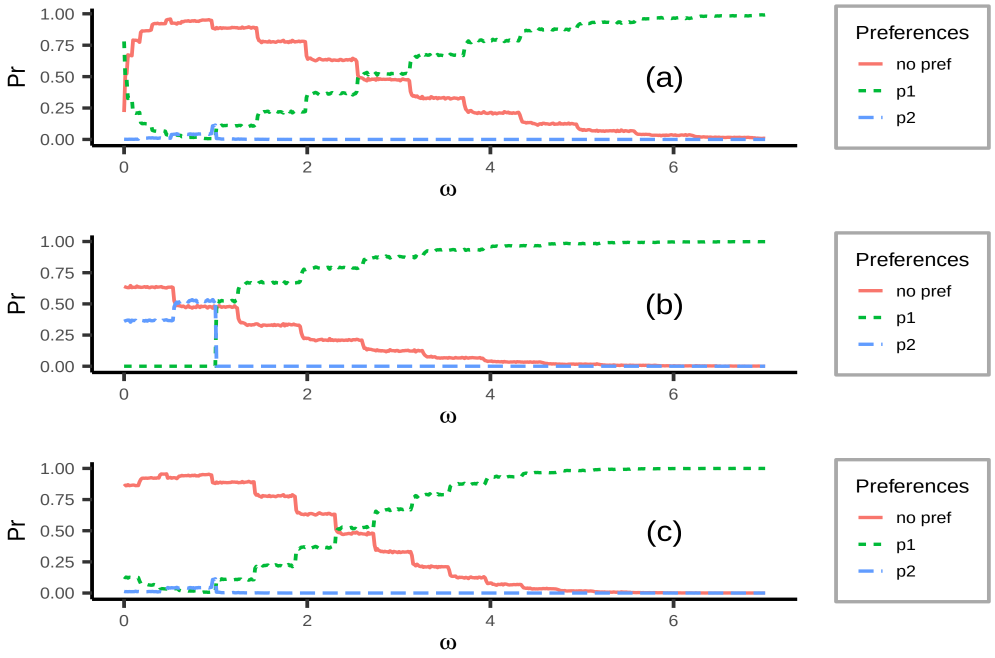


Figure 7.12: Multiple Bins Multiple Probabilities case: each plot shows the probability of each possible outcome (solid no preference, dotted preference for $\mathbf{p}_1$, dashed preference for $\mathbf{p}_2$) as a function of $\omega = \mathbf{p}_1/\mathbf{p}_2$ for $\omega \in (10^{-3}, 7)$. The log (a), the pairwise gambling (b) and the full gambling (c) scores are considered. We set $\mathbf{p}^*$ equal to 5-year adaptively-smoothed Italy forecast, $\mathbf{p}_1 = \mathbf{p}^*$, $\mathbf{p}_2 = \omega\mathbf{p}^*$, and reference model for the pairwise gambling score $\mathbf{p}_0 = 5\mathbf{p}^*$.

the comparisons for $k > 2$ may be biased. This difficulties may be circumvented by using any proper scoring rule that allows for multi-forecast comparison. In fact, maps of this kind may be produced by reporting the score difference between a forecast and the average one. Furthermore, given that proper scores are scale invariant, we can re-scale the score values to be between -1 and 1. In this way, we can visualize which bins has a positive or negative contribution to the average score difference.

The need to compare forecasts defined on different set of bins comes from the design of the forecasting experiment. In the RELM experiment (Zechar et al., 2013) modelers were allowed to choose a subset of bins to include in their forecast, referred to as masking. Modellers involved in the RELM experiment provided forecasts with very different masks; some issued forecasts for the entire California region (Bird and Liu, 2007; Helmstetter et al., 2007; Holliday et al., 2007), some for only Southern California (Ward, 2007; Shen et al., 2007; Kagan et al., 2007), while others used irregular masks (Ebel et al., 2007). The parimutuel gambling score addressed these differences using the gambling metaphor. Each forecaster is a gambler which plays a certain number of rounds (bets) corresponding to the bins. A forecaster does not have to make a forecast for every bin - they can just sit out this round. The forecasters are ranked by their total reward (i.e. the sum of the rewards for each bin). We argue that this solution is still problematic. First, the parimutuel gambling score needs at least two forecasts for each bin to be computed. If only one forecaster *plays* in a bin we can not calculate the parimutuel gambling score for that bin. Second, consider two bins for which different sets of forecasters provided a forecast; in this situation the models are rewarded with respect to different odds. This becomes problematic when we attempt to interpret the observed result because in each bin the reference model is given by a potentially different combination of models.

The Brier score can also be used to assess masked forecasts. The maximum Brier score value obtainable by a forecast is zero and it is achieved by the *perfect* forecast which assumes $p = 1$ when $x = 1$ and $p = 0$ when $x = 0$. Any other forecast obtains a negative Brier score. Therefore, the Brier score of a forecast can be seen as the Brier score difference between the perfect forecast and the forecast under evaluation. Given two models, the one with the highest average Brier score is the one *closest* (on average) to the perfect forecast. If the models provide forecasts on two different sets of bins, B_1 and B_2, we can still compare the forecasts in terms of their average Brier score. Suppose the first model achieves an higher average Brier score, we can conclude that, on average, the first model in B_1 is closer to the perfect forecast than the second model in B_2. We can make the above comparison also when B_1 and B_2 have zero bins in common because the Brier score requires only a forecast and the observation to be calculated.

In this paper, we have used confidence intervals to asses the statistical significance of observed score differences. Analytically determining these confidence intervals may be too complex and a basic approximative Gaussian approach may fail, as highlighted by the Brier score example. Problems of this type come from the fact that the score differences per bin are treated as independent and identically distributed. This assumption is false, especially when considering space-time bins which depend on each other both in time and space due to clustering of earthquakes. A possible solution to relax the independence assumption is to consider a Diebold-Mariano test (Diebold and Mariano, 2002) on the score differences which takes into account the correlation structure of the score differences sequence.

7.2.7 Conclusions

The parimutuel gambling score, commonly applied to compare earthquake forecasts, is improper when the number of forecasts being tested is greater than two. In the special case of two competing forecasts, the score is proper, and can return results similar to alternate proper scoring methods, but even then it can be used improperly. In the common testing scenario of multiple forecasts being compared simultaneously, or when multiple forecasts are compared against a reference model, the parimutuel gambling score provides a biased assessment of the skill of a forecast when it is tested against a given outcome. This is fundamentally a problem of the gambling analogy itself; the betting strategy of maximizing the expected reward (score) does not have to be consistent with the data generating model (in the case where this is known) and, therefore, gamblers (modellers) are not encouraged to provide forecasts resembling the data generating model. This is because the score for a given forecast is dependent on all the forecasts taking part in the competition, not just on the observed data; one can therefore change the ranking of two models by changing one of the other models in the pool. This introduces the undesirable property that one can potentially game the system to prefer a specific model. Further, if we only have access to the forecasts and the data, it is impossible to know if the parimutuel gambling score results will be biased or not. Moreover, the only case in which they are correct is when one of the competing forecasts is the data generating model, which is highly unlikely. These findings are sufficiently clear for us to discourage the use of the parimutuel gambling score in distinguishing between multiple competing forecasts, and for care to be taken even in the case where only two are being compared.

We recommend that alternative scores that do not suffer from these shortcomings should be used instead to assess the skill of prospective earthquake forecasts in a formal testing environment. The Brier and log scores are both proper, and require no new information beyond what was used to calculate the parimutuel gambling score, so switching existing analyses to a proper score should be simple to implement. We recommend testing for properness when introducing new scoring rules, either analytically or via simulations using a known model to generate testing data.

7.3 Conclusion

This paper shows that scores that may sound appealing may be biased and that a fair comparison of the performance of a score should always be made before using it to rank earthquake forecasts. I have shown that this can be done with simulated data for which the data-generation model is known and we know the *right answer* that the score should provide when the data-generating model is compared to alternative models. This method can also be used to calculate the power of the test or to estimate relevant quantities for the experimental design such as the number of events or the number of bins used for the validation stage. I believe it would be helpful for the community to establish *sanity checks* that a score has to pass before being used in CSEP experiments and that also the metrics already employed in CSEP should be tested in this sense. The next chapter discusses this matter more in detail.

Chapter 8

Discussion and Conclusion

8.1 Discussion

In this thesis, I have proposed a novel technique to perform Bayesian inference on the parameters of the ETAS model. The proposed technique is general and can be applied to Hawkes process models different than ETAS. I build on the work done in Bayliss et al. (2020) and Bayliss et al. (2022) which used LGCP models for time-independent models of seismicity with tha ability to incorporate the information from a vast set of available covariates, and that has proven to be competitive with best-performing time-independent models such as (Helmstetter et al., 2007). This approach can be used effectively to produce long-term maps of seismicity as the ones used in PSHA, but it lacks the ability to describe the evolution of a sequence in real-time crucial for any short-term OEF strategy. Indeed, one of the limitations mentioned by the authors, and common at most time-independent models, is the impossibility of explicitly model the clustering process of earthquakes and therefore the inability to accurately describe the time evolution of the sequences. Here, I have shown that the method used in `inlabru` for LGCP models and effective to produce long-term models of seismicity described in Chapter 3 can be adapted to autoregressive processes such as Hawkes process models and therefore, the ETAS model. This is an advance in terms of methodology extending the number of classes supported by the `inlabru` approach and implementable through INLA, and in terms of application extending the number of seismicity analyses implementable through `inlabru`. To make our implementation openly available I have made an R-package called `ETAS.inlabru` to perform analyses of seismicity with our approach. The approach has been used to provide daily forecasts of seismicity for a real-time loss forecasting experiment which are described in Chapter 6. Furthermore, I will submit the model for the next Italian CSEP forecasting experiment starting in 2023.

The proposed method is different from the MCMC methods (Rasmussen, 2013; Ross, 2021) commonly used to perform Bayesian inference on the ETAS parameters. MCMC methods are based on sampling the posterior distribution many times and reconstructing the distribution from the sample; the INLA method, instead, on which this work relies on, is based on a deterministic approximation of the posterior distribution. This makes the INLA method sensibly faster than MCMC methods and enables more complex analyses, with more data in less time. The advantages are particularly evident with models with a large number of parameters presenting a strong correlation structure, a situation that indeed makes MCMC methodology usually inefficient. In Chapter 4, I have compared results obtained with the `ETAS.inlabru` package with the `bayesianETAS` package and shown that `ETAS.inlabru` is up to 10 faster on simulated catalogues with more than 2000 events providing similar results in terms of the number of events, the temporal evolution of the number of events, and a measure of goodness-of-fit. Chapter 5 shows the performance of our approach on various synthetic examples exploring potential biases in parameter estimates due to using sequences that do not represent well the process under study. We have explored the effect of the length of the temporal window considered, the number of large earthquakes in the sequence, the effect of considering a pre-conditioning period, and the effect of data incompleteness. The efficiency of our methodology enables users to fit synthetic models on many different datasets representing difficult situations in order to investigate which estimation problems may be resolved and provide guidance on *good practices* to be followed to reduce the bias in parameter estimates.

Chapter 6 extends the approach used for the temporal ETAS model to the spatio-temporal case with spatially varying background rate, and shows that the method supports also the approach used by Adelfio and Chiodi (2021) and Chiodi et al. (2021) to introduce covariates in modelling the expected number of aftershocks. The authors relies on a frequentist method to estimate the parameters (Chiodi and Adelfio, 2011), and the method I described in this thesis is the first Bayesian implementation of the ETAS model supporting covariates. Using the definition of spatially varying background rate provided in Chapter 6, the proposed method can be used to study the effect of covariates in two ways: in modelling the spatial variation of the background rate and in modelling the expected number of aftershocks. The former in fact is obtained from an LGCP model as the ones used in (Bayliss et al., 2020) and (Bayliss et al., 2022). More formally, the only constraint for the spatial variation of the background rate is to integrate to one. This means that any time-independent or time-dependent map of seismicity can be used and offers a way to introduce the information provided by long-term PSHA maps into short-term models useful for OEF. Similarly, the expected number of aftershocks is modelled as a log-linear function of the covariates and the coefficients can be studied as it is done in a Generalized Linear Model framework, which is commonly employed to study the effect of available covariates on phenomena of interest in many different fields. This allows us to apply to seismic data techniques and ideas borrowed from different fields where hypothesis testing based on observed data is routinely done. This gives the possibility to build more complex models leveraging the additional information provided by covariates and investigate if their use leads to more skillful models in a rigorous Bayesian statistical framework offering a more accurate description of the uncertainty around the results.

The proposed method can be used both for operational tasks or for more explorative analyses. Regarding the former, the approach can be used to construct models to produce forecasts of future seismicity for any OEF or PSHA task. In fact, the spatially-varying background rate can be used as long-term PSHA model, and the full model can be used to decluster a catalogue as it is done in Zhuang et al. (2002). The full model with aftershock can also be used as short-term OEF model. In this situation, the Bayesian framework is not only useful to describe the uncertainty around the parameters but its way to update the information on the parameters can be exploited from an operational point of view. For example, suppose a large earthquake strikes tomorrow, and suppose to already have chosen a model, which parameters should be used? I have shown in Chapter 6 that different sequences occurring in the same area (like the 2009 L'Aquila sequence and the 2016 Amatrice sequence) are best described by a different set of parameters. Which one is better to describe the new sequence? The Bayesian approach offers a way to answer the question. The past data along with expert judgments can be used to construct the prior distribution, and as new data arrives we can fit the model and update our knowledge. Then, we can construct new priors using the information provided by the posterior. Using the posterior as prior as it is may lead to underestimate the uncertainty because the updated posterior variance will likely be smaller than the prior variance, so repeating this process many times will force the posterior variance to shrink toward zero. However, we can set a prior with mean equal to the posterior mean and inflated variance and start the iterative method from the last iteration used to determine the posterior. In this way, the variance will not shrink, it may indeed increase, and the parameters will *move* from the value provided only if the data suggests so. In order to do this, we need a Bayesian framework allowing a certain freedom in specifying the prior and fast enough to be used effectively in near real-time. The approach I propose in this thesis has both. The copula transformation method described in Chapter 3 grants freedom in choosing the prior, it only requires a quantile function to be specified but this is easily obtainable. Furthermore, the method is faster than alternative MCMC methods, scales better increasing the amount of data, and can be parallelised so that is faster on more powerful machines.

On the explorative side, the proposed approach can be used to explore many different covariates in a variety of formats. We have only shown how to use event-specific information like the depth and information regarding the fault network, however, there are many more possibilities that we have not explored yet. For example, more complex fault representations can be used to better associate each event with the corresponding fault and different fault characteristics may be used. In the same way, other spatially varying covariates may be used such as strain rate maps, heat-flow maps, or maps representing historical seismicity. Also, time-varying covariates such as GPS displacement data or maps obtained with Coulomb rate-and-state models may be used and their value as a precursory signal may be tested. Furthermore, each parameter can be considered as spatially or temporally (or both) varying without the use of covariates, but assuming it is described by a Gaussian Markov Random Field. For example, GMRF with Matérn covariance function implemented through the SPDE approach may be used to consider parameters continuously varying in time or in space; Conditional autoregressive models (CAR, Besag, 1974) or the Besag-York-Mollier model (BYM, Besag et al., 1991) can be used in situations where the space is partitioned in sub-regions, the parameter is assumed to be constant in each region and the value of the parameter in adjacent sub-regions are correlated. These models can be used in combination with maps partitioning the space in seismic zones to have discretely varying parameters with similar values in adjacent regions without recurring to data partitioning. All of these different structured random effects are already available in `inlabru` which provides functions to use them efficiently.

Our approach sparked interest from other research groups and we are involved in different parallel projects. We are involved in a study about real-time loss forecasting in a Rapid Loss Assessment framework. The study has the aim of providing a proof of concept on the utility of using data from sensors installed in each building to monitor the evolution of their structural health and therefore of the risk of collapsing, during a seismic sequence. In this study, our isotropic space-time ETAS model with no additional covariates is used to produce catalogue-based forecasts, composed of 10000 synthetic catalogues, of seismicity for the L'Aquila and Amatrice sequences. Our forecasts are then used to estimate the level of ground shaking perceived by each building and are used to predict structural response within the first seconds of ground shaking. Each synthetic catalogue is produced using a different set of parameters sampled from the joint posterior distribution of the parameters which ensures a fair representation of the epistemic uncertainty around the value of the parameter. Additionally, models produced with the proposed apporach will be submitted to the next Italian CSEP forecasting experiment and will be validated prospectively against future seismicity.

Another essential aspect of earthquake forecasting is forecast validation. On one hand we need to be able to build increasingly complex models of seismicity, and I have shown that this can be done with the proposed methodology, and, on the other hand, we need validation metrics with the ability to distinguish between increasingly complex models. This can be done using scoring rules which assign a quantity measuring the ability (if positively oriented) of a model to describe the given data, model can then be ranked accordingly. From the theory around scoring rules, it is known that a score needs to be *proper* in order to provide a fair ranking of the competing models based on observed data and Chapter 7 illustrates the consequences of using improper scoring rules to rank earthquake forecasts and proves that the Parimutuel Gambling score is proper only in a specific situation (when two forecasts are compared against each other) and not in general. The work is relevant because the Parimutuel Gambling score was used improperly to rank earthquake forecasts in previous studies and because it shows how the performance of different scores can be evaluated using synthetic data. The approach of simulating data from a known model and studying the ability of a score in distinguishing between forecasts generated by the data-generating model and alternatives model is a simple way of investigating potential biases in the validation process. It also offers a way to understand which models are distinguishable with a score and to define which score should be used for each situation. The design of the experiment can be decided depending on the competing models. In fact, using simulated data from the model is possible to estimate the amount of data or the binning required to distinguish between the competing models with a given probability level. Scores can also be used to construct new tests using the same idea of CSEP consistency tests employing the score as test statistic. In general, simulation from known models can be used to design *sanity-checks* for validation techniques before introducing them in the next CSEP experiments and proper scores can be used to define new metrics for consistency and comparison tests.

8.2 Future work

The approach I have presented in this thesis has the potential of enabling a deeper and more accessible way to formalize and test scientific hypotheses on the earthquake-generation process as in other fields of science like medicine or psychology where hypotheses testing is done routinely including *blind* (and *double-blind*) testing protocols. To make the presented approach available to a large number of potential users we have already made the `ETAS.inlabru` (Naylor and Serafini, 2023). For the time being, the `ETAS.inlabru` package only supports the temporal ETAS model, and therefore a natural first step is to include also the spatio-temporal ETAS model. We plan to start by including the basic version of the model illustrated in Chapter 6 which includes a spatially varying background field, isotropic spatial kernel, and the linear predictor is based on the magnitude only.

The second step would be to include the possibility to account for temporally varying incompleteness in the catalogue. In doing this, I plan to incorporate temporal incompleteness coming from two sources: the quality of the seismometers network and the censoring induced by large earthquakes. The first one can be accounted having a temporally varying cutoff magnitude. With our approach, in principle, any observation can be associated to a specific cutoff magnitude reflecting the quality of the seismometers network at the time of the observation. This will allow to use catalogues dating further back in time and account for historical seismicity in periods where we have only high magnitude (M5+) observations. The second should reflect the inability of detecting earthquakes in periods of time just after a large earthquake. This can be done assuming to be known the temporal evolution of the rate at which events are missing. An example is the function used in Helmstetter et al. (2006b) and used in Chapter 5 to artificially introduce incompleteness in synthetic catalogues. Knowing the rate at which events are missing it is possible to retrieve the probability of detecting an event. The approach proposed in this thesis can then be modified to work with a modified intensity which is a product of the classical ETAS intensity and the detection probability. I am already doing synthetic experiments to implement this with encouraging results.

A further step would be to include the possibility for the user to introduce covariates. This can be done in two ways with our approach. The covariates can be included in the formulation of the spatial variation of the background field and in the linear predictor present in the expression of the logarithm of the number of aftershocks. We plan to give the user the possibility of using an alternative catalogue to fit the LGCP model determining the spatial variation of the background field so that the redundancy of using the same data for the background and the other ETAS parameters is removed. This will also allow the user to account for historic seismicity in the background, or to use a declustered catalogue for this and a non-declustered one for the other ETAS parameters. This will allow using the output of models such as the one presented in (Bayliss et al., 2020) and (Bayliss et al., 2022) to determine the spatial variation of the background field. Furthermore, including time-varying covariates in the spatial variation of the background rate yields a spatio-temporally varying background rate which is rarely part of earthquake models. This can also be done considering a spatio-temporal GMRF as a structured random effect which is already supported by INLA and `inlabru` as described in Chapter 7 of Blangiardo and Cameletti (2015).

The second way in which covariates can be introduced is using the approach described in Chapter 6 and proposed by Adelfio and Chiodi (2021) and Chiodi et al. (2021). In this context, there are different ways in which this approach can be extended. The first and most natural one is to consider more and different covariates. In doing this, we can consider spatially varying covariates like strain rates maps, heatflow maps, fault characteristics, and GPS measurements quantifying the earth displacement rate, or alternatively, we can use the output from other models, such as spatio-temporally varying Coulomb maps and test in a statistically rigorous way if they provide additional information and increase our ability to forecast future seismicity. There is also the possibility to use maps dividing the area in different seismic zones such as the ZS9 map for Italy (Stucchi et al., 2004). In this regard, using random effects varying discretely over space such as the CAR Besag (1974) of the BYM Besag et al. (1991) model it would be possible to include in the linear predictor a spatially varying coefficient constant in each zone and with correlated values in adjacent zones. This can be done also using fault maps in order to have fault-specific parameters with similar values for adjacents faults. In the same way, products such as Peak Ground Acceleration (PGA) maps which are routinely produced for seismic hazard analyses can be incorporated into the linear predictor. A further option would be to consider non-linear functions of the covariates which also are supported by `inlabru`. For example, we can estimate a non-linear function of a covariate using the SPDE approach described in Chapter 3 including in the linear predictor a one-dimensional GMRF with Matérn covariance function on the covariate domain. This will produce a smooth function of the covariate because the correlation between values of the GMRF for different values of the covariate is a function of the distance on the covariate dimension so that similar values of the covariate have similar values of the effect.

Another important aspect on which the approach proposed in this thesis can be generalized is to allow for an anisotropic spatial kernel. Ideally, our plan is to use an approach similar to the one illustrated in Grimm et al. (2022a) in which they consider an isotropic spatial kernel for events below a certain magnitude and an anisotropic one for events above a certain magnitude. This is can be seen as generalizing the approach used to introduce covariates in the modelling of the expected number of aftershocks. Essentially, we are substituting parameters with functions of covariates (e.g. the magnitude). But this idea, in principle, can be used also for other parameters. For example, it can be assumed that the material of the lithosphere at the depth at which an event occurs will influence the characteristics of its aftershocks. Therefore, it would interesting to study the effect of this covariate not only on the number of aftershocks but also on their spatio-temporal distribution. Our plan is to extend the approach used to introduce covariates in modelling the expected number of aftershocks also to other ETAS parameters. This will give the possibility to use our approach to study non-isotropic, non-stationary processes and if they add values in forecasting future earthquakes.

Regarding the CSEP testing procedures, we plan to write an article revising the M-test and comparing its performances with alternative methods. I explain the issue on the GitHub page of the `pyCSEP` python library which can be found at https://github.com/SECcode/pycsep/issues/196. Basically, the issue is that the M-test as presented in Savran et al. (2020) is biased when a forecast overpredicts the total number of events. I have proposed a solution and also designed a modified multinomial likelihood score test that does not suffer from this problem and is more powerful than the M-test. We plan to write an article assessing the performance of the M-test similarly to what is done in Khawaja et al. (2023) for the S-test.

8.3 Conclusion

In this thesis, I have presented a new approximation method to perform inference on Bayesian Hawkes process models with applications to seismicity and the ETAS model. The method is general and has the potential to be applied to different Hawkes processes than the ETAS one although we have only considered this case here. I have shown that the proposed methodology produces similar results to alternative Bayesian methods (MCMC) but is faster and scales more efficiently increasing the number of events per catalogue. I have also shown how the proposed technique can be generalized to include covariates in modelling the expected number of aftershocks.

The proposed methodology has the potential of greatly simplifying the process of incorporating hypotheses on the earthquake process in a model and then, testing them against observed data. This is fundamental to building more flexible models of seismicity and increasing our level of knowledge on the earthquake generation process. I envision continuing to work on extending this approach and making it available to the wider public as a freely available R-package could be relevant for many applied researchers in this field. The final goal is to provide researchers with a playground in which they can easily formulate, implement and test hypotheses on the earthquake-generation process without the burden of coding ad-hoc algorithms and limiting the number of subjective choices that they need to make. This will increase the reliability and reproducibility of the results.

To do that, another fundamental part is the testing metrics. I believe that much work is still needed to improve the existing ones and propose new ones accounting for other aspects of the process but we are on the right path. The efforts made by CSEP in the next several years should be the basis to deepen the discussion around testing metrics for earthquake forecasts and to design new prospective experiments.

References

Giada Adelfio and Marcello Chiodi. Alternated estimation in semi-parametric space-time branching-type point processes with application to seismic catalogs. *Stochastic Environmental Research and Risk Assessment*, 29(2):443–450, 2015.

Giada Adelfio and Marcello Chiodi. Including covariates in a space-time point process with application to seismicity. *Statistical Methods & Applications*, 30(3):947–971, 2021.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Keiiti Aki. Maximum likelihood estimate of b in the formula logn= a-bm and its confidence limits. *Bull. Earthquake Res. Inst. Univ. Tokyo*, 43:237–239, 1965.

Linda Altieri, Alessio Farcomeni, and Danilo Alunni Fegatelli. Continuous time-interaction processes for population size estimation, with an application to drug dealing in italy. *Biometrics*, 2022.

Hartwig Anzt, Felix Bach, Stephan Druskat, Frank Löffler, Axel Loewe, Bernhard Y Renard, Gunnar Seemann, Alexander Struck, Elke Achhammer, Piush Aggarwal, et al. An environment for sustainable research software in germany and beyond: current state, open challenges, and call for action. *F1000Research*, 9, 2020.

Krishna B Athreya, Peter E Ney, and PE Ney. *Branching processes*. Courier Corporation, 2004.

Shahriar Azizpour, Kay Giesecke, and Gustavo Schwenkler. Exploring the sources of default clustering. *Journal of Financial Economics*, 129(1):154–183, 2018.

Christoph Bach and Sebastian Hainzl. Improving empirical aftershock modeling based on additional source information. *Journal of Geophysical Research: Solid Earth*, 117(B4), 2012.

Fabian E Bachl, Finn Lindgren, David L Borchers, and Janine B Illian. inlabru: an r package for bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766, 2019.

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

Jack Baker, Brendon Bradley, and Peter Stafford. *Seismic hazard and risk analysis*. Cambridge University Press, 2021.

Haakon Bakka, Håvard Rue, Geir-Arne Fuglstad, Andrea Riebler, David Bolin, Janine Illian, Elias Krainski, Daniel Simpson, and Finn Lindgren. Spatial modeling with r-inla: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):e1443, 2018.

Haakon Bakka, Jarno Vanhatalo, Janine B Illian, Daniel Simpson, and Håvard Rue. Non-stationary gaussian models with physical barriers. *Spatial statistics*, 29:268–288, 2019.

WH Bakun, B Aagaard, B Dost, WL Ellsworth, JL Hardebeck, RA Harris, C Ji, MJS Johnston, J Langbein, JJ Lienkaemper, et al. Implications for prediction and hazard assessment from the 2004 parkfield earthquake. *Nature*, 437(7061):969–974, 2005.

William H Bakun and Allan G Lindh. The parkfield, california, earthquake prediction experiment. *Science*, 229(4714):619–624, 1985.

Earvin Balderama, Frederic Paik Schoenberg, Erin Murray, and Philip W Rundel. Application of branching models in the study of invasive species. *Journal of the American Statistical Association*, 107(498):467–476, 2012.

Roberto Basili, Pierfrancesco Burrato, Giorgio Maria De Santis, Umberto Fracassi, Francesco Emanuele Maesano, Gabriele Tarabusi, Mara Monica Tiberti, Gianluca Valensise, Roberto Vallone, Paola Vannoli, et al. Database of individual seismogenic sources (diss), version 3.3. 0: A compilation of potential sources for earthquakes larger than m 5.5 in italy and surrounding areas. 2021.

K. Bayliss, M. Naylor, F. Kamranzad, and I. Main. Pseudo-prospective testing of 5-year earthquake forecasts for california using inlabru. *Natural Hazards and Earth System Sciences*, 22(10):3231–3246, 2022. doi: 10.5194/nhess-22-3231-2022. URL https://nhess.copernicus.org/articles/22/3231/2022/.

Kirsty Bayliss, Mark Naylor, Janine Illian, and Ian G Main. Data-driven optimization of seismicity models using diverse data sets: Generation, evaluation, and ranking using inlabru. *Journal of Geophysical Research: Solid Earth*, 125(11):e2020JB020226, 2020.

JA Bayona, W Savran, Anne Strader, S Hainzl, Fabrice Cotton, and Danijel Schorlemmer. Two global ensemble seismicity models obtained from the combination of interseismic strain measurements and earthquake-catalogue information. *Geophysical Journal International*, 224(3):1945–1955, 2021.

Bernice Bender. Maximum likelihood estimation of b values for magnitude grouped data. *Bulletin of the Seismological Society of America*, 73(3):831–851, 1983.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20, 1991.

Peter Bird and Zhen Liu. Seismic hazard inferred from tectonics: California. *Seismological Research Letters*, 78(1):37–48, 2007.

Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

Marta Blangiardo, Michela Cameletti, Gianluca Baio, and Håvard Rue. Spatial and spatio-temporal models with r-inla. *Spatial and spatio-temporal epidemiology*, 4:33–49, 2013.

Margaret S Boettcher, Art McGarr, and Malcolm Johnston. Extension of gutenberg-richter distribution to mw- 1.3, no lower limit in sight. *Geophysical Research Letters*, 36(10), 2009.

David Bolin and Kristin Kirchner. The rational spde approach for gaussian random fields with general smoothness. *Journal of Computational and Graphical Statistics*, 29(2):274–285, 2020.

SJ Bourne, SJ Oates, and J Van Elk. The exponential rise of induced seismicity with increasing stress levels in the groningen gas field and its implications for controlling seismic risk. *Geophysical Journal International*, 213(3):1693–1700, 2018.

Andrew Bray and Frederic Paik Schoenberg. Assessment of point process models for earthquake forecasting. *Statistical science*, 28(4):510–520, 2013.

Jonas Brehmer, Tilmann Gneiting, Martin Schlather, and Kirstin Strokorb. Using scoring functions to evaluate point process forecasts. *arXiv preprint arXiv:2103.11884*, 2021.

Pierre Brémaud. *Point processes and queues: martingale dynamics*, volume 50. Springer, 1981.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

RJ Budnitz, G Apostolakis, and David M Boore. Recommendations for probabilistic seismic hazard analysis: guidance on uncertainty and use of experts. Technical report, US Nuclear Regulatory Commission (NRC), Washington, DC (United States). Div . . . , 1997.

Mauro Buttinelli, Lorenzo Petracchini, Francesco Emanuele Maesano, Chiara D'Ambrogi, Davide Scrocca, Maurizio Marino, Franco Capotorti, Sabina Bigi, Gian Paolo Cavinato, Maria Teresa Mariucci, et al. The impact of structural complexity, fault segmentation, and reactivation on seismotectonics: Constraints from the upper crust of the 2016–2017 central italy seismic sequence area. *Tectonophysics*, 810:228861, 2021.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Camilla Cattania, Maximilian J Werner, Warner Marzocchi, Sebastian Hainzl, David Rhoades, Matthew Gerstenberger, Maria Liukis, William Savran, Annemarie Christophersen, Agnès Helmstetter, et al. The forecasting skill of physics-based seismicity models during the 2010–2012 canterbury, new zealand, earthquake sequence. *Seismological Research Letters*, 89(4):1238–1250, 2018.

Wen-Hao Chiang, Xueying Liu, and George Mohler. Hawkes process modeling of covid-19 with mobility leading indicators and spatial covariates. *International journal of forecasting*, 38(2):505–520, 2022.

Marcello Chiodi and Giada Adelfio. Forward likelihood-based predictive approach for space–time point processes. *Environmetrics*, 22(6):749–757, 2011.

Marcello Chiodi, Orietta Nicolis, Giada Adelfio, Nicoletta D'angelo, and Alex Gonzàlez. Etas space–time modeling of chile triggered seismicity using covariates: Some preliminary results. *Applied Sciences*, 11(19):9143, 2021.

Kim Christensen and Zeev Olami. Variation of the gutenberg-richter b values and nontrivial temporal correlations in a spring-block model for earthquakes. *Journal of Geophysical Research: Solid Earth*, 97(B6):8729–8735, 1992.

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

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Charles B Connor, Neil A Chapman, and Laura J Connor. *Volcanic and tectonic hazard assessment for nuclear facilities*. Cambridge University Press, 2009.

C Allin Cornell. Engineering seismic risk analysis. *Bulletin of the seismological society of America*, 58(5):1583–1606, 1968.

P Cosentino, V Ficarra, and D Luzio. Truncated exponential frequency-magnitude relationship in earthquake statistics. *Bulletin of the Seismological Society of America*, 67(6): 1615–1623, 1977.

David R Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(2):129–157, 1955.

Helen Crowley, Julian J Bommer, Rui Pinho, and Juliet Bird. The impact of epistemic uncertainty on an earthquake loss model. *Earthquake engineering & structural dynamics*, 34(14):1653–1685, 2005.

Daryl J Daley and David Vere-Jones. Scoring probability forecasts for point processes: The entropy score and information gain. *Journal of Applied Probability*, 41(A):297–312, 2004.

Daryl J Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*. Springer, 2008.

Daryl J Daley, David Vere-Jones, et al. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.

James E Daniell, B Khazai, F Wenzel, and A Vervaeck. The catdat damaging earthquakes database. *Natural Hazards and Earth System Sciences*, 11(8):2235–2251, 2011.

Scott D Davis and Cliff Frohlich. Single-link cluster analysis of earthquake aftershocks: Decay laws and regional variations. *Journal of Geophysical Research: Solid Earth*, 96(B4): 6335–6350, 1991.

Francis X Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–1, 2015.

Francis X Diebold and Robert S Mariano. Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144, 2002.

Sophie Donnet, Vincent Rivoirard, and Judith Rousseau. Nonparametric bayesian estimation for multivariate hawkes processes. *The Annals of Statistics*, 48(5):2698–2727, 2020.

John Douglas. Earthquake ground motion estimation using strong-motion records: a review of equations for the estimation of peak ground acceleration and response spectral ordinates. *Earth-Science Reviews*, 61(1-2):43–104, 2003.

John E Ebel. The importance of small earthquakes. *Seismological Research Letters*, 79(4): 491–493, 2008.

John E Ebel, Daniel W Chambers, Alan L Kafka, and Jenny A Baglivo. Non-poissonian earthquake clustering and the hidden markov model as bases for earthquake forecasting in california. *Seismological Research Letters*, 78(1):57–65, 2007.

David AJ Eberhard, J Douglas Zechar, and Stefan Wiemer. A prospective earthquake forecast experiment in the western pacific. *Geophysical Journal International*, 190(3):1579–1592, 2012.

Hossein Ebrahimian and Fatemeh Jalayer. Robust seismicity forecasting based on bayesian parameter estimation for epidemiological spatio-temporal aftershock clustering models. *Scientific reports*, 7(1):1–15, 2017.

Hossein Ebrahimian, Fatemeh Jalayer, Domenico Asprone, Anna Maria Lombardi, Warner Marzocchi, Andrea Prota, and Gaetano Manfredi. Adaptive daily forecasting of seismic aftershock hazard. *Bulletin of the Seismological Society of America*, 104(1):145–161, 2014.

Hossein Ebrahimian, Fatemeh Jalayer, Behnam Maleki Asayesh, Sebastian Hainzl, and Hamid Zafarani. Improvements to seismicity forecasting based on a bayesian spatio-temporal etas model. *Scientific Reports*, 12(1):20970, 2022.

Zuhair H El-Isa and David W Eaton. Spatiotemporal variations in the b-value of earthquake magnitude–frequency distributions: Classification and causes. *Tectonophysics*, 615:1–11, 2014.

SA Fedotov. Regularities of the distribution of strong earthquakes in kamchatka, the kuril islands, and northeastern japan. *Tr. Inst. Phys. Earth Acad. Sci. USSR*, 36:66–93, 1965.

Edward H Field. Overview of the working group for the development of regional earthquake likelihood models (relm). *Seismological Research Letters*, 78(1):7–16, 2007.

Edward H Field, Ramon J Arrowsmith, Glenn P Biasi, Peter Bird, Timothy E Dawson, Karen R Felzer, David D Jackson, Kaj M Johnson, Thomas H Jordan, Christopher Madden, et al. Uniform california earthquake rupture forecast, version 3 (ucerf3)—the time-independent model. *Bulletin of the Seismological Society of America*, 104(3):1122–1180, 2014.

Edward H Field, Kevin R Milner, Jeanne L Hardebeck, Morgan T Page, Nicholas van der Elst, Thomas H Jordan, Andrew J Michael, Bruce E Shaw, and Maximilian J Werner. A spatiotemporal clustering model for the third uniform california earthquake rupture forecast (ucerf3-etas): Toward an operational earthquake forecast. *Bulletin of the Seismological Society of America*, 107(3):1049–1081, 2017.

Edward H Field, Kevin R Milner, Morgan T Page, William H Savran, and Nicholas Van Der Elst. Improvements to the third uniform california earthquake rupture forecast etas model (ucerf3-etas). *The Seismic Record*, 1(2):117–125, 2021.

Vladimir Filimonov and Didier Sornette. Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Physical Review E*, 85(5):056108, 2012.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

Tobias Fissler and Johanna F Ziegel. Higher order elicitability and osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.

Chiara Forlani, Samir Bhatt, Michela Cameletti, Elias Krainski, and Marta Blangiardo. A joint bayesian space–time model to integrate spatially misaligned air pollution data in r-inla. *Environmetrics*, 31(8):e2644, 2020.

Arthur D Frankel, Mark D Petersen, Charles S Mueller, Kathleen M Haller, Russell L Wheeler, EV Leyendecker, Robert L Wesson, Stephen C Harmsen, Chris H Cramer, David M Perkins, et al. Documentation for the 2002 update of the national seismic hazard maps. *US Geological Survey open-file report*, 420:39, 2002.

David A Freedman and Philip B Stark. What is the chance of an earthquake. *NATO Science Series IV: Earth and Environmental Sciences*, 32:201–213, 2003.

Rafael M Frongillo and Ian A Kash. Elicitation complexity of statistical properties. *Biometrika*, 108(4):857–879, 2021.

Geir-Arne Fuglstad, Finn Lindgren, Daniel Simpson, and Håvard Rue. Exploring a new class of non-stationary spatial gaussian random fields with varying local anisotropy. *Statistica Sinica*, pages 115–133, 2015.

JK Gardner and Leon Knopoff. Is the sequence of earthquakes in southern california, with aftershocks removed, poissonian? *Bulletin of the seismological society of America*, 64(5): 1363–1367, 1974.

Michele Garetto, Emilio Leonardi, and Giovanni Luca Torrisi. A time-modulated hawkes process to model the spread of covid-19 and the impact of countermeasures. *Annual reviews in control*, 51:551–563, 2021.

GM Geffers, IG Main, and Mark Naylor. Biases in estimating b-values from small earthquake catalogues: how high are high b-values? *Geophysical Journal International*, 229(3):1840–1855, 2022.

Robert J Geller. Shake-up time for japanese seismology. *Nature*, 472(7344):407–409, 2011.

Robert J Geller, Francesco Mulargia, and Philip B Stark. Why we need a new paradigm of earthquake occurrence. *Subduction dynamics: from mantle flow to mega disasters*, pages 183–191, 2015.

Matthew C Gerstenberger and David A Rhoades. New zealand earthquake forecast testing centre. In *Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II*, pages 23–38. Springer, 2010.

MC Gerstenberger, DA Rhoades, MW Stirling, R Brownrigg, and A Christophersen. Continued development of the new zealand earthquake forecast testing centre. *GNS Science Consultancy Report 2009*, 182, 2009.

Grove Karl Gilbert. A theory of the earthquakes of the great basin, with a practical application. *American Journal of Science*, 3(157):49–53, 1884.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Thomas HW Goebel, Grzegorz Kwiatek, Thorsten W Becker, Emily E Brodsky, and Georg Dresen. What allows seismic events to grow big?: Insights from b-value and fault roughness analysis in laboratory stick-slip experiments. *Geology*, 45(9):815–818, 2017.

MICHAEL Goldstein. Observables and models: exchangeability and the inductive argument. *Bayesian Theory and Its Applications*, pages 3–18, 2013.

Virgilio Gómez-Rubio. *Bayesian inference with INLA*. CRC Press, 2020.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society, Ser. B*, pages 107–114, 1952.

IJ Goodd and Ray A Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

D Gospodinov and R Rotondi. Statistical analysis of triggered seismicity in the kresna region of sw bulgaria (1904) and the umbria-marche region of central italy (1997). *Pure and applied geophysics*, 163(8):1597–1615, 2006.

Christian Grimm, Sebastian Hainzl, Martin Käser, and Helmut Küchenhoff. Solving three major biases of the etas model to improve forecasts of the 2019 ridgecrest sequence. *Stochastic Environmental Research and Risk Assessment*, pages 1–20, 2022a.

Christian Grimm, Martin Käser, Sebastian Hainzl, Marco Pagani, and Helmut Küchenhoff. Improving earthquake doublet frequency predictions by modified spatial trigger kernels in the epidemic-type aftershock sequence (etas) model. *Bulletin of the Seismological Society of America*, 112(1):474–493, 2022b.

ISIDe Working Group. Italian seismological instrumental and parametric database, 2007.

CPTI Gruppo di Lavoro. Catalogo parametrico dei terremoti italiani, versione 2004 (cpti04). ingv, bologna, 2004a.

MPS Gruppo di Lavoro. Redazione della mappa di pericolosità sismica prevista dall'ordinanza pcm 3274 del 20 marzo 2003. *Rapporto Conclusivo per il Dipartimento della Protezione Civile, INGV, Milano-Roma*, 5, 2004b.

Zhenqi Guo and Yosihiko Ogata. Statistical relations between the parameters of aftershocks in time, space, and magnitude. *Journal of Geophysical Research: Solid Earth*, 102(B2): 2857–2873, 1997.

Beno Gutenberg and Charles F Richter. Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, 34(4):185–188, 1944.

Beno Gutenberg and Charles Francis Richter. Magnitude and energy of earthquakes. *Annals of Geophysics*, 9(1):1–15, 1956.

Sebastian Hainzl. Rate-dependent incompleteness of earthquake catalogs. *Seismological Research Letters*, 87(2A):337–344, 2016a.

Sebastian Hainzl. Rate-Dependent Incompleteness of Earthquake Catalogs. *Seismological Research Letters*, 87(2A):337–344, 02 2016b. ISSN 0895-0695. doi: 10.1785/0220150211. URL https://doi.org/10.1785/0220150211.

Sebastian Hainzl. Apparent triggering function of aftershocks resulting from rate-dependent incompleteness of earthquake catalogs. *Journal of Geophysical Research: Solid Earth*, 121(9):6499–6509, 2016c. doi: https://doi.org/10.1002/2016JB013319. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JB013319.

Sebastian Hainzl. Etas-approach accounting for short-term incompleteness of earthquake catalogs. *Bulletin of the Seismological Society of America*, 112(1):494–507, 2022.

Sebastian Hainzl and David Marsan. Dependence of the omori-utsu law parameters on main shock magnitude: Observations and modeling. *Journal of Geophysical Research: Solid Earth*, 113(B10), 2008.

Sebastian Hainzl, A Christophersen, and B Enescu. Impact of earthquake rupture extensions on parameter estimations of point-process models. *Bulletin of the Seismological Society of America*, 98(4):2066–2072, 2008.

Sebastian Hainzl, Olga Zakharova, and David Marsan. Impact of aseismic transients on the estimation of aftershock productivity parameters. *Bulletin of the Seismological Society of America*, 103(3):1723–1732, 2013.

Jaana I Halonen, Anna L Hansell, John Gulliver, David Morley, Marta Blangiardo, Daniela Fecht, Mireille B Toledano, Sean D Beevers, Hugh Ross Anderson, Frank J Kelly, et al. Road traffic noise is associated with increased cardiovascular morbidity and mortality and all-cause mortality in london. *European heart journal*, 36(39):2653–2661, 2015.

Thomas C Hanks, Norm A Abrahamson, David M Boore, Kevin J Coppersmith, and Nichole E Knepprath. Implementation of the sshac guidelines for level 3 and 4 pshas—experience gained from actual applications. *US Geological Survey Open-File Report*, 1093:66, 2009.

Thomas C Hanks, Gregory C Beroza, and Shinji Toda. Have recent earthquakes exposed flaws in or misunderstandings of probabilistic seismic hazard analysis? *Seismological Research Letters*, 83(5):759–764, 2012.

Theodore Edward Harris et al. *The theory of branching processes*, volume 6. Springer Berlin, 1963.

David Harte and David Vere-Jones. The entropy score and its uses in earthquake forecasting. *Pure and Applied Geophysics*, 162(6):1229–1253, 2005.

DS Harte. Bias in fitting the etas model: a case study based on new zealand seismicity. *Geophysical Journal International*, 192(1):390–412, 2013.

DS Harte. Log-likelihood of earthquake models: evaluation of models and forecasts. *Geophysical Journal International*, 201(2):711–723, 2015.

DS Harte. Model parameter estimation bias induced by earthquake magnitude cut-off. *Geophysical Journal International*, 204(2):1266–1287, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Egill Hauksson, Wenzheng Yang, and Peter M Shearer. Waveform relocated earthquake catalog for southern california (1981 to june 2011). *Bulletin of the Seismological Society of America*, 102(5):2239–2244, 2012.

Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971a.

Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971b.

Alan G Hawkes. Hawkes processes and their applications to finance: a review. *Quantitative Finance*, 18(2):193–198, 2018.

Matthew J Heaton, Abhirup Datta, Andrew O Finley, Reinhard Furrer, Joseph Guinness, Rajarshi Guhaniyogi, Florian Gerber, Robert B Gramacy, Dorit Hammerling, Matthias Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019.

Claudio Heinrich-Mertsching, Thordis L Thorarinsdottir, Peter Guttorp, and Max Schneider. Validation of point process predictions with proper scoring rules. *arXiv preprint arXiv:2110.11803*, 2021.

Agnes Helmstetter, Yan Y Kagan, and David D Jackson. Comparison of short-term and time-independent earthquake forecast models for southern california. *Bulletin of the Seismological Society of America*, 96(1):90–106, 2006a.

Agnes Helmstetter, Yan Y Kagan, and David D Jackson. High-resolution time-independent grid-based forecast for m $\geq$ 5 earthquakes in california. *Seismological Research Letters*, 78(1):78–86, 2007.

Agnès Helmstetter, Yan Y. Kagan, and David D. Jackson. Comparison of Short-Term and Time-Independent Earthquake Forecast Models for Southern California. *Bulletin of the Seismological Society of America*, 96(1):90–106, 2006b. ISSN 0037-1106. doi: 10.1785/0120050067.

Ernest J Henley and Hiromitsu Kumamoto. Probabilistic risk assessment and management for engineers and scientists. *IEEE Press (2nd Edition)*, 1996.

José Hernández-Orallo, Peter Flach, and César Ferri Ramírez. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869, 2012.

Marcus Herrmann and Warner Marzocchi. Inconsistencies and lurking pitfalls in the magnitude–frequency distribution of high-resolution earthquake catalogs. *Seismological Research Letters*, 92(2A):909–922, 2021.

Marcus Herrmann, Ester Piegari, and Warner Marzocchi. Revealing the spatiotemporal complexity of the magnitude distribution and b-value during an earthquake sequence. *Nature communications*, 13(1):1–10, 2022.

Andrew J Holbrook, Charles E Loeffler, Seth R Flaxman, and Marc A Suchard. Scalable bayesian inference for self-excitatory stochastic processes applied to big american gunfire data. *Statistics and computing*, 31(1):1–15, 2021.

James R Holliday, Kazuyoshi Z Nanjo, Kristy F Tiampo, John B Rundle, and Donald L Turcotte. Earthquake forecasting and its verification. *arXiv preprint cond-mat/0508476*, 2005.

James R Holliday, Chien-chih Chen, Kristy F Tiampo, John B Rundle, Donald L Turcotte, and Andrea Donnellan. A relm earthquake forecast based on pattern informatics. *Seismological Research Letters*, 78(1):87–93, 2007.

M Holschneider, C Narteau, P Shebalin, Z Peng, and Danijel Schorlemmer. Bayesian analysis of the modified omori law. *Journal of Geophysical Research: Solid Earth*, 117(B6), 2012.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

Mireille Huc and Ian G Main. Anomalous stress diffusion in earthquake triggering: Correlation length, time dependence, and directionality. *Journal of Geophysical Research: Solid Earth*, 108(B7), 2003.

John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

Allen Husker, Maximilian J Werner, José A Bayona, Miguel Santoyo, and Raul Daniel Corona-Fernandez. A test of the earthquake gap hypothesis in mexico: The case of the guerrero gap. *Bulletin of the Seismological Society of America*, 113(1):468–479, 2023.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Iunio Iervolino, Eugenio Chioccarelli, Massimiliano Giorgio, Warner Marzocchi, Giulio Zuccaro, Mauro Dolce, and Gaetano Manfredi. Operational (short-term) earthquake loss forecasting in italy. *Bulletin of the Seismological Society of America*, 105(4):2286–2298, 2015.

Iunio Iervolino, Eugenio Chioccarelli, and Massimiliano Giorgio. Aftershocks' effect on structural design actions in italy. *Bulletin of the Seismological Society of America*, 108(4): 2209–2220, 2018.

David D Jackson and Yan Y Kagan. The 2004 parkfield earthquake, the 1985 prediction, and characteristic earthquakes: Lessons for the future. *Bulletin of the Seismological Society of America*, 96(4B):S397–S409, 2006.

David D Jackson and Mitsuhiro Matsu'Ura. A bayesian approach to nonlinear inversion. *Journal of Geophysical Research: Solid Earth*, 90(B1):581–591, 1985.

Abdollah Jalilian. Etas: An r package for fitting the space-time etas model to earthquake data. *Journal of Statistical Software, Code Snippets*, 88(1):1–39, 2019. doi: 10.18637/jss.v088.c01.

Harold Jeffreys. *The theory of probability*. OUP Oxford, 1998.

IT Jolliffe and DB Stephenson. Forecast verification. chichester, england and hoboken, 2003.

Lucile M Jones. Foreshocks and time-dependent earthquake hazard assessment in southern california. *Bulletin of the Seismological Society of America*, 75(6):1669–1679, 1985.

T Jordan, Y-T Chen, Paolo Gasparini, Raul Madariaga, Ian Main, Warner Marzocchi, Gerassimos Papadopoulos, K Yamaoka, J Zschau, et al. Operational earthquake forecasting: State of knowledge and guidelines for implementation. *Annals of Geophysics*, 2011.

TH Jordan, W Marzocchi, AJ Michael, and MC Gerstenberger. Operational earthquake forecasting can enhance earthquake preparedness, 2014.

Thomas H Jordan. Earthquake predictability, brick by brick. *Seismological Research Letters*, 77(1):3–6, 2006.

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

Yan Kagan, David D Jackson, Robert J Geller, et al. Characteristic earthquake model, 1884–2011, rip. *arXiv preprint arXiv:1207.4836*, 2012.

Yan Y Kagan. Likelihood analysis of earthquake catalogues. *Geophysical journal international*, 106(1):135–148, 1991.

Yan Y Kagan. Seismic moment distribution revisited: I. statistical results. *Geophysical Journal International*, 148(3):520–541, 2002.

Yan Y Kagan. Short-term properties of earthquake catalogs and models of earthquake source. *Bulletin of the Seismological Society of America*, 94(4):1207–1228, 2004.

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

Yan Y Kagan. *Earthquakes: models, statistics, testable forecasts*. John Wiley & Sons, 2013.

Yan Y Kagan and L Knopoff. Statistical short-term earthquake prediction. *Science*, 236 (4808):1563–1567, 1987.

Yan Y Kagan and F Schoenberg. Estimation of the upper cutoff parameter for the tapered pareto distribution. *Journal of Applied Probability*, 38(A):158–175, 2001.

Yan Y Kagan and David Vere-Jones. Problems in the modelling and statistical analysis of earthquakes. In *Athens Conference on Applied Probability and Time Series Analysis*, pages 398–425. Springer, 1996.

Yan Y Kagan, David D Jackson, and Yufang Rong. A testable five-year forecast of moderate and large earthquakes in southern california based on smoothed seismicity. *Seismological Research Letters*, 78(1):94–98, 2007.

YY Kagan. Statistical methods in the study of seismic processes. *Bull. Int. Stat. Inst*, 45 (3):437–453, 1973.

YY Kagan. Universality of the seismic moment-frequency relation. In *Seismicity patterns, their statistical significance and physical meaning*, pages 537–573. Springer, 1999.

YY Kagan and DD Jackson. New seismic gap hypothesis: Five years after. *Journal of Geophysical Research: Solid Earth*, 100(B3):3943–3959, 1995.

Immanuel Kant. *Geschichte und Naturbeschreibung der merkewürdigsten Vorfälle des Erdebens welches an dem Ende des 1755sten Jahres einen grossen Theil der Erde erschüttert hat.* gedruckt und verlegt von Joh. Heinrich Hartung, 1756.

Amato Kasahara, Yuji Yagi, and Bogdan Enescu. etas_solve: A robust program to estimate the etas model parameters. *Seismological Research Letters*, 87(5):1143–1149, 2016.

Asim M Khawaja, Sebastian Hainzl, Danijel Schorlemmer, Pablo Iturrieta, Jose A Bayona, William H Savran, Maximilian Werner, and Warner Marzocchi. Statistical power of spatial earthquake forecast tests. *Geophysical Journal International*, page ggad030, 2023.

Carl Kisslinger. The stretched exponential function as an alternative model for aftershock decay rate. *Journal of Geophysical Research: Solid Earth*, 98(B2):1913–1921, 1993.

Carl Kisslinger and Lucile M Jones. Properties of aftershock sequences in southern california. *Journal of Geophysical Research: Solid Earth*, 96(B7):11947–11958, 1991.

Ryota Kobayashi and Renaud Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

Kossobokov. Earthquake prediction: basics, achievements, perspectives. *Acta Geodaetica et Geophysica Hungarica*, 39(2-3):205–221, 2004.

Kossobokov. Testing earthquake prediction methods: The west pacific short-term forecast of earthquakes with magnitude mwhr $\geq$ 5.8. *Tectonophysics*, 413(1-2):25–31, 2006.

RB Kulkarni, RR Youngs, and KJ Coppersmith. Assessment of confidence intervals for results of seismic hazard analysis. In *Proceedings of the eighth world conference on earthquake engineering*, volume 1, pages 263–270, 1984.

Shinyoung Kwag and Abhinav Gupta. Probabilistic risk assessment framework for structural systems under multiple hazards using bayesian statistics. *Nuclear Engineering and Design*, 315:20–34, 2017.

Patrick J Laub, Young Lee, and Thomas Taimre. The elements of hawkes processes, 2021.

PA W Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

Xiaoting Li, Christian Genest, and Jonathan Jalbert. A self-exciting marked point process model for drought analysis. *Environmetrics*, 32(8):e2697, 2021.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Finn Lindgren, David Bolin, and Håvard Rue. The spde approach for gaussian and non-gaussian fields: 10 years and still running. *Spatial Statistics*, page 100599, 2022.

Barbara Lolli, Daniele Randazzo, Gianfranco Vannucci, and Paolo Gasperini. The homogenized instrumental seismic catalog (horus) of italy from 1960 to present. *Seismological Research Letters*, 91(6):3208–3222, 2020.

AM Lombardi and W Marzocchi. The assumption of poisson seismic-rate variability in csep/relm experiments. *Bulletin of the Seismological Society of America*, 100(5A):2293–2300, 2010a.

Anna Maria Lombardi. Estimation of the parameters of etas models by simulated annealing. *Scientific reports*, 5(1):1–11, 2015.

Anna Maria Lombardi. The epistemic and aleatory uncertainties of the etas-type models: an application to the central italy seismicity. *Scientific reports*, 7(1):1–9, 2017.

Anna Maria Lombardi. Anomalies and transient variations of b-value in italy during the major earthquake sequences: what truth is there to this? *Geophysical Journal International*, 232 (3):1545–1555, 2022.

Anna Maria Lombardi and Warner Marzocchi. The etas model for daily forecasting of italian seismicity in the csep experiment. *Annals of Geophysics*, 2010b.

Edward N Lorenz. A study of the predictability of a 28-variable atmospheric model. *Tellus*, 17(3):321–333, 1965.

Amanda C Lough, Douglas A Wiens, C Grace Barcheck, Sridhar Anandakrishnan, Richard C Aster, Donald D Blankenship, Audrey D Huerta, Andrew Nyblade, Duncan A Young, and Terry J Wilson. Seismic detection of an active subglacial magmatic complex in marie byrd land, antarctica. *Nature Geoscience*, 6(12):1031–1035, 2013.

Brad Luen, Philip B Stark, et al. Testing earthquake predictions. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 302–315. Institute of Mathematical Statistics, 2008.

David J Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10 (4):325–337, 2000.

Ian Main. Long odds on prediction. *Nature*, 385(6611):19–20, 1997.

Simone Mancini, Margarita Segou, Maximilian Jonas Werner, and Tom Parsons. The predictive skills of elastic coulomb rate-and-state aftershock forecasts during the 2019 ridgecrest, california, earthquake sequence. *Bulletin of the Seismological Society of America*, 110(4): 1736–1751, 2020.

Benoit B Mandelbrot and Benoit B Mandelbrot. *The fractal geometry of nature*, volume 1. WH freeman New York, 1982.

David Marsan and Olivier Lengline. Extending earthquakes' reach through cascading. *Science*, 319(5866):1076–1079, 2008.

Thiago G Martins, Daniel Simpson, Finn Lindgren, and Håvard Rue. Bayesian computing with inla: new features. *Computational Statistics & Data Analysis*, 67:68–83, 2013.

W Marzocchi, D Schorlemmer, and S Wiemer. An earthquake forecast experiment in italy. *Annals of geophysics*, 53, 2010.

Warner Marzocchi and Thomas H Jordan. Testing for ontological errors in probabilistic forecasting models of natural systems. *Proceedings of the National Academy of Sciences*, 111(33):11973–11978, 2014.

Warner Marzocchi and Anna Maria Lombardi. Real-time forecasting following a damaging earthquake. *Geophysical Research Letters*, 36(21), 2009.

Warner Marzocchi and Laura Sandri. A review and new insights on the estimation of the b-valueand its uncertainty. *Annals of geophysics*, 2003.

Warner Marzocchi and J Douglas Zechar. Earthquake forecasting and earthquake prediction: different approaches for obtaining the best model. *Seismological Research Letters*, 82(3): 442–448, 2011.

Warner Marzocchi and Jiancang Zhuang. Statistics between mainshocks and foreshocks in italy and southern california. *Geophysical Research Letters*, 38(9), 2011.

Warner Marzocchi, J Douglas Zechar, and Thomas H Jordan. Bayesian forecast evaluation and ensemble earthquake forecasting. *Bulletin of the Seismological Society of America*, 102(6):2574–2584, 2012.

Warner Marzocchi, Anna Maria Lombardi, and Emanuele Casarotti. The establishment of an operational earthquake forecasting system in italy. *Seismological Research Letters*, 85 (5):961–969, 2014.

Warner Marzocchi, Matteo Taroni, and Jacopo Selva. Accounting for epistemic uncertainty in psha: Logic tree and ensemble modeling. *Bulletin of the Seismological Society of America*, 105(4):2151–2159, 2015.

Warner Marzocchi, Matteo Taroni, and Giuseppe Falcone. Earthquake forecasting during the complex amatrice-norcia seismic sequence. *Science Advances*, 3(9):e1701239, 2017.

Warner Marzocchi, Ilaria Spassiani, Angela Stallone, and Matteo Taroni. How to be fooled searching for significant variations of the b-value. *Geophysical Journal International*, 220 (3):1845–1856, 2020.

Bertil Matérn. *Spatial Variation: Stochastic Models and Their Application to Some Problems in Forst Survey and Other Sampling Investigations*. Esselte, 1960.

Robert McCaffrey. The next great earthquake. *science*, 315(5819):1675–1676, 2007.

Dan P McKenzie and Robert L Parker. The north pacific: an example of tectonics on a sphere. *Nature*, 216(5122):1276–1280, 1967.

Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

Carlo Meletti, Warner Marzocchi, Vera D'amico, Giovanni Lanzano, Lucia Luzi, Francesco Martinelli, Bruno Pace, Andrea Rovida, Matteo Taroni, Francesco Visini, et al. The new italian seismic hazard model (mps19). *Annals of Geophysics*, 64(1), 2021.

Paul-André Meyer. Démonstration simplifiée d'un théorème de knight. *Séminaire de probabilités de Strasbourg*, 5:191–195, 1971.

Sebastian Meyer, Leonhard Held, and Michael Höhle. Spatio-temporal analysis of epidemic phenomena using the r package surveillance. *arXiv preprint arXiv:1411.0416*, 2014.

Andrew J Michael and Maximilian J Werner. Preface to the focus section on the collaboratory for the study of earthquake predictability (csep): New results and future directions. *Seismological Research Letters*, 89(4):1226–1228, 2018.

Maddalena Michele, Raffaele Di Stefano, Lauro Chiaraluce, Marco Cattaneo, Pasquale De Gori, Giancarlo Monachesi, Diana Latorre, Simone Marzorati, Luisa Valoroso, Chiara Ladina, et al. The amatrice 2016 seismic sequence: a preliminary look at the mainshock. *Annals of Geophysics*, 2016.

Arnaud Mignan. Seismicity precursors to large earthquakes unified in a stress accumulation framework. *Geophysical research letters*, 39(21), 2012.

Arnaud Mignan and Jochen Woessner. Estimating the magnitude of completeness for earthquake catalogs. *Community Online Resource for Statistical Seismicity Analysis*, pages 1–45, 2012.

Leila Mizrahi, Shyam Nandan, and Stefan Wiemer. Embracing data incompleteness for better earthquake forecasting. *Journal of Geophysical Research: Solid Earth*, 126(12): e2021JB022379, 2021.

Kiyoo Mogi. On the time distribution of aftershocks accompanying the recent major earthquakes in and near japan. *Bulletin of the Earthquake Research Institute, University of Tokyo*, 40(1):107–124, 1962.

George Mohler. Modeling and estimation of multi-source clustering in crime and security data. *The Annals of Applied Statistics*, pages 1525–1539, 2013.

George Mohler, Jeremy Carter, and Rajeev Raje. Improving social harm indices with a modulated hawkes process. *International Journal of Forecasting*, 34(3):431–439, 2018.

George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

George Molchan. Space—time earthquake prediction: The error diagrams. In *Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II*, pages 53–63. Springer, 2010.

George M Molchan. Structure of optimal strategies in earthquake prediction. *Tectonophysics*, 193(4):267–276, 1991.

Christian Molkenthin, Christian Donner, Sebastian Reich, Gert Zöller, Sebastian Hainzl, Matthias Holschneider, and Manfred Opper. Gp-etas: semiparametric bayesian inference for the spatio-temporal epidemic type aftershock sequence model. *Statistics and Computing*, 32(2):1–25, 2022.

Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

W Jason Morgan. Rises, trenches, great faults, and crustal blocks. *Journal of Geophysical Research*, 73(6):1959–1982, 1968.

Robert Muir-Wood. From global seismotectonics to global seismic hazard. 1993.

Francesco Mulargia, Philip B Stark, and Robert J Geller. Why is probabilistic seismic hazard analysis (psha) still used? *Physics of the Earth and Planetary Interiors*, 264:63–75, 2017.

Allan H Murphy. What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and forecasting*, 8(2):281–293, 1993.

Roger Musson. On the nature of logic trees in probabilistic seismic hazard assessment. *Earthquake Spectra*, 28(3):1291–1296, 2012.

Shyam Nandan, Guy Ouillon, Didier Sornette, and Stefan Wiemer. Forecasting the full distribution of earthquake numbers is fair, robust, and better. *Seismological Research Letters*, 90(4):1650–1659, 2019.

Shyam Nandan, Yavor Kamer, Guy Ouillon, Stefan Hiemer, and Didier Sornette. Global models for short-term earthquake forecasting and predictive skill assessment. *The European Physical Journal Special Topics*, 230(1):425–449, 2021a.

Shyam Nandan, Sumit Kumar Ram, Guy Ouillon, and Didier Sornette. Is seismicity operating at a critical point? *Physical Review Letters*, 126(12):128501, 2021b.

KZ Nanjo, H Tsuruoka, S Yokoi, Y Ogata, G Falcone, N Hirata, Y Ishigaki, TH Jordan, K Kasahara, K Obara, et al. Predictability study on the aftershock sequence following the 2011 tohoku-oki, japan, earthquake: First results. *Geophysical Journal International*, 191 (2):653–658, 2012.

C Narteau, P Shebalin, and M Holschneider. Temporal limits of the power law aftershock decay rate. *Journal of Geophysical Research: Solid Earth*, 107(B12):ESE–12, 2002.

FA Nava, VH Márquez-Ramírez, FR Zúñiga, L Ávila-Barrientos, and CB Quinteros. Gutenberg-richter b-value maximum likelihood estimation and sample size. *Journal of Seismology*, 21(1):127–135, 2017.

Mark Naylor and Francesco Serafini. edinburgh-seismicity-hub/etas.inlabru: Temporal hawkes, 2023. URL https://doi.org/10.5281/zenodo.7515785.

Mark Naylor, Francesco Serafini, Finn Lindgren, and Ian Main. Bayesian modelling of the temporal evolution of seismicity using the etas. inlabru r-package. *arXiv preprint arXiv:2212.06077*, 2022.

Jerzy Neyman and Elizabeth L Scott. Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(1):1–29, 1958.

Cecilia I Nievas, Helen Crowley, Yves Reuland, Graeme Weatherill, Georgios Baltzopoulos, Kirsty Bayliss, Eleni Chatzi, Philippe Guéguen, Mark Naylor, Mabel Orlacchio, et al. Exploration of state-dependent rapid loss assessment and event-based operational earthquake loss forecasting incorporating structural health monitoring: An open-source tool. In *SECED 2023 Conference: Earthquake Engineering & Dynamics for a Sustainable Future*, 2023.

Erhard Oeser. *Historical earthquake theories from Aristotle to Kant*. Geologische Bundesanstalt, 1992.

Y. Ogata. *Statistical Analysis of Seismicity, Updated Version (SASeis 2006)*. Computer science monographs. Institute of Statistical Mathematics, 2006. URL https://books.google.co.uk/books?id=HH5EGwAACAAJ.

Yosihiko Ogata. Estimation of the parameters in the modified omori formula for aftershock frequencies by the maximum likelihood procedure. *Journal of Physics of the Earth*, 31(2): 115–124, 1983.

Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83(401):9–27, 1988.

Yosihiko Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998.

Yosihiko Ogata. Seismicity analysis through point-process modeling: A review. In *Seismicity patterns, their statistical significance and physical meaning*, pages 471–507. Springer, 1999.

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

Yosihiko Ogata and Koichi Katsura. Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogues. *Geophysical Journal International*, 113(3):727–738, 1993.

Yosihiko Ogata and Jiancang Zhuang. Space–time etas models and an improved extension. *Tectonophysics*, 413(1-2):13–23, 2006.

Emile A Okal and Seth Stein. Observations of ultra-long period normal modes from the 2004 sumatra–andaman earthquake. *Physics of the Earth and Planetary Interiors*, 175 (1-2):53–62, 2009.

Takahiro Omi, Yosihiko Ogata, Yoshito Hirata, and Kazuyuki Aihara. Estimating the etas model from an early aftershock sequence. *Geophysical Research Letters*, 41(3):850–857, 2014.

Takahiro Omi, Yosihiko Ogata, Yoshito Hirata, and Kazuyuki Aihara. Intermediate-term forecasting of aftershocks from an early aftershock sequence: Bayesian and ensemble forecasting approaches. *Journal of Geophysical Research: Solid Earth*, 120(4):2561–2578, 2015.

Fusakichi Omori. On the after-shocks of earthquakes. *J. Coll. Sci., Imp. Univ., Japan*, 7: 111–200, 1894.

Fusakichi Omori. *On the after-shocks of earthquakes*. PhD thesis, The University of Tokyo, 1895.

Norimitsu Onishi. Seawalls offered little protection against tsunami's crushing waves. *New York Times*, 13, 2011.

Nina Opitz, Caroline Marcon, Anja Paschold, Waqas Ahmed Malik, Andrew Lithio, Ronny Brandt, Hans-Peter Piepho, Dan Nettleton, and Frank Hochholdinger. Extensive tissue-specific transcriptomic plasticity in maize primary roots upon water deficit. *Journal of Experimental Botany*, 67(4):1095–1107, 2016.

Guy Ouillon and Didier Sornette. Magnitude-dependent omori law: Theory and empirical study. *Journal of Geophysical Research: Solid Earth*, 110(B4), 2005.

Marco Pagani, Julio Garcia-Pelaez, Robin Gee, Kendra Johnson, Valerio Poggi, Vitor Silva, Michele Simionato, Richard Styron, Daniele Viganò, Laurentiu Danciu, et al. The 2018 version of the global earthquake model: hazard component. *Earthquake Spectra*, 36 (1_suppl):226–251, 2020.

Frederic Paik Schoenberg. Nonparametric estimation of variable productivity hawkes processes. *Environmetrics*, 33(6):e2747, 2022.

Wendy S Parker. Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change*, 4(3):213–223, 2013.

Roger D Peng, Frederic Paik Schoenberg, and James A Woods. A space–time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association*, 100(469):26–35, 2005.

Zhigang Peng and Peng Zhao. Migration of early aftershocks following the 2004 parkfield earthquake. *Nature Geoscience*, 2(12):877–881, 2009.

Patrizio Petricca, Eugenio Carminati, and Carlo Doglioni. The decollement depth of active thrust faults in italy: implications on potential earthquake magnitude. *Tectonics*, 38(11): 3990–4009, 2019.

Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria., 2003.

Karl R Popper. The logic of scientific discovery. *Central Works of Philosophy v4: Twentieth Century: Moore to Popper*, 4:262, 2015.

Jakob Gulddahl Rasmussen. Bayesian inference for hawkes processes. *Methodology and Computing in Applied Probability*, 15(3):623–642, 2013.

Paul Reasenberg. Second-order moment of central california seismicity, 1969–1982. *Journal of Geophysical Research: Solid Earth*, 90(B7):5479–5495, 1985.

Alex Reinhart and Joel Greenhouse. Self-exciting point processes with spatial covariates: modelling the dynamics of crime. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1305–1329, 2018.

DA Rhoades, M Liukis, A Christophersen, and MC Gerstenberger. Retrospective tests of hybrid operational earthquake forecasting models for canterbury. *Geophysical Journal International*, 204(1):440–456, 2016.

David A Rhoades, Danijel Schorlemmer, Matthew C Gerstenberger, Annemarie Christophersen, J Douglas Zechar, and Masajiro Imoto. Efficient testing of earthquake forecasting models. *Acta Geophysica*, 59(4):728–747, 2011.

David A Rhoades, Annemarie Christophersen, Matthew C Gerstenberger, Maria Liukis, Fabio Silva, Warner Marzocchi, Maximilian J Werner, and Thomas H Jordan. Highlights from the first ten years of the new zealand earthquake forecast testing center. *Seismological Research Letters*, 89(4):1229–1237, 2018.

Andrea Riebler, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165, 2016.

Brian D Ripley. The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266, 1976.

Brian D Ripley. Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):172–192, 1977.

Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

Natalia C Roos, Adriana R Carvalho, Priscila FM Lopes, and M Grazia Pennino. Modeling sensitive parrotfish (labridae: Scarini) habitats along the brazilian coast. *Marine Environmental Research*, 110:92–100, 2015.

David B Rosen. How good were those probability predictions? the expected recommendation loss (erl) scoring rule. In *Maximum Entropy and Bayesian Methods*, pages 401–408. Springer, 1996.

Gordon J Ross. Bayesian estimation of the etas model for earthquake occurrences. *Bulletin of the Seismological Society of America*, 111(3):1473–1480, 2021.

Gordon J Ross and Aleksandar A Kolev. Semiparametric bayesian forecasting of spatiotemporal earthquake occurrences. *The Annals of Applied Statistics*, 16(4):2083–2100, 2022.

Andrea Rovida, Mario Locati, Romano Camassi, Barbara Lolli, and Paolo Gasperini. The italian earthquake catalogue cpti15. *Bulletin of Earthquake Engineering*, 18(7):2953–2984, 2020.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

Håvard Rue, Andrea Riebler, Sigrunn H Sørbye, Janine B Illian, Daniel P Simpson, and Finn K Lindgren. Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017.

Eva Santermans, Emmanuel Robesyn, Tapiwa Ganyani, Bertrand Sudre, Christel Faes, Chantal Quinten, Wim Van Bortel, Tom Haber, Thomas Kovac, Frank Van Reeth, et al. Spatiotemporal evolution of ebola virus disease at sub-national level during the 2014 west africa epidemic: model scrutiny and data meagreness. *PloS one*, 11(1):e0147172, 2016.

William H Savran, Maximilian J Werner, Warner Marzocchi, David A Rhoades, David D Jackson, Kevin Milner, Edward Field, and Andrew Michael. Pseudoprospective evaluation of ucerf3-etas forecasts during the 2019 ridgecrest sequence. *Bulletin of the Seismological Society of America*, 110(4):1799–1817, 2020.

William H Savran, José A Bayona, Pablo Iturrieta, Khawaja M Asim, Han Bao, Kirsty Bayliss, Marcus Herrmann, Danijel Schorlemmer, Philip J Maechling, and Maximilian J Werner. pycsep: A python toolkit for earthquake forecast developers. *Seismological Society of America*, 93(5):2858–2870, 2022.

Mark J Schervish. *Theory of statistics*. Springer Science & Business Media, 2012.

Max Schneider and Peter Guttorp. Final technical report bayesian etas: Towards improved operational aftershock forecasting: Collaborative research with uw and usgs.

Max Schneider, Robert Clements, David Rhoades, and Danijel Schorlemmer. Likelihood- and residual-based evaluation of medium-term earthquake forecast models for california. *Geophysical Journal International*, 198(3):1307–1318, 2014.

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

Frederic Paik Schoenberg. Testing separability in spatial-temporal marked point processes. *Biometrics*, pages 471–481, 2004.

Frederic Paik Schoenberg. Facilitated estimation of etas. *Bulletin of the seismological Society of America*, 103(1):601–605, 2013.

Frederic Paik Schoenberg, Annie Chu, and Alejandro Veen. On the relationship between lower magnitude thresholds and bias in epidemic-type aftershock sequence parameter estimates. *Journal of Geophysical Research: Solid Earth*, 115(B4), 2010.

D Schorlemmer, MC Gerstenberger, S Wiemer, DD Jackson, and DA Rhoades. Earthquake likelihood model testing. *Seismological Research Letters*, 78(1):17–29, 2007a.

D Schorlemmer, MC Gerstenberger, S Wiemer, DD Jackson, and DA Rhoades. Earthquake likelihood model testing. *Seismological Research Letters*, 78(1):17–29, 2007b.

Danijel Schorlemmer and MC Gerstenberger. Relm testing center. *Seismological Research Letters*, 78(1):30–36, 2007a.

Danijel Schorlemmer and MC Gerstenberger. Relm testing center. *Seismological Research Letters*, 78(1):30–36, 2007b.

Danijel Schorlemmer, Stefan Wiemer, and Max Wyss. Variations in earthquake-size distribution across different stress regimes. *Nature*, 437(7058):539–542, 2005.

Danijel Schorlemmer, J Douglas Zechar, Maximilian J Werner, Edward H Field, David D Jackson, and Thomas H Jordan. First results of the regional earthquake likelihood models experiment. In *Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II*, pages 5–22. Springer, 2010.

Danijel Schorlemmer, Maximilian J Werner, Warner Marzocchi, Thomas H Jordan, Yosihiko Ogata, David D Jackson, Sum Mak, David A Rhoades, Matthew C Gerstenberger, Naoshi Hirata, et al. The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismological Research Letters*, 89(4):1305–1313, 2018.

Birgit Schrödle and Leonhard Held. A primer on disease mapping and ecological regression using `INLA`. *Computational statistics*, 26(2):241–258, 2011a.

Birgit Schrödle and Leonhard Held. Spatio-temporal disease mapping using inla. *Environmetrics*, 22(6):725–734, 2011b.

Bernd Schurr, Günter Asch, Sebastian Hainzl, Jonathan Bedford, Andreas Hoechner, Mauro Palo, Rongjiang Wang, Marcos Moreno, Mitja Bartsch, Yong Zhang, et al. Gradual unlocking of plate boundary controlled initiation of the 2014 iquique earthquake. *Nature*, 512(7514):299–302, 2014.

David P Schwartz and Kevin J Coppersmith. Fault behavior and characteristic earthquakes: Examples from the wasatch and san andreas fault zones. *Journal of Geophysical Research: Solid Earth*, 89(B7):5681–5698, 1984.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Stefanie Seif, Arnaud Mignan, Jeremy Douglas Zechar, Maximilian Jonas Werner, and Stefan Wiemer. Estimating etas: The effects of truncation, missing data, and model assumptions. *Journal of Geophysical Research: Solid Earth*, 122(1):449–469, 2017.

Francesco Serafini, Finn Lindgren, and Mark Naylor. Approximation of bayesian hawkes process models with inlabru. *arXiv preprint arXiv:2206.13360*, 2022a.

Francesco Serafini, Mark Naylor, Finn Lindgren, Maximilian J Werner, and Ian Main. Ranking earthquake forecasts using proper scoring rules: Binary events in a low probability environment. *Geophysical Journal International*, 230(2):1419–1440, 2022b.

Robert Shcherbakov. Bayesian confidence intervals for the magnitude of the largest aftershock. *Geophysical Research Letters*, 41(18):6380–6388, 2014.

Robert Shcherbakov. Statistics and forecasting of aftershocks during the 2019 ridgecrest, california, earthquake sequence. *Journal of Geophysical Research: Solid Earth*, 126(2):e2020JB020887, 2021. doi: https://doi.org/10.1029/2020JB020887. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JB020887. e2020JB020887 2020JB020887.

Peter M Shearer. *Introduction to seismology*. Cambridge university press, 2019.

Zheng-Kang Shen, David D Jackson, and Yan Y Kagan. Implications of geodetic strain rate for future earthquakes, with a five-year forecast of m5 earthquakes in southern california. *Seismological Research Letters*, 78(1):116–120, 2007.

Vitor Silva, Desmond Amo-Oduro, Alejandro Calderon, Catarina Costa, Jamal Dabbeek, Venetia Despotaki, Luis Martins, Marco Pagani, Anirudh Rao, Michele Simionato, et al. Development of a global seismic risk model. *Earthquake Spectra*, 36(1_suppl):372–394, 2020.

Ansie Smit, Alfred Stein, and Andrzej Kijko. Bayesian inference in natural hazard analysis for incomplete and uncertain data. *Environmetrics*, 30(6):e2566, 2019.

G Solomos, A Pinto, and S Dimova. A review of the seismic hazard zonation in national building codes in the context of eurocode 8. *JRC Scientific and Technical reports*, 2008.

Didier Sornette. *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media, 2006.

Simon C Stähler, Amir Khan, W Bruce Banerdt, Philippe Lognonné, Domenico Giardini, Savas Ceylan, Mélanie Drilleau, A Cecilia Duran, Raphaël F Garcia, Quancheng Huang, et al. Seismic detection of the martian core. *Science*, 373(6553):443–448, 2021.

Philip B Stark. Earthquake prediction: the null hypothesis. *Geophysical Journal International*, 131(3):495–499, 1997.

Philip B Stark. Pay no attention to the model behind the curtain. *Pure and Applied Geophysics*, pages 1–25, 2022.

Sandy Steacy, Matt Gerstenberger, Charles Williams, David Rhoades, and Annemarie Christophersen. A new hybrid coulomb/statistical model for forecasting aftershock rates. *Geophysical Journal International*, 196(2):918–923, 2014.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

Seth Stein, Robert J Geller, and Mian Liu. Why earthquake hazard maps often fail and what to do about it. *Tectonophysics*, 562:1–25, 2012.

Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Conference on Learning Theory*, pages 482–526. PMLR, 2014.

Tom Stindl and Feng Chen. Spatiotemporal etas model with a renewal main-shock arrival process. *arXiv preprint arXiv:2112.07861*, 2021.

MW Stirling and CJN Wilson. Development of a volcanic hazard model for new zealand: first approaches from the methods of probabilistic seismic hazard analysis. *Bulletin of the New Zealand Society for Earthquake Engineering*, 35(4):266–277, 2002.

Anne Strader, Max Schneider, and Danijel Schorlemmer. Prospective and retrospective evaluation of five-year earthquake forecast models for california. *Geophysical Journal International*, 211(1):239–251, 2017.

Anne Strader, Maximilian Werner, José Bayona, Philip Maechling, Fabio Silva, Maria Liukis, and Danijel Schorlemmer. Prospective evaluation of global earthquake forecast models: 2 yrs of observations provide preliminary support for merging smoothed seismicity with geodetic strain rates. *Seismological Research Letters*, 89(4):1262–1271, 2018.

M Stucchi, C Meletti, V Montaldo, A Akinci, E Faccioli, P Gasperini, L Malagnini, and G Valensise. Pericolosità sismica di riferimento per il territorio nazionale mps04 [data set]. *Istituto Nazionale di Geofisica e Vulcanologia (INGV)*, 2004.

Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

John A Swets. The relative operating characteristic in psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science*, 182(4116):990–1000, 1973.

Lynn R Sykes. Aftershock zones of great earthquakes, seismicity gaps, and earthquake prediction for alaska and the aleutians. *Journal of Geophysical Research*, 76(32):8021–8041, 1971.

M Taroni, JD Zechar, and W Marzocchi. Assessing annual global m 6+ seismicity forecasts. *Geophysical Journal International*, 196(1):422–431, 2014.

Matteo Taroni, Warner Marzocchi, and Pamela Roselli. Assessing 'alarm-based cn'earthquake predictions in italy. *Annals of Geophysics*, 59(6):S0648–S0648, 2016.

Matteo Taroni, Warner Marzocchi, Danijel Schorlemmer, Maximilian Jonas Werner, Stefan Wiemer, Jeremy Douglas Zechar, Lukas Heiniger, and Fabian Euchner. Prospective csep evaluation of 1-day, 3-month, and 5-yr earthquake forecasts for italy. *Seismological Research Letters*, 89(4):1251–1261, 2018.

Benjamin M Taylor and Peter J Diggle. Inla or mcmc? a tutorial and comparative evaluation for spatial prediction in log-gaussian cox processes. *Journal of Statistical Computation and Simulation*, 84(10):2266–2284, 2014.

T Team. Pandas development pandas-dev/pandas: Pandas. *Zenodo*, 21:1–9, 2020. URL `10.5281/zenodo.3509134`.

Jiaqi Teng, Shuzhen Ding, Huiguo Zhang, Kai Wang, and Xijian Hu. Bayesian spatiotemporal modelling analysis of hemorrhagic fever with renal syndrome outbreaks in china using r-inla. *Zoonoses and Public Health*, 2022.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86, 1986.

Theresa Tormann, Stefan Wiemer, Sabrina Metzger, A Michael, and Jeanne L Hardebeck. Size distribution of parkfield's microearthquakes reflects changes in surface creep rate. *Geophysical Journal International*, 193(3):1474–1478, 2013.

S. Touati, M. Naylor, I. G. Main, and M. Christie. Masking of earthquake triggering behavior by a high background rate and implications for epidemic-type aftershock sequence inversions. *Journal of Geophysical Research: Solid Earth*, 116(B3), 2011. doi: https://doi.org/10.1029/2010JB007544. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JB007544.

Sarah Touati, Mark Naylor, and Ian G Main. Origin and nonuniversality of the earthquake interevent time distribution. *Physical Review Letters*, 102(16):168501, 2009.

Sarah Touati, Mark Naylor, and Ian G. Main. Statistical Modeling of the 1997–1998 Colfiorito Earthquake Sequence: Locating a Stationary Solution within Parameter Uncertainty. *Bulletin of the Seismological Society of America*, 104(2):885–897, 2014. ISSN 0037-1106. doi: 10.1785/0120130270.

Hiroshi Tsuruoka, Naoshi Hirata, Danijel Schorlemmer, Fabian Euchner, Kazuyoshi Z Nanjo, and Thomas H Jordan. Csep testing center and the first results of the earthquake forecast testing experiment in japan. *Earth, planets and space*, 64(8):661–671, 2012.

T Utsu. Introduction to seismicity. *Mathematical Seismology*, 7:139–157, 1992.

Tokuji Utsu. A relation between the area of after-shock region and the energy of main-shock. *J. Seismol. Soc. Jpn., 2*, 7:233–240, 1955.

Tokuji Utsu. Magnitudes of earthquakes and occurrence of their aftershocks. *Zisin, Ser. 2*, 10:35–45, 1957.

Tokuji Utsu. A statistical significance test of the difference in b-value between two earthquake groups. *Journal of Physics of the Earth*, 14(2):37–40, 1966.

Tokuji Utsu. Representation and analysis of the earthquake size distribution: a historical review and some new approaches. In *Seismicity patterns, their statistical significance and physical meaning*, pages 509–535. Springer, 1999.

Tokuji Utsu and Hiromu Okada. Anomalies in seismic wave velocity and attenuation associated with a deep earthquake zone (2). *Journal of the Faculty of Science, Hokkaido University. Series 7, Geophysics*, 3(2):65–84, 1969.

Tokuji Utsu, Yosihiko Ogata, et al. The centenary of the omori formula for a decay law of aftershock activity. *Journal of Physics of the Earth*, 43(1):1–33, 1995.

Nicholas J. van der Elst. B-positive: A robust estimator of aftershock magnitude distribution in transiently incomplete catalogs. *Journal of Geophysical Research: Solid Earth*, 126(2): e2020JB021027, 2021a. doi: https://doi.org/10.1029/2020JB021027.

Nicholas J van der Elst. B-positive: A robust estimator of aftershock magnitude distribution in transiently incomplete catalogs. *Journal of Geophysical Research: Solid Earth*, 126(2): e2020JB021027, 2021b.

Nicholas J van der Elst and Morgan T Page. Nonparametric aftershock forecasts based on similar sequences in the past. *Seismological Research Letters*, 89(1):145–152, 2018.

Nicholas J van der Elst, Jeanne L Hardebeck, Andrew J Michael, Sara K McBride, and Elizabeth Vanacore. Prospective and retrospective evaluation of the us geological survey public aftershock forecast for the 2019–2021 southwest puerto rico earthquake and aftershocks. *Seismological Society of America*, 93(2A):620–640, 2022.

Janet van Niekerk and Haavard Rue. Correcting the laplace method with variational bayes. *arXiv preprint arXiv:2111.12945*, 2021.

Janet Van Niekerk, Elias Krainski, Denis Rustand, and Haavard Rue. A new avenue for bayesian inference with inla. *Computational Statistics & Data Analysis*, page 107692, 2023.

N Vargas and Tilmann Gneiting. Bayesian point process modelling of earthquake occurrences. Technical report, Technical Report, Ruprecht-Karls University Heidelberg, Heidelberg, Germany . . . , 2012.

Alejandro Veen and Frederic P Schoenberg. Estimation of space–time branching process models in seismology using an em–type algorithm. *Journal of the American Statistical Association*, 103(482):614–624, 2008.

D Vere-Jones. Probabilities and information gain for earthquake forecasting. *Computational Seismology*, 30:249–263, 1998.

David Vere-Jones. Stochastic models for earthquake occurrence. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(1):1–45, 1970.

Sean Wallis. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics*, 20(3): 178–208, 2013.

Wang, Yu Yue, and Faraway. *Bayesian Regression Modeling with INLA*. Chapman & Hall, UK United Kingdom, 2018. ISBN 9781498727259.

Qi Wang, David D Jackson, and Jiancang Zhuang. Missing links in earthquake clustering models. *Geophysical Research Letters*, 37(21), 2010.

Steven N Ward. Methods for evaluating earthquake potential and likelihood in and around california. *Seismological Research Letters*, 78(1):121–133, 2007.

Sumio Watanabe. A widely applicable bayesian information criterion. *The Journal of Machine Learning Research*, 14(1):867–897, 2013.

Alfred Wegener. Die entstehung der kontinente. *Geologische Rundschau*, 3(4):276–292, 1912.

Zachary D Weller, Jennifer A Hoeting, and Joseph C von Fischer. A calibration capture–recapture model for inferring natural gas leak population characteristics using data from google street view cars. *Environmetrics*, 29(7):e2519, 2018.

Maximilian J Werner and Didier Sornette. Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments. *Journal of Geophysical Research: Solid Earth*, 113(B8), 2008.

Maximilian J Werner, Agnès Helmstetter, David D Jackson, Yan Y Kagan, and Stefan Wiemer. Adaptively smoothed seismicity earthquake forecasts for italy. *arXiv preprint arXiv:1003.4374*, 2010.

Maximilian J Werner, Agnès Helmstetter, David D Jackson, and Yan Y Kagan. High-resolution long-term and short-term earthquake forecasts for california. *Bulletin of the Seismological Society of America*, 101(4):1630–1648, 2011.

Peter Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

Stefan Wiemer and Max Wyss. Mapping spatial variability of the frequency-magnitude distribution of earthquakes. In *Advances in geophysics*, volume 45, pages 259–V. Elsevier, 2002.

Frank Wilcoxon. Individual comparisons by ranking methods (1945). *Breakthroughs in Statistics*, pages 196–202.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

Samuel Stanley Wilks. Mathematical statistics. Technical report, J. Wiley, 1964.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Mathilde Wimez and WB Frank. A recursive matched-filter to systematically explore volcanic long-period earthquake swarms. *Geophysical Journal International*, 231(2):912–920, 2022.

J Woessner, Sebastian Hainzl, W Marzocchi, MJ Werner, AM Lombardi, F Catalli, B Enescu, M Cocco, MC Gerstenberger, and S Wiemer. A retrospective comparative forecast test on the 1992 landers sequence. *Journal of Geophysical Research: Solid Earth*, 116(B5), 2011.

Jochen Woessner and Stefan Wiemer. Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty. *Bulletin of the Seismological Society of America*, 95(2):684–698, 2005.

J Douglas Zechar and Thomas H Jordan. Testing alarm-based earthquake predictions. *Geophysical Journal International*, 172(2):715–724, 2008.

J Douglas Zechar and Thomas H Jordan. The area skill score statistic for evaluating earthquake predictability experiments. In *Seismogenesis and Earthquake Forecasting: The Frank Evison Volume II*, pages 39–52. Springer, 2010.

J Douglas Zechar and Jiancang Zhuang. Risk and return: evaluating reverse tracing of precursors earthquake predictions. *Geophysical Journal International*, 182(3):1319–1326, 2010.

J Douglas Zechar and Jiancang Zhuang. A parimutuel gambling perspective to compare probabilistic seismicity forecasts. *Geophysical Journal International*, 199(1):60–68, 2014.

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

J Douglas Zechar, Danijel Schorlemmer, Maria Liukis, John Yu, Fabian Euchner, Philip J Maechling, and Thomas H Jordan. The collaboratory for the study of earthquake predictability perspective on computational earthquake science. *Concurrency and Computation: Practice and Experience*, 22(12):1836–1847, 2010b.

J Douglas Zechar, Danijel Schorlemmer, Maximilian J Werner, Matthew C Gerstenberger, David A Rhoades, and Thomas H Jordan. Regional earthquake likelihood models i: First-order results. *Bulletin of the Seismological Society of America*, 103(2A):787–798, 2013.

Lizhong Zhang, Maximilian J Werner, and Katsuichiro Goda. Spatiotemporal seismic hazard and risk assessment of aftershocks of m 9 megathrust earthquakesspatiotemporal seismic hazard and risk assessment of aftershocks of m 9 megathrust earthquakes. *Bulletin of the Seismological Society of America*, 108(6):3313–3335, 2018.

Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649. PMLR, 2013.

Mu Zhu and Arthur Y Lu. The counter-intuitive non-informative prior for the bernoulli family. *Journal of Statistics Education*, 12(2), 2004.

Shixiang Zhu, Shuang Li, Zhigang Peng, and Yao Xie. Imitation learning of neural spatio-temporal point processes. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Jiancang Zhuang. Gambling scores for earthquake predictions and forecasts. *Geophysical Journal International*, 181(1):382–390, 2010.

Jiancang Zhuang. Next-day earthquake forecasts for the japan region generated by the etas model. *Earth, planets and space*, 63(3):207–216, 2011.

Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97(458): 369–380, 2002.

Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research: Solid Earth*, 109(B5), 2004.

Jiancang Zhuang, Yosihiko Ogata, and Ting Wang. Data completeness of the kumamoto earthquake sequence in the jma catalog and its influence on the estimation of the etas parameters. *Earth, Planets and Space*, 69(1):1–12, 2017.

Jiancang Zhuang, Yosihiko Ogata, and Ting Wang. Data completeness of the Kumamoto earthquake sequence in the JMA catalog and its influence on the estimation of the ETAS parameters. *Earth, Planets and Space*, 69(1):36, feb 2017. doi: 10.1186/s40623-017-0614-6.

Gert Zöller and Matthias Holschneider. The earthquake history in a fault zone tells us almost nothing about m max. *Seismological Research Letters*, 87(1):132–137, 2016.